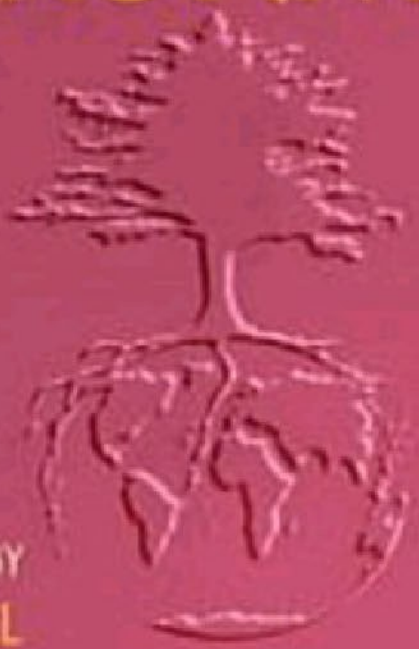
ENCYCLOPEDIA OF SOILS IN THE ENVIRONMENT



EDITED BY
DANIEL HILLEL
CYNTHIA ROSENZWEIG
DAVID POWLSON
KATE SCOW
MICHAIL SINGER
DONALD SPARKS

VOLUME ONE

ENCYCLOPEDIA OF SOILS IN THE ENVIRONMENT

FOUR-VOLUME SET

by Daniel Hillel (Editor-in-Chief)

Hardcover: 2200 pages
Publisher: Academic Press; 1 edition (November 8, 2004)
Language: English
ISBN-10: 0123485304
ISBN-13: 978-0123485304

Book Description

More than ever before, a compelling need exists for an encyclopedic resource about soil the rich mix of mineral particles, organic matter, gases, and soluble compounds that foster both plant and animal growth. Civilization depends more on the soil as human populations continue to grow and increasing demands are placed upon available resources.
The Encyclopedia of Soils in the Environmentis a comprehensive and integrated consideration of a topic of vital importance to human societies in the past, present, and future.

This important work encompasses the present knowledge of the world's variegated soils, their origins, properties, classification, and roles in the biosphere. A team of outstanding, international contributors has written over 250 entries that cover a broad range of issues facing today's soil scientists, ecologists, and environmental scientists.
This four-volume set features thorough articles that survey specific aspects of soil biology, ecology, chemistry and physics. Rounding out the encyclopedia's excellent coverage, contributions cover cross-disciplinary subjects, such as the history of soil utilization for agricultural and engineering purposes and soils in relation to the remediation of pollution and the mitigation of global climate change.

This comprehensive, yet accessible source is a valuable addition to the library of scientists, researchers, students, and policy makers involved in soil science, ecology, and environmental science.

Also available online via ScienceDirect featuring extensive browsing, searching, and internal cross-referencing between articles in the work, plus dynamic linking to journal articles and abstract databases, making navigation flexible and easy. For more information, pricing options and availability visit www.info.sciencedirect.com.

* A distinguished international group of editors and contributors
* Well-organized encyclopedic format providing concise, readable entries, easy searches, and thorough cross-references
* Abundant visual resources — photographs, figures, tables, and graphs — in every entry
* Complete up-to-date coverage of many important topics — essential information for scientists, students and professionals alike

EDITOR-IN-CHIEF

Daniel Hillel
Columbia University
New York, NY
USA

EDITORS

Jerry L Hatfield
National Soil Tilth Laboratory
Ames, IA
USA

Kate M Scow
University of California
Davis, CA
USA

David S Powlson
Rothamsted Research
Harpenden
UK

Michael J Singer
University of California
Davis, CA
USA

Cynthia Rosenzweig
NASA Goddard Institute for Space Studies
New York, NY
USA

Donald L Sparks
University of Delaware
Newark, DE
USA

EDITORIAL ADVISORY BOARD

R Bardgett
Lancaster University
Lancaster
UK

J L Boettinger
Utah State University
Logan, UT
USA

G Gee
Pacific Northwest National Laboratory
Richland, WA
USA

R Keren
The Volcani Center
Bet Dagan
Israel

J Kimble
USDA Natural Resources Conservation Service
Lincoln, NE
USA

M B Kirkham
Kansas State University
Manhattan, KS
USA

M Kutilek
Prague
Czech Republic

D Martens
Southwest Watershed Research Center
Tucson, AZ
USA

K Mengel
Justus Leibig University
Giessen
Germany

K Reichardt
Center for Nuclear Energy in Agriculture
Piracicaba
Brazil

K Ritz
Cranfield University
Silsoe
UK

R Schulin
Swiss Federal Institute of Technology Zurich
Schlieren
Switzerland

N Senesi
Università di Bari
Bari
Italy

J T Sims
University of Delaware
Newark, DE
USA

K Smith
University of Edinburgh
Edinburgh
UK

R L Tate
Rutgers University
New Brunswick, NJ
USA

N van Breemen
Wageningen Agricultural University
Wageningen
The Netherlands

W H van Riemsdijk
Department of Soil Quality
Wageningen
The Netherlands

FOREWORD

The *Encyclopedia of Soils in the Environment* is a vitally important scientific publication and an equally important contribution to global public policy. The *Encyclopedia* brings together a remarkable range of cutting-edge scientific knowledge on all aspects of soil science, as well as the links of soils and soil science to environmental management, food production, biodiversity, climate change, and many other areas of significant concern. Even more than that, the *Encyclopedia* will immediately become an indispensable resource for policy makers, analysts, and students who are focusing on one of the greatest challenges of the 21st century. With 6.3 billion people, our planet is already straining to feed the world's population, and is failing to do so reliably in many parts of the world. The numbers of chronically poor in the world have been stuck at some 800 million in recent years, despite long-standing international goals and commitments to reduce that number by several hundred million. Yet the challenge of food production will intensify in coming decades, as the human population is projected to rise to around 9 billion by mid-century, with the increased population concentrated in parts of the world already suffering from widespread chronic under-nourishment.

Unless the best science is brought to these problems, the situation is likely to deteriorate sharply. Food production systems are already under stress, for reasons often related directly to soils management. In Africa, crop yields are disastrously low and falling in many places due to the rampant depletion of soil nutrients. This situation needs urgent reversal, through increasing use of agro-forestry techniques (e.g. inter-cropping cereals with leguminous nitrogen-fixing trees) and increasing the efficient applications of chemical fertilizers. In other impoverished, as well as rich, parts of the planet, decades of intensive agriculture under irrigation have led to salinization, water-logging, eutrophication of major water bodies, dangerous declines of biodiversity and other forms of environmental degradation. These enormous strains are coupled with the continuing pressures of tropical deforestation and the lack of new promising regions for expanding crop cultivation to meet the needs of growing populations. Finally, there looms the prospect of anthropogenic climate change. Global warming and associated complex and poorly understood shifts in precipitation extremes and other climate variables all threaten the world's natural ecosystems and food production systems in profound yet still imperfectly understood ways. The risks of gradual or abrupt climate change are coupled with the risks of drastic perturbations to regional and global food supplies.

The *Encyclopedia* offers state-of-the-art contributions on each of these challenges, as well as links to entries on the fundamental biophysical processes that underpin the relevant phenomena. The world-scale and world-class collaboration that stands behind this unique project signifies its importance for the world community. It is an honor and privilege for me to introduce this path-breaking endeavor.

Jeffrey D Sachs
Director
The Earth Institute at Columbia University
Quetelet Professor of Sustainable Development
Columbia University, New York, USA

PREFACE

The term 'soil' refers to the weathered and fragmented outer layer of our planet's land surfaces. Formed initially through the physical disintegration and chemical alteration of rocks and minerals by physical and biogeochemical processes, soil is influenced by the activity and accumulated residues of a myriad of diverse forms of life. As it occurs in different geologic and climatic domains, soil is an exceedingly variegated body with a wide range of attributes.

Considering the height of the atmosphere, the thickness of the earth's rock mantle, and the depth of the ocean, one observes that soil is an amazingly thin body – typically not much more than one meter thick and often less than that. Yet it is the crucible of terrestrial life, within which biological productivity is generated and sustained. It acts like a composite living entity, a home to a community of innumerable microscopic and macroscopic plants and animals. A mere fistful of soil typically contains billions of microorganisms, which perform vital interactive biochemical functions. Another intrinsic attribute of the soil is its sponge-like porosity and its enormous internal surface area. That same fistful of soil may actually consist of several hectares of active surface, upon which physicochemical processes take place continuously.

Realizing humanity's utter dependence on the soil, ancient peoples, who lived in greater intimacy with nature than many of us today, actually revered the soil. It was not only their source of livelihood, but also the material from which they built their homes and that they learned to shape, heat, and fuse into household vessels and writing tablets (ceramic, made of clayey soil, being the first synthetic material in the history of technology). In the Bible, the name assigned to the first human was Adam, derived from 'adama,' meaning soil. The name given to that first earthling's mate was Hava (Eve, in transliteration), meaning 'living' or 'life-giving.' Together, therefore, Adam and Eve signified quite literally 'Soil and Life.'

The same powerful metaphor is echoed in the Latin name for the human species – Homo, derived from humus, the material of the soil. Hence, the adjective 'human' also implies 'of the soil.' Other ancient cultures evoked equally powerful associations. To the Greeks, the earth was a manifestation of Gaea, the maternal goddess who, impregnated by Uranus (god of the sky), gave birth to all the gods of the Greek pantheon.

Our civilization depends on the soil more crucially than ever, because our numbers have grown while available soil resources have diminished and deteriorated. Paradoxically, however, even as our dependence on the soil has increased, most of us have become physically and emotionally detached from it. Many of the people in the so-called 'developed' countries spend their lives in the artificial environment of a city, insulated from direct exposure to nature, and some children may now assume as a matter of course that food originates in supermarkets.

Detachment has bred ignorance, and out of ignorance has come the delusion that our civilization has risen above nature and has set itself free of its constraints. Agriculture and food security, erosion and salination, degradation of natural ecosystems, depletion and pollution of surface waters and aquifers, and decimation of biodiversity – all of these processes, which involve the soil directly or indirectly – have become abstractions to many people. The very language we use betrays disdain for that common material underfoot, often referred to as 'dirt.' Some fastidious parents prohibit their children from playing in the mud and hurry to wash their 'soiled' hands when the children nonetheless obey an innate instinct to do so. Thus soil is devalued and treated

as unclean though it is the terrestrial realm's principal medium of purification, wherein wastes are decomposed and nature's productivity is continually rejuvenated.

Scientists who observe soil closely see it in effect as a seething foundry in which matter and energy are in constant flux. Radiant energy from the sun streams onto the field and cascades through the soil and the plants growing in it. Heat is exchanged, water percolates through the soil's intricate passages, plant roots extract water and transmit it to their leaves, which transpire it back to the atmosphere. Leaves absorb carbon dioxide from the air and synthesize it with soil-derived water to form the primary compounds of life. Oxygen emitted by the leaves makes the air breathable for animals, which consume and in turn fertilize plants.

Soil is thus a self-regulating bio-physio-chemical factory, processing its own materials, water, and solar energy. It also determines the fate of rainfall and snowfall reaching the ground surface – whether the water thus received will flow over the land as runoff, or seep downward to the subterranean reservoir called groundwater, which in turn maintains the steady flow of springs and streams. With its finite capacity to absorb and store moisture, and to release it gradually, the soil regulates all of these phenomena. Without the soil as a buffer, rain falling over the continents would run off entirely, producing violent floods rather than sustained river flow.

Soil naturally acts as a living filter, in which pathogens and toxins that might otherwise accumulate to foul the terrestrial environment are rendered harmless. Since time immemorial, humans and other animals have been dying of all manner of disease and have then been buried in the soil, yet no major disease is transmitted by it. The term *antibiotic* was coined by soil microbiologists who, as a consequence of their studies of soil bacteria and actinomycetes, discovered streptomycin (an important cure for tuberculosis and other infections). Ion exchange, a useful process of water purification, also was discovered by soil scientists studying the passage of solutes through beds of clay.

However unique in form and function, soil is not an isolated body. It is, rather, a central link in the larger chain of interconnected domains and processes comprising the terrestrial environment. The soil interacts both with the overlying atmosphere and the underlying strata, as well as with surface and underground bodies of water. Especially important is the interrelation between the soil and the climate. In addition to its function of regulating the cycle of water, it also regulates energy exchange and surface temperature.

When virgin land is cleared of vegetation and turned into a cultivated field, the native biomass above the ground is often burned and the organic matter within the soil tends to decompose. These processes release carbon dioxide into the atmosphere, thus contributing to the earth's greenhouse effect and to global warming. On the other hand, the opposite act of reforestation and soil enrichment with organic matter, such as can be achieved by means of conservation management, may serve to absorb carbon dioxide from the atmosphere. To an extent, the soil's capacity to store carbon can thus help to mitigate the greenhouse effect.

Thousands of years are required for nature to create life-giving soil out of sterile bedrock. In only a few decades, however, unknowing or uncaring humans can destroy that wondrous work of nature. In various circumstances, mismanaged soils may be subject to erosion (the sediments of which tend to clog streambeds, estuaries, lakes, and coastal waters), to leaching of nutrients with attendant loss of fertility and eutrophication of water bodies, to waterlogging and impaired aeration, or to an excessive accumulation of salts that may cause a once-productive soil to become entirely sterile. Such processes of soil degradation, sometimes called 'desertification,' already affect large areas of land.

We cannot manage effectively and sustainably that which we do not know and thoroughly understand. That is why the tasks of developing and disseminating sound knowledge of the soil and its complex processes have assumed growing urgency and importance. The global environmental crisis has created a compelling need for a concentrated, concise, and definitive source of information – accessible to students, scientists, practitioners, and the general public – about the soil in all its manifestations – in nature and in relation to the life of humans.

Daniel Hillel
Editor-in-Chief
May 2004

INTRODUCTION

The *Encyclopedia of Soils in the Environment* contains nearly 300 articles, written by the world's leading authorities. Pedologists, biologists, ecologists, earth scientists, hydrologists, climatologists, geographers, and representatives from many other disciplines have contributed to this work. Each of the articles separately, and all of them in sequence and combination, serve to summarize and encapsulate our present knowledge of the world's variegated soils, their natural functions, and their importance to humans.

Concise articles surveying specific aspects of soils (soil genesis, soil chemistry and mineralogy, soil physics and hydrology, and soil biology) are complemented by articles covering transdisciplinary aspects, such as the role of soils in ecology, the history of soil utilization for agricultural and engineering purposes, the development of soil science as a discipline, and the potential or actual contributions of soils to the generation, as well as to the mitigation, of pollution and of global climate change.

This comprehensive reference encompasses both the fundamental and the applied aspects of soil science, interfacing in general with the physical sciences and life sciences and more specifically with the earth sciences and environmental sciences.

The *Encyclopedia of Soils in the Environment* manifests the expanding scope of modern soil science, from its early sectarian focus on the utilitarian attributes of soils in agriculture and engineering, to a wider and much more inclusive view of the soil as a central link in the continuous chain of processes constituting the dynamic environment as a whole. Thus it both details and integrates a set of topics that have always been of vital importance to human societies and that are certain to be even more so in the future.

Daniel Hillel
Editor-in-Chief
May 2004

CONTENTS

Contents are given as follows: CHAPTER NAME Author(s) Page number

VOLUME 1

A

ACID RAIN AND SOIL ACIDIFICATION L Blake 1

ACIDITY N Bolan, D Curtin and D Adriano 11

AERATION D E Rolston 17

AGGREGATION

Microbial Aspects S D Frey 22

Physical Aspects J R Nimmo 28

AGROFORESTRY P K R Nair 35

AIR PHASE see AERATION; DIFFUSION

ALBEDO see ENERGY BALANCE; RADIATION BALANCE

ALLOPHANE AND IMOGOLITE see AMORPHOUS MATERIALS

ALLUVIUM AND ALLUVIAL SOILS J L Boettinger 45

ALUMINUM SPECIATION D R Parker 50

AMMONIA D E Kissel and M L Cabrera 56

AMORPHOUS MATERIALS J Harsh 64

ANAEROBIC SOILS P W Inglett, K R Reddy and R Corstanje 72

ANION EXCHANGE see CATION EXCHANGE

APPLICATIONS OF SOILS DATA P J Lawrence 78

ARCHAEA J E T McLain 88

ARCHEOLOGY IN RELATION TO SOILS J A Homburg 95

B

BACTERIA

Plant Growth-Promoting Y Bashan and L E de-Bashan 103

Soil L J Halverson 115

BACTERIOPHAGE M Radosevich, K E Williamson and K E Wommack 122

BIOCONTROL OF SOIL-BORNE PLANT DISEASES C E Pankhurst and J M Lynch 129

BIODIVERSITY D H Wall 136

BUFFERING CAPACITY B R James 142

BULK DENSITY see POROSITY AND PORE-SIZE DISTRIBUTION

C

CALCIUM AND MAGNESIUM IN SOILS N Bolan, P Loganathan and S Saggar 149

CAPILLARITY D Or and M Tuller 155

CARBON CYCLE IN SOILS

Dynamics and Management C W Rice 164

Formation and Decomposition C A Cambardella 170

CARBON EMISSIONS AND SEQUESTRATION K Paustian 175

CATION EXCHANGE L M McDonald, V P Evangelou and M A Chappell 180

CHEMICAL EQUILIBRIA A P Schwab 189

CHEMICAL SPECIATION MODELS see SURFACE COMPLEXATION MODELING

CHERNOZEMS see GRASSLAND SOILS

CHILDS, ERNEST CARR E G Youngs 195

CIVILIZATION, ROLE OF SOILS D Hillel 199

CLASSIFICATION OF LAND USE see LAND-USE CLASSIFICATION

CLASSIFICATION OF SOILS R W Arnold 204

CLASSIFICATION SYSTEMS

Australian R W Fitzpatrick 211

FAO F O Nachtergaele 216

Russian, Background and Principles M Gerasimova 223

Russian, Evolution and Examples D Konyushkov 227

USA D J Brown 235

CLAY MINERALS D G Schulze 246

CLIMATE CHANGE IMPACTS P Bullock 254

CLIMATE MODELS, ROLE OF SOIL P Smith 262

COLD-REGION SOILS C-L Ping 268

COLLOID-FACILITATED SORPTION AND TRANSPORT R Kretzschmar 276

COMPACTION J J H van den Akker and B Soane 285

COMPOST T L Richard 294

CONDITIONERS R E Sojka, J A Entry and W J Orts 301

CONSERVATION see EROSION: Water-Induced; Wind-Induced; SUSTAINABLE SOIL AND LAND MANAGEMENT; TERRACES AND TERRACING

CONSERVATION TILLAGE M R Carter 306

COVER CROPS L Edwards and J Burney 311

CROP ROTATIONS C A Francis 318

CROP WATER REQUIREMENTS L S Pereira and I Alves 322

CROP-RESIDUE MANAGEMENT D C Reicosky and A R Wilts 334

CRUSTS

Biological J Belnap 339

Structural R L Baumhardt and R C Schwartz 347

CULTIVATION AND TILLAGE M R Carter and E McKyes 356

D

DARCY'S LAW D Swartzendruber 363

DEGRADATION C J Ritsema, G W J van Lynden, V G Jetten and S M de Jong 370

DENITRIFICATION D A Martens 378

DESERTIFICATION D Hillel and C Rosenzweig 382

DIFFUSION T Addiscott and P Leeds-Harrison 389

DISINFESTATION A Gamliel and J Katan 394

DISPERSION see FLOCCULATION AND DISPERSION

DISSOLUTION PROCESSES, KINETICS K G Scheckel and C A Impellitteri 400

DRAINAGE, SURFACE AND SUBSURFACE N R Fausey 409

DRYLAND FARMING G A Peterson 414

E

EARTHWORMS see FAUNA

EDAPHOLOGY A L Ulery 419

ELECTRON PARAMAGNETIC RESONANCE see ELECTRON-SPIN RESONANCE SPECTROSCOPY

ELECTRON-SPIN RESONANCE SPECTROSCOPY N Senesi and G S Senesi 426

ELECTROSTATIC DOUBLE-LAYER see CATION EXCHANGE

ENERGY BALANCE M Fuchs 438

ENVIRONMENTAL MONITORING P J Loveland and P H Bellamy 441

ENZYMES IN SOILS R P Dick and E Kandeler 448

EROSION

Irrigation-Induced G A Lehrsch, D L Bjorneberg and R E Sojka 456

Water-Induced J E Gilley 463

Wind-Induced T M Zobeck and R S Van Pelt 470

ESSENTIAL ELEMENTS E A Kirkby 478

EUTROPHICATION A J Gold and J T Sims 486

EVAPORATION OF WATER FROM BARE SOIL C W Boast and F W Simmons 494

EVAPOTRANSPIRATION G Stanhill 502

F

FACTORS OF SOIL FORMATION

Biota A H Jahren 507

Climate O C Spaargaren and J A Deckers 512

Human Impacts J Sandor, C L Burras and M Thompson 520

Parent Material K R Olson 532

Time E F Kelly and C M Yonker 536

FAUNA T Winsome 539

VOLUME 2

FERTIGATION U Kafkafi and S Kant 1

FERTILITY J L Havlin 10

FERTILIZERS AND FERTILIZATION H W Scherer 20

FIELD CAPACITY see WATER CYCLE

FLOCCULATION AND DISPERSION I Shainberg and G J Levy 27

FLUORESCENCE SPECTROSCOPY N Senesi and V D'Orazio 35

FOLIAR APPLICATIONS OF NUTRIENTS M Tagliavini and M Toselli 53

FOOD-WEB INTERACTIONS P C de Ruiter and J C Moore 59

FORENSIC APPLICATIONS W F Rowe 67

FOREST SOILS J R Boyle 73

FOURIER TRANSFORM INFRARED SPECTROSCOPY D Peak 80

FRACTAL ANALYSIS Y Pachepsky and J W Crawford 85

FREEZING AND THAWING

Cycles B Sharratt 98

Processes G N Flerchinger, G A Lehrsch and D K McCool 104

FUNGI K Ritz 110

G

GEOGRAPHICAL INFORMATION SYSTEMS J Bo¨hner, T Selige and R Ko¨the 121

GERMINATION AND SEEDLING ESTABLISHMENT A Hadas 130

GLOBAL WARMING see CARBON EMISSIONS AND SEQUESTRATION; CLIMATE CHANGE IMPACTS; GREENHOUSE GAS EMISSIONS

GRASSLAND SOILS J A Mason and C W Zanner 138

GREEN MANURING see COVER CROPS

GREENHOUSE GAS EMISSIONS K A Smith 145

GROUNDWATER AND AQUIFERS Y Bachmat 153

GROUNDWATER POLLUTION see POLLUTION: Groundwater

H

HEAT AND MOISTURE TRANSPORT R Horton and A Globus 169

HEAT CAPACITY see THERMAL PROPERTIES AND PROCESSES

HEAT FLOW see THERMAL PROPERTIES AND PROCESSES

HEAVY METALS D C Adriano, N S Bolan, J Vangronsveld and W W Wenzel 175

HILGARD, EUGENE WOLDEMAR R Amundson 182

HOOGHOUDT, SYMEN BAREND P A C Raats and R R van der Ploeg 188

HUMIFICATION T C Balser 195

HYDRAULIC PROPERTIES, TEMPERATURE EFFECTS S A Grant 207

HYDRIC SOILS G W Hurt 212

HYDROCARBONS P Kostecki, R Morrison and J Dragun 217

HYDRODYNAMIC DISPERSION see SOLUTE TRANSPORT

HYDRODYNAMICS IN SOILS T P A Ferre´ and A W Warrick 227

HYSTERESIS J H Dane and R J Lenhard 231

I

IMMISCIBLE FLUIDS R J Lenhard, J H Dane and M Oostrom 239

INCEPTISOLS A Palmer 248

INDUSTRIAL POLLUTION see POLLUTION: Industrial

INFILTRATION T P A Ferre´ and A W Warrick 254

INFRARED SPECTROSCOPY see FOURIER TRANSFORM INFRARED SPECTROSCOPY

IRON NUTRITION K Mengel and H Kosegarten 260

IRRIGATION

Environmental Effects S Topcu and C Kirda 267

Methods D L Bjorneberg and R E Sojka 273

ISOTOPES IN SOIL AND PLANT INVESTIGATIONS K Reichardt and O O S Bacchi 280

ISOTROPY AND ANISOTROPY T-C J Yeh, P Wierenga, R Khaleel and R J Glass 285

J

JENNY, HANS R Amundson 293

K

KELLOGG, CHARLES J D Helms 301

KINETIC MODELS P M Jardine 307

KIRKHAM, DON D R Nielsen and R R van der Ploeg 315

L

LAMINAR AND TURBULENT FLOW see HYDRODYNAMICS IN SOILS

LANDFILLS see WASTE DISPOSAL ON LAND: Municipal

LAND-USE CLASSIFICATION J A LaGro Jr 321

LAWES, JOHN BENNET AND GILBERT, JOSEPH HENRY A E Johnston 328

LEACHING PROCESSES B E Clothier and S Green 336

LIEBIG, JUSTUS VON R R van der Ploeg, W Bo¨hm and M B Kirkham 343

LIMING E J Kamprath and T J Smyth 350

LIPMAN, JACOB G. J C F Tedrow 358

LOESS A J Busacca and M R Sweeney 364

LOWDERMILK, WALTER CLAY J D Helms 373

LYSIMETRY T A Howell 379

M

MACRONUTRIENTS C W Wood, J F Adams and B H Wood 387

MACROPORES AND MACROPORE FLOW, KINEMATIC WAVE APPROACH P F Germann 393

MAGNESIUM IN SOILS see CALCIUM AND MAGNESIUM IN SOILS

MANURE MANAGEMENT J T Sims and R O Maguire 402

MARBUT, CURTIS FLETCHER J P Tandarich 410

MATRIC POTENTIAL see HYDRODYNAMICS IN SOILS; WATER POTENTIAL; WATER RETENTION AND CHARACTERISTIC CURVE

MEDITERRANEAN SOILS J Torrent 418

METAL OXIDES A C Scheinost 428

METALS AND METALLOIDS, TRANSFORMATION BY MICROORGANISMS S M Glasauer, T J Beveridge, E P Burford, F A Harper and G M Gadd 438

METALS, HEAVY see HEAVY METALS

MICROBIAL PROCESSES

Environmental Factors P G Hartel 448

CommunityAnaly sis C H Nakatsu 455

Kinetics N S Panikov 463

MICRONUTRIENTS L M Shuman 479

MINERAL–ORGANIC–MICROBIAL INTERACTIONS P M Huang, M C Wang and M K Wang 486

MINERALS, PRIMARY P M Huang and M K Wang 500

MINERALS, SECONDARY see CLAY MINERALS

MINIMUM TILLAGE see CONSERVATION TILLAGE

MISCIBLE DISPLACEMENT see SOLUTE TRANSPORT

MORPHOLOGY P R Owens and E M Rutledge 511

MULCHES C L Acharya, K M Hati and K K Bandyopadhyay 521

MYCORRHIZAL FUNGI L M Egerton-Warburton, J I Querejeta, M F Allen and S L Finkelman 533

VOLUME 3

N

NEMATODES D A Neher and T O Powers 1

NEUTRON SCATTERING M J Fayer and G W Gee 6

NITROGEN IN SOILS

- Cycle M S Coyne and W W Frye 13
- Nitrates D S Powlson and T M Addiscott 21
- Nitrification J I Prosser 31
- Plant Uptake A Hodge 39
- Symbiotic Fixation J I Sprent 46

NITROGEN FERTILIZERS see FERTILIZERS AND FERTILIZATION

NUCLEAR WASTE DISPOSAL G W Gee, P D Meyer and A L Ward 56

NUTRIENT AVAILABILITY N K Fageria and V C Baligar 63

NUTRIENT MANAGEMENT G D Binford 71

O

ORGANIC FARMING C A Francis 77

ORGANIC MATTER

- Principles and Processes M Schnitzer 85
- Genesis and Formation K M Haider and G Guggenberger 93
- Interactions with Metals N Senesi and E Loffredo 101

ORGANIC RESIDUES, DECOMPOSITION A J Franzluebbers 112

ORGANIC SOILS D L Mokma 118

OVERLAND FLOW T S Steenhuis, L Agnew, P Ge´rard-Marchant and M T Walter 130

OXIDATION–REDUCTION OF CONTAMINANTS C J Matocha 133

P

PADDY SOILS C Witt and S M Haefele 141

PARENT MATERIAL see PEDOLOGY: Basic Principles; FACTORS OF SOIL FORMATION: Parent Material

PEDOLOGY

- Basic Principles M J Singer 151
- Dynamic F C Ugolini 156

PEDOMETRICS I O A Odeh and A B McBratney 166

PENMAN, HOWARD LATIMER J L Monteith 176

PENMAN–MONTEITH EQUATION R Allen 180

PERCOLATION see HYDRODYNAMICS IN SOILS

PERMAFROST see POLAR SOILS

PERMEABILITY see HYDRODYNAMICS IN SOILS

PERSISTENT ORGANIC POLLUTANTS (POPS) see POLLUTANTS: Persistent Organic (POPs)

PESTICIDES R H Bromilow 188

PETROLEUM see HYDROCARBONS

pH N Bolan and K Kandaswamy 196

PHOSPHORUS IN SOILS

Overview J T Sims and P A Vadas 202

Biological Interactions M D Mullen 210

PHYTOTOXIC SUBSTANCES IN SOILS M Qadir, S Schubert and D Steffens 216

PLANT–SOIL–WATER RELATIONS R A Feddes and J C van Dam 222

PLANT–WATER RELATIONS C Gimenez, M Gallardo and R B Thompson 231

POISEUILLE'S LAW see HYDRODYNAMICS IN SOILS

POLAR SOILS J C F Tedrow 239

POLLUTANTS

Biodegradation P B Hatzinger and J W Kelsey 250

Effects on Microorganisms M E Fuller 258

Persistent Organic (POPs) D Johnson 264

POLLUTION

Groundwater H Rubin 271

Industrial S P McGrath 282

POLYMERS AND MICROORGANISMS M C Rillig 287

POORLY CRYSTALLINE ALLUMINOSILICATES see AMORPHOUS MATERIALS

POROSITY AND PORE-SIZE DISTRIBUTION J R Nimmo 295

POTASSIUM IN SOILS P M Huang, J M Zhou, J C Xie and M K Wang 303

PRECIPITATION, WATERSHED ANALYSIS J V Bonta 314

PRECIPITATION–DISSOLUTION PROCESSES W P Robarge 322

PRECISION AGRICULTURE see SITE-SPECIFIC SOIL MANAGEMENT

PREFERENTIAL FLOW see UNSTABLE FLOW; MACROPORES AND MACROPORE FLOW, KINEMATIC WAVE APPROACH

PRODUCTIVITY D L Karlen 330

PROFILE see MORPHOLOGY

PROTOZOA W Foissner 336

Q

QUALITY OF SOIL B J Wienhold, G E Varvel and J W Doran 349

R

RADIATION BALANCE J L Hatfield, T J Sauer and J H Prueger 355

RADIONUCLIDES see ISOTOPES IN SOIL AND PLANT INVESTIGATIONS

RAINFED FARMING see DRYLAND FARMING

RANGE MANAGEMENT G L Anderson 360

RECYCLING OF ORGANIC WASTES see POLLUTANTS: Biodegradation

REDISTRIBUTION see WATER CYCLE

REDOX POTENTIAL R D DeLaune and K R Reddy 366

REDOX REACTIONS, KINETICS P S Nico and S Fendorf 372

REMEDIATION OF POLLUTED SOILS E Lombi and R E Hamon 379

REMOTE SENSING

Organic Matter D K Morris, C J Johannsen, S M Brouder and G C Steinhardt 385

Soil Moisture T J Jackson 392

RHIZOSPHERE A C Kennedy and L Z de Luna 399

RICHARDS, LORENZO A. W R Gardner 407

ROOT ARCHITECTURE AND GROWTH L E Jackson 411

ROOT EXUDATES AND MICROORGANISMS B-J Koo, D C Adriano, N S Bolan and C D Barton 421

S

SALINATION PROCESSES I Shainberg and G J Levy 429

SALINITY

Management D Hillel 435

Physical Effects D Russo 442

SALT BALANCE OF SOILS see SALINATION PROCESSES

SALT-AFFECTED SOILS, RECLAMATION R Keren 454

SAND DUNES H Tsoar 462

SATURATED AND UNSATURATED FLOW see HYDRODYNAMICS IN SOILS; VADOSE ZONE: Hydrologic Processes

SCALING

Physical Properties and Processes G Sposito 472

Transport Processes R P Ewing 477

SEPTIC SYSTEMS R L Lavigne 485

SHIFTING CULTIVATION R Lal 488

SITE-SPECIFIC SOIL MANAGEMENT C J Johannsen and P G Carter 497

SLASH AND BURN AGRICULTURE see SHIFTING CULTIVATION

SLUDGE see WASTE DISPOSAL ON LAND: Liquid; Municipal

SODIC SOILS G J Levy and I Shainberg 504

SOIL–PLANT–ATMOSPHERE CONTINUUM J M Norman and M C Anderson 513

SOLUTE TRANSPORT M C Sukop and E Perfect 521

SORPTION

Metals D L Sparks 532

Organic Chemicals B Xing and J J Pignatello 537

Oxyanions C P Schulthess, H Wijnja and W Yang 548

SORPTION–DESORPTION, KINETICS D L Sparks 556

SPATIAL PATTERNS J H Go¨rres and J A Amador 562

VOLUME 4

SPATIAL VARIATION, SOIL PROPERTIES R Webster 1

SPECIFIC SURFACE AREA K D Pennell 13

STATISTICS IN SOIL SCIENCE R Webster 19

STERILIZATION see DISINFESTATION

STOCHASTIC ANALYSIS OF SOIL PROCESSES D Russo 29

STRESS–STRAIN AND SOIL STRENGTH S K Upadhyaya 38

STRUCTURE V A Snyder and M A Va´zquez 54

SUBSOILING R L Raper 69

SULFUR IN SOILS

- Overview M A Tabatabai 76
- Biological Transformations S D Siciliano and J J Germida 85
- Nutrition M A Tabatabai 91

SURFACE COMPLEXATION MODELING S Goldberg 97

SUSTAINABLE SOIL AND LAND MANAGEMENT J L Berc 108

SWELLING AND SHRINKING D Smiles and P A C Raats 115

T

TEMPERATE REGION SOILS E A Nater 125

TEMPERATURE REGIME see THERMAL PROPERTIES AND PROCESSES

TENSIOMETRY T K Tokunaga 131

TERMITES see FAUNA

TERRA ROSSA see MEDITERRANEAN SOILS

TERRACES AND TERRACING G R Foster 135

TESTING OF SOILS A P Mallarino 143

TEXTURE G W Gee 149

THERMAL PROPERTIES AND PROCESSES D Hillel 156

THERMODYNAMICS OF SOIL WATER P H Groenevelt 163

TILLAGE see CONSERVATION TILLAGE; CULTIVATION AND TILLAGE; ZONE TILLAGE

TILTH D L Karlen 168

TIME-DOMAIN REFLECTOMETRY G C Topp and T P A Ferre´ 174

TROPICAL SOILS

Arid and Semiarid H C Monger, J J Martinez-Rios and S A Khresat 182

Humid Tropical S W Buol 187

U

UNSTABLE FLOW T S Steenhuis, J-Y Parlange, Y-J Kim, D A DiCarlo, J S Selker, P A Nektarios, D A Barry and F Stagnitti 197

URBAN SOILS J L Morel, C Schwartz, L Florentin and C de Kimpe 202

V

VADOSE ZONE

Hydrologic Processes J W Hopmans and M Th van Genuchten 209

Microbial Ecology P A Holden and N Fierer 216

VIRUSES see BACTERIOPHAGE

VOLCANIC SOILS G Uehara 225

W

WAKSMAN, SELMAN A. H B Woodruff 233

WASTE DISPOSAL ON LAND

Liquid C P Gerba 238

Municipal D A C Manning 247

WATER AVAILABILITY see PLANT–SOIL–WATER RELATIONS

WATER CONTENT AND POTENTIAL, MEASUREMENT G S Campbell and C S Campbell 253

WATER CYCLE D K Cassel and B B Thapa 258

WATER EROSION see EROSION: Water-Induced

WATER HARVESTING D Hillel 264

WATER MANAGEMENT see CROP WATER REQUIREMENTS

WATER POTENTIAL D Or, M Tuller and J M Wraith 270

WATER REQUIREMENTS see CROP WATER REQUIREMENTS

WATER RETENTION AND CHARACTERISTIC CURVE M Tuller and D Or 278

WATER TABLE see GROUNDWATER AND AQUIFERS

WATER, PROPERTIES D Hillel 290

WATER-REPELLENT SOILS J Letey 301

WATERSHED MANAGEMENT M D Tomer 306

WATER-USE EFFICIENCY M B Kirkham 315

WEED MANAGEMENT D D Buhler 323

WETLANDS, NATURALLY OCCURRING E K Hartig 328

WIDTSOE, JOHN A. AND GARDNER, WILLARD G S Campbell and W H Gardner 335

WIND EROSION see EROSION: Wind-Induced

WINDBREAKS AND SHELTERBELTS E S Takle 340

WOMEN IN SOIL SCIENCE (USA) M J Levin 345

WORLD SOIL MAP H Eswaran and P F Reich 352

Z

ZERO-CHARGE POINTS J Chorover 367

ZONE TILLAGE J L Hatfield and A T Jeffries 373

Table of Contents, Volume 1

A **1**

Acid Rain and Soil Acidification **1**

Introduction 1

Acid Rain 1

Atmospheric Transport, Secondary Chemistry, and Acid Rain 2

Dry oxidation of sulfur and nitrogen oxides 2

Wet oxidation of sulfur 2

Acid Soil 3

Sources of Acids in Soils 4

Internal Acid Production in Soils 4

Carbonic and organic acid production 4

Nitrogen accumulation and transformations 5

Assimilation of Nutrients by Vegetation 5

Deposition of Nitrogen Species and Nitrogen Transformations 5

Addition of Dissolved Strong Acids Through Acid Deposition 5

Observations of Changes in Soil Acidification due to Acid Deposition 5

Proton Sinks and Buffer Ranges 6

Carbonate and Silicate Buffers 7

Cation Exchange Buffer 7

Aluminum and Iron Buffer 8

Proton Adsorption Reactions 9

Acid Sources and Sinks: Input-Output Relationships and the Calculation of Proton Balances 9

Acidification, Ecosystem Stability and Global Change 9

Further Reading 11

Acidity **11**

Introduction 11

Processes Generating Acidity in Soils 12

Acid Drainage 12

Acid Precipitation 12

Elemental Cycling 13

Fertilizer Reactions 14

Measurement and Effects of Soil Acidity 15

Amelioration of Soil Acidity 16

Further Reading 17

Aeration **17**

Introduction 17

Soil-Air Composition 17

Gas Exchange Mechanisms 18

Diffusion 18

Convection 18

Gas Reactions 19

Respiration 19

Oxidation-Reduction Processes 19

Production and Consumption of Other Gases 19

Aeration Requirements 20

Plants 20

Remediation of Contaminated Soils 20

Summary 20

Further Reading 21

Aggregation **22**

Microbial Aspects 22

Introduction 22

Microorganisms and Aggregate Formation 22

Soil Aggregates as Habitats for Microbiota 25
Soil Structure and Microorganism Interactions 26
Protection of Organic Matter Conferred by Soil Aggregates 26
Soil as a Spatially Continuous Medium 27
List of Technical Nomenclature 28
Further Reading 28
Physical Aspects 28
Forces on Soil Particles 29
Aggregate Physical Properties and their Measurement 31
Basic Properties 31
Size and Stability 32
Fundamentals 32
Commonly used methods 32
Miscellaneous methods 33
Issues of general importance 33
Representation and Interpretation 33
Summary 34
Further Reading 34
Agroforestry 35
Introduction 35
Emergence of Agroforestry as a Land-Use Approach 35
Agroforestry Systems 36
Examples of Common Agroforestry Systems 38
Improved Fallow 38
Alley Cropping 38
Homegarden, Shaded-Perennial, and Multistrata Systems 39
Silvopastoral Systems 40
Agroforestry Applications for Soil Conservation and Reclamation 41
Other Agroforestry Applications 41
Agroforestry and Ecosystem Services 42
Soil Productivity and Protection 42
Carbon Sequestration 42
Agroforestry Research 43
Future Directions 43
Acknowledgments 43
List of Technical Nomenclature 44
Further Reading 44
Alluvium and Alluvial Soils 45
Introduction 45
Alluvium 45
Composition 45
Texture 45
Landforms 46
Alluvial Soils 47
Recent Alluvial Soils 47
Older Alluvial Soils 47
Classification 48
Human Use of Alluvial Soils 48
List of Technical Nomenclature 49
Further Reading 49
Aluminum Speciation 50
Introduction 50
Inorganic Aluminum Species in Soil Solution 50
Organic Complexes of Aluminum 51
Determining Aluminum Speciation in Experimental Samples 53
Computational Methods 53

Fractionation Based on Size, Charge, and Reactivity 53
Speciation Using More Direct Analytical Methods 54
Further Reading 56
Ammonia 56
Introduction 56
Chemical Reactions of Ammoniacal N 57
Ammonia-Ammonium Equilibrium 57
Retention of Ammonia by Soils 57
Retention and Loss of Surface-Applied Ammoniacal N in Soils 59
Reactions of Urea in Soils 59
Environmental Conditions and Ammonia Formation 60
Urea concentration 60
Soil pH and H+ buffering capacity 61
Factors affecting the rate of urea hydrolysis 61
Soil urease concentration 61
Urea concentration 62
Soil temperature 62
Soil water content 62
Soil pH 62
Reactions of NH4+-N Applied to Calcareous Soils 62
List of Technical Nomenclature 63
Further Reading 63
Amorphous Materials 64
Introduction 64
Occurrence in Soils 64
Identification 65
Structure and Charge 67
Chemisorption and Ligand Exchange 68
Reaction with Soil Organic Matter 70
Aluminum Solubility 70
Physical Properties 71
Further Reading 71
Anaerobic Soils 72
Physical Characteristics 72
Biological Characteristics 73
Chemical Characteristics 73
Morphological Characteristics of Wetland Soils 74
Biogeochemical Characteristics 74
Microbial Respiration 75
Agronomic, Ecologic, and Environmental Significance 77
Further Reading 78
Applications of Soils Data 78
Introduction 78
Development of Interpretations in the USA 78
Interpretations Beyond Agriculture 80
Today's Interpretive Categories 81
National Inventory Groupings 81
Land-Use Planning 81
Farmland 81
Rangeland 81
Forest Land 81
Windbreaks 81
Recreation 81
Wildlife Habitat 83
Construction Materials 83
Building Sites 83

Waste Disposal 83
Water Management 83
Uses of Soils Data and Interpretations 83
Federal and State Programs 83
State and Local Planning 84
Natural Resource Management 84
Land Appraisal and Assessment 84
Engineering and Construction 86
Hazardous Waste, Brownfields, and Remediation 86
Research and Analyses 86
Conclusion 88
Further Reading 88
Archaea 88
Introduction 88
Archaeal Evolution 89
The Archaeal Cell 89
Three Archaeal Groups 90
Extreme Halophiles 90
Methanogens 91
Hyperthermophiles 92
Psychrophiles 93
Thermoplasma 93
Biotechnological Use of Extremoenzymes 93
Further Reading 94
Archeology in Relation to Soils 95
Introduction 95
Pedostratigraphy and the Archeological Record 95
Archeological Soil Chemistry 97
Soil as a Resource 98
Further Reading 101
B 103
Bacteria 103
Plant Growth-Promoting 103
Introduction 103
Plant Growth-Promoting Bacteria in Agriculture and the Environment 103
Endophytic PGPB 108
Biocontrol of Phytopathogens 108
Mechanisms Employed by Biocontrol-PGPB to Control Phytopathogens 109
Production of Antibiotics 109
Production of Siderophores 110
Production of Small Molecules 110
Production of Enzymes 110
Competition and Displacement of Pathogens 111
Modification of Plant Metabolism 112
Induced and Acquired Systemic Resistance 112
Modification of Plant Ethylene Levels 112
Prospects for Improving PGPB by Genetic Manipulation 112
PGPB Inoculants 112
Conclusions and Prospects 113
Further Reading 114
Soil 115
Introduction 115
Bacterial Diversity 115
The Microbe as the Scale of Study 116
Two Basic Survival Strategies 117

Avoidance 117
Stress Tolerance 118
Water Availability 118
Extracellular Polysaccharide Production 118
Biofilms 119
Competition for Resources 119
Cooperation Among Bacteria 120
Siderophores 120
Quorum-Sensing 120
Death 121
Summary 121
Further Reading 122
Bacteriophage 122
Introduction 122
Why Soil Bacteriophage? 122
Ecologic Role of Viruses in Microbial Communities 123
Bacteriophage Contain Genetic Elements that Can Increase Host Fitness: Bacteriophage Conversion 123
The Role of Phage in Horizontal Gene Transfer and Bacterial Evolution 123
Evidence for Natural Transduction in Microbial Communities 124
Cultivation-Independent Investigation of Soil Viral Communities 125
Microscopy of Soil Viruses 125
Genomic Approaches to Viral Diversity 127
Summary 128
List of Technical Nomenclature 128
Further Reading 128
Biocontrol of Soil-Borne Plant Diseases 129
Introduction 129
Enhancement of Biocontrol by Agronomic Practices 130
Crop Rotations 130
Organic Amendments 130
Biocontrol with Introduced Microbial Biocontrol Agents 131
Selection, Evaluation, and Delivery of Microbial Biocontrol Agents 131
Mode of Action of Biocontrol Agents 133
Competition 133
Antibiosis 133
Parasitism 133
Induced systemic resistance 134
Plant growth-promoting rhizobacteria 134
Future Prospects 134
Conclusions 135
Further Reading 135
Biodiversity 136
Introduction 136
Diversity and Abundance 136
Distribution 138
Soil Biodiversity and Ecosystem Functioning 139
Global Change and Impacts on Soil Biodiversity 140
Sustaining Soil Biodiversity 141
Further Reading 141
Buffering Capacity 142
Introduction 142
The Samovar Analogy for Buffering in Soils 142
Buffering of Soil Acidity 143
Dissolution and Precipitation of CaCO3 and Al(OH)3 144

Cation Exchange Reactions and pH Buffering....144
pH-Dependent Surface Charge and pH Buffering....145
Buffering of Soil Oxidation-Reduction Status....146
Buffering of Ion Activities via Dissolution-Precipitation, Ion Exchange, and Ligand Complexation....146
Further Reading....147
C....149
Calcium and Magnesium in Soils....149
Introduction....149
Input to Soils....149
Reactions in Soils....151
Modeling the Dissolution Reactions of Calcium and Magnesium Compounds in Soils....151
Plant Requirements and Deficiency Symptoms....152
Plant Uptake....153
Interactions with Other Nutrients....153
Animal Requirements....153
Further Reading....154
Capillarity....155
Introduction....155
Liquid Properties....155
Surface Tension....155
Contact Angle....155
Curved Surfaces and Capillarity....156
The Capillary Rise Model....157
Capillarity in Soils....158
Capillarity in Angular Pores....158
Dynamic Aspects of Capillarity....159
Dynamics of Capillary Rise....159
Dynamic Contact Angle....160
Heterogeneous Surfaces and Microscale Hysteresis....161
Contact Angle on Chemically Heterogeneous and Rough Surfaces....161
Hysteresis....163
Further Reading....163
Carbon Cycle in SoilsDynamics and Management....164
Dynamics and Management....164
Introduction....164
Fluxes....166
Soil-Forming Factors....166
Functions....166
Regulation of Soil Organic Carbon Dynamics....167
Composition of Soil Organic Carbon....168
Management....168
Further Reading....170
Formation and Decomposition....170
Introduction....170
Sources of Carbon....170
Aerobic Decomposition....171
Anaerobic Decomposition....172
Factors Controlling Decomposition....173
Soil Organic Matter Formation and Turnover....174
Further Reading....175
Carbon Emissions and Sequestration....175
Introduction....175
Soils and Carbon-Cycling....175
Controls on Emissions and Sequestration....176

Land Use and Land-Use Change 177
Management Practices for Soil C Sequestration 178
Regional and Country Estimates of C-Sequestration Potential 179
List of Technical Nomenclature 179
Further Reading 180
Cation Exchange 180
Introduction 180
Cation Exchange Capacity 181
Qualitative Description of Cation Exchange 182
Properties of Ions 182
Properties of Surfaces 183
Properties of Solvents 183
Quantitative Description of Cation Exchange 184
Homovalent Binary Exchange 184
Heterovalent Binary Exchange 185
The Future of Cation Exchange 187
List of Technical Nomenclature 188
Further Reading 188
Chemical Equilibria 189
Introduction 189
Chemical Equilibrium 189
Kinetic Definition 189
Free-Energy Definition 189
Applicability of Chemical Equilibrium Principles to Soils 190
Limitations to Equilibrium in Soils 190
Soil Systems in Which Equilibrium Concepts May Apply 191
Reasonable Expectations from Application of Equilibrium 191
Predict endpoints of reactions 191
Determine ionic activities in soil solution 191
Calculating final conditions for systems that attain equilibrium rapidly 191
Experimental Approach to Equilibrium in Soils 191
System Variables 191
Reactions and Associated Equilibrium Constants 191
Equilibrium Modeling of Soil Systems 192
Geochemical models 192
Model inputs 192
Application of Chemical Equilibrium to Soils: Examples 192
Iron Oxides in Reduced Systems 192
Manganese(III) Phosphate 193
Calcite in Soils: Equilibrium versus Nonequilibrium 193
The Outlook for Studying Chemical Equilibrium in Soils 193
List of Technical Nomenclature 194
Further Reading 194
Childs, Ernest Carr 195
Further Reading 198
Civilization, Role of Soils 199
Human Management of the Soil 199
Historical Attitudes Toward the Soil 199
Human Origins 200
The Paleolithic Transformation 201
The Agricultural Transformation 201
Soil Husbandry and Ceramics 202
From Rainfed to Irrigated Farming 203
Silt and Salt in Ancient Mesopotamia 203
The Sustainability of Egyptian Agriculture 203
Further Reading 204

Classification of Soils **204**
Principle of Purpose 205
Principle of Domain 205
Principle of Identity 206
Principle of Differentiation 206
Principle of Prioritization 207
Principle of Diagnostics 209
Principle of Membership 209
Principle of Certainty 210
Summary 210
List of Technical Nomenclature 210
Further Reading 210
Classification Systems **211**
Australian 211
Introduction 211
History of General-Purpose National Soil Classifications 211
The New National Classification 212
Soil Classifications Used on Australian State and Regional Scales 213
The Basis for State and Regional Soil-Classification Systems 213
State and Regional Examples 213
Special-Purpose Soil-Classification Systems 214
Viticultural Soils 214
Engineering Applications: Laying of Telecommunication Optical Fiber Cables 214
Future Directions 216
Further Reading 216
FAO 216
Introduction 216
The Development of the FAO Legend 1960-—1981 216
Set 1 218
Set 2 218
Set 3 218
Set 4 218
Set 5 219
Set 6 219
Set 7 219
Set 8 219
The FAO Revised Legend 1981-1990 219
The World Reference Base for Soil Resources 1981-1998 220
Summary 222
Further Reading 222
Russian, Background and Principles 223
Introduction 223
Principles, Categories, and Classes in Major Classification Systems Developed in Russia 223
Early Systems 223
NonOfficial Systems in the Second Half of the Twentieth Century 224
Official Systems 224
Classification and diagnostics of soils of the USSR - 1977 225
Russian soil classification system - 1997 225
Conclusion 227
Further Reading 227
Russian, Evolution and Examples 227
Introduction 227
Pre-Dokuchaev Period 228
Agroproductive Soil Groups, Folk Soil Nomenclature, and First Scientific

Soil Classifications....228
Dokuchaev's Period....229
Genetic Approach to Soil Classification....229
Post-Dokuchaev Period....230
Diversification of Classification Decisions....230
Official Soil Classification and Unofficial (Authors') Systems....232
The Problem of Basic Substantive-Genetic-Soil Classification and the new Russian Classification....232
Further Reading....235
USA....235
Introduction....235
Soil Classification Criteria....236
Soil Composition....236
Genesis and Soil Taxonomy....237
Climate Zones in Soil Taxonomy....238
Structure and Nomenclature....240
How it Works....240
Example Profile....242
Closing Thoughts....245
List of Technical Nomenclature....245
Further Reading....245
Clay Minerals....246
Introduction....246
Major Element Composition of Clay Minerals....246
Basic Structural Concepts....246
Tetrahedra and Octahedra....246
Representing Crystal Structures....247
Tetrahedral and Octahedral Sheets....247
Tetrahedral Sheet....247
Octahedral Sheet....247
Phyllosilicate Minerals Common In Soil Clays....248
1:1-Type Minerals....248
1:1 Layer structure....248
Kaolinite....248
Halloysite....248
2:1-Type Minerals....249
2:1 Layer structure....249
Pyrophyllite....249
Micas....249
Vermiculites....251
Smectites....251
Chlorites....251
Hydroxy-interlayered vermiculite and smectite....251
Interstratification in phyllosilicates....252
Palygorskite and sepiolite....252
Other Minerals that Occur in Soil Clays....252
Zeolites....253
Allophane and Imogolite....253
Aluminum Hydroxide Minerals....253
Iron Oxide Minerals....253
Manganese Oxide Minerals....253
Titanium Oxide Minerals....253
Carbonates, Sulfates, and Soluble Salts....253
List of Technical Nomenclature....254
Further Reading....254
Climate Change Impacts....254

Introduction 254
Climate Change Predictions 254
Timescale for Change....255
Climate Change Impacts on Soil Water and Soil Temperature 255
Soil Water....255
Soil Temperature....257
Changes in Soil-Forming Processes and Properties....257
Soil Organic Matter....257
Soil Structure....258
Soil Fauna and Soil Flora....258
Acidification and Nutrient Status 258
Changes in Land-Degradation Processes....259
Soil Erosion....259
Desertification 259
Salinization....260
Climate Change Impacts on Soil Functions....260
Further Reading....262
Climate Models, Role of Soil 262
Introduction....262
Approaches to Modeling SOC Dynamics for the Study of Climate Change ...262
Process-Based, Multicompartment Models of SOC Dynamics 263
Cohort Models Describing Decomposition as a Continuum....263
Food-Web Models....263
Factors Affecting the Turnover of Soil Organic Carbon in Models....263
Model Performance 264
Models Used to Study the Impacts of Climate Change 264
Climate Models with Coupled Description of Biospheric Carbon Feedbacks .264
List of Technical Nomenclature 267
Further Reading....267
Cold-Region Soils 268
Introduction....268
Definition of Cold Soils....268
Characteristics and Genesis of Cold-Region Soils....269
Physical Properties....269
Boreal soils 269
Arctic soils....269
Alpine and high arctic soils 269
Morphological Properties....270
Biological and Chemical Properties....271
Alpine and Plateau Regions....272
Vegetation and Fire 273
Land Use of Cold-Region Soils....273
Engineering....274
Agriculture Development....274
Further Reading....275
Colloid-Facilitated Sorption and Transport 276
Introduction....276
Nature and Stability of Soil Colloids....276
Colloid Transport in Porous Media 279
Factors Influencing Colloid Release 280
Factors Influencing Colloid Deposition 281
Colloid-Facilitated Transport of Contaminants....283
Further Reading....284
Compaction 285
Introduction....285
Factors and Processes Affecting Distribution and Intensity of Compaction ...285

Compaction under Running Gear....285
Loading Characteristics of Individual Wheels or Caterpillar Tracks....285
Field Traffic Intensity and Distribution....286
Soil Compactibility....287
Effects on Soil Physical and Mechanical Properties....287
Bulk Density, Porosity, and Packing State....287
Hydraulic Properties....287
Aeration Characteristics....288
Strength Characteristics....288
Compaction in Crop Production....288
Effects on Germination and Establishment....289
Effects on Root Growth and Distribution....289
Effects on Plant Growth and Yield....289
Interactive Crop Responses to Compaction and Fertilizer Application....289
Crop Responses to Subsoil Compaction....289
Modeling of Crop Responses to Soil Compaction....290
Effects on Environmental Components....290
Techniques for the Reduction of Compaction....292
Amelioration of Compacted Soils....293
Further Reading....293
Compost....294
Introduction....294
Compost Production....294
Compost Application and Incorporation....295
Compost-Soil Interactions....296
Physical Effects....296
Chemical Effects....297
Biological Effects....299
List of Technical Nomenclature....300
Further Reading....300
Conditioners....301
Introduction....301
Early Use of Mineral and Organic Materials....301
Use of Waste Materials as Conditioners....302
Advent of Synthetic Conditioners....302
Hydrogels and Super Water-Absorbent Polymers....303
Recent Advances Using Polyacrylamide....304
Further Reading....305
Conservation Tillage....306
Introduction....306
Evolution of Conservation Tillage Systems....306
Defining Conservation Tillage....307
Variants of Conservation Tillage....307
Common Types of Conservation Tillage Practices....308
Constraints to the Adoption of Conservation Tillage Practices....308
Climate and Soil....308
High Levels of Crop Residue....308
Mixed Cropping Systems....309
Beneficial Influence of Conservation Tillage....309
Improving Soil Quality and Health....309
Enhancing Soil Organic Matter Storage....309
Reducing Soil Erosion and Environmental Risk....310
Tillage Costs....310
Further Reading....310
Cover Crops....311
Introduction....311

Agronomy 312
Species Selection....312
Seeding and Husbandry....313
Role in Soil Quality Improvement 314
Role in Soil Erosion Control....315
Nature of Soil Erosion 315
Runoff and Erosion Control....316
Perspective....316
Further Reading....317
Crop Rotations 318
Introduction....318
Soil Fertility and Crop Yields....319
Cultural Management 320
Pest Management with Rotations....320
Economic Diversity 321
Future Perspectives....321
Further Reading....322
Crop Water Requirements....322
Introduction....322
Crop Evapotranspiration....323
Crop Coefficients....324
The Single-Crop Coefficient Approach....325
Kc Adjustment for Climate....330
Kc Adjustment for Nonpristine Conditions....330
The Dual-Crop Coefficient Approach 330
Kc for Nonpristine and Unknown Conditions....332
Irrigation Water Requirements....332
Summary 332
List of Technical Nomenclature 333
Further Reading....334
Crop-Residue Management....334
Introduction....334
Conservation, Carbon Cycle, Soil Organic Matter and Carbon Sinks 334
Crop Residues and Nutrient Cycling....335
Factors Controlling Residue Decomposition and Soil Quality 336
Crop Residues: Social and Environmental Benefits 336
Crop Residues, Research and Global Change....337
Summary 337
Further reading 338
Crusts....339
Biological 339
Introduction 339
Species Composition and Growth Forms....339
Ecological Roles....341
Carbon Fixation 341
Nitrogen Fixation....341
Dust Trapping 343
Effects on Vascular Plants....343
Nutrient Levels in Vascular Plants....343
Soil Hydrology and Stabilization....343
Effects of Disturbance 344
Species Composition....344
Water Erosion....344
Wind Erosion 344
Nutrient Cycles 345
Albedo....345

Recovery from Disturbance....345
Natural Recovery Rates....345
Assisted Recovery....345
Conclusion....345
Further Reading....347
Structural....347
Introduction....347
Crust Types and Morphology....348
Factors Governing Crust Formation....349
Rainstorm Characteristics....349
Soil Properties....351
Hydrologic Impact of Crusts....351
Infiltration - Developed Crusts....352
Infiltration - Developing Crusts....352
Numerical Infiltration Models for Developing Crusts....353
Agronomic Importance of Crusts....354
Further Reading....355
Cultivation and Tillage....356
Introduction....356
Goals of Soil Tillage....356
Main Types of Tillage Tools....357
Moldboard Plow....357
Applications....357
Disadvantages and limitations....357
Chisel Plow....358
Applications....358
Disk Harrow....358
Toothed Cultivator....359
Subsoiler....359
No-tillage....359
Basic Tillage Operations....360
Types of Tillage Systems....360
List of Technical Nomenclature....361
Further Reading....361
D....363
Darcy's Law....363
Introduction....363
Water-Saturated Soil Conditions....363
Validity of Darcy's Equation....364
Solving Saturated-Flow Problems....366
Water in Unsaturated Soils....366
Solving Unsaturated-Flow Problems....368
List of Technical Nomenclature....368
Further Reading....369
Degradation....370
Introduction....370
Factors and Processes Affecting Degradation of Soils....370
Types of Soil Degradation....371
Assessment of Degradation....371
Approaches....371
Potentials and Limitations....373
Degree and Impact of Degradation....374
Degree of Degradation....374
Rate of Soil Degradation....375
Impact of Degradation....375
Preventing and Combating Degradation....375

Further Reading....377
Denitrification....378
Introduction....378
Definitions and Pathways....378
Organisms and Substrates....378
Denitrification....378
Nitrifier Denitrification....379
Environmental Factors....380
Organic C Availability....380
Controls on O2 Availability....380
Nitrate Supply....381
Temperature....381
pH....381
Conclusion....381
Further Reading....382
Desertification....382
Introduction....382
The Role of Drought....383
Primary Production and Carrying Capacity....384
Soil Degradation and Rehabilitation....385
Potentialities and Problems of Irrigation....385
Social Factors....386
Monitoring Desertification....386
The Role of Climate Change....387
Overview....387
Further Reading....388
Diffusion....389
Introduction....389
How Diffusion Occurs....389
Fick's Laws....390
Initial and Boundary Conditions....390
Components of the Diffusion Coefficient in Soil....390
Diffusion and Geometry....391
Some Useful Analytical Solutions for Various Geometries....391
Diffusion from a Semiinfinite Medium....391
Diffusion in a Cylinder....392
Diffusion in a Sphere....392
Simple Model for Various Geometries....392
Applications in Soil Science....393
Plant Nutrition....393
Kinetics of Potassium Release From Clays....393
Gaseous Diffusion....393
Leaching....393
List of Technical Nomenclature....393
Further Reading....394
Disinfestation....394
Introduction....394
Principles of Soil Disinfestation....395
Methods of Soil Disinfestation....395
Physical....395
Chemical....396
Methyl bromide....396
Metham sodium....396
Dazomet (3,5-dimethyltetrahydro-1,3,5,(2H)-thiodiazinothione)....396
Other fumigants and mixtures....396
Application....396

Soil Solarization 397
Principles of soil solarization....397
Soil heating: simulation models for the prediction of soil temperatures....398
Combining Disinfestation Methods 398
Benefits and Limitations....399
The MBr Crisis and Its Implications 399
Further Reading....399
Dissolution Processes, Kinetics....400
Introduction....400
Rates and Limits of Element-Cycling....401
Kinetics of Proton- and Ligand-Promoted Mineral Dissolution 401
Metal-Promoted Mineral Dissolution....403
Reductive Dissolution of Minerals....404
Dissolution of Metal Surface Precipitates 404
The Good and Bad of Dissolution Kinetics 407
List of Technical Nomenclature 408
Further Reading....409
Drainage, Surface and Subsurface....409
Introduction....409
Hydrology and Drainage....410
Accelerated Drainage 410
Surface Drainage Principles and Practices....410
Subsurface Drainage Principles and Practices 411
Recent Innovations in Drainage Practice....412
Further Reading....413
Dryland Farming....414
Introduction....414
Definition....414
World Scope of Dryland Farming....414
Major Dryland Farming Issues....415
Dryland Farming Management Techniques....415
Fallowing 415
Water Capture Management....416
Water Retention Management 416
Efficient Water Use by Crop Plants....417
Conclusion....417
Further Reading....417
E 419
Edaphology....419
Introduction....419
Edaphic Properties 419
Edaphic Factors Affecting Water and Air 420
Edaphic Factors Affecting Plant Nutrients....421
Edaphic Factors Affecting Root Growth 423
Research 424
List of Technical Nomenclature 424
Further Reading....425
Electron-Spin Resonance Spectroscopy....426
Introduction....426
Basic Principles 426
The Resonance Phenomenon 426
Spectral Parameters....426
Instrumentation and Methodology 427
The Instrument and the Experiment....427
Sample Preparation 428
Sensitivity and Resolution 429

Determination and Interpretation of Spectral Parameters....429
Whole Soil....429
Organic Free Radicals in Soil Humic Substances....430
ESR Spectra and Parameters....430
Interpretation of ESR Parameters....431
Factors Affecting ESR Parameters of Humic Substances....431
Structural Implications....431
Interactions of Humic Free Radicals with Organic Chemicals and Metal Ions....432
Paramagnetic Transition Metal Ions in Soil Constituents....432
ESR Spectra and ESR Parameters....432
Ferric iron....433
Divalent copper....434
Vanadyl ion....435
Divalent manganese....435
Spin-Derivatization Studies....435
Spin-Labeling and Spin-Trapping....435
Metal Spin Probes....435
Complexation chemistry of humic substances....435
Ion exchange on layer silicates....436
Chemisorption on mineral surfaces....436
Thermodynamic constants from ESR parameters....436
List of Technical Nomenclature....436
Further Reading....437
Energy Balance....438
Introduction....438
The Energy Input From the Sun....438
Energy Balance of a Planet Without Atmosphere....438
Energy Balance of the Earth's Surface....438
Anthropogenic Change of the Energy Balance....441
Further Reading....441
Environmental Monitoring....441
Introduction....441
Why Monitor Soils?....442
The Starting Point....443
Soil-Monitoring Networks....444
Protocols....444
Monitoring Sites....444
Soil Sampling....446
Quality Control....447
Summary....447
Further Reading....447
Enzymes in Soils....448
Introduction....448
Spatial Distribution of Enzymes in Soils....449
Extracellular Enzyme Stabilization in Soil Matrix....449
Microscale Distribution....449
Macroscale Distribution....450
Methods of Studying Enzyme Activities in Soils....450
Enzyme Kinetics....451
Soil-Quality Indicators and Technologies....452
Sensitivity to Soil Management and Ecosystem Stress....452
Ecological Dose Value....454
Detoxification of Polluted Soils....454
Summary....455
Further Reading....455

Erosion **456**
Irrigation-Induced 456
Introduction 456
Significance 457
Erosion Under Sprinkler Irrigation 458
Processes Causing Soil Loss 458
Practices Controlling Soil Loss 459
Irrigation practices 459
Soil and crop management practices 459
Erosion with Surface Irrigation 460
Processes Causing Soil Loss 460
Practices Controlling Soil Loss 460
Irrigation practices 460
Runoff management practices 461
Soil and crop-management practices 462
Summary 462
Prospects for Future Control 463
Further Reading 463
Water-Induced 463
Introduction 463
Impacts of Erosion 464
Types of Erosion 464
Rill Erosion 464
Interrill Erosion 464
Gully Erosion 464
Ephemeral Gully Erosion 465
Stream Channel Erosion 465
Factors Affecting Erosion 465
Rainfall Characteristics 465
Soil Factors 465
Topography 465
Climate 466
Land Use 466
Cropland areas 466
Rangelands and pastures 466
Forest areas 466
Construction sites 467
Erosion Control Measures 467
Erosion Control Measures on Agricultural Areas 467
Contouring 467
Strip cropping 467
Conservation tillage 467
Terraces 467
Buffer strips 468
Grassed waterways 468
Polyacrylamide 469
Erosion Control Measures on Forest Areas 469
Erosion Control Measures on Construction Sites 469
Erosion Control Measures in Stream Channels 469
Further Reading 469
Wind-Induced 470
Introduction 470
Wind Profile 470
Modes of Transport 470
Soil Surface Conditions 471
Vegetation 473

Wind Erosion Modeling 473
Wind Erosion Sampling 474
Use of Anthropogenic Radioisotope Tracers....475
Effects of Wind Erosion 475
Wind Erosion Control 476
List of Technical Nomenclature 477
Further Reading 477
Essential Elements 478
Introduction 478
Classification of Essential Elements 478
Nutrient Deficiency Symptoms....478
Nitrogen 479
Sulfur 480
Phosphorus....480
Potassium 481
Calcium....481
Magnesium 482
Iron....482
Manganese 483
Zinc 483
Copper 483
Molybdenum 484
Boron 484
Chlorine 485
Nickel 485
Further Reading....485
Eutrophication 486
Introduction 486
Aquatic Eutrophication....486
Causes 486
Limiting Nutrients....486
Consequences of Aquatic Eutrophication....489
Stratification and Oxygen Depletion 489
Changes in Species Composition 489
Sea-Grass Destruction 490
Degradation of Corals 491
Modeling Eutrophication 491
Phosphorus-Limited Freshwater Systems....491
Nitrogen-Limited Coastal Marine Systems 491
Terrestrial Eutrophication....491
Watershed Export and Models 492
Controlling Eutrophication: Future Trends 493
List of Technical Nomenclature 494
Further Reading....494
Evaporation of Water from Bare Soil 494
Introduction 494
Evaporation Rate and Cumulative Evaporation....494
Processes Involved in Evaporation, and Latent Heat of Evaporation....495
Net Radiation 495
Energy Balance 495
Stages of Evaporation 496
Evaporation under Conditions of Unlimited Potential Evaporation 497
Evaporation Below the Soil Surface 497
Methods for Measuring Evaporation....498
Weigh the Loss of Water Mass in the Soil....498
Detect the Arrival of Water Vapor Within the Air 499

Quantify the Upward Transport of Water Vapor in the Air....499
Measure Enough Elements in the Energy Balance to Calculate LeE....499
Compare the Soil of Interest to a Completely Dry Soil....500
List of Technical Nomenclature....501
Further Reading....502
Evapotranspiration....502
Introduction....502
Development of the Concept....502
Methods of Measurement....503
Magnitude of Evapotranspiration....503
Control of Evapotranspiration....504
Climate Change and Evapotranspiration....505
Further Reading....506
F....507
Factors of Soil Formation....507
Biota....507
Introduction....507
The Biosequence....507
The Biotic Factor....508
Vegetation as a Soil-Forming Factor....508
Grassland-Forest Biosequence....510
Conifer Forest-Deciduous Forest Biosequence....511
List of Technical Nomenclature....512
Further Reading....512
Climate....512
Introduction....512
Soils in Tropical Climates....514
Soils in Dry Climates....515
Soils in Warm Temperate Climates....516
Soils in Cool Temperate Climates....517
Soils in the Polar Climates....518
Soils as Climatological Archive....518
Summary....519
Further Reading....519
Human Impacts....520
Introduction....520
Scales and Scope of Human Impacts on Soil Formation....520
Evaluating Human Impacts on Soil Formation....522
Changes in Soil Properties and Processes Resulting From Human Activities....523
Agriculture....523
Cities and Industry....525
Mining....527
War....529
Climate Change and Soil Formation....529
Significance and Future of Human Impacts on Soil Formation....530
List of Technical Nomenclature....531
Further Reading....531
Parent Material....532
Introduction....532
Defining Soil Parent Material and Role in Soil Genesis....532
Soil-Forming Factors....532
Weathering....533
Effects of Parent Material on Soil Properties....534
Influence of Parent Materials in Soil Genesis....534
Parent Material and Soil Classification....535

Summary 535
List of Technical Nomenclature 535
Further Reading 535
Time 536
Introduction 536
Key Soil Properties that Vary with Time 536
Studies of Time and Soil Formation 538
List of Technical Nomenclature 538
Further Reading 539
Fauna 539
Introduction 539
Characterization of Soil Macrofauna 539
Saprovores and Microbivores 539
Herbivores and Granivores 541
Predators and Omnivores 541
The Role of Soil Macrofauna in Ecosystem Function 542
Decomposition and Nutrient Cycling 542
Soil Modification 544
Participation in Belowground and Aboveground Food Webs 548
Further Reading 548

A

ACID RAIN AND SOIL ACIDIFICATION

L Blake, Rothamsted Research, Harpenden, UK

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

During the last three decades of the twentieth century, several studies in northern, central, and western Europe and in North America demonstrated, almost invariably, increasing acidity in soils beneath natural and seminatural ecosystems. Several ecosystem and soil processes may decrease soil pH, particularly changes in land use or soil management; thus the interpretation of acidifying sources can be difficult and complex. However, long-term studies of changes in deposition and soil chemistry have shown that inputs of atmospherically derived acidity cause soils to acidify rapidly, depending on the intensity of the acid input and the acid-buffering capacity of the soil. This acidification process is characterized by specific changes in soil and soil-solution chemistry.

Acid Rain

Rain is naturally acidic, usually in the pH range 5.0–5.5, mainly because of dissolved atmospheric carbon dioxide (CO_2) forming carbonic acid (H_2CO_3):

$$CO_2 \rightarrow CO_2 \cdot H_2O \rightarrow HCO_3^- + H^+ \qquad [1]$$

This 2.5×10^{-5} H^+ represents only a minute quantity of acid and is probably beneficial, even essential, to such processes as the weathering of soil from rocks and the release of insoluble nutrients from soil minerals. Only when the concentration of protons is increased by factors of 10–100 over 'natural' pH values by the presence of stronger acids do the detrimental effects of 'acid rain' occur.

The precursors of stronger acids in the atmosphere are pollutants attributable to (1) the combustion of fossil fuels and (2) the smelting of sulfide ores. However, these sources can be significantly augmented by ammonia, resulting from the management practices of intensive agriculture, particularly from livestock wastes.

There are essentially two kinds of pollutants of concern, gaseous and particulate, each of which can be subdivided by their modes of formation: primary pollutants are produced directly from industrial and domestic activities, and secondary pollutants are created in the atmosphere by chemical processes acting on primary pollutants. These distinctions enable four groups of pollutant to be identified:

1. Gaseous, primary:
   Sulfur dioxide (SO_2);
   Nitric oxide (NO) and nitrogen dioxide (NO_2); (collectively designated NO_x);
   Nitrous oxide (N_2O);
   Ammonia (NH_3);
   Carbon dioxide (CO_2);
   Hydrocarbons;
2. Gaseous, secondary:
   NO_2 from oxidation of NO;
   Ozone (O_3) and other photochemical oxidants formed in the lower atmosphere by the action of sunlight on mixtures of NO_x and hydrocarbons;
   Nitric acid (HNO_3) formed from the oxidation of NO_x;
3. Particulate, primary: fuel ash and metallic particles, from smelting and heavy industry;
4. Particulate, secondary: the reaction products of sulfuric acid and nitric acid with other atmospheric constituents, notably NH_3 (NH_4HSO_4, NH_4NO_3, etc.). H_2SO_4 and HNO_3 formed by the oxidation of SO_2 and NO_x, respectively.

The primary source of oxidized sulfur from human activity is the burning of coal. Much of the world's coal used for energy purposes contains 2% S, half of which is present as pyrite (Fe_2S) and the remainder is organic. SO_2 is produced readily on burning:

$$4FeS_2 + 11O_2 \rightarrow 2Fe_2O_3 + 8SO_2 \qquad [2]$$

NO and NO_2 enter the atmosphere from natural and pollutant sources, such as the burning of fossil fuels in motor vehicles or stationary furnaces. The

formation of NO from the reaction of N_2 and O_2 occurs at high temperatures:

$$N_2 + O_2 \rightarrow 2NO \quad [3]$$

Once NO has been formed, rapid cooling of exhaust gases prevents further reaction and traps the oxide in the atmosphere. NO is also formed naturally in the atmosphere through reaction of O_2 and N_2 caused by lightning.

N_2O is formed mainly from biological agents and is an essential intermediate in the N cycle in soils; for example, in the process of denitrification:

$$NO_3 \rightarrow NO_2 \rightarrow NO \rightarrow N_2O \rightarrow N_2(g) \quad [4]$$

This process is often observed in organic-rich sediments and flooded soils and is facilitated by microorganisms growing anaerobically. Significant emissions of nitrous oxide also arise in aerobic or partially aerobic soils from the rapid oxidation of organic matter or ammonium fertilizers by the process of nitrification:

$$\begin{array}{l} NH_4^+ \rightarrow NH_2OH \rightarrow (NOH) \rightarrow NO_2 \rightarrow NO_3^- \\ \qquad\quad \downarrow \\ \qquad\quad N_2O \end{array} \quad [5]$$

A very important source of acidification in ecosystems is the production of protons from NH_4^+ ions by the nitrification process:

$$NH_4^+ + 2O_2 \rightarrow NO_3 + H_2O + 2H^+ \quad [6]$$

Impacts from atmospheric NH_3 and ammonium have been shown to be especially important in forest ecosystems, particularly in the Netherlands, where almost 50% of the acidic deposition at the 'edges' of Dutch forests is considered to be derived from volatilized ammonia. The ammonium ion is a conspicuous constituent of areas suffering impacts from acidic deposition due to the rapid nitrification (oxidation) of ammonium salts; sulfate, nitrate, and bicarbonate are typical counterions:

$$(NH_4)_2SO_4 + 4O_2 \rightarrow 2NO_3^- + 4H^+ + 2H_2O + SO_3^{2-} \quad [7]$$

Atmospheric Transport, Secondary Chemistry, and Acid Rain

The concentrations of primary pollutants and secondary pollutants determine the nature and pattern of atmospheric deposition over a region. These concentrations are related to the chemical processes that occur simultaneously with transport and dispersion through the atmosphere. Emitted primary pollutants such as SO_2 and NO_x, which have not already been deposited directly by dry deposition, are oxidized to secondary pollutants and converted to acidic, water-soluble aerosols. For example, SO_2 is oxidized to H_2SO_4 and NO_2 to HNO_3.

Atmospheric cycling of these pollutants is a complex process (Figure 1). Atmospheric oxidation is mediated by chains of free-radical reactions. The oxidation proceeds in a system of very dilute reactants using an external energy source provided by solar radiation. Within the troposphere, the most important reactions – gas-phase reactions – are those initiated by the hydroxyl radical, OH. This is known to react with many pollutants and it controls the tropospheric concentrations of many of them. The OH radical is often regarded as the natural cleaning agent of the atmosphere. In turn, trace gases influence atmospheric concentrations of OH. The two other key oxidizing species in the production of atmospheric acidity are ozone (O_3) and hydrogen peroxide (H_2O_2).

Dry oxidation of sulfur and nitrogen oxides In the absence of clouds, the major route for the production of H_2SO_4 from SO_2 is via reaction with OH radicals in the gas phase:

$$OH + SO_2 \rightarrow HSO_3 \quad [8]$$

$$HSO_3 + O_2 \rightarrow HO_2 + SO_3 \quad [9]$$

$$SO_3 + H_2O \rightarrow H_2SO_4 \quad [10]$$

In daylight, the major gas-phase route for the production of HNO_3 is via reaction of NO_2 with OH:

$$OH + NO_2 \rightarrow HNO_3 \quad [11]$$

At night a further route is available via the formation of the nitrate radical, which is photolytically unstable in sunlight:

$$NO_2 + O_3 \rightarrow NO_3 + O_2 \quad [12]$$

$$NO_2 + NO_3 \rightarrow N_2O_5 \quad [13]$$

$$N_2O_5 + H_2O \rightarrow 2HNO_3 \quad [14]$$

Wet oxidation of sulfur SO_2 dissolves in atmospheric water droplets to form the bisulfite ion (HSO_3^-):

$$SO_2 \rightarrow SO_2 \cdot H_2O \rightarrow HSO_3^- + H^+ \quad [15]$$

Oxidation of HSO_3^- yields a sulfate ion (SO_4^{2-}) and a further proton:

$$HSO_3^- + \text{Oxidant} \rightarrow SO_4^{2-} + H^+ \quad [16]$$

The major oxidants for the production of dissociated SO_4^{2-} and H^+ ions are H_2O_2 and O_3. The H^+ ions are deposited as wet-deposited acidity.

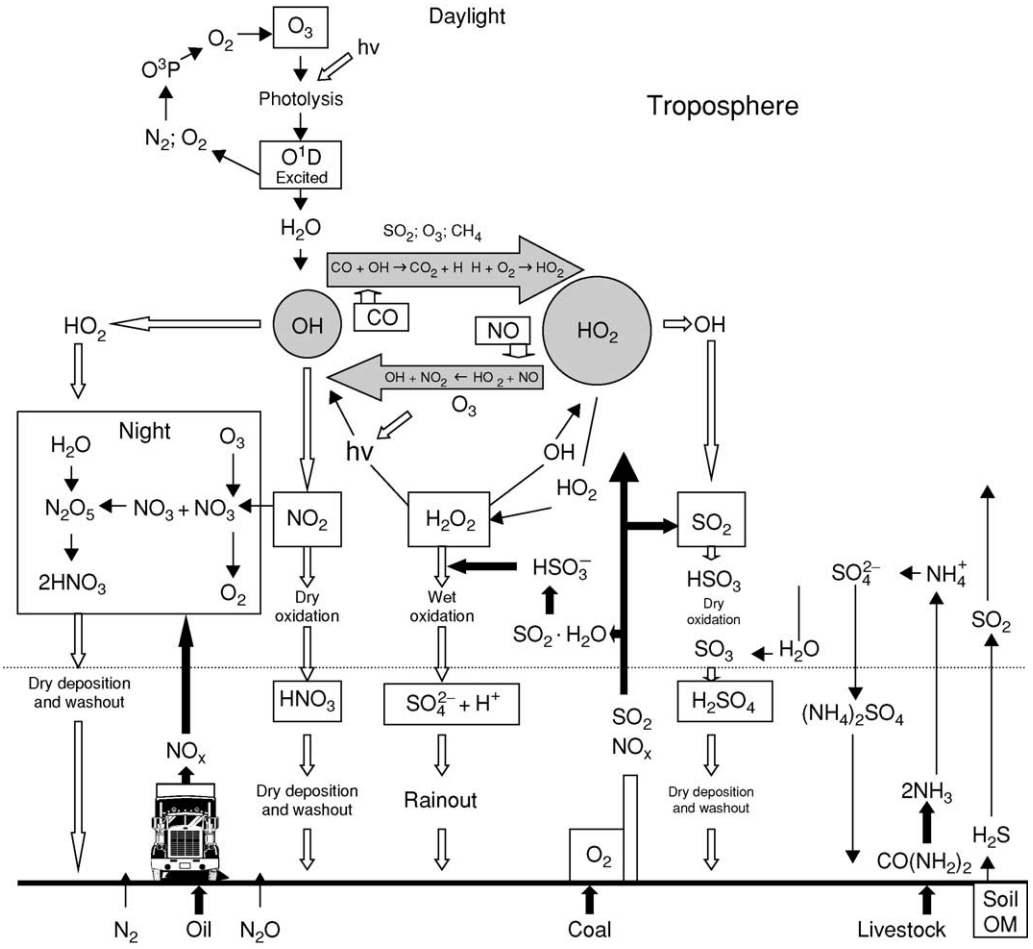


Figure 1 Summary of the processes governing acid deposition. The recycling of the hydroxyl radical is outlined in gray. The dotted line represents the boundary layer. OM, organic matter; hV, UV protons from solar radiation; O^3P, ground state atomic oxygen.

Deposition of both 'wet' and 'dry' products of S and N reactions occurs. Dry deposition is the direct transfer of gases and particulates from the atmosphere. The primary gaseous species of S, SO_2, may be deposited directly from the pollution source, as may the N products from the dry oxidation stages of the secondary reactions, NO_2 and HNO_3 (most N_2O and NO emissions are oxidized). Gaseous species also directly deposited are NH_3, HCl, and O_3.

Dry deposition of particles and gases occurs by sedimentation and surface adsorption; it has greatest significance close to pollutant areas. Dry-deposited species themselves react with surface moisture (in soil or on plant foliage) to produce acidity by the dissociation of H^+ ions and mobile conjugate anions, although some H^+ tends to be neutralized by NH_3, resulting in the formation of NH_4^+ ions.

Wet deposition – the precipitation of scavenged pollutants from the atmosphere – is the dominant removal mechanism remote from the pollutant source. Wet deposition involves the removal of primary SO_2 by reaction with atmospheric water and oxidation with H_2O_2 or O_3 and the removal of secondary NO_2, as HNO_3, by reaction with the hydroxyl radical. Wet deposition is characterized by two processes: (1) rainout – whereby atmospheric species are associated with cloud phenomena; (2) washout – where species are removed by falling precipitation. Figure 2 shows the long-term trend in the deposition of acidity (H^+) and the precursors of acidity in southern England since 1853.

Acid Soil

Soil acidification is defined as a decrease in acid-neutralizing capacity (ANC) or an increase in base-neutralizing capacity (BNC), resulting in an increase in acid strength as represented by a decrease in soil pH:

$$\text{Soil acidity}(+\Delta\text{BNC}) = -(\text{soil alkalinity}) = -\Delta\text{ANC} \quad [17]$$

Soil acidification processes in aerated soils are a consequence of: (1) production of various acids in the

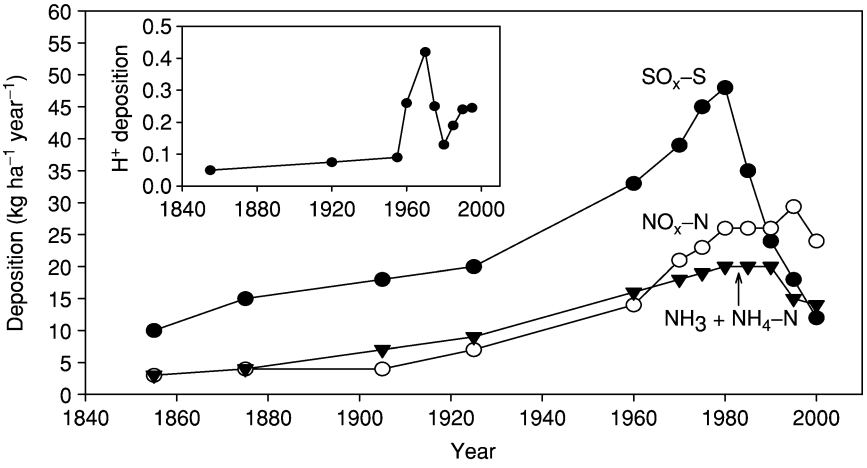


Figure 2 Wet-deposited acidity (H^+; the truly acid rain) and the estimated total deposition (wet plus dry deposition) of the precursors of acidity at Rothamsted Research in southern England, UK. Wet deposition has been measured at Rothamsted since 1853. Dry deposition of NO_3^- and NH_4^+ measured since 1990, and of SO_4^{2-} since 1980, calculated in other years from the ratio of wet/dry since 1990.

soil, (2) the effect of ion uptake by biota, (3) nitrogen inputs and transformations, and (4) the addition of dissolved strong acids (acid deposition).

Sources of Acids in Soils

Several ecosystem and soil processes decrease soil pH through internal acid production. These processes can also be directly or indirectly influenced by atmospheric deposition.

Substances which react in soils as proton donors (acidity) are as follows:

1. In the solid phase:
   - Cations forming weak hydroxides (mainly Al, Fe, and Mn, either as exchangeable cations or bound within clay minerals or associated with organic matter – M_a cations). For Al:

     $$Al^{3+} + H_2O \leftrightarrow AlOH^{2+} + H^+,$$
     $$Al^{3+} + 2H_2O \leftrightarrow Al(OH)^{2+} + 2H^+ \quad [18]$$

   - Mineralization of organically bound N followed by nitrification:

     $$R{-}C{-}NH_2 \rightarrow NH_3 + H^+ \rightarrow NH_4^+ + 2O_2 \rightarrow NO_3 + H_2O + 2H^+ \quad [19]$$

   - Undissociated acidic groups on (1) clay minerals (pH-dependent charge) and (2) organic matter (R-OOH);
   - Aluminum hydroxysulfates and sulfate sorbed on aluminum hydroxides;
   - Exchangeable NH_4^+ (nitrification or assimilation of the ammonium cation);
   - Organically bound S (organic S $\rightarrow$ H_2SO_4);
2. In solution:
   - $CO_2 \cdot H_2O \rightarrow HCO_3^- + H^+$ (carbonic acid); [20]
   - NH_4^+: ($NH_4^+ \leftrightarrow NH_3 + H^+$; $NH_4^+ \rightarrow$ organic $N + H^+$; assimilation); [21]
   - cations forming weak acids (as for the first item in the solid-phase list);
   - Organic acids (R-COOH):

     $$CO_2 + R{-}CH_2OH \rightarrow RCOO^- + H_2O + H^+ \quad [22]$$

Internal Acid Production in Soils

In order to isolate the specific impact that acid deposition has on soil pH, the effects of internal acid production in soils need to be assessed.

Carbonic and organic acid production The production of weak acids such as $CO_2 \cdot H_2O$ or dissolved organic acids by plants and by decomposition of organic substances does not result in soil acidification, because the weathering of silicates releases alkali and earth alkali (M_b) cations into solution. However, ANC decreases if M_b cations are removed from the soil by drainage. Thus, dissolution of Na, Ca, and Mg from soil minerals by $CO_2 \cdot H_2O$, followed by leaching with HCO_3^-, is the dominant soil-acidification process in nature (pH >5).

Stronger organic acids replace CO_2 as the acidifying agent in acidic soils of pH 4–5 (eqn [22]) and are the dominant acidifying agent in such acid soils under natural conditions. Low-molecular-weight species of humic acids as well as aliphatic acids are produced in the A horizon of soils and are transported downward toward the B horizon. Humic acids contain phenolic and carboxylic groups, the latter readily binding metal ions such as Al and Fe. The metal ions link the organic molecules, producing larger molecules with lower solubility, and release protons to the soil

solution of the B horizon. The production of organic acids results largely from microbial activity in the A horizon; chemical precipitation reactions take place in the B horizon. As a result, this wholly natural process of acidification is restricted to the A and B horizons of soils beneath typically podzolizing plant communities such as heather or spruce stands. Podzolization cannot explain the acidification of soils below the B horizon of freely draining soils.

Nitrogen accumulation and transformations The accumulation of organic N in soils is potentially a large proton source for acidity because of nitrification following mineralization. Changing conditions of organic matter decomposition and humus formation strongly influence this powerful proton source. Thus, changes to ecosystems which influence the turnover on organic matter can result in considerable internal proton production. These are largely the effect of anthropogenic changes in ecosystem management such as increased rates of agricultural or forest production, and also to atmospheric N enrichment. The capacity for proton production by this process is only limited by the amount of organic N accumulated in the mineral soil during ecosystem development, but is almost entirely restricted to the surface soil horizons (rooting zone) where this accumulation occurs.

Assimilation of Nutrients by Vegetation

In ecosystems where efficient N cycling occurs, N species are unimportant in terms of soil acidification. In such ecosystems, however, plants take up more basic components (M_b cations) than acidifying components. To maintain electroneutrality, plants excrete protons as counterions to cation uptake. As a consequence, ecosystems that increase in biomass acidify the soil. For natural ecosystems in steady state, with no export of cations in biomass, mineralization equals assimilation and the soil does not acidify by assimilation of nutrient cations. However, the export of biomass (e.g., timber) results in the removal of bases stored in the vegetation, leading to a decline in ANC and further soil acidification.

Deposition of Nitrogen Species and Nitrogen Transformations

Atmospheric inputs of N species influence soil acidification either (1) directly through the deposition of acid nitrogen compounds or (2) indirectly through biological N transfers within the ecosystem. N transfers cause proton fluxes if there is a net input or output of NH_4^+ or NO_3^- from the soil. These fluxes result from deposition-derived NH_4^+ or NO_3^- and uptake by vegetation (which causes an equivalent release of H^+ or HCO_3^- from the roots, respectively) or nitrification of NH_4^+ (eqn [19]) followed by leaching of NO_3^-. For soils with high inputs of NH_4^+ from the atmosphere, this can be the dominant form of acidity.

Addition of Dissolved Strong Acids Through Acid Deposition

HNO_3 and H_2SO_4 are the major components of directly derived atmospheric acidity. After their addition ANC is decreased rapidly when basic cations (M_b) released from soil minerals by strong acid weathering are leached with sulfate or nitrate in drainage. The relative contribution of HNO_3 and H_2SO_4 to acidification can be assessed from comparison of the input/output (deposition/seepage) balance of SO_4-S/Cl and NO_3-N/Cl (chloride is not bound in soil, so output equals input). ANC is also decreased if H_2SO_4 is retained in the soil, by sulfate adsorption or by precipitation of sulfates of Fe and Al.

Observations of Changes in Soil Acidification due to Acid Deposition

To assess the causes and effects of acid deposition on ecosystems involves measuring changes determined by soil measurement over a period of time, or making measurements on soils that have been sampled chronologically and stored. Relatively little information on the analysis of long-term data is available, but numerous measurements based on lysimeter studies have been made since the 1970s.

There are four prime sources that show the results of reexamining soils in polluted regions and the long-term (more than 30 years) changes involved:

1. Eastern USA: A reexamination of soils in the Adirondack Mountains, New York, after a lapse of 50 years, revealed that the pH of organic horizons initially at pH 4.0 (Al buffer) remained unchanged, but the pH of originally less acid horizons had declined by 0.3–0.5. Acidic deposition and natural leaching were considered to be the major contributors to these pH changes;
2. Sweden: Long-term pH changes in southern Sweden were documented by repeating studies from 1927 on the Tonnersjoheden experimental area. Two trends were detected: (a) the pH of the humus layer (O horizon) and others decreased as stands of Norway spruce increased in age; and (b) there was a significant pH decrease in all soil horizons, including the deepest C horizon, where pH was independent of stand age. The decreases ranged between pH 0.31 and 0.65. A combination of acid rain and biological acidification was the cause of the changes;

3. Scotland: Soils were resampled in 1987 from 15 sites in Alltcailleach Forest, first sampled in 1949–1950. The surface horizons at 12 sites had decreased by 0.07 units to pH 1.28, while decreases of 0.16–0.54 units were found in the deeper mineral horizons. All of these decreases were also associated with decreased base saturation and markedly increased amounts of extractable Al;

4. England: The most detailed study of long-term (1876–1991) changes in soil chemistry in response to acidic deposition used archived soil samples taken from woodland and grassland at Rothamsted Research in southern England. The changes were assessed in light of the measured increases in acid deposition at this location, as shown in Figure 2, and are used here to illustrate the long-term effects of acid deposition and soil acidification.

Proton Sinks and Buffer Ranges

The major changes in the chemistry of soils subject to acidification result from proton-buffering processes. Substances which react in soils as proton acceptors (basicity) are responsible for soil ANC. In the solid phase these are: (1) carbonates, (2) silicates, and (3) alkali and earth alkali (M_b) cations which are bound or exchangeable on weakly acidic groups of soil minerals and organic matter.

In the solution phase, OH^- and $R\text{-}COO^-$ are usually at too low a concentration to be of importance. HCO_3^- is the dominant base, but only at pH >5:

$$HCO_3^- + H^+ \rightarrow CO_2 + H_2O \qquad [23]$$

The soil phases accepting protons release M_a and M_b cations through dissolution reactions, which change with increasing acid strength. This leads to a sequence of buffer reactions as pH decreases. In addition to dissolution reactions, proton adsorption also takes place on negatively charged surfaces associated with clay minerals and organic matter. These processes vary according to the composition of the buffer substances and the reaction products. Figure 3 shows how proton sinks have influenced soil pH in silty clay loam soils at Rothamsted and illustrates two important aspects of soil acidification. First, the pH value does not change in a simple, constant manner with the quantity of H^+ input. This is due to stronger

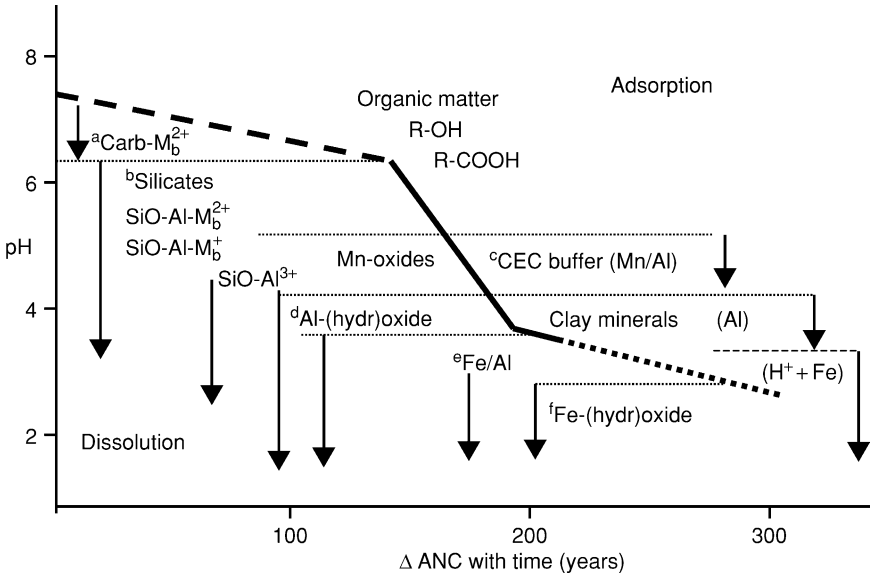


Figure 3 pH range of action of proton sinks in a silty clay loam soil, based on the analysis of archived soils from three long-term (more than 100 years) field experiments receiving acidifying inputs at Rothamsted, UK: thick dashed line, woodland heavily limed in 1881 subject to acid deposition; thick continuous line, unlimed woodland subject to acid deposition; thick dotted line, grassland strongly acidified by annual applications of ammonium sulfate (144 kg N ha^{-1} per year; since 1856). Reaction products: a, carbonate buffer: $Ca(HCO_3)_2$ in solution, leaching of Ca and basicity; b, silicate buffer: Basic cations (M_b) in solution. At pH <5, Al species in solution (monomeric Al-hydroxy cations, e.g., $Al(OH)^{2+}$, Al^{3+}, polymeric Al-hydroxy cations: $n(Al(OH)_x^{3-x})^+$; Mn^{2+} from oxides. Overlaps with cation exchange capacity (CEC), Al and Fe buffers; c, cation exchange (CEC) buffer: nonexchangeable polymeric Al-hydroxy cations (blockage of permanent charge – reduction of CEC). Exchangeable Mn^{2+}, monomeric Al-hydroxy cations, and Al^{3+} reduce base saturation by displacement of M_b cations. At pH <4.5, amphoteric Al-hydroxy cations, (blocks and/or changes permanent charge to pH dependent charge – reduction of CEC). Formation of Al-hydroxy sulfates at pH >4.2; d, aluminum buffer: Al^{3+} in solution (from interlayer Al, Al-hydroxy sulfates, and Al hydroxides). H^+ adsorption – permanent charge to pH-dependent charge – reduction of CEC; e, Al/Fe buffer: exchangeable H^+ and Fe (displacement of Al); f, iron buffer: exchangeable H^+ and Fe (mineral disruption). ANC, acid-neutralizing capacity.

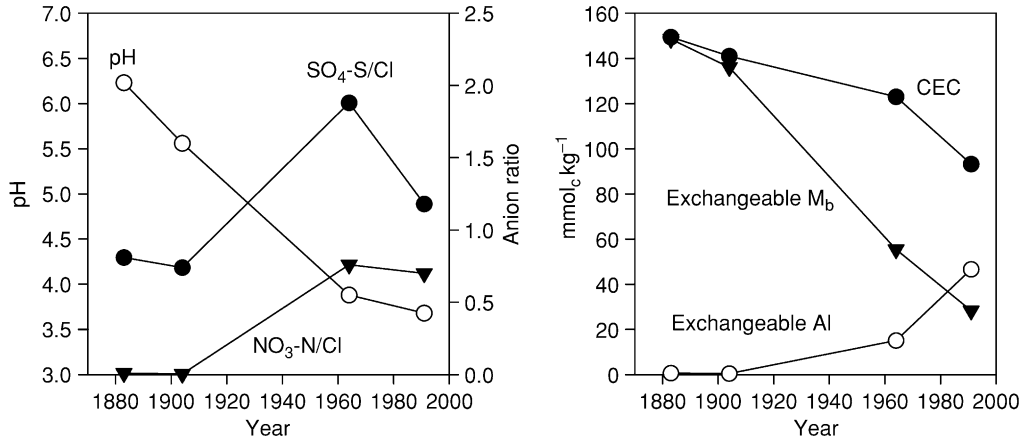


Figure 4 Long-term (1883–1991) changes in soil chemistry in archived soil samples (0–23 cm depth) taken from woodland (Geescroft wilderness) subject to acid deposition at Rothamsted, UK. CEC, cation exchange capacity.

buffering in certain pH ranges: by carbonates at near-neutral pH values, and by hydroxy Al and Fe compounds and clay minerals at pH 4.2–3.2. Second, the rapid addition of strong mineral acids from acid deposition depresses the pH much more rapidly than slower addition from most internal sources. This is particularly characterized by a relatively rapid decline of the CEC buffer. Major long-term changes in soil chemistry through soil acidification beneath woodland at Rothamsted are shown in Figure 4. A schematic representation of changes through the soil profile is shown in Figure 5.

Carbonate and Silicate Buffers

The dissolution of Ca and Mg carbonates is driven by the protonation of CO_3^{2-} at pH values greater than 6.2; the reaction products are soluble bicarbonates which are leached through the soil:

$$CaCO_3 + H_2CO_3 \leftrightarrow Ca^{2+} + 2HCO_3^- \qquad [24]$$

As soon as $CaCO_3$ is used up, the M_b cations leached with bicarbonate originate from silicate weathering.

The weathering of primary silicates is the dominant buffer reaction in soils at pH <6. For relatively abundant SiO–Al–M groups, there is a two-step dissolution process. At pH values less than 6.5, SiO–Al–M_b groups become progressively protonated, liberating M^+ and M^{2+}, while SiO–Al groups remain intact. The products of acid weathering of silicates are clay minerals with permanent negative charge and mobile cations. Thus, stable ANC present in silicates is transferred to labile-mobilizable ANC as exchangeable base cations adsorbed on clay minerals. When pH has decreased to approximately pH 5, the SiO–Al become progressively protonated, releasing Al species into solution. The mobilized reaction products are: (1) monomeric Al-hydroxy cations and Al^{3+}, which are adsorbed on to exchange surfaces; and (2) non-exchangeable polymeric Al-hydroxy cations, which lead to a reduction in CEC. Mn oxides (if present) are also readily protonated at pH <5 to such an extent that the acidification front is characterized by a high proportion of exchangeable Mn^{2+}.

Cation Exchange Buffer

As acidity increases at pH values of less than 5, Al^{3+} in solution and on exchangeable surfaces increases. Al^{3+} saturation on exchange surfaces can reach high values at low Al^{3+} concentrations in solution, because of highly selective bonding of Al^{3+} on exchange surfaces. As a result, M_b cations bound to the exchange sites are leached together with the anion of the acid-generating proton and the output of anions comes close to the input. This results in a progressive decline in base saturation. Where the dominating anion leached is SO_4^{2-}, the main source of acidity is H_2SO_4 from acidic deposition.

In most cases, the continued weathering of silicates results in substantially more Al^{3+} than M_b cations and is a major source of acidity, because Al^{3+} and hydrolyzed Al ions are intermediaries of protons (BNC; eqn [18]). As a result, acidity moves down the soil profile with the Al species in the leachate. The CEC partially buffers this potential acidity through adsorption on to exchange surfaces. If the acid loading is H_2SO_4, the formation of Al-hydroxy sulfates occurs at the lower pH range of the CEC buffer:

$$Al(OH)_3 + H_2SO_4 \leftrightarrow AlOHSO_4 + 2H_2O \qquad [25]$$

This gives the soil the additional property to react as a sink for H_2SO_4. At the same time the CEC is strongly reduced by the covering of clay surfaces

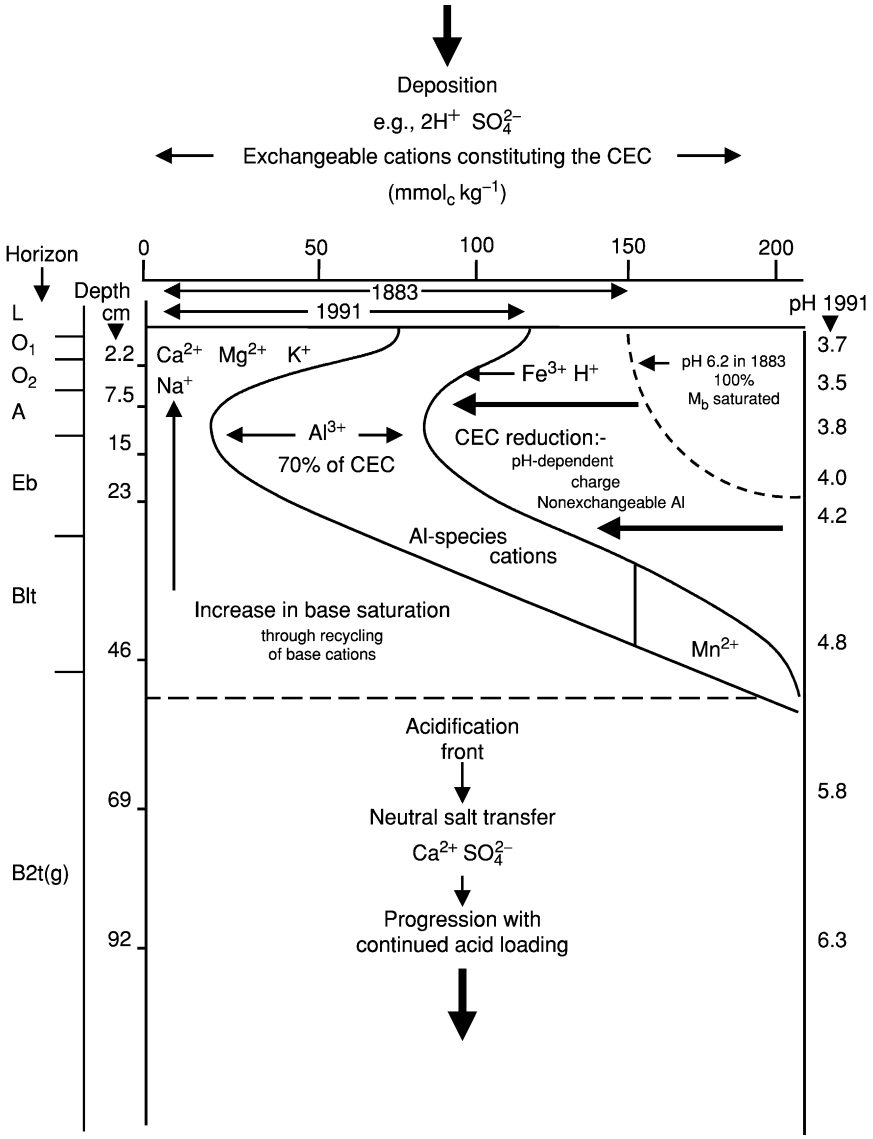


Figure 5 A schematic model of soil acidification beneath a typical deciduous woodland (Geescroft wilderness) receiving long-term acid deposition. The model is based on the analysis of archived soils (1883–1991) taken in incrementing depths to a total depth of 92 cm. CEC, cation exchange capacity.

with amphoteric-Al hydroxy cations, blocking adsorption sites and 'changing' permanent charge into pH-dependent charge.

The degree of saturation of the CEC with Al species indicates the remaining capacity of the solid phase for proton binding. Acid mineral subsoils with Al-saturated exchange surfaces (less than 5% M_b) characterize soils subject to long-term acid deposition.

Aluminum and Iron Buffer

At pH <4.2, Al-hydroxy compounds and interlayer Al become increasingly weathered to an extent that Al^{3+} becomes the dominant cation in the soil solution and exchange surfaces become saturated with Al^{3+}. Al-hydroxy cations and sulfates accumulated in the CEC buffer range decrease through dissolution:

$$AlOHSO_4 + H^+ \rightarrow Al^{3+} + SO_4^{2-} + H_2O \qquad [26]$$

Leached (output) SO_4^{2-}, H^+, and Al^{3+} levels show a tendency to exceed SO_4^{2-} and H^+ inputs where acidic deposition is dominated by H_2SO_4.

At pH <3.8, Fe (hydr)oxides become increasingly protonated. This buffer mechanism runs parallel with the Al buffer to pH 3.2 and is accompanied by a

progressive shift in the exchangeable cation composition from Al^{3+} to Fe^{3+} and H^+ ions, with Al^{3+} being desorbed.

Proton Adsorption Reactions

The adsorption of H^+ between pH 5 and 4 is through negatively charged surfaces on organic matter (R-COO^-) and (hydr)oxides (CEC buffer). Such storage is reversible through exchange with M_b cations. At pH >4, only adsorption sites that exert a high preference for protons contribute significantly to H^+ adsorption on to clay minerals. The direct protonation of adsorption sites on clay minerals contributes to ANC most at pH <4 and permanent charge at mineral surfaces becomes progressively pH-dependent. Isomorphous substitution may also be restricted through mineral disruption and the inclusion of M_a cations. These processes reduce the CEC.

The permanent effect of dissolution and adsorption reactions on soil clay may be reflected by X-ray diffraction analysis in a reduction in expansion of swelling mineral (where present) upon solvation. This can be accounted for by the introduction of interlayer Al species.

Acid Sources and Sinks: Input–Output Relationships and the Calculation of Proton Balances

In order to quantify the impact of acidic deposition on changes in soil pH in relation to other soil acidification processes, all important H^+ fluxes need to be assessed. Protons, acid precursors, and bases entering or transferred within the soil (inputs) leave the system as solutes or in biomass (outputs). Protons can be produced or consumed within the system by the buffering processes described above. From the principle of electroneutrality, for a given soil compartment, the storage of cations ($M_a = M_b = M^+$) equals the storage of anions (A^-) expressed in moles ion charge (equivalents):

$$M^+ + H^+ = A^- \qquad [27]$$

The change in protons (H^+) can therefore be calculated from the change in storage fluxes of cations and anions within soil compartments:

$$H^+ = A^- - M^+ \qquad [28]$$

On this basis the total proton load (TPL) can be estimated by the determination of fluxes within the individual compartments (H^+ sources and sinks) as follows:

$$\underset{\text{ion flux (weathering)}}{\text{Input}-\text{output}} \quad \underset{\text{production}}{\text{Internal proton}}$$

$$H^+ = \underbrace{(H^+_{in} - H^+_{out})}_{\text{Protons}} + \underbrace{[(M^+_{in} - M^+_{out}) - (A^-_{in} - A^-_{out})]}_{\text{Cations}\qquad\text{Anions}}$$

$$+ \underbrace{[(M^+_{up} - A^-_{up}) + (HCO^-_{3\,in} - HCO^-_{3\,out})}_{\text{Ion uptake}\qquad\text{Dissociation}}$$

$$+ \underbrace{(R{-}COO^-_{in} + RCOO^-_{out})]}_{\text{Protolysis}}$$

$$+ \underbrace{[(NH^+_{4\,in} - NO^-_{3\,in})(NH^+_{4\,out} - NO^-_{3\,out})]}_{\text{N-transformations: deposition + organic N}} = TPL \qquad [29]$$

The weakness of this approach is the difficulty involved in determining the rates of these processes. Cation and anion weathering and organic acid dissociation can be estimated by the differences between the sum of measured cations and anions in soil solution or drainage using standard lysimeters or suction-cup lysimeters. If the amount of biomass and its cation and anion content are known, the cation excess can be calculated; the proton production due to the development and export of biomass is equal to the cation excess. The kind of N nutrition cannot be measured directly; its effect on the soil proton flux is included in the production and consumption of NH_4^+ and NO_3^- in the input–output balance of the soil; this assesses the whole effect of the N cycle, which includes N mineralization as well as N uptake.

Table 1 shows examples of proton budgets determined for four of the 'classic' ecosystem studies of acidification with the necessary calculation steps. Note that proton sources should balance proton sinks. The isolation of internal and external proton sources highlights the profound effect that acid deposition has had on the ecosystems at Solling (Germany) and Rothamsted (England) and the much lesser effect of internal proton production at Cedar River (USA). The predominant source of acidity at these locations is also identified by the anion composition (SO_4-S:Cl and NO_3-N:Cl ratios) in the soil solution or leachate, providing clear evidence of the contribution of acid deposition to recent soil acidification.

Acidification, Ecosystem Stability and Global Change

The influence of acid deposition on soil acidity can be evaluated by looking for the characteristic effects that it has. The transfer of soil horizons beneath the rooting zone of nonpodzolized soils into the Al-buffer range only occurs through long-term acid deposition because

Table 1 Processes which have to be considered for the estimation of proton budgets, with four examples from different locations. Values are in kilomoles of charge per hectare per year

	Solling, Germany[g]		*Geescroft wilderness, UK*[h] *deciduous woodland*	*Cedar river, USA Douglas fir*[i]
	Spruce	*Beech*		
Source				
Wet deposition of H^+[a]	↓	↓	0.2	↓
$SO_2 + NO_2$ deposition[a]	3.7	1.7	2.5	0.3
N transformations[b]	1.1	0.8	2.8	0.1
Base cation uptake[c]	2.1	1.5	0.9	0.5
Carbonic acid and/or organic acid dissociation[d]	0.6	0.5	0.1	0.4
Anion weathering[e]	0.2	0.1	0	0.3
Total	7.7	4.6	6.5	1.6
Depth of soil profile (M)	1.0	1.0	0.46	0.45
pH of soil solution/output	4.0	4.2	4.0	6.7
SO_4-S/Cl deposition	2.0	1.6	1.3[f]	–
output	3.0	2.0	0.9	0.7
NO_3-N/Cl deposition	0.40	0.33	1.6	–
output	0.35	0.05	0.3	0.04
Sink				
Leaching of H^+	0.4	0.5	1.1	0.0
N transformations	0.0	0.5	0	0.1
Anion accumulation	0.6	0.3	0	0.1
Cation weathering	6.1	3.0	5.9	1.3
Total	7.1	4.3	7.0	1.5

[a]Wet-deposited acidity and dry deposition of acid precursors such as SO_2 and NO_x.
[b]N transformation refers to the input–output balance of NH_4^+ and NO_3^-. There is no proton production if $(NH_{4\,in}^+ - NH_{4\,out}^+) + (NO_{3\,out}^- - (NO_{3\,in}^-) > 0$.
[c]Cation and anion accumulation refer to the accumulation in biomass and humus of cations and anions other than NH_4^+ and NO_3^-.
[d]Estimated from the difference between the sum of measured cations and the sum of measured anions in soil solution or drainage water.
[e]The weathering of individual cations and anions is estimated from the following mass balance: $W = O - I = \Delta B$, where O is the output in drainage water; I is the input through deposition; ΔB is accumulation in biomass and humus.
[f]Reflects local decline in S deposition since 1980 (see also **Figure 4**).
[g]Source: van Breeman N, Driscoll CT, and Mulder J (1984) Acidic deposition and internal proton sources in acidification of soils and waters. *Nature* 307: 599–604.
[h]Source: Blake L, Goulding KWT, Mott CJB, and Johnston AE (1999) Changes in soil chemistry accompanying soil acidification over more than 100 years under woodland and grass at Rothamsted Experimental Station, UK. *European Journal of Science* 50: 401–412.
[i]Source: van Miegroet H (1986). *Role of N Status and N Transformations in H^+ Budget, Cation Loss and S Retention Mechanisms in Adjacent Douglas-Fir and Red Alder Forests*. PhD Thesis. Washington, DC: University of Washington.

nitrification of organic N is restricted to the rooting zone. Once the rooted soil is acidified, the release of M_a cations compounds the buildup of BNC and the acidifying front passes into the deeper horizons together with leached M_b cations as sulfates and/or nitrates. Under continued acid deposition loading, this process continues through the soil profile beneath the rooting zone with soluble Al-species, sulfate, and/or nitrate accumulating in the soil solution and seepage waters.

Since 1980 the deposition of S from the atmosphere has decreased throughout Europe and North America as the result of pollution abatement strategies based on the quantitative determinations of critical loads of acidifying pollutants. For many natural ecosystems to maintain soils at their present pH, this approach requires deposition values to be restricted to preindustrial emissions of less than 5 kg ha^{-1} per year each of total S and N. The deposition of acidifying N-species in many locations still considerably exceeds this value. Deacidification can only occur if acid deposition tends to zero and the release of M_b cations from silicates and organic matter can replenish base saturation. However, if acidification has been allowed to proceed to a point where the CEC is nearly empty of nutrient cations (Al buffer), the time for the soil system to repair itself naturally to a state that is ecologically acceptable will be in the order of centuries.

The acidification of ecosystems has a considerable influence on soil biological and chemical processes. Al mobilization and release to surface waters are of most concern, because Al is highly toxic to terrestrial plants, and to humans and animals drinking the water. Soil acidification also causes solubilization of heavy metals in soils, both those naturally contained in minerals and any that entered the soil from pollution arising from waste disposal. As with Al, metals such as Pb, Cu, Zn, or Cd can be toxic to plants or, if they enter the food chain, to humans and animals.

See also: **Acidity**; **Buffering Capacity**; **Liming**; **pH**; **Sulfur in Soils:** Overview

Further Reading

Berden M, Nilsen SI, Rosen K, and Tyler G (1987) *Soil Acidification: Extent, Causes and Consequences*. National Swedish Environment Protection Board Report 3292. Solna, Sweden: National Swedish Environmental Protection Board Information Section.

Blake L, Goulding KWT, Mott CJB, and Johnston AE (1999) Changes in soil chemistry accompanying acidification over more than 100 years under woodland and grass at Rothamsted Experimental Station, UK. *European Journal of Soil Science* 50: 401–412.

Billet MF, Parker Javis F, Fitzpatrick EA, and Cresser MS (1990) Forest soil chemical changes between 1949/50 and 1987. *Journal of Soil Science* 41: 133–145.

Bredemeier M (1989) Nature and potential of ecosystem – internal acidification processes in relation to acid deposition. In: Longhurst JWS (ed.) *Acid Deposition*, pp. 197–212. London, UK: British Library Technical Communications.

Hallbacken L and Tamm CO (1986) Changes in soil acidity from 1927 to 1982–84 in a forest area of southwest Sweden. *Scandinavian Journal of Forest Research* 1: 219–232.

Helyar KR and Porter WM (1989) Soil acidification, its measurement and the processes involved. In: Robson AD (ed.) *Soil Acidity and Plant Growth*, pp. 61–101. Sydney, Australia: Academic Press.

Johnson DW (1987) A discussion of changes in soil acidity due to natural processes and acid deposition. In: Hutchinson TC and Meema K (eds) *Effects of Acidic Deposition on Forests, Wetlands and Agricultural Ecosystems*, pp. 333–346. New York: Springer-Verlag.

Kennedy IR (1992) *Acid Soil and Acid Rain*. New York: Research Studies Press, LTD/John Wiley.

Last FT and Watling R (eds) (1991) Acidic deposition its nature and impacts. *Royal Society of Edinburgh Proceedings. Section B Biological Sciences* 97: 17–34, 81–116, 155–168, 169–191.

Reuss JO and Johnson DW (1986) Acid deposition and the acidification of soils and waters. *Ecological Studies Analysis and Synthesis,* vol. 59. New York: Springer-Verlag.

Reuss JO and Walthall PM (1989) Soil reaction and acidic deposition. In: Norton SA, Lindberg SE, and Page AL (eds) *Acidic Precipitation,* vol. 4, pp. 1–31. *Soils, Aquatic Processes and Lake Acidification*. New York: Springer-Verlag.

Ulrich B (1983) Effects of acid deposition. In: Beilke S and Elshout AJ (eds) *Acid Deposition*, pp. 31–41. Dordrecht, The Netherlands: Reidel.

Ulrich B (1983) A concept of forest ecosystem stability and of acid Deposition as a driving force for destabilization. In: Ulrich B and Pankrath J (eds) *Effects of Accumulation of Air Pollutants in Forest Ecosystems*, pp. 1–29. Dordrecht, The Netherlands: Reidel.

Ulrich B and Sumner ME (eds) (1991) *Soil Acidity*. Berlin, Germany: Springer-Verlag.

van Breeman N, Driscoll CT, and Mulder J (1984) Acidic deposition and internal proton sources in acidification of soils and waters. *Nature* 307: 599–604.

ACIDITY

N S Bolan, Massey University, Palmerston North, New Zealand
D Curtin, New Zealand Institute for Crop and Food Research, Christchurch, New Zealand
D C Adriano, University of Georgia, Aiken, SC, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Soil acidification is a natural process that can either be accelerated by certain plants and human activities or slowed down by careful management practices. Industrial and mining activities lead to soil acidification due to acid produced from pyrite oxidation and from acid precipitation caused by the emission of sulfur (S) and nitrogen (N) gases. In managed ecosystems, soil acidification is mainly caused by the release of protons (H^+) during the transformation and cycling of carbon (C), N, and S, and fertilizer reactions. Soil acidification caused by these processes can have adverse impacts where soils are unable to buffer against further pH decrease. For example, in parts of North America and Europe, soil acidification caused by acid precipitation has resulted in forest decline and, in some parts of Australia, continuous legume cultivation and inappropriate use of N fertilizer have generated sufficient soil acidity that cereal crop cultivation has had to be abandoned due to aluminum (Al) and manganese (Mn) toxicity.

Historically, liming is the most common management practice used to neutralize soil acidity. Most plants grow well in the pH range 5.5–6.5, and the usual objective of liming programs is to maintain pH in this range. Liming enhances the physical, chemical, and biological properties of soil through its direct effect in ameliorating soil acidity and through its

indirect effects in mobilizing plant nutrients, immobilizing toxic heavy metals, and improving soil physical conditions. In variable-charge soils, liming can be used as a management tool to manipulate the surface charge, thereby controlling the reactions of nutrient ions and heavy metals. Liming provides optimum conditions for a number of biological processes, including N_2 fixation and mineralization of N, phosphorus (P), and S.

Processes Generating Acidity in Soils

Processes generating acidity in soils can be broadly grouped into two categories: (1) those occurring in natural ecosystems through industrial activities, and (2) those occurring in managed ecosystems because of farming activities. The various reactions involved in these processes are given in Table 1.

The two important acid-generating processes resulting from industrial activities in natural ecosystems are acid drainage following pyrite oxidation and deposition of acidity in precipitation. While the first process occurs at a local level, the second process can lead to acidification far away from the source of acid production and hence an international effort is needed to combat it.

Acid Drainage

Acid drainage has both anthropogenic and natural origins, including mining of coal and sulfide-containing metal ores, land disturbances (e.g., rice cultivation), and industrial activities (mineral processing, manufacturing or recycling of batteries, electronic equipment, wood pulp, tanneries and textile manufacturing, and food processing).

Pyrite is commonly associated with coal and metal ores, as well as mine deltas, wetlands, and rice fields. Exposure of pyrite to the atmosphere leads to its oxidation and the production of extremely acidic drainage water. Pyrite oxidation includes biological and electrochemical (abiotic) reactions (Table 1). Abiotic oxidation of pyrite is pH-sensitive and it is extremely slow in very acidic conditions. *Thiobacillus ferrooxidans* and *T. thiooxidans* are mainly responsible for the oxidation of pyrite, especially in acid soils.

Acid Precipitation

Carbon dioxide combines with water to form a dilute solution of carbonic acid (H_2CO_3) with an equilibrium pH of approximately 5.6. For this reason, acid precipitation is arbitrarily defined as precipitation with a pH value of less than 5.6. Natural sources of acid

Table 1 Proton generation and consumption processes in the soil–plant system, including those associated with acid precipitation, pyrite oxidation, and the C, N, and S biogeochemical cycles

Process	*Reaction equation*	H^+ ($mol_c mol^{-1}$)	*Eqn no.*
Acid precipitation			
Oxidation of sulfur dioxide	$2SO_2 + O_2 \rightarrow 2SO_3$	0	1.1
Hydrolysis of sulfur trioxide	$SO_3 + H_2O \rightarrow H_2SO_4 \rightarrow SO_4^{2-} + 2H^+$	+2	1.2
Photochemical oxidation of nitric oxide	$O_3 + NO \rightarrow N_2O + O_2$	0	1.3
Hydrolysis of nitrogen dioxide	$2NO_2 + H_2O \rightarrow HNO_3 + HNO_2 \rightarrow NO_3 + H^+$	+1	1.4
Pyrite oxidation			
Pyrite oxidation by oxygen	$2FeS_2 + 7O_2 + 2H_2O \rightarrow 2Fe^{2+} + 4SO_4^{2-} + 4H^+$	+2	2.1
Ferrous iron oxidation	$4Fe^{2+} + O_2 + 4H^+ \rightarrow 4Fe^{3+} + 2H_2O$	−1	2.2
Ferric iron precipitation	$Fe^{3+} + 3H_2O \rightarrow Fe(OH)_3 + 3H^+$	+3	2.3
Pyrite oxidation by ferric iron	$FeS_2 + 14Fe^{3+} + H_2O \rightarrow 15Fe^{2+} + 2SO_4^{2-} + 16H^+$		2.4
Carbon cycle			
Dissolution of carbon dioxide	$CO_2 + H_2O \rightarrow H_2CO_3 \rightarrow H^+ + HCO_3^-$	+1	3.1
Synthesis of organic acid	Organic C $\rightarrow$ RCOOH $\rightarrow$ $RCOO^- + H^+$	+1	3.2
Nitrogen cycle			
N fixation	$2N_2 + 2H_2O + 4R{\cdot}OH \rightarrow 4R\text{-}NH_2 + 3O_2$	0	3.3
Mineralization of organic N	$R\text{-}NH_2 + H^+ + H_2O \rightarrow R\text{-}OH + NH_4^+$	−1	3.4
Urea hydrolysis	$(NH_2)_2CO + 3H_2O \rightarrow 2NH_4^+ + 2OH^- + CO_2$	−1	3.5
Ammonium assimilation	$NH_4^+ + R{\cdot}OH \rightarrow R\text{-}NH_2 + H_2O + H^+$	+1	3.6
Ammonia volatilization	$NH_4^+ + OH^- \rightarrow NH_3\uparrow + H_2O$	+1	3.7
Nitrification	$NH_4^+ + 2O_2 \rightarrow NO_3^- + H_2O + 2H^+$	+2	3.8
Nitrate assimilation	$NO_3^- + 8H^+ + 8e^- \rightarrow NH_3 + 2H_2O + OH^-$	−1	3.9
Denitrification	$4NO_3^- + 4H^+ \rightarrow 2N_2 + 5O_2 + 2H_2O$	−1	3.10
Sulfur cycle			
Mineralization of organic S	$2\text{Organic S} + 3O_2 + 2H_2O \rightarrow 2SO_4^{2-} + 4H^+$	+2	3.11
Assimilation of sulfate	$SO_4^{2-} + 8H^+ + 8e^- \rightarrow SH_2 + 2H_2O + 2OH^-$	−2	3.12
Oxidation of S^0	$2S^0 + 2H_2O + 3O_2 \rightarrow 2SO_4^{2-} + 4H^+$	+2	3.13

precipitation include geologic weathering, volcanic eruptions, anaerobic decomposition of organic matter, air-borne sea-salt sprays, and production of N oxides (NO_x) during lightning storms. The increased acid precipitation burden in recent decades has been attributed to anthropogenic sources that include combustion of fossil fuels (especially coal and oil), certain industrial processes (especially smelting of ores), exhausts from internal combustion engines, and N fertilization of agricultural and forest lands.

Widespread occurrence of acid precipitation (i.e., both wet and dry deposition) results in large part from industrial emissions of oxides of S (SO_x) and N (NO_x). These compounds are transformed in the atmosphere to sulfuric and nitric acids (**Table 1**), which can be transported over great distances before deposition on vegetation, soils, surface waters, and building structures. The average annual ratio of sulfuric to nitric acids in North America is approximately 2:1, but nitric acid is becoming progressively important relative to sulfuric acid because of the installation of flue gas desulfurization (FGD) systems in coal-fired power stations. The major impact of acid precipitation has been on freshwater ecosystems, while its impact on terrestrial ecosystems is more controversial. Forest health may be one casualty of acid precipitation, but increasing tropospheric ozone may also be a factor in forest decline.

Elemental Cycling

The most significant proton (H^+)-generating processes in soil are associated with the cycling of C, N, and S (**Table 1**) and these processes can be grouped into two main categories: plant-induced (i.e., uptake and assimilation) and soil-induced (i.e., transformation) processes.

In the case of the C cycle, the metabolism of photosynthates results in the synthesis of organic acids and, at the cytoplasmic pH of the plants (pH ≈7.2–7.4), some of the carboxyl groups of these acids dissociate to produce H^+ ions. Cytoplasmic pH regulation is generally achieved by transport of excess H^+ ions out of the cytoplasm. Some terrestrial plant species counteract the change in cytoplasm pH by excreting H^+ ions into the soil solution and, at the same time, taking in a nutrient basic cation to balance the charge.

Plants take up N in three main forms – as an anion (nitrate, NO_3^-), as a cation (ammonium, NH_4^+), or as a neutral N_2 molecule (N_2 fixation). Depending upon the form of N taken up and the mechanism of assimilation in the plant, excessive cation or anion uptake may occur. To maintain charge balance during the uptake process, H^+, OH^-, or bicarbonate (HCO_3^-) ions must pass out of the root into the surrounding soil. While the uptake of NH_4^+ and N_2 fixation result in a net release of H^+ ions, uptake of NO_3^- can result in a net release of OH^- ions. In the case of N_2 fixation, acidity is generated even when no ionic species of N are taken up by the plant. This is because basic cations are imported into the legume in exchange for H^+ ions generated during C assimilation into carboxylic acids. The amount of H^+ ion released during N_2 fixation is a function of C assimilation and therefore it depends mainly on the nature and amounts of amino acids and organic acids synthesized by the plant. When ammonium (NH_4^+) assimilation occurs in the root, deprotonation of NH_4^+ to R-NH_2 releases one mole of H^+ per mole of NH_4^+ (eqn [3.6]; **Table 1**). When plants take up NO_3^-, the NO_3^- ion is first reduced to NH_4^+, which is subsequently assimilated into amino acids. This NO_3^- reduction produces one mole of OH^- ion for every mole of NO_3^- reduced to NH_4^+ (eqn [3.9]; **Table 1**).

Sulfate (SO_4^{2-}) is assimilated into sulfur-containing amino acids (cysteine, cystine, and methionine) in the form of sulfhydryl (−SH) groups. This reduction process produces two net moles of OH^- for each mole of S assimilated (eqn [3.12]). On decomposition of sulfhydryl-containing amino acids, two moles of H^+ ions are generated for each mole of −SH oxidized to sulfate. Since plants require roughly 10 times less S than N, assimilation of SO_4^{2-} will have only a minor effect on H^+ balance in plants.

As microorganisms decompose soil organic matter, they respire CO_2, which upon hydrolysis forms H_2CO_3 (eqn [3.1]), eventually dissociating to H^+ and HCO_3^- ions (eqn [3.2]). The continuous production of CO_2 through soil and root respiration increases the concentration of CO_2 in the soil air spaces. Soil microorganisms also produce organic acids when they decompose plant litter that is rich in organic compounds but low in charge-balancing basic cations (eqn [3.2]). In general, conifer litter tends to produce more organic acids than does leaf fall from deciduous woodlands.

Nitrification and ammonia (NH_3) volatilization result in the release of H^+ ions. While the ammonification process (conversion of organic forms of N to NH_4^+-N) results in the release of OH^- ions, the subsequent oxidation of NH_4^+ to NO_3^- (nitrification) results in the release of H^+ ions. Combined ammonification (eqn [3.4]) and nitrification (eqn [3.8]) of organic N compounds, including urea, in theory generates one net mole of H^+ for every mole of N transformed. In alkaline media NH_4^+ ions dissociate into gaseous NH_3, which is subject to volatilization (eqn [3.7]), resulting in a net decrease in pH due to the consumption of OH^- ions (or release of H^+ ions) as NH_4^+ is converted to NH_3.

In aerobic soils, H^+ ions are produced during the mineralization and subsequent oxidation of S in soil organic matter (eqn [3.11]). As soil bacteria and fungi decompose plant litter and soil organic matter rich in C but low in S, soil solution SO_4^{2-} may be immobilized. In this case eqn [3.11] (**Table 1**) is reversed and it becomes an H^+-consuming reaction, because SO_4^{2-} is assimilated into microbial biomass. Under anaerobic conditions, some bacteria have the capacity to use SO_4^{2-} as a terminal electron acceptor, resulting in H^+ consumption as SO_4^{2-} is reduced along a chain of intermediate compounds to hydrogen sulfide (H_2S). The H_2S reacts with metal ions to precipitate as metal sulfides, which is an H^+-consuming process. However, when these metal sulfides are reoxidized, they generate H^+, acidifying the soil (acid drainage).

Fertilizer Reactions

Fertilizer application in ecosystems managed for agricultural production is a major contributor to soil acidification. The acidifying effects of fertilizers commonly used in agricultural production are presented in **Table 2**.

Application of N fertilizers such as urea and ammonium sulfate to soils produces H^+ by two processes: nitrification (eqn [3.8]) and NO_3^- leaching. Part of the H^+ produced is neutralized by OH^- released by plants during the subsequent uptake of the NO_3^- ions (eqn [3.9]). The depletion of basic cations (Ca, K, Mg, and Na) during the leaching of NO_3^- ions (i.e., as ion pair) and product removal (i.e., silage) accelerates the acidification process. In the case of ammonium sulfate, the concomitant leaching of SO_4^{2-} and NO_3^- ions causes greater depletion of basic cations. With urea, the initial conversion of amide N (R-NH_2) to NH_4^+ (ammonification) releases OH^- (eqn [3.5]), which neutralizes part of the H^+ produced during the subsequent oxidation of NH_4^+ to NO_3^-, explaining why urea-based N fertilizers are less acidifying than the NH_4^+-based fertilizers (**Table 2**).

Superphosphates, the most common phosphate fertilizers, contain monocalcium phosphate (MCP) as the principal P component. Dissolution of MCP in soils results in the formation of dicalcium phosphate with a release of phosphoric acid, which subsequently dissociates into phosphate anions ($H_2PO_4^-$) and H^+. Part of H^+ is subsequently neutralized by the OH^- released during the adsorption of the $H_2PO_4^-$ by soil particles. Compared with weakly adsorbred SO_4^{2-} and NO_3^- ions, $H_2PO_4^-$-induced leaching of basic cations is unlikely to occur, since the $H_2PO_4^-$ ions are strongly adsorbed by most soils.

In legume-based pasture and crop-production systems, P fertilizers are added to promote N_2 fixation by the legumes. Regardless of P fertilizer source, application of P to legume-based systems promotes N_2 fixation, thereby indirectly causing soil acidification.

Table 2 Acidity production by various fertilizers and its effect on pH of two soils with contrasting buffer capacities

Fertilizer	*Acidity equivalence*[a]	*Acidity produced (kmol H^+ ha^{-1})*[b]	*Number of years required to reduce soil pH by 1 unit*[c]	
			Tokomaru silt loam	*Egmont clay loam*
Ammonium sulfate	110	2.60	8	26
Ammonium chloride	93			
Ammonium nitrate	60			
Diammonium phosphate	74	2.06	10	33
Monoammonium phosphate	55			
Urea	79	0.86	25	78
Potassium nitrate	−23			
Calcium nitrate	−50			
Sodium nitrate	−29			
Nitrogen fixation	70–250	–	–	–
Single superphosphate	8	0.48	45	140
Triple superphosphate	15	0.50	43	135
North Carolina phosphate rock	−50			
Calcium sulfate	−57			
Potassium sulfate	−64			
Elemental sulfur (S^0)	310	1.55	14	43

[a]Acidity equivalence is the number of parts by weight of pure lime (calcium carbonate) required to neutralize the acidity caused by 100 parts of the fertilizer. Negative values indicate the liming value (kilograms $CaCO_3$/100 kg) of the fertilizer.
[b]Ammonium sulfate, diammonium phosphate, and urea added at the rate of 25 kg N ha^{-1} $year^{-1}$; single and triple superphosphate at the rate of 30 kg P ha^{-1} $year^{-1}$; and S^0 at the rate of 30 kg S ha^{-1} $year^{-1}$.
[c]pH-buffering capacity (kilomoles H^+ per hectare) of 21.7 and 67.5 for the Tokomaru and the Egmont soil, respectively.

The amount of acidity produced indirectly by N_2 fixation depends mainly on the extent of NO_3^- leaching and it can be greater than that produced directly by the dissolution of MCP in superphosphate fertilizer granules (Table 2).

Elemental sulfur (S^0) is used as an acidifying agent, as a slow-release S fertilizer or, in a finely divided form, as a fungicide. When S^0 is added to soils, it is oxidized to sulfuric acid, which dissociates into SO_4^{2-} and H^+ ions (eqn [3.13]).

Measurement and Effects of Soil Acidity

The acidity of a soil is assessed from the activity of H^+ ions in the soil solution. Since there is a huge range of $[H^+]$ in soil solution, a logarithmic scale (known as pH scale; $pH = -\log [H^+]$) is used to quantify acidity. In soils, acidification is indicated by a decrease in either pH or anion neutralizing capacity. For a given input of acidity, the extent to which pH decreases depends mainly on the pH buffering capacity of the soil. Various soil constituents, such as organic matter, Fe and Al oxides, and $CaCO_3$ (in calcareous soils) contribute to the pH buffering of soils at different pH values. Depending on organic matter content, texture, and the nature of the clay mineralogy, the amount of acidity needed to reduce pH of topsoil (noncalcareous) by 1 unit is normally in the 1- to 8-cmol (H^+) kg^{-1} range. These values translate to buffer capacities of approximately 10–70 kmol (H^+) ha^{-1} for the top 7.5 cm soil layer. The amounts of acidity produced by selected fertilizers and the number of years required to reduce soil pH by 1 unit are shown in Table 2 for two soils with contrasting pH buffer capacities. It is clear that annual application of ammonium sulfate or ammonium phosphates could acidify soil within a short time (10 years), particularly when soil pH-buffering capacity is low.

Acidification affects the transformation and biogeochemical cycling of both nutrients and heavy metals through its effect on the physical, chemical, and biological characteristics of soils. Soil pH can be viewed as the master variable of all the driving factors, because it can affect the surface charge and subsequent adsorption of solutes by variable-charge soil components. In addition to its effect on the sorption of metal cations and anions in soils, it also influences metal speciation, complexation of metals with organic matter, precipitation/dissolution reactions, redox reactions, mobility and leaching, dispersion of colloids and, ultimately, the bioavailability of trace metals.

Soil pH affects the surface charge through the supply of H^+ for adsorption on to the metal oxides and the dissociation of the functional groups in the soil organic matter. A decrease in pH decreases the net negative charge (often referred to as cation exchange capacity or CEC) and increases the net positive charge (often referred to as anion exchange capacity or AEC).

Because acidity governs the type, number, and activity of microorganisms in soil, it influences the rate of organic matter mineralization and availability of elements such as N, S, and P. Because of suboptimal microbial activity in highly acid conditions, organic matter accumulates, giving rise to a storehouse of nutrients that can be exploited by liming. Likewise, nitrification is markedly reduced below pH 6, whereas the ammonification reaction is relatively insensitive to acidity, resulting in the accumulation of NH_4^+ in acidic soil where nitrification is inhibited. Thus, for certain crops unable to use NH_4^+, acidification can result in restricted uptake of N and it may promote NH_4^+ toxicity.

Acidity has a deleterious effect on the symbiotic relationship between *Rhizobia* and legumes, and nodulation and N_2 fixation are generally poor below pH 6. Physiological reasons for this include: (1) inhibition of rhizobial infection of legume roots, decreasing nodule formation; (2) inhibition of nitrogenase enzyme activity in the nodule due to modification of the nitrogenase iron protein; (3) decrease in bacterial membrane potential and the inhibition of the leghemoglobin; (4) decrease in the supply of photosynthate to the rhizobium due to the poor supply of major nutrients, such as P; and (5) poor supply of Mo and Ca, which are essential for N_2 fixation.

A decrease in soil pH will increase the concentrations of Fe and Al in soil solution and this may lead to adsorption/precipitation of phosphate. In variable-charge soils, a decrease in pH increases the AEC, increasing the retention of phosphate. However, at very low pH, solubilization of P compounds may result in an increase in the concentration of P in soil solution. Under acid conditions, weathering liberates K from micaceous and feldspar minerals, enabling it to enter the soluble and exchangeable pools; but, in variable-charge soils, increasing acidity decreases CEC, reducing the ability of the soil to retain K, and this may cause K to be more prone to leaching.

In many soils, organic matter is the main source of S and, since mineralization of organic matter may be slowed by acidity, the release of S for plant uptake may decline when soil is acidified. Furthermore, in highly weathered acid soils, SO_4^{2-} may be adsorbed by sesquioxide surfaces, precipitated as Al-OH-SO_4-type minerals, such as alunite and basulminite, and/or held as an exchangeable anion on positively charged sites on sesquioxides.

One of the major consequences of acidification is a decline in basic cations such as Ca^{2+} and Mg^{2+},

leading to potential deficiency of these cations for plant growth. Furthermore, at low pH, the bioavailability of Ca may be restricted due to antagonistic effects of soluble Al. With increasing soil acidification, smaller amounts of Mg^{2+} remain in exchangeable form due to reduction in negative charge. Since Mg^{2+} is a poor competitor with Al^{3+} and Ca^{2+} for the exchange sites, it tends to accumulate in the solution phase and is therefore more prone to leaching.

Acidification affects the transformation of metal ions through: (1) changes in surface charge in variable-charge soils; (2) changes in metal speciation; and (3) shifts in the oxidation–reduction reactions of metals. In general, the solubility and phytoavailability of metals such as Cu and Zn are inversely related to soil pH. Since Cu is largely complexed with humic substances, the greater solubility of humic compounds at high pH may result in more soluble Cu when pH is raised, though this organically complexed Cu may not have low bioavailability. One of the major consequences of soil acidification is the increase in concentrations of soluble Al and Mn, both of which are highly toxic to plant growth. A primary aim of liming soils for agricultural production is to decrease the concentration of these elements. While Mn toxicity is directly related to the metabolic requirements of plants, the effect of Al toxicity appears to be largely manifested as malformation and malfunction of the root system, a syndrome which is exacerbated by low levels of solution Ca in acid soils.

Under acid conditions, boron (B) occurs in soil solution as the uncharged H_3BO_3 molecule and thus it is readily available in acid soil. Unlike other anions, Mo is highly insoluble in low pH conditions and it may become limiting when soil is acidified.

Cadmium (Cd) has been identified as a potential human health hazard, which enters the food chain through plant uptake. Adsorption of Cd^{2+} by soil decreases with decreasing pH for the following reasons: (1) in variable-charge soils, a decrease in pH results in a decrease in cation adsorption and (2) a decrease in soil pH is likely to result in an increase in free Cd^{2+} at the expense of the more strongly adsorbed hydroxy-cadmium species. In general, Cd uptake by plants increases with decreasing pH and consequently it is recommended that soil pH be maintained at pH 6.5 or greater in land receiving Cd-rich biosolids.

The effect of soil acidity on the adsorption of metalloids such as arsenic (As) and selenium (Se) is controlled by two interacting factors – pH-related changes in negative potential in the plane of adsorption and pH-induced changes in amount of negatively charged ionic species present in soil solution. While the first factor results in a decrease in As(V) adsorption when pH is raised, the latter factor is likely to cause the opposite effect. Thus, the effect of pH on As(V) adsorption depends on the nature of the mineral surface. In soils with low oxide content, increasing the pH will have little effect on As(V) adsorption, while in highly oxidic soils, adsorption will decrease with increasing pH.

Amelioration of Soil Acidity

Three approaches can be taken to minimize the rate and impacts of soil acidification: (1) reduce the amount of H^+ ions generated, (2) minimize the extent to which the H^+ and OH^- ions generating processes are uncoupled, and (3) neutralize the acidity produced. In managed ecosystems, the rate of acid generation can be altered by applying nutrients in forms that produce less acidity, selecting plant species that accumulate less cation, and reducing the loss of C, N, and S from soil.

Traditionally liming has been the most common practice used to alleviate soil acidification. A range of liming materials is available, including calcite ($CaCO_3$), burnt lime (CaO), slaked lime ($Ca(OH)_2$), dolomite ($CaMg(CO_3)_2$), and slag ($CaSiO_3$). Recently, the liming potential of other Ca-containing compounds has been evaluated. Some of these materials include phosphate rocks (PRs), FGD gypsum, fluidized bed boiler ash, fly ash, and lime-stabilized organic composts. The amount of liming material required to rectify soil acidity depends on the neutralizing value of the liming material and pH-buffering capacity of the soil.

Unlike soluble P fertilizers, PRs can have a liming value, because they contain some free $CaCO_3$, which itself can act as a liming agent and, secondly, the dissolution process of the P mineral component (i.e., apatite) consumes H^+, thereby reducing the soil acidity. While the free $CaCO_3$ in PRs dissolves reasonably fast, providing a small, immediate liming effect, the apatite generally dissolves at a slower (but variable) rate and has liming value over a longer period of time.

Alkaline-stabilized biosolids are increasingly being used as agricultural lime substitutes and soil amendments. Alkaline stabilization of biosolid utilizes a combination of high pH, heat, and composting to kill pathogens and stabilize organic matter. A range of alkaline materials are used for this purpose, including cement-kiln dust, lime-kiln dust, lime, limestone, alkaline coal fly ash, FGD, other coal-burning ashes, and wood ash. To minimize metal mobility and

bioavailability in biosolid-amended soils, the US Environmental Protection Agency (USEPA) recommends the application of alkaline-stabilized biosolids and other liming agents to increase the soil pH to 6.5 or more.

The primary purpose of liming arable soils is to overcome the chemical problems associated with soil acidity that include high concentrations of acid ions (H^+ and Al^{3+}) and toxic elements (Mn^{2+}), and low concentrations of basic cations (Ca and Mg) and other nutrient ions such as Mo and P. The hydrolysis of the basic cations in lime produces OH^- ions, which neutralize the H^+ ions, thereby decreasing the activity and bioavailability of Al and Mn. Liming also increases the solubility of Mo and P, thereby increasing their availability. Lime provides the basic nutrient cations (Ca^{2+} and Mg^{2+}) and also reduces the solubility of heavy metals, thereby minimizing their bioavailability and mobility in soils.

Further Reading

Adams F (1984) *Soil Acidity and Liming*. Madison, WI: Soil Science Society of America Publishing.

Evangelou VP (1995) *Pyrite Oxidation and its Control*. New York, NY: CRC Press.

Haynes RJ (1984) Lime and phosphate in the soil–plant system. *Advances in Agronomy* 37: 249–467.

Longhurst JWS (1991) *Acid Deposition: Origin, Impacts and Abatement Strategies*. New York, NY: Springer-Verlag.

Marschner H (1995) *Mineral Nutrition of Higher Plants*, 2nd edn. London, UK: Academic Press.

Rengel Z (ed.) (2003) *Handbook of Soil Acidity*. New York, NY: Marcel Dekker.

Robson AD (ed.) (1989) *Soil Acidity and Plant Growth*. New York, NY: Academic Press.

Ulrich B and Sumner ME (eds) (1991) *Soil Acidity*. New York, NY: Springer-Verlag.

Wright RJ, Baligar VC, and Murrmann RP (eds) (1990) *Plant–Soil Interactions at Low pH*. Dordrecht, The Netherlands: Kluwer Academic Publishers.

AERATION

D E Rolston, University of California–Davis, Davis, CA, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

In a general sense, aeration is the interchange of various gases between the atmosphere and the Earth and the various reactions that either consume or produce gases in the soil. The interchange results from concentration gradients established within soil by respiration of microorganisms and plant roots, by production of gases associated with biological reactions such as fermentation, nitrification, and denitrification, reduction–oxidation reactions of soil chemicals, and by soil incorporation of materials such as fumigants, anhydrous ammonia, pesticides, and various volatile organic chemicals from toxic waste sites. The two major gases associated with aeration are oxygen (O_2) and carbon dioxide (CO_2), where O_2 moves from the atmosphere to soil and is consumed, and CO_2 is produced in soil and moves from the soil to the atmosphere. Figure 1 indicates the general direction of the flow of gases within soil profiles. Soil aeration has been reviewed extensively over the years.

Soil-Air Composition

The amount of air or soil-air content is directly related to the bulk density of the soil and the amount of water in the soil profile. The bulk density of natural soil varies from approximately 1.0 $Mg\,m^{-3}$ to 1.7–1.8 $Mg\,m^{-3}$. Thus, the relative amount of void or pore space in the soil varies between approximately 30 and 60%. The soil pores or voids can be filled with either air or water. Therefore the soil-air content or air-filled porosity can vary between approximately 30 and 60%.

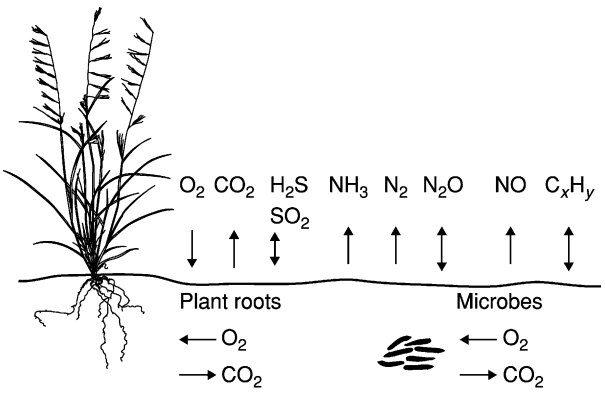


Figure 1 Schematic indicating the general flow directions of important gases within soil.

The composition of soil air depends on the relative magnitude of both the sources and the sinks of the various gas components, the interchange between soil air and atmospheric air, and the partitioning of the gases between the gaseous, liquid, and solid (mineral and organic matter) phases of the soil. If a soil were completely 'aerated' the concentrations of the gases in the soil air would be similar to that in the atmosphere. Oxygen concentrations in the soil air will be somewhat below that in the atmosphere (approximately 20% by volume), since O_2 is consumed in soil by plant root and microbial respiration and through chemical reactions. Under some conditions, O_2 concentrations can fall to zero and the soil becomes anaerobic (anoxic). It is now widely accepted that under some conditions soil profiles do not have to be either fully aerated or fully anaerobic but may be partially aerobic and partially anaerobic. Anaerobic pockets or 'hot spots' may exist within the soil due to pockets of very high O_2 consumption such as around incorporated carbon materials and/or due to very slow diffusion to regions of O_2 consumption. For example, the interior of large aggregates may be anaerobic for these reasons.

CO_2 concentrations in the soil air can be as high as 10 times more than in the atmosphere (0.036% by volume). Since nitrogen gas (N_2) is more abundant than other gases in the atmosphere (approx. 78%) and there are generally no sources or sinks for N_2 in the soil (except N_2 absorbed during nitrogen fixation or produced during denitrification), the concentration of N_2 in the soil air will be similar to that in the atmosphere, varying only slightly depending on the production and consumption of other soil gases. The soil air will also contain varying amounts of nitric oxide and nitrous oxide (from nitrification and denitrification); methane, hydrogen sulfide, and ethylene (from anaerobic processes); water vapor; and trace amounts of inert gases such as argon (**Figure 1**). Human activities also result in the accidental or intentional introduction of gases in the soil profile such as fumigants, anhydrous ammonia, pesticides, and various volatile organic chemicals that exist partially in the vapor phase.

Gas Exchange Mechanisms

Diffusion

Diffusion is considered to be the principal mechanism in the exchange of gases between the soil and the atmosphere. The diffusion velocities of gas mixtures in soil are related to each other in a complex manner dependent upon the mole fraction of each gas, the molar fluxes of each gas, and the binary diffusion coefficient of each gas pair. General equations for steady transport of a multicomponent mixture of gases have been developed based on gas kinetic theory. If gravity effects are ignored or diffusion occurs only horizontally, the well-known Stefan–Maxwell equations provide the theoretical framework for diffusion of gases in soils. Ficks law is generally applicable for only a few special cases. One of these cases is for the diffusion of a trace gas in a binary mixture, meaning that the mole fraction of the tracer gas is small. Since the binary diffusion coefficients of N_2 in air and O_2 in air are very similar and CO_2 may be considered to exist in trace amounts, the diffusion of the two major gases associated with aeration may come under this case. Diffusion of some of the other gases existing in soil may not meet this criterion, however.

Assuming that the special case conditions are met, Ficks law is given by:

$$\frac{M_g}{At} = f_{g,d} = -D_p \frac{dC_g}{dx} \qquad [1]$$

where M_g is the amount of gas diffusing (kilograms of gas), A is the cross-sectional area of the soil (square meters of soil), t is time (seconds), $f_{g,d}$ is the gas flux density (kilograms of gas per square meter of soil per second) due to molecular diffusion, C_g is concentration in the gaseous phase (kilograms of gas per cubic meter of soil air), x is distance (meters of soil), and D_p is the soil-gas diffusion coefficient (cubic meters of soil air per meter of soil per second).

The soil-gas diffusion coefficient is the main variable controlling the degree of soil aeration. It is highest when the soil is dry and approaches zero as the soil becomes very wet or saturated. The D_p has been related both empirically and theoretically to the soil-air content by a number of authors. Since discrepancies between measured soil-gas diffusion coefficients stemming from eqn [1] and those calculated from the many empirical equations occur, it is often desirable to measure the soil-gas diffusion coefficient for particular situations. Laboratory methods using soil cores and field methods for measuring soil-gas diffusion coefficients have been developed.

Convection

In addition to molecular diffusion processes, soil gases may also exchange with the atmosphere through convection (advection). Convective flow of gases means that the whole air parcel is moving through soil pores due to a pressure difference (gradient) between the soil and the atmosphere. Pressure gradients may develop due to barometric pressure changes in the atmosphere; wind blowing across the soil surface or against a hill or other landscape

feature; infiltration of rainfall or irrigation water into the soil, soil-water redistribution, and evaporation; temperature differences across the upper part of the soil profile; and density differences due to high concentrations of gases that have densities much different from air.

Convective gas flow is described by a form of Darcy's Law:

$$f_{g,c} = -\frac{K_a}{\mu}\left[\frac{dP}{dx}\right] \quad [2]$$

where $f_{g,c}$ is the flux density of gas due to convection (cubic meters of gas per square meter of soil per second), K_a is the air permeability (cubic meters of gas per meter of soil), μ is the viscosity of the gas mixture (pascal-seconds), P is pressure (pascals), and x is distance (meters of soil).

Air permeability is strongly influenced by soil-water content. Air permeability is a maximum in dry soil and decreases as the soil becomes wet, until it reaches zero at saturation. This is caused by progressive blockage of the soil pores by water. Several field and laboratory methods for measuring K_a have been developed, involving either steady-state or non-steady-state flow, though steady-state measurements of gas permeability are preferred. Most field methods are based on the same principles as the laboratory methods, i.e., they involve measuring the flow rate of air through a soil column under known pressure differences across the column.

Convective processes occur rapidly and sporadically. Thus, it is very difficult to observe, measure, and predict the gas exchange that occurs by convection. During infiltration into soil, there is ample evidence that air pressure increases ahead of the water wetting front, and convection of gases occurs. It is often assumed that diffusion is the main gas exchange process overall, because it is operating continuously, whereas convection occurs episodically. Of the convective processes, rainfall and irrigation may contribute the most to aeration of soils, with estimates that rainfall may account for 7–9% of total aeration. Adequately quantifying these processes and being able to predict and model convective flow processes is a goal yet to be fully attained.

Gas Reactions

Respiration

O_2 is continuously consumed and CO_2 produced by plant roots and by soil microorganisms. Even plants such as rice that grow best in water-submerged soil transport O_2 from the leaves to the roots for respiration. The rates of O_2 consumption and CO_2 production are directly related to the rate of plant and microbial growth, which in turn is related to several environmental factors, including air and soil temperature, substrate availability, and soil moisture. For aerobic conditions, the amount of CO_2 produced and O_2 consumed tends to be about equal. For anaerobic conditions, the CO_2 production will tend to be larger than the O_2 consumption because other reactions are occurring.

The concentrations of O_2 and CO_2 that occur in the soil pore space vary widely, especially for CO_2, and depend on the rate of consumption and production and upon the rate that the soil is able to exchange these gases between the soil and the atmosphere through diffusion and convection. The diurnal and annual variability in the soil-gas concentrations are generally much greater for clayey soils than for sandy soils owing to the ability of sandy soils to transmit gases at a higher rate (larger soil-gas diffusion coefficients and air permeabilities) and maintain more constant concentrations.

Oxidation–Reduction Processes

Oxidation and reduction are connected with the transfer of electrons from soil organic matter (or organic contaminants) to oxidized inorganic compounds catalyzed by enzymes produced by soil microorganisms. For well-aerated conditions (aerobic), O_2 is the electron acceptor. When O_2 becomes limiting (anaerobic), other substances will accept electrons or be reduced. Examples of compounds that can be reduced under anaerobic conditions are nitrate (denitrification), manganic manganese, ferric iron, sulfate, and perchlorate (a natural and anthropogenic contaminant). The reduction of nitrate and sulfate results in N_2 (and N_2O) and hydrogen sulfide gases, respectively. Another gas produced from reduction processes is methane (CH_4). Both CH_4 and N_2O are strong greenhouse gases and contribute to global warming. In waterlogged or very wet soils and sediments that are unable to transmit O_2 sufficiently fast through the profile, O_2 is the first to disappear, followed by nitrate and then sequentially by the reduction of manganese, iron, and sulfate. Perchlorate is reduced at about the same point as nitrate. Several toxic organic chemicals also undergo redox reactions that greatly affect the kind and toxicity of the reaction products.

Production and Consumption of Other Gases

There are a few gases produced in soils that are not necessarily associated with redox reactions. When fertilizer materials such as urea and ammonium salts are applied to the soil, reactions can occur, produce

ammonia gas (NH_3), particularly in soils with a pH of more than about 7.5 or 8, that fairly large emissions of NH_3 can occur, for instance, if urea is applied to the soil surface, since urea hydrolysis results in a large increase in pH near the fertilizer granules, with NH_3 being emitted. The deeper the material is placed, the less NH_3 will be emitted. Ammonia may also be produced in soil from the incorporation of animal waste.

Under aerobic conditions, ammonium-based materials either from fertilizer or mineralization of organic materials will be oxidized to nitrite and then to nitrate by microbial processes (nitrification). During nitrification, some N_2O and nitric oxide (NO) can be produced and emitted to the atmosphere. Nitric oxide enters into the tropospheric ozone cycle and can contribute to very small particle (less than 10 μm) generation in the atmosphere.

Besides O_2, gases including hydrocarbons, N_2O, NO, CH_4, and some sulfur gases may move from the atmosphere into the soil and be consumed by biological processes. Thus, the soil may act as a significant sink for atmospheric gases under some circumstances.

Aeration Requirements

Plants

The plant response to inadequate aeration or lack of O_2 is highly dependent upon the plant species, stage of growth, and upon several soil and environmental conditions such as temperature, water relations, and occurrence of toxic by-products of anaerobic conditions. The response of plants to inadequate aeration may be due to either direct or indirect effects. The direct effect is because of the lack of oxygen for root respiration. The indirect effect is due to changes in redox conditions that affect nutrient and water availability, soil pH, buildup of toxic concentrations of metabolites and metals, and the viability of pests and diseases.

The direct effect of lack of O_2 or slow diffusion of O_2 to plant roots has been characterized by a measurement called the oxygen diffusion rate (ODR). The ODR is measured by placing a cylindrical platinum electrode into the soil. Oxygen is reduced at the electrode surface and creates a current that can be measured. Diffusion of O_2 to the water-covered electrode is meant to mimic the diffusion through the water film around roots. Many studies have attempted to relate the ODR to plant response. ODR values smaller than approximately 0.2 μg cm^{-2} min^{-1} indicate potentially poor aeration for many plant species. Values of more than approximately 0.4 μg cm^{-2} min^{-1} are indicative of relatively good aeration for growth of most plants.

The plant symptoms of poor aeration are poor root growth or death, negative geotropism of roots, depressed shoot growth, wilting, leaf senescence, abortion of flowers, and termination of shoot apex growth. Poor aeration may also result in decreased transpiration; accumulation of ethylene, ethanol, and other metabolites; and decreased nutrient uptake. Reduced conditions in soil may also result in increased solubility of some chemicals that are toxic to plants.

Remediation of Contaminated Soils

With the large use of chemicals in modern society, both inorganic and organic chemicals, either intentionally or accidentally, end up in soil as contaminants with varying degrees of toxicity. Large amounts of petroleum products end up contaminating soil, both as point and nonpoint sources of pollution. The hydrocarbons in crude petroleum include alkanes, cycloalkanes, aromatics, polycyclic aromatics, asphaltines, and resins. In addition, chlorinated solvents, pesticides, detergents, metals, and other kinds of chemicals may pollute soil. Microbial processes in soil can degrade many of the organic compounds, but the biodegradability of various compounds is greatly influenced by their physical state and toxicity, as well as soil environmental conditions, including soil water content, soil pH, temperature, levels of inorganic nutrients, levels of electron acceptors, and aeration. Organic contaminants provide a source of carbon to microorganisms or they provide electrons, which the organisms can use to obtain energy. Metal contaminants may undergo redox reactions affecting their solubility and toxicity, which are also dependent upon the aeration status of the soil.

In most cases, biodegradation of organic contaminants depends on the activities of aerobic organisms. Thus, the presence of an adequate supply of O_2 in the soil is essential for biodegradation or bioremediation to occur. In general, it takes 2–3 kg of O_2 to degrade 1 kg of petroleum hydrocarbon. On the other hand, some compounds like the highly chlorinated chemicals are not easily degraded and may be broken down more effectively under anaerobic conditions. Some chemicals, perchlorate for instance, are only broken down under anaerobic conditions. Table 1 gives a list of some chemicals, indicating whether they are degradable by aerobic or anaerobic processes. For more information on biodegradation and bioremediation of contaminated soils, *see* **Pollutants:** Biodegradation in this encyclopedia.

Summary

Aeration is the interchange of various gases between the atmosphere and soil and the various reactions that

Table 1 Organic chemicals and their biodegradability

Chemical class	*Examples*	*Biodegradability*
Aromatic hydrocarbons	Benzene, toluene	Aerobic and anaerobic
Ketones and esters	Acetone, methylethyl ketone	Aerobic and anaerobic
Petroleum hydrocarbons	Fuel oil	Aerobic
Chlorinated solvents	Trichloroethylene, perchloroethylene	Aerobic (methanotrophs), anaerobic (reductive dechlorination)
Polyaromatic hydrocarbons	Anthracene, benzo[a]pyrene, creosote	Aerobic
Polychlorinated biphenyls	Arochlors	?
Organic cyanides		Aerobic

Adapted from Baker KH and Herson DS (1994) *Bioremediation.* New York: McGraw-Hill. © 1994 with permission.

either consume or produce gases in the soil. The composition of soil air depends on the relative magnitude of both the sources and sinks of the various gas components, the interchange between soil air and atmospheric air, and the partitioning of the gases between the gaseous, liquid, and solid (mineral and organic matter) phases of the soil. The two major gases associated with aeration are O_2 and CO_2, where O_2 moves from the atmosphere to soil and is consumed by plant roots and microorganisms and CO_2 moves from the soil, where it is produced by plant and microbial respiration, to the atmosphere. In addition to root and microbial respiration, O_2 may also be consumed by reaction with metals and other compounds. The two transport mechanisms that result in aeration of soils are molecular diffusion and convection. Although convection can result in significant transport under certain situations, such as infiltration of rainfall or irrigation water into soil, diffusion is considered to be the dominant mechanism of exchange over the long term. The plant symptoms of poor aeration are poor root growth or death, negative geotropism of roots, depressed shoot growth, wilting, leaf senescence, abortion of flowers, and termination of shoot apex growth. Aeration is also important for the soil's ability to degrade pollutants. In most cases, biodegradation of organic contaminants depends on the activities of aerobic organisms. Thus, the presence of an adequate supply of O_2 in the soil is essential for biodegradation or bioremediation to occur. On the other hand, some chemicals will only degrade under anaerobic conditions. Knowledge of a chemical's biodegradability and whether the chemical will degrade under aerobic or anaerobic conditions plays a major role in design of effective bioremediation schemes.

See also: **Anaerobic Soils**; **Carbon Emissions and Sequestration**; **Diffusion**; **Greenhouse Gas Emissions**; **Hydrocarbons**; **Oxidation–Reduction of Contaminants**; **Pollutants:** Biodegradation; **Remediation of Polluted Soils**; **Vadose Zone:** Hydrologic Processes

Further Reading

Baker KH and Herson DS (1994) *Bioremediation*. New York: McGraw-Hill.

Dane JH and Topp GC (eds) (2002) *Methods of Soil Analysis*, part 4, *Physical Methods*. Soil Science Society of America Book Series 5. Madison, WI: Soil Science Society of America.

Eweis JB, Ergas SJ, Chang DPY, and Schroeder ED (1998) *Bioremediation Principles*. Boston: McGraw-Hill WCB.

Gerstl Z, Chen Y, Mingelgrin U, and Yaron B (eds) (1989) *Toxic Organic Chemicals in Porous Media*. New York: Springer-Verlag.

Glinski J and Steniewski W (1985) *Soil Aeration and its Role for Plants*. Boca Raton, FL: CRC Press.

Hillel D (1998) *Environmental Soil Physics*. San Diego, CA: Academic Press.

Jury WA, Gardner WR, and Gardner WH (1991) *Soil Physics*. New York: John Wiley.

Scott HD (2000) *Soil Physics, Agricultural and Environmental Applications*. Ames, IA: Iowa State University Press.

AGGREGATION

Contents
Microbial Aspects
Physical Aspects

Microbial Aspects

S D Frey, University of New Hampshire, Durham, NC, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Aggregates are the basic unit of soil structure, consisting of primary particles (sand, silt, and clay), organic materials in various stages of decay, and living organisms all bound together in clusters ranging in size from less than 2 μm to greater than 2 mm. The size, arrangement, and stability of aggregates is of primary importance in determining many soil physical, chemical, and biological properties, including soil water and air relations, microbial activity, the turnover of soil organic matter, and the release of plant-available nutrients. Good soil structure depends on the presence of aggregates that remain stable when disturbed by wetting–drying and freezing–thawing cycles and that contain a range of pore sizes. There should be sufficient small pores (0.2–30 μm) to retain water for plant and microbial growth, yet enough large pores (>30 μm) to promote rapid oxygen diffusion, water infiltration, and drainage.

The activities of soil organisms are, to a large degree, responsible for the formation and stabilization of soil aggregates. During the breakdown of organic matter, decomposer organisms produce 'microbial glues' that bind soil particles together into aggregates of various sizes and stabilities, depending on the type and persistence of the binding agent. At the same time, soil structure regulates the movement and growth of soil organisms by affecting the size and degree of continuity of the soil pores within which soil organisms live. Thus there are strong interactions and feedbacks between soil organisms and soil structure. Research related to soil aggregation has significantly increased in the past decade as it has become increasingly clear that integrating soil structure into microbiological and ecologic studies is necessary to understand fully how soils function and how they can be managed sustainably.

Microorganisms and Aggregate Formation

The improvement of soil physical conditions brought about by the addition and incorporation of organic materials has been appreciated for centuries. However, organic matter additions have little effect on aggregation, unless microorganisms are present and actively decomposing the added materials. Fungi physically entangle soil particles in their hyphal networks, and both fungi and bacteria produce extracellular polysaccharides and other by-products of growth that cement soil particles together.

Studies with pure and mixed cultures of soil microorganisms indicate that a wide range of bacteria, actinomycetes, and fungi influence the formation and stabilization of aggregates in the presence of a suitable carbon source. The important role that microorganisms play in aggregate formation has been substantiated in a number of ways. Some of the earliest studies confirmed that aggregation could be brought about by adding microbial cells or the products of microbial biosynthesis to soil. For example, soil aggregation can be increased by adding bacterial or fungal extracellular polysaccharides that have been isolated from pure cultures or from soil. Another approach has been to observe changes in aggregation after soil has been treated with sodium periodate, a chemical that selectively oxidizes polysaccharides. Periodate treatment has been reported in numerous studies to reduce the quantity of stable aggregates, suggesting that microbial extracellular polysaccharides play a key role in aggregate stabilization.

When soil is sieved (e.g., <250 μm) to remove large aggregates, amended with organic matter, and incubated in the laboratory under optimal moisture and temperature conditions, a significant number of new aggregates form rapidly, reaching a maximum within days to a few weeks. These newly formed aggregates often remain stable for a considerable period of time, even while the amount of microbial biomass declines (Figure 1), providing further support that by-products that remain after microbial growth subsides act as stabilizing agents. If a bacteriocide or fungicide

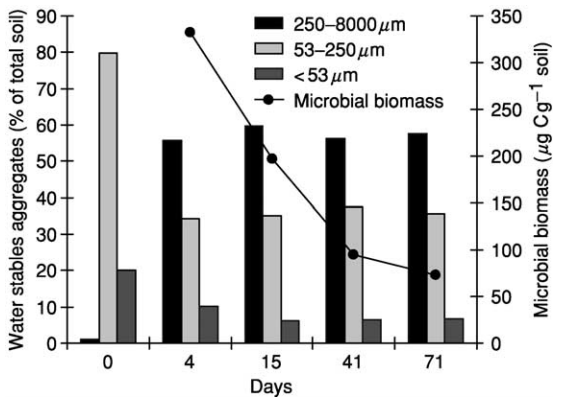


Figure 1 Percentage of water-stable aggregates following amendment of microaggregates (53–250 μm) with starch and incubation for 71 days. The closed circles represent the amount of microbial biomass in the macroaggregates (250–8000 μm) that formed during the incubation. Data from Guggenberger G, Elliott ET, Frey SD, Six J, and Paustian K (1999) Microbial contributions to the aggregation of a cultivated grassland soil amended with starch. *Soil Biology and Biochemistry* 31: 407–419.

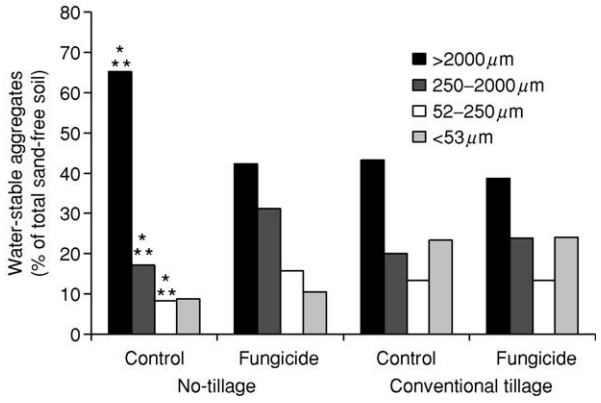


Figure 2 Effects of a fungicide on the percentage of water-stable aggregates in surface soils (0–5 cm) of no-tillage and conventional tillage plots at the long-term tillage comparison experiment in Horseshoe Bend, Georgia. Asterisks indicate significant differences ($P < 0.05$) among aggregate-size classes within a tillage treatment (**) and between tillage treatments within an aggregate-size class (*). Data from Beare MH, Hu S, Coleman DC, and Hendrix PF (1997) Influences of mycelial fungi on soil aggregation and organic matter storage in conventional and no-tillage soils. *Applied Soil Ecology* 5: 211–219.

is applied to the soil to inhibit bacterial or fungal growth, aggregate formation is suppressed. When already well-aggregated field soils are treated with a fungicide to inhibit fungal activity and growth, the level of aggregation is significantly reduced, especially in soils where fungi contribute significantly to the total microbial biomass, such as in no-tillage agroecosystems (Figure 2).

Aggregates are subjected to a variety of forces such as freeze–thaw and wetting–drying cycles that can lead to aggregate disintegration. Rapid wetting of dry aggregates, which occurs frequently at the soil surface during rain or irrigation events, can be particularly damaging. As water enters the soil pores, air becomes entrapped in the interior of the aggregate, leading to a pressure buildup which can cause the aggregate to fall apart (i.e., slake). Stable aggregates resist slaking and are thus termed 'water-stable.' Microbial exudates may influence aggregate stability by enhancing the ability of aggregates to withstand the pressure caused by slaking and/or by increasing the water repellency of the aggregates such that water entry into dry aggregates is significantly reduced or at least sufficiently slowed to allow entrapped air to escape.

Stable aggregates are not a random arrangement of soil particles, but are formed by a complex process involving a variety of binding mechanisms interacting across a range of spatial scales. According to the aggregate hierarchy concept, aggregate formation and stabilization occur in four distinct stages, with each stage dependent on a different type of organic binding agent and resulting in aggregates of increasing size (Table 1). At the lowest level of organization, individual clay plates flocculate into stable particles <2 μm in diameter in response to van der Waal's forces and hydrogen bonding. The presence of persistent binding agents consisting of inorganic and organic coatings on the clay surfaces may enhance attraction between clay plates and stabilize the particles against dispersion. These flocculated particles attach to and encrust living bacterial cells, fungal hyphae, and microbial debris (dead cells and extracellular products), forming stable particles of 2–20 μm in diameter. Electron

Table 1 Aggregate size classes and their primary binding agents

Aggregate class	*Particle size (μm)*	*Primary binding agents*
Macroaggregates	>250	Plant roots and fungal hyphae
Microaggregates	20–250	Plant and microbial cells and by-products (e.g., polysaccharides) encrusted with clay particles
Silt-sized particles	2–20	Bacterial and fungal debris encrusted with clay particles
Flocculated clay	<2	Amorphous aluminosilicates, oxides, and organic polymers on clay surfaces; electrostatic bonding

Adapted from Tisdall JM and Oades JM (1982) Organic matter and water-stable aggregates in soils. *Journal of Soil Science* 33: 141–163.

micrographs of soils clearly show that bacterial cells and fungal hyphae produce and are surrounded by an extracellular polysaccharide coating to which clay particles firmly attach (Figure 3a, b).

Microaggregates, particles 20–250 μm in diameter, are highly stable owing to their small size and to the persistent nature of the binding agents that hold them together. They resist destruction when subjected to rapid wetting or the disturbance caused by cultivation. This aggregate size class is critically important for the protection of organic matter sequestered inside.

Microaggregates are enmeshed and bound together into macroaggregates (greater than 250 μm) by a network of plant roots and fungal hyphae (Figure 3c, d). This largest class of aggregates is often separated into small (250–2000 μm) and large (greater than 2000 μm) macroaggregate fractions. The binding agents responsible for macroaggregate formation (roots and hyphae) are considered transient because they are readily decomposed by microorganisms. The temporary nature of macroaggregate binding makes this aggregate class especially susceptible to disturbance; therefore, the proportion of macroaggregates found in a particular soil is highly dependent on how that soil is managed. Soils that are regularly plowed generally have significantly fewer macroaggregates than uncultivated soils or cultivated soils where reduced tillage practices are employed.

The aggregate hierarchy model has been found generally to apply in temperate region soils where soil organic matter is a main binding agent. Aggregate hierarchy may be less evident in soils dominated by 1:1-type clays, and iron and aluminum oxides, where mineral interactions are the dominant stabilizing force. If aggregate hierarchy does exist, the 'porosity exclusion principle' proposed by Dexter should apply. Namely, large aggregates will have a greater total porosity compared with small aggregates, because they will contain

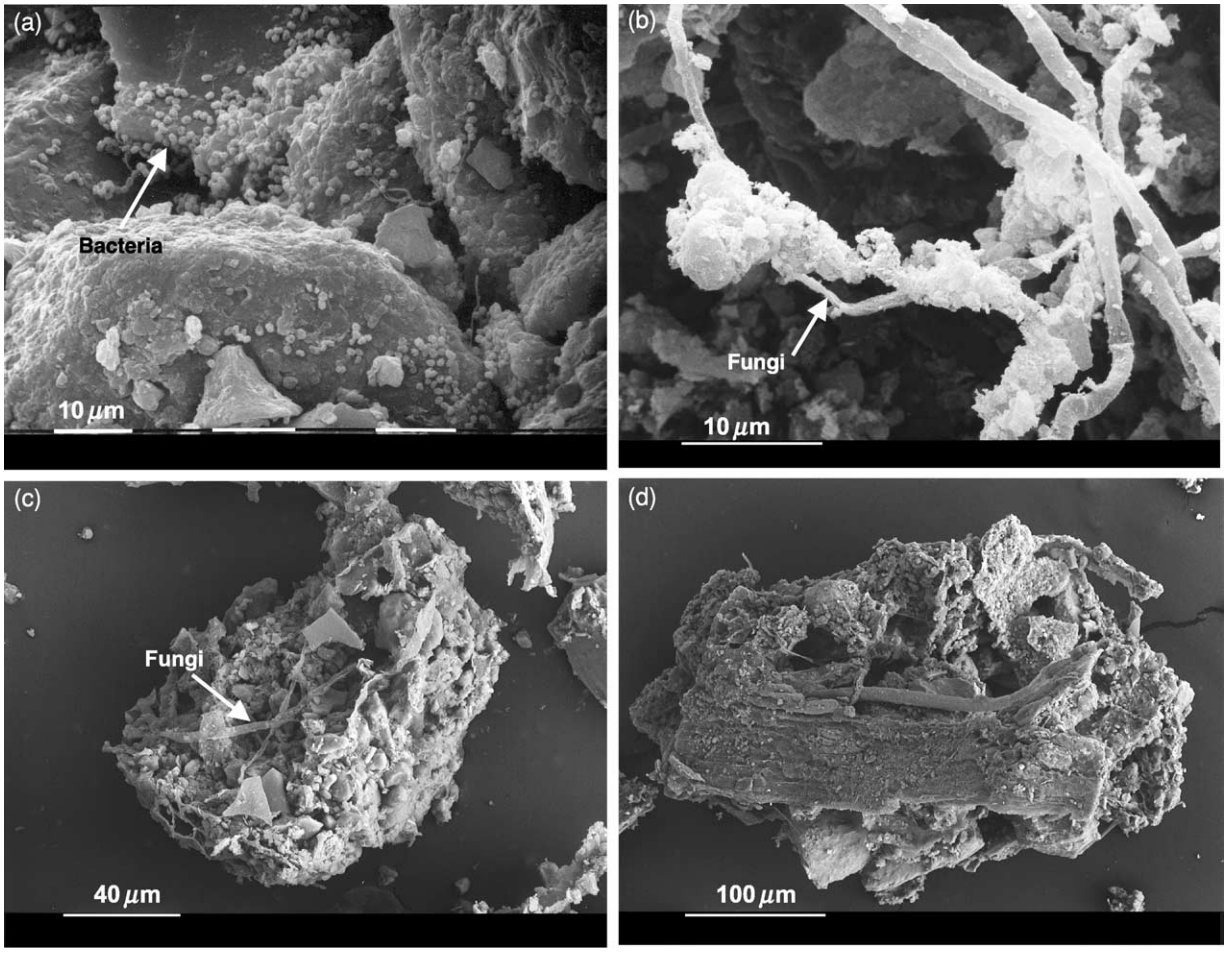


Figure 3 Scanning electron micrographs of (a) a colony of bacteria adhered to particle surfaces; (b) fungal hyphae encrusted with clay particles; (c) a microaggregate between 53 and 250 μm in size; and (d) particulate organic matter and soil particles bound together by roots, fungal hyphae, and microbial exudates into a macroaggregate (>250 μm). Courtesy of V.V.S.R. Gupta, CSIRO Land and Water, Glen Osmond, SA, Australia.

pores both within and between the smaller aggregates that comprise them.

Soil Aggregates as Habitats for Microbiota

The shape and arrangement of soil mineral and organic particles are such that pores of various shapes and sizes are created during the aggregate formation process. In a well-aggregated soil, up to 60% of the total soil volume will be comprised of pores that are either air- or water-filled depending on the moisture conditions. These pores may be open and connected to adjoining pores or closed and isolated from the surrounding soil. Thus soil at a microscopic scale is not dissimilar to a large cave system, with its complex maze of underground tunnels and chambers.

There are four categories of pores: micropores ranging in size from less than 0.2 to 10 μm and found inside microaggregates; pores of between 10 and 100 μm, located between microaggregates but within macroaggregates; pores between macroaggregates; and macropores created by roots and earthworms or by abiotic processes such as cracking or the shrinking and swelling of certain clay minerals when exposed to drying–wetting cycles. The size distribution and degree of continuity of soil pores depend to a large degree on soil texture. In clay soils, nearly 50% of the pores, which are often poorly interconnected, are less than 0.2 μm in size. Pores of 6–30 μm are most abundant in sandy soils and less than 20% are smaller than 0.2 μm. Less than 10% of pores in all soil types are greater than 150 μm.

Pores of different shapes, sizes, and degree of continuity provide a mosaic of microbial habitats with very different physical, chemical, and biological characteristics, resulting in an uneven distribution of soil organisms. Since soil organisms themselves vary in size, structural heterogeneity determines where a particular organism can reside, the degree to which its movement is restricted, and its interactions with other organisms. In this regard, it is the diameter of pore necks, or pore openings, rather than the enlarged section of pores that determines the location of soil organisms.

Every aggregate is a microcosm containing a highly variable microbial community of hundreds to thousands of different species of bacteria, actinomycetes, fungi, protozoa, and algae. The numbers and types of organisms vary from aggregate to aggregate and even between pores within a given aggregate. Mycelial fungi, unlike the other microbial groups, do not require a water film for growth and movement, and can therefore extend their hyphae across air-filled macropores, connecting aggregates and binding them together. Bacteria and protozoa, however, require water for motility and are thus largely restricted from movement between aggregates, since there is typically a discontinuous water film in the pore network, except when the soil becomes saturated following a precipitation or irrigation event.

Soil bacteria, which typically average 0.2–1.0 μm in size, can occupy both large and small pores; however, more than 80% of bacteria are thought to reside preferentially in small pores. The maximum diameter of pores most frequently colonized by bacteria is estimated to range from 2.5 to 9 μm for fine- and coarse-textured soils, respectively. There is a positive correlation between bacterial biomass and pores with a mean diameter of 1.2 μm. Few bacteria have been observed to reside in pores less than 0.8 μm in diameter, which means that 20–50% of the total soil pore volume, depending on soil texture and the pore-size distribution, cannot be accessed and utilized by the microbial community. Electron microscopy has revealed that bacteria often occur as isolated cells or small colonies (less than 10 cells) associated with decaying organic matter; however, larger colonies of several hundred cells have been observed on the surface of aggregates isolated from a clayey pasture soil and in soils under native vegetation (**Figure 3a**). Bacterial cells are often embedded in mucilage, a sticky substance of bacterial origin to which clay particles attach. Clay encapsulation and residence in small pores may provide bacteria with protection against desiccation, predation, bacteriophage attack, digestion during travel through an earthworm gut, and the deleterious effects of introduced gases such as ethylene bromide, a soil fumigant.

Fungi, protozoa, and algae are mainly found in pores larger than 5 μm. Fungi are commonly observed on aggregate surfaces (**Figure 3c**) and typically do not enter small microaggregates (less than 30 μm). Like bacteria, fungal hyphae are often sheathed in extracellular mucilage, which not only serves as protection against predation and desiccation, but also is a gluing agent in the soil aggregation process (**Figure 3b**). Mycelial fungi develop extensive hyphal networks and, as they grow through the soil and over aggregate surfaces, they bind soil particles and microaggregates (53–250 μm) together, thereby playing an important role in the formation and stabilization of macroaggregates (greater than 250 μm).

The relative abundance and distribution of bacteria and fungi vary across aggregate size classes within a given soil and between soils differing in clay content. Macroaggregates have been observed to contain higher total microbial biomass and higher fungal biomass in particular than microaggregates. Aggregates isolated from sandy soils tend to have a more even

distribution of microorganisms than those from clayey soils. That is, bacterial cells and fungal hyphae occur on both the surfaces and inside aggregates from sandy soils, while microbes are largely concentrated on the surface of aggregates in clayey soils, with few microbes being present inside. This is probably due to the differential pore-size distribution of soils with different textures and especially the preponderance of pores less than 0.2 μm in diameter in clay soils.

The heterogeneous nature of the soil-pore network plays a fundamental role in determining microbial abundance, activity, and community composition by affecting the relative proportion of air- versus water-filled pores, which in turn regulates water and nutrient availability, gas diffusion, and biotic interactions such as competition and predation. Microbial activity, measured as respiratory output (i.e., CO_2 evolution), is maximized when approximately 60% of the total soil-pore space is water-filled. As soil moisture declines below this level, pores become poorly interconnected, water circulation becomes restricted, and dissolved nutrients which are carried by the soil solution become less available for microbial utilization. Soil drying leads to a reduction of microbial biomass, particularly in the larger pores, where organisms are subjected to more frequent alterations between desiccation and wetting.

At the other extreme, when most or all of the pores are filled with water, oxygen becomes limiting, since diffusion rates are significantly greater in air than through water. Gas diffusion into micropores is particularly slow, since small pores often retain water even under dry conditions. Restricted oxygen diffusion into micropores combined with biological oxygen consumption during the decomposition of organic matter can lead to the rapid development and persistence of anaerobic conditions. Thus survival of bacteria residing in small pores depends on their ability to carry out anaerobic respiration (e.g., denitrification), replacing oxygen with an alternative electron acceptor (e.g., nitrate, NO_3^-). More than 85% of the potential denitrifying activity of whole soil has been attributed to the microaggregate fraction, where most micropores are located.

Soil Structure and Microorganism Interactions

Ecologic interactions between soil organisms, such as competition and predation, regulate the flow of nutrients in ecosystems. For example, the release of plant-available nutrients such as nitrogen is stimulated when bacterial cells are consumed by protozoa. It is estimated that as much as 30% of the inorganic nitrogen released into the soil solution from decomposing organic matter is due to protozoan predation of bacteria; and more nitrogen is taken up by plants in soils containing protozoa compared with those without protozoa. The release of carbon from soils as CO_2 is also often enhanced in the presence of protozoa.

Soil heterogeneity indirectly influences nutrient-cycling dynamics by restricting organism movement and thereby modifying the interactions between organisms. For example, small pores influence trophic relationships and nutrient mineralization by providing refuges and protection for smaller organisms, particularly bacteria, against attack from larger predators (e.g., protozoa) that are typically unable to enter smaller pores. The location of bacteria within the pore network is a key factor in their survival and activity. Bacterial populations are consistently high in small pores, but highly variable in large pores, where they are vulnerable to being consumed. This may explain, in part, why introduced bacteria (e.g., *Rhizobium* and biocontrol organisms) often exhibit poor survival relative to indigenous bacteria. When they are introduced in such a way as to be transported by water movement into small, protected pores, their ability to persist is enhanced. This example stresses the importance of integrating structural aspects into soil microbiological studies.

Protozoa can consume 2000–12 000 bacteria per protozoan cell division and, since bacteria often occur as individual cells or small colonies, predation is greatest under conditions in which protozoa can readily move between and access a large number of pores. Thus predation rates are high if protozoan numbers are high and they are present in a large number of pores; whereas predation is low if protozoa are restricted to a few large pores. Under typical soil-moisture conditions, protozoan movement is restricted, since they require a continuous water film for migration between pores. Only at high soil-moisture contents (more than 60% of soil water-holding capacity) are protozoa free to move from pore to pore and perhaps from aggregate to aggregate. This explains, in part, why nutrient release is stimulated following a rain or irrigation event.

Protection of Organic Matter Conferred by Soil Aggregates

Bacteria and fungi are the primary decomposers in terrestrial ecosystems and most organic materials entering the soil must eventually pass through the microbial equivalent of the 'eye of the needle' on their way to becoming soil organic matter. Soil microorganisms also recycle the products that they themselves produce during microbial growth and

biosynthesis, including cell debris from dead and decaying cells and substances exuded into the soil environment by living microorganisms (e.g., extracellular polysaccharides). They are also responsible for the further processing and turnover of various soil organic matter pools. These activities lead to the loss of carbon from the soil as carbon dioxide (CO_2), which ultimately diffuses out of the soil, contributing significantly to atmospheric CO_2 levels.

Soil structure is a dominant control on microbial decomposition processes and thus indirectly influences the amount of carbon and other nutrients released from decomposing plant material and soil organic matter. During decomposition, particulate organic materials, colonized by bacteria and fungi, become encrusted with microbial exudates and soil particles, initiating the aggregate formation process (Figure 3). In turn, newly formed and stabilized aggregates protect organic matter from further decomposition by controlling microbial access and activity.

That aggregates protect organic matter from decomposition is indirectly supported by experiments where CO_2 and nutrient release is monitored during the incubation of disturbed versus undisturbed aggregates. When macroaggregates are crushed, for example, there is a flush of microbial activity and the net release of CO_2 and nitrogen (Table 2), suggesting that previously protected organic materials are made more accessible for microbial attack when aggregates are disturbed. Aggregates collected from soils that were previously relatively undisturbed (e.g., soils under native vegetation or no-tillage agricultural management) generally exhibit the greatest response to crushing. Air drying of soils followed by rapid rewetting and incubation also results in a greater release of CO_2 compared with soils kept continuously moist, due partially to aggregate disruption caused by slaking.

More than 25% of the carbon stored in arable soils worldwide has been lost due to cultivation over the past century. The greatest impact of cultivation on aggregation is typically the loss of macroaggregates greater than 250 μm in size. Macroaggregate disruption and the subsequent release and decomposition of organic material once protected within the aggregate structure is one important mechanism by which carbon is lost from soils. There is growing interest in determining to what extent this trend can be reversed. By implementing management practices (e.g., no-tillage, cover crops, perennial vegetation) that promote the formation and stabilization of soil aggregates, it is thought that soils can sequester carbon and thereby mitigate, to some degree, the rapid accumulation of CO_2 in the atmosphere.

Soil as a Spatially Continuous Medium

Much of what is known about the relationships between soil structure, microorganisms, and microbial-mediated processes is based on studies where soil has been broken down into aggregates of different sizes. The isolated size fractions represent those aggregates that are resistant to the method of disruption employed and as such are arbitrary structures. Intact soil is a continuum of soil particles, pore spaces, organic

Table 2 Effect of crushing on the mineralization of carbon and nitrogen from macroaggregates

		Carbon mineralized (mg kg^{-1} soil)		*Nitrogen mineralized (mg kg^{-1} soil)*	
Collection site	*Incubation time (days)*	*Intact*	*Crushed*	*Intact*	*Crushed*
Sidney, NE, USA[a]					
Native grassland	20	1085	1129	68	94
Cultivated field	20	414	493	17	38
Saskatchewan, Alberta, Canada[b]					
Native prairie	14	508	581	45	64
Cultivated field	14	311	326	23	28
Horseshoe Bend, GA, USA[c]					
No-tillage	20	2186	2804	207	305
Conventional tillage	20	1361	1402	150	162

[a]Aggregates (300–2000 μm) were isolated from soil samples collected to a depth of 20 cm and incubated intact or crushed for 20 days (Elliott ET (1986) Aggregate structure and C, N and P in native and cultivated soils. *Soil Science Society of America Journal* 50: 627–633).

[b]Soil samples were collected to a depth of 15 cm at 10 points across each site. Macroaggregates (250–8000 μm) were incubated intact or crushed for 14 days (Gupta VVSR and Germida JJ (1988) Distribution of microbial biomass and its activity in different soil aggregate size classes as affected by cultivation. *Soil Biology & Biochemistry* 20: 777–786, with permission).

[c]Intact soil cores (0–5 cm) were separated into aggregate size fractions and the three macroaggregate size classes (>2000, 250–2000, and 106–250 μm) were incubated intact or crushed for 20 days. Data for the fraction >2000 μm are shown here and expressed on a sand-free aggregate basis (Beare MH, Cabrera ML, Hendrix PF, and Coleman DC (1994) Aggregate-protected and unprotected organic matter pools in conventional- and no-tillage soils. *Soil Science Society of America Journal* 58: 787–795).

materials, and organisms rather than a collection of discrete aggregates. Traditional aggregate-isolation techniques remove aggregates and their associated microbial communities from their spatial context and fail to capture the heterogeneity and connectivity of the pore network within which soil organisms live. Methods which combine soil thin-sectioning with image analysis, geostatistical tools, and mathematical models are now available to describe and quantify the spatial distribution of microbial cells in relation to soil particles and pore spaces. Such nondestructive approaches should provide a more complete understanding of soil microbial communities and the ecosystem processes they mediate.

List of Technical Nomenclature

CO_2	Carbon dioxide
NO_3^-	Nitrate

See also: **Cultivation and Tillage**; **Factors of Soil Formation:** Biota; **Structure**; **Tilth**

Further Reading

Degens BP (1997) Macro-aggregation of soils by biological bonding and binding mechanisms and the factors affecting these: a review. *Australian Journal of Soil Research* 35: 431–459.

Dexter AR (1988) Advances in characterization of soil structure. *Soil Tillage Research* 11: 199–238.

Elliott ET (1986) Aggregate structure and C, N and P in native and cultivated soils. *Soil Science Society of America Journal* 50: 627–633.

Elliott ET and Coleman DC (1988) Let the soil work for us. *Ecological Bulletins* 39: 23–32.

Gupta VVSR and Germida JJ (1988) Distribution of microbial biomass and its activity in different soil aggregate size classes as affected by cultivation. *Soil Biology & Biochemistry* 20: 777–786.

Lynch JM and Bragg E (1985) Microorganisms and soil aggregate stability. *Advances in Soil Science* 2: 133–171.

Miller RM and Jastrow JD (1990) Hierarchy of root and mycorrhizal fungal interactions with soil aggregation. *Soil Biology & Biochemistry* 22: 579–584.

Ranjard L and Richaume A (2001) Quantitative and qualitative microscale distribution of bacteria in soil. *Research in Microbiology* 152: 707–716.

Tisdall JM (1994) Possible role of soil microorganisms in aggregation in soils. *Plant and Soil* 159: 115–121.

Tisdall JM and Oades JM (1982) Organic matter and water-stable aggregates in soils. *Journal of Soil Science* 33: 141–163.

van Veen JA and Kuikman PJ (1990) Soil structural aspects of decomposition of organic matter by micro-organisms. *Biogeochemistry* 11: 213–233.

Vargas R and Hattori T (1986) Protozoan predation of bacterial cells in soil aggregates. *FEMS Microbiology Ecology* 38: 233–242.

Young IM and Ritz K (1998) Can there be a contemporary ecological dimension to soil biology without a habitat? *Soil Biology & Biochemistry* 30: 1229–1232.

Physical Aspects

J R Nimmo, US Geological Survey, Menlo Park, CA, USA

Published by Elsevier Ltd.

A soil aggregate is "a group of primary soil particles that cohere to each other more strongly than to other surrounding particles." Soil aggregates form through the combined action of aggregation and fragmentation processes. That is, attractive and disruptive forces act on the particles in the soil to cause greater cohesion among some individual particles and groups of particles than others. Most soils break up naturally into some form of aggregate, as shown in **Figure 1**. Important physical aspects of aggregates include their size, density, stability, structure, and effect on the transport of fluids, solutes, particles, and heat.

The analysis of soil aggregation is important in a variety of applications. Aggregation is a major influence on the growth and effectiveness of roots. Aggregate stability and size information may be used to evaluate or predict the effects of various agricultural techniques such as tillage or addition of organic matter. Aggregate analysis is often used in experiments where various tillage methods are applied and then evaluated by examining the stable aggregates that result. Because of their direct relation to cohesive forces, aggregate size and stability are important to the understanding of soil erosion and surface sealing. Analysis of dry aggregates is logically related to wind-erosion effects, while wet analysis may be more appropriate to evaluate or predict erosion due to rainfall impact and runoff. The stability of wet aggregates can be related to surface-seal development and field infiltration, as water-stable cohesion among particles may lead to restriction of water entry and formation of surface seals. Through these erosion and sealing effects, as well as the relation between aggregation

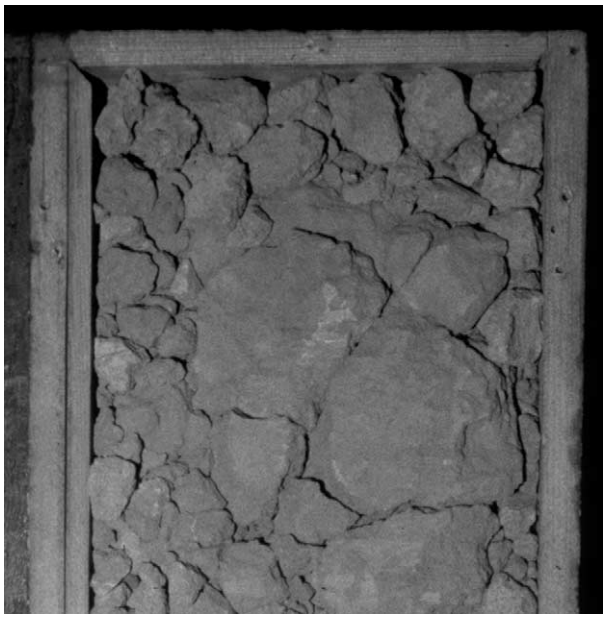


Figure 1 Soil with aggregates partially separated, in a tray. (From the Historic Russian Soil Collection of the St. Petersburg Academy of Forestry, provided by Jennifer Harden.)

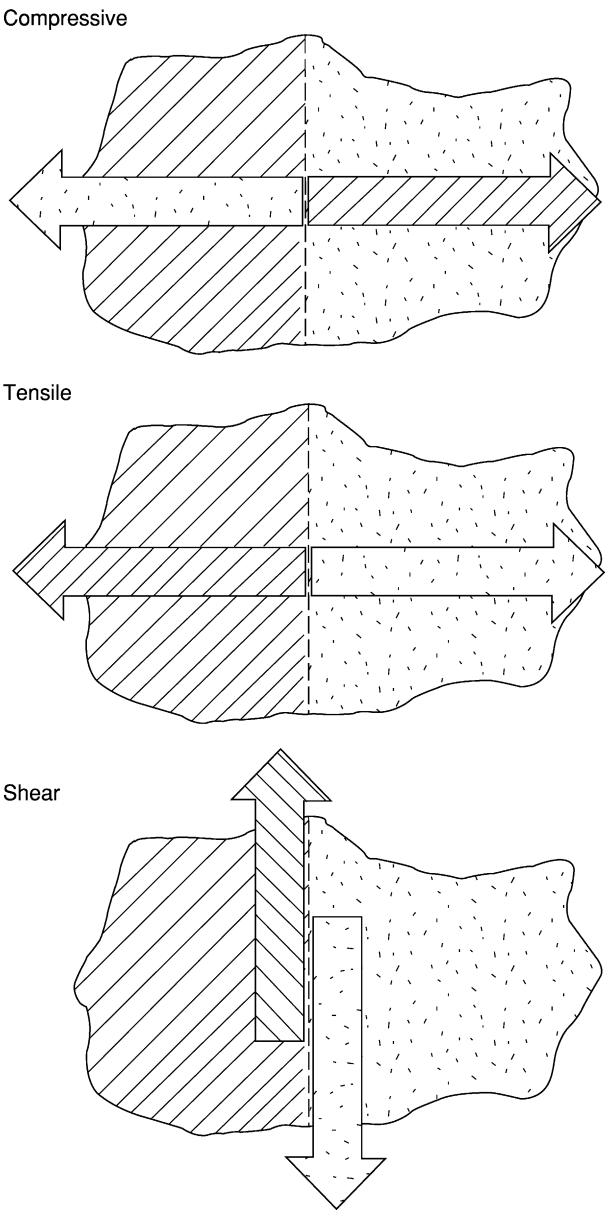


Figure 2 Types of stresses on an aggregate. Stresses are defined with respect to a selected plane, shown as the dashed line in this cross-sectional diagram. Each arrow indicates the force acting on the portion of the aggregate with the same type of shading, at the selected plane.

and structural features such as macropores, aggregate analysis may help in the understanding of most aspects of soil water behavior, including runoff, infiltration, and redistribution, as well as soil aeration. Increasingly, aggregate properties are used in models that predict soil hydraulic properties, including water retention and unsaturated hydraulic conductivity.

Closely related terms include 'ped,' 'clod,' and 'crumb.' A ped is an aggregated unit representative of the innate structural classification of the soil. It has a characteristic shape related to structural designations such as prismatic, columnar, and blocky. The term 'clod' applies to an aggregate separated from the bulk soil by artificial means such as digging or plowing. A crumb is a small aggregate, typically less than 5 mm in diameter.

Forces on Soil Particles

The strength of interparticle cohesion depends on a variety of soil physical, chemical, and biological influences, some of the most important being air–water surface tension, intermolecular attractive forces between water and solids, cementation by precipitated solutes, entanglement by roots and fungal hyphae, and various chemical phenomena. The forces of soil cohesion depend strongly on water content and other conditions.

Fundamentally, the forces important in aggregation can be considered as stresses, that is, force per unit area acting on a given cross-sectional plane within the aggregate. These categorize as compressive, tensile, and shear stresses (Figure 2). Compressive stresses act to push soil particles closer together, as for example by the weight of soil above a given horizontal plane. Tensile stresses pull apart; for example, forces from soil shrinkage. Shear stresses act along a plane parallel to the direction of force,

as in an aggregate at the edge of a zone of compaction. Tensile and shear stresses tend to disrupt aggregates; compressive stresses tend to consolidate aggregates, except that, when uneven across a plane, they lead to shear stresses that disrupt.

Several types of forces act to hold soil particles together. Water in the soil does this through surface tension, as well as through the attractiveness of water molecules for soil solids and for each other. Dissolved ions are important, especially in terms of the electrical double layer. The tendency of soil particles to have a negative surface charge means that water close to them is rich in positive ions, which in turn attracts other particles, in a process of flocculation. Because clay particles are especially sensitive to flocculating influences, higher clay content of a soil generally makes for more aggregation. Chemicals that precipitate or otherwise turn into cementing agents also play an obvious role. Various other chemicals, especially certain organic compounds such as polysaccharides, attract soil particles. Some organic materials just happen to be attractive; others, like roots and fungal hyphae, adhere to soil as part of their natural function. Because aggregation strongly affects air, water, roots within the soil, and other factors affecting plant growth, one may assume that plants have evolved so as to generate decay products that affect aggregation in ways favorable to the plant. A major part of this influence is to promote aggregation. Bacteria and other microorganisms contribute similarly to aggregation. To increase aggregation artificially by adding organic material to the soil is not straightforward. Not just the type and quantity of organic compounds but also their microscale distribution is critically important. Typically particles within an aggregate may be held together by a sort of glue made up of water, clay, and organic materials (**Figure 3**).

There is a similar variety of mechanisms that pull soil particles apart, either directly or by a decrease in attractive force. Some of the most common are associated with the addition of water. The breakup of aggregates that results from this, especially from sudden immersion, is called slaking. Increased water content can dissolve cementing precipitates and can decrease flocculation, while the resultant dilution weakens the effects of electrical double layers. As water infiltrates an aggregate, the expansion of trapped air, as well as the release of adsorbed air from newly wetted surfaces, can generate substantial disruptive force. Other disruptive mechanisms include the expansion of water upon freezing, impacts of rain or falling objects, and vibrations – either natural or artificial such as ultrasound or jostling on a sieve. Mechanisms associated mainly with compressive force disrupt by the generation of shear stresses;

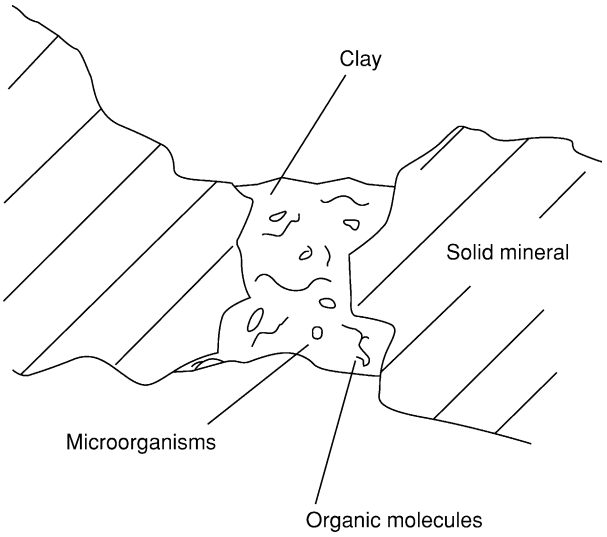


Figure 3 The microscopic region between solid grains within an aggregate.

examples include foot or wheel traffic, which is always to some degree uneven across the land surface, and gravity acting on an uneven mass distribution of soil or on an aggregate unevenly supported from below.

Aggregates are less stable as they get bigger. This generalization applies within a given soil and should not be confused with the idea that soils forming larger aggregates have greater aggregate stability. One reason large aggregates are less stable is simply that interparticle forces vary, and the bigger the aggregate, the greater likelihood that it contains a plane-like region of low tensile strength where it breaks in response to stress. Similarly, the bigger the aggregate, the greater likelihood that it contains an expanding root or other agent that breaks it apart. Some disruptive stresses increase as the size of an aggregate increases, for example its weight increases with size. Forces due to shrinking and swelling are cumulative so that their net effect within the volume of a single aggregate is greater for a larger aggregate. These effects can be generalized by noting that attractive forces (cementing, intermolecular attraction, etc.) are predominantly short-range, whereas disruptive forces (which are mechanically transmitted through the soil fabric) are predominantly longer-range. Thus, as long as the same basic soil material is considered, the balance of forces within an aggregate increasingly favors disruption as larger aggregates are considered. This fundamental linkage between aggregate size and aggregate stability is a crucial factor in virtually any assessment of soil aggregation.

Whether a given sample of soil tends toward relatively large or relatively small aggregates depends chiefly on the interparticle attractive forces. This is

because these forces vary more from soil to soil than do disruptive effects such as gravity and surface traffic, and therefore they dominate the issue of how the attractive and repulsive forces balance out. The soil's characteristic response to shear stresses – the extent to which it undergoes plastic in contrast to brittle deformation – is also important. Both interparticle attraction and plasticity depend strongly on soil texture and the composition of the soil–plant–water system. The aggregates of a fine-textured soil with much organic matter are likely to be larger than those of a sandy soil. The net interparticle attractive force in a sand can easily be so small that the characteristic aggregate size it indicates is smaller than the individual soil particles, so aggregates do not occur.

Aggregate Physical Properties and their Measurement

Basic Properties

Some physical properties of aggregates can be determined directly, especially where it is possible physically to isolate individual aggregates. Aggregate size and shape can be determined optically, by comparison with a ruled grid or by analysis of digital images. Because aggregates have irregular shape, their size cannot be indicated by a single linear dimension. Thus a choice must be made whether to indicate size by greatest dimension, average dimension, diameter of an equivalent sphere, or some other measure. The volume and bulk density of an aggregate can be measured by the clod method, using the aggregate's weight in air and in a liquid of known density, after coating it to prevent liquid intrusion. Alternatively, to reduce or eliminate the need for coating, a fine granular material of known bulk density may be used instead of a liquid. The strength of an aggregate can be measured, at least operationally, by breaking it with a known mechanical force applied by impact or by gradual increase in magnitude.

In most applications, attributes such as size, density, and strength are important as characterizations of the bulk soil, rather than of particular aggregates. Then, if the measurements are performed on individual aggregates, they must be made on enough aggregates to enable the determination of representative property values by statistical techniques. Alternatively, there is a wide range of methods that can be applied on aggregates in bulk.

Another important aspect is the internal structure of an aggregate. Soil within an aggregate may be more homogeneous than within a greater volume of soil, but, like any body of soil, it is not perfectly homogeneous. At one extreme, it might have a monolithic character not readily subdivided into units larger than individual particles. Alternatively, an aggregate may comprise smaller aggregates held by greater forces within themselves than between each other. In this way, each subdivision of aggregates may comprise a smaller subdivision of aggregates, down to a limit as the subdivided aggregates approach the size of particles. This sort of structure has led some scientists to propose fractal models for the structure of individual aggregates. Discussions below, on aggregate density dependence on aggregate size, and on mathematical representations of aggregate size distribution, explore this issue further.

Aggregate density and how it correlates with aggregate size can provide evidence for or against hypothesized forms of structure and data for predicting or correlating with other soil properties. Figure 4 shows an example of such data for soils of three different textures. For two of the soils, the smaller aggregates have greater density, which more closely approximates the particle density of soil minerals and indicates a tighter, more compact, and probably more stable structure. The material labeled 'sand' or 'quartz sand with pebbles' in the original figure has an aggregate density that has little size dependence and differs little from the particle density of pure quartz, approximately $2.65\,g\,cm^{-3}$. This indicates that this sand essentially has no aggregates, except

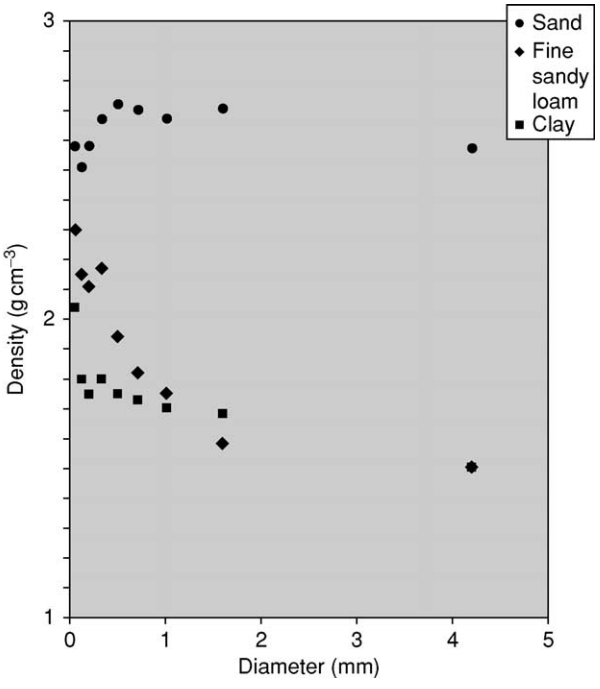


Figure 4 Measured aggregate density as a function of mean aggregate size for soils of three textures. (Source: Chepil WS (1950) Methods of estimating apparent density of discrete soil grains and aggregates. *Soil Science* 70: 351–362.)

possibly for some small aggregates (approx. 0.1 mm in diameter) of the smallest particles.

Size and Stability

Fundamentals Measuring physical aspects of aggregation of a coherent volume of soil containing a representative number of aggregates is the most common way to characterize the size distribution or stability of the soil's aggregates, but it is complicated by the interrelationships of aggregate properties. Especially important is that aggregate size and net cohesive force are conceptually inseparable. Measured sizes depend on the disruptive force applied to separate the aggregates, so force and size cannot be measured independently; all techniques characterizing these for a bulk soil involve some combination of both. For example, size determination by sieving cannot be done without the disruptive force of collisions between the aggregates and the sieve. Methods tend to be called stability methods or size methods depending on which of these gets more emphasis. This difference may be in the technique itself, for example a size method relying on a specified disruptive force or a stability method relying on the effect of force on a given size of aggregates. Alternatively, it may be in the interpretation of its results, for example whether a measured size distribution is presented as a distribution function or as a single index (such as an average aggregate size) that is considered to indicate stability.

Ideally, the size and strength of aggregates would be defined on a fundamental physical basis, in the way that hydraulic conductivity can be defined in terms of flux and potential gradients. In that case, any given measurement technique would provide an approximation to the defined ideal. Improved methods would produce results that are increasingly close approximations of the ideal. For aggregates, the definition would have to encompass both size and force, but the difficulty of quantifying the force prevents use of such a definition. Research on disruptive and cohesive forces may eventually solve this problem. Present characterizations, however, have to rely on operational definitions that endorse the result of a particular procedure with a particular apparatus, so the method cannot be separated from the definition.

In choosing a method for obtaining aggregate size and stability information, either stability or size distribution, and either wet or dry aggregates can be focused upon. The needs of the application should guide this choice. Erosion applications, for example, usually relate more directly to stability, while hydraulic and gas transport properties may relate more directly to the size distribution. The choice of wet or dry aggregates for measuring may depend on which condition most resembles the field situation, or on such considerations as reproducibility or consistency with other measurements.

Commonly used methods The forces applied to fragment or separate aggregates of the main bulk of soil are fundamentally artificial, though they may in some ways resemble forces in the natural setting. In the laboratory it is impossible to exert disruptive forces that exactly oppose the microscopic forces of cohesion, so practical methods rely on the variable and poorly known forces in a process that usually involves some combination of sieving, grinding, or vibration. Some methods make use of other phenomena that break aggregates apart, especially the forces involved when water or another liquid is introduced in relatively dry soil. In specifying the procedures that effectively define the aggregate characteristics, there are three realms of variables: the disrupting force or energy applied, the distribution of aggregates and particles, and the conditions of testing.

Some methods afford ways of quantifying some aspect of the applied force or energy. For example, the rupture-threshold approach considers one aggregate at a time, squeezing the aggregate between parallel plates while measuring both the applied force and the linear displacement. The drop-shatter method considers a known mass of nominally undisturbed soil, which is dropped from a known height on to a hard surface. The difference in potential energy associated with the distance of fall serves as an index of the energy applied to break apart the aggregated soil. The translation of this energy into energy that promotes sample rupture is imperfect, though it may be a useful index of the applied disruptive energy.

Some stability methods produce their own stability index as a result of the specified procedures and data analysis techniques, for example the fraction of soil weight that comprises stable aggregates. Stability interpretations may also be derived from aggregate-size distributions, usually by mathematically converting the tabular data or parameterized distribution formula to an average or other simple index. In this way the mathematical representation of size distribution not only serves as a convenience, but also provides the link, where needed, between size distribution and stability. In relating aggregate size to stability, the basic idea is that bigger aggregates imply greater stability. The most widely used index for this purpose is the mean weight diameter, defined as the sum of the weighted mean diameters of all size classes, the weighting factor of each class being its proportion of the total sample weight. Ideally this would be determined from integration of the cumulative abundance of aggregates as a function of diameter. The geometric mean diameter can also serve

as an aggregate size index, though in recent literature it appears less than the mean weight diameter.

Miscellaneous methods Many techniques involve deliberate wetting or immersion of the sample. Wet compared with dry measurements on aggregates effectively measure different physical properties of the soil. It is not only the degree of wetness that is important, but also the means by which water has been applied. Wetting the soil in a vacuum, for example, reduces the disruptive forces associated with trapped air and thus produces larger aggregates.

Fast wetting with no vacuum involves immersion of air-dried aggregates in water for a period of time before beginning the mechanical sieving process. This type of wetting of dry aggregates produces disintegration and slaking, which may be undesirable. High-vacuum fast wetting involves de-airing aggregates in a vacuum chamber under high vacuum, then instantaneously wetting the aggregates in the chamber. It generally produces minimal disruption. Slow aerosol wetting, in which samples on screens are wet by vapor from below, produces little disintegration. Stabilities are higher and more reproducible with this type of wetting, in contrast to vacuum wetting. Wetting by slow wicking with or without a vacuum allows aggregates to draw moisture in from moist filter paper. As an alternative to water in initial wetting or other procedures, organic solvents such as methanol may reduce aggregate disintegration by slaking and may better preserve aggregate structure in drying.

A method derived from the older, 'high-energy moisture characteristic' method is based on differences in the water retention curves for fast-wetted and slow-wetted aggregates from replicate soil samples. Soil water retention is measured for thin beds of sieved aggregates on sintered glass. The wettest portion of the retention curve is considered, since the matric pressure range relatively close to zero is most affected by interaggregate pores. The key principle is that the less stable the aggregates, the more vulnerable they will be to disintegration during fast wetting, so the greater will be the difference between retention curves for the fast-wetted and slow-wetted samples. The method produces an index of aggregate stability on a dimensionless zero-to-one scale that is easily compared among different soils but is not directly comparable with other stability indices.

Ultrasonic dispersion can supply the disruptive force to associate with aggregate stability. The energy level that achieves a plateau in the quantity of aggregates remaining intact serves as an index of stability.

Stability is sometimes considered operationally in terms of the fraction of sample weight remaining after a prescribed sieving operation. Other methods measure the energy needed to break aggregates by crushing with parallel plates, as described above in connection with quantification of the energy of rupture. The results for a significant number of aggregates need to be reduced to a statistical representation indicative of the properties of the bulk sample. The energy required per increase in aggregate surface area can serve this purpose, as can a distribution function that indicates the probability of failure for a given applied rupture energy.

Issues of general importance The size distribution and stability of aggregates depend on numerous secondary factors, apart from the soil type. An obvious consideration as it is for spatial variability, expected to be substantial like most other soil properties. Aggregate stability can increase with storage time of the sample; it can also increase with increasing salt content of the water and is likely to decrease with temperature.

Soils with concretions (assemblages of primary particles that cannot be broken apart by the disaggregation processes of the chosen method) must be analyzed with respect to the application. In some cases the concretions may be treated as indivisible particles, because they are stable under normal cultivation practices; or in other cases as stable aggregates, because they usually have porosity, internal surface area, and substantial exchange capacity.

Some soils, especially from humid regions, may be nearly 100% stable in terms of the fraction remaining after prescribed sieving. Greater disruptive force, achievable by increasing the duration and amplitude of sieving, or by a more disruptive wetting technique, enables the detection of differences among highly stable soils, though with the drawback of precluding comparison with results obtained by more standard procedures on less-stable soils. The use of multiple methods increases the likelihood that comparisons will be possible among diverse soils.

Representation and Interpretation

To represent size distributions, the fraction of material at particular values of effective aggregate diameter can be graphed directly or cumulatively, as in **Figure 5**. For convenience in representation or for further mathematical development, these data can be fitted to a specific mathematical form.

Of various mathematical functions that have been used to fit the aggregate data, the lognormal distribution is one of the most useful and reasonably fits data from a variety of soils. Being a normal (Gaussian) distribution on a log scale, this distribution is skewed toward the small-diameter end of the range covered. It also has appropriate tapering-off of abundance at

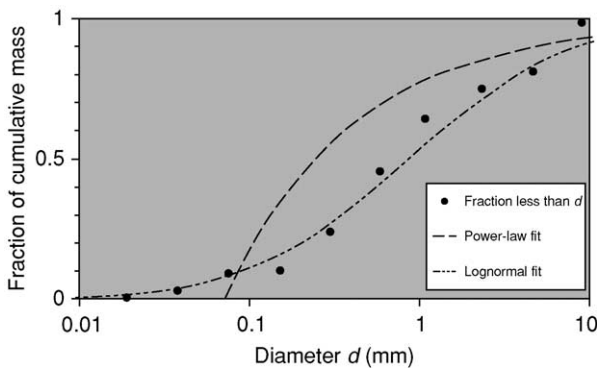


Figure 5 Measured and fitted aggregate size distribution for Sharpsburg silty clay loam. (Source: Wittmuss HD and Mazurak AP (1958) Physical and chemical properties of soil aggregates in a brunizem soil. *Soil Science Society of America Proceedings* 22: 1–5.)

both the small- and large-diameter extremes. The lognormal representation has also been used in some of the recent hydraulic property models that are based on aggregate properties.

Fractal interpretations have been applied to both aggregate stability and size distribution. A fractal characterization is valid if each subunit of the system is structurally identical (at a reduced scale) to the whole system. This idea is attractive for aggregates, since they are not made of primary soil particles on a fully equal basis. Larger aggregates may be thought of as being made of smaller aggregates that are more strongly bound internally than to each other. One attribute of fractal representation is that the cumulative number-size distribution can be represented as a power law; the cumulative abundance of objects greater than a given size is proportional to that size raised to some exponent. Like the lognormal model, a fractal model with an appropriate fractal dimension has a distribution skewed toward the small diameters. By fractal theory the power-law exponent is directly related to the mass fractal dimension, so this dimension may be known once the value of the exponent is established. Geometrically, the fractal dimension depends on the shape of the objects and the extent of fragmentation. Tests with the density, shape, and relative diameter variables represented fractally do not always show the degree of consistency over different scales that the most straightforward fractal models would predict. Natural aggregates may tend toward a monolithic internal structure or otherwise to deviate from true fractal character. Another shortcoming of fractal models is that they are fundamentally unrealistic at extremes of the range. Even so, fractal models remain useful for relating disruptive force to aggregate size, and in general for the modeling of relationships between mechanical properties and other soil properties and conditions.

Summary

The concept of an aggregate arises simply because some particles in the soil adhere more strongly than others. Physical aspects of aggregation are fundamental to the character, function, and behavior of soil. They give insight, with the possibility of much-needed quantitative insight, into soil structure.

The characterization of physical aspects of aggregation requires crucial tradeoffs. Little can be learned quantitatively about aggregates without operational definitions and criteria. Thus the specifics of measurement techniques are more closely bound to the consideration of aggregate physical properties than to many other aspects of soil. The choice between widely used, informally standardized methods and more novel ones often involves a substantial tradeoff between the need for consistency and the ultimate appropriateness of the method. The more standardized methods facilitate comparability, but the quantitative indices they generate may not give the sort of aggregate characterization most pertinent to the application at hand. Reliance on operational definitions also makes it awkward to incorporate ongoing scientific advances in conceptualizations and techniques. Research that leads to standardization of the specified force and a fundamental physical definition would help in allowing aggregate-measurement technology to advance without loss of comparability. The difficulty of this undertaking parallels the difficulty of bringing the general concept of soil structure into an objective, quantitative realm, but the potential benefits of even partial success justify much effort.

See also: **Aggregation:** Microbial Aspects; **Flocculation and Dispersion**; **Fractal Analysis**; **Structure**; **Swelling and Shrinking**

Further Reading

Ashman MR and Puri G (2002) *Essential Soil Science*. Oxford, UK: Blackwell.

Currie JA (1966) The volume and porosity of soil crumbs. *Journal of Soil Science* 17(1): 24–35.

Dickson JB, Rasiah V, and Groenevelt PH (1991) Comparison of four prewetting techniques in wet aggregate stability determination. *Journal of Soil Science* 71: 67–72.

Holden NM (1995) Temporal variation in ped shape in an old pasture soil. *Catena* 24: 1–11.

Kemper WD and Chepil WS (1965) Size distribution of aggregates. In: Black CA (ed.) *Methods of Soil Analysis*, part 1, *Agronomy*, pp. 499–510. Madison, WI: American Society of Agronomy.

Kosugi K and Hopmans JW (1998) Scaling water retention curves for soils with lognormal pore-size distribution. *Soil Science Society of America Journal* 62: 1496–1505.

Nimmo JR and Perkins KS (2002) Aggregate stability and size distribution. In: Dane JH and Topp GC (eds) *Methods of Soil Analysis*, part 4, *Physical Methods*, pp. 317–328. Soil Science Society of America Book Series No. 5. Madison, WI: Soil Science Society of America.

Oades JM and Waters AG (1991) Aggregate hierarchy in soils. *Australian Journal of Soil Research* 29: 815–828.

Rieu M and Sposito G (1991a) Fractal fragmentation, soil porosity, and soil water properties. I. Theory. *Soil Science Society of America Journal* 55: 1233–1238.

Rieu M and Sposito G (1991b) Fractal fragmentation, soil porosity, and soil water properties. II. Applications. *Soil Science Society of America Journal* 55: 1239–1244.

AGROFORESTRY

P K R Nair, University of Florida, Gainesville, FL, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Agroforestry, a loosely defined term, involves the deliberate growing of trees and shrubs with crops and/or animals in interacting combinations for a variety of objectives. Such farming practices have been used throughout the world for a long time; but scientific attention was focused on them, and thus they attained prominence as a land-use practice, only since the late 1970s. Since then, substantial progress has been made in the science and practice of agroforestry. Today, acting as an interface between agriculture and forestry, agroforestry is considered to be a promising and sustainable approach to land use. The objective here is to present the essential features and forms of agroforestry and review the problems and prospects of its future development.

Emergence of Agroforestry as a Land-Use Approach

Why and how did this old form of land use attain prominence lately? During the mid-to-late 1900s, agriculture and forestry in the industrialized nations were transformed into commercial enterprises with emphasis on the production of single commodities. Certainly, it paid rich dividends. This apparently successful model was embraced by many newly independent nations of the developing world that were faced with the problem of feeding their millions. Several food-production technologies were developed according to this model and tried in the tropics. Some of them resulted in substantial increases in agricultural production, especially through the so-called Green Revolution of the 1970s. The traditional, mixed production systems of raising food crops, trees, and animals together, as well as exploiting a multiple range of products from natural woodlots, did not fit into the single-commodity paradigm, and were discouraged.

Serious doubts began to be expressed, however, about the relevance of single-commodity strategies and policies promoting them. In particular, there was concern that the basic needs of the poorest farmers, especially those in rural areas, were neither being considered nor adequately addressed. Soon it became clear that many of the technologies that contributed to the Green Revolution were not affordable to the poor farmer and that most tropical soils were unable to withstand the impact of high-input technology. The disastrous consequences of increasing rates of deforestation in the world's tropical regions also became a matter of serious concern. It was recognized that a major cause of deforestation was the search for more land to provide food and fuelwood for rapidly increasing populations.

The search for appropriate strategies to address these problems – strategies that would be socially acceptable, would enhance the sustainability of the production base, and would meet the need for production of multiple outputs – led to studies of age-old practices based on combinations involving trees, crops, and livestock on the same land unit. The inherent advantages of traditional land-use practices involving trees, such as sustained yield, environmental conservation, and multiple outputs, were quickly recognized. Agroforestry thus began to come of age in the late 1970s; the trend was institutionalized with the establishment of the International Council (Centre, since 1991) for Research in Agroforestry (ICRAF) in 1977, in Nairobi, Kenya. Since then, ICRAF has been the world leader in tropical agroforestry research.

In the temperate region, agroforestry has had a slower adoption rare than in the tropics. During the

1970s and 1980s, agroforestry as a concept had very little support in the industrialized countries. But the situation changed gradually and agroforestry gained better acceptance in the 1990s. For example, faced with the environmental consequences of agricultural and forestry practices that focused on the economic bottom line, the American public is now demanding greater environmental accountability for land-use practices and the application of more ecologically and socially friendly management approaches. Agroforestry fits well into that mold.

Agroforestry Systems

An agroforestry system normally involves two or more species of plants (or plants and animals), at least one of which is a woody perennial. The system has two or more outputs and a production cycle of more than 1 year, and both its ecology and economics are more complex than a monocultural system of agriculture or forestry. The essence of agroforestry can be expressed in four key 'I' words: intentional, intensive, interactive, and integrated. The term 'intentional' implies that systems are intentionally designed and managed as a whole unit, and 'intensive' means that the systems are intensively managed for productive and protective benefits. The biological and physical interactions among the system's components (tree, crop, and animal) are implied in the term 'interactive,' and 'integrative' refers to the structural and functional combinations of the components as an integrated management unit. It is often emphasized that all agroforestry systems are characterized by three basic sets of attributes: productivity (production of preferred commodities as well as productivity of the land's resources), sustainability (conservation of the production potential of the resource base), and adoptability (acceptance of the practice by the farming community or other targeted clientele).

In addition to agroforestry, several other 'forestry' terms became prominent in the 1980s, reflecting the increasing global interest in tree-planting activities. These include community forestry, social forestry, and farm forestry. Although these terms have also not been defined precisely, they emphasize people's participation in tree-planting activities. Social forestry refers to using trees and/or tree planting specifically to pursue social objectives, usually betterment of the poor, through delivery of the benefits of trees and/or tree planting to local people. Community forestry, a form of social forestry, refers to tree-planting activities undertaken by a community on communal lands or the so-called common lands; it is based on local people's direct participation in the process, either by growing trees themselves, or by processing the tree products. Farm forestry, a term used mainly in Asia, refers to tree planting on farms. The major distinction between agroforestry and these other systems is that, while agroforestry emphasizes the interactive association between woody perennials and agricultural crops and/or animals for multiple products and services, the other terms refer to tree planting, often as woodlots. In practice, however, all these labels directly or indirectly refer to growing and using trees to provide food, fuelwood, fodder, medicines, building materials, and cash income. Therefore, in common land-use parlance, these different terms are often used as synonyms, and sometimes even out of context.

A large number of traditional agroforestry systems have been recognized from different parts of the world. A major characteristic of these systems is their location-specificity. Each is a specific local example of the association or combination of the components, characterized by the plant species and their arrangement and management, and environmental and socioeconomic factors. In spite of the large variations among them, broad similarities can be recognized among the systems. For example, all these systems are composed of three basic groups of components or constituents: woody perennials (trees), herbaceous and other agricultural species (crops), and/or livestock. Agrisilvicultural systems involve trees and crops; silvopastoral systems involve trees and pasture/animals, and agrosilvopastoral systems involve all three major groups of components (crops, trees, and pasture/animals). Another way of describing the systems is based on the temporal arrangements of the components. Thus, 'simultaneous system' is used when the trees and crops exist together on the same unit of land during the same period of time; similarly, when the trees and crops are separated in time (such as when one alternates with the other), it is a 'sequential system.'

Because of the indigenous and traditional nature of agroforestry systems, there is some ambiguity in the terms used to describe them. All these systems consist, however, of a few specific patterns of component arrangements in space and time; these patterns are sometimes called agroforestry practices. In other words, although there are several hundred agroforestry systems, they all consist of a relatively few agroforestry practices. The same or similar such practices are found in various systems in different situations; the common practices and their brief descriptions are given in Table 1. Both the systems and the practices are known by similar names, but the systems are related to the specific locality or the region where they exist, or other descriptive characteristics that are specific to it. For example, growing

Table 1 Major agroforestry practices in the tropics and the temperate regions

Agroforestry practice	*Brief description*
Tropical agroforestry	
Alley cropping (hedgerow intercropping)	Fast-growing, preferably leguminous woody species grown in crop production fields; woody species are periodically pruned at low height (<1.0 m) to reduce shading of crops; prunings are applied as mulch into alleys as a source of organic matter and nutrients, or used as animal fodder
Taungya	Growing agricultural crops during the early stages of establishment of forestry (timber) plantations
Homegardens	Intimate multistorey combinations of a large number of various trees and crops in homesteads; livestock may or may not be present
Improved fallow	Fast-growing, preferably leguminous, woody species planted and left to grow during the fallow phase of shifting cultivation; the woody species cause site improvement and may yield economic products
Trees in soil conservation and reclamation	Trees on terraces and raisers with or without grass strips; use of trees for reclamation of saline, acidic, or otherwise degraded lands
Multipurpose trees (MPTs) on farms and rangelands	Fruit trees and other MPTs scattered haphazardly or according to some systematic planting arrangements in crop or animal production fields; trees provide fruits, fuelwood, fodder, and timber
Pasture under plantations (a form of silvopasture)	Cattle grazing on pasture under widely spaced rows of plantation species
Shaded perennial-crop systems	Integrated multistorey mixtures of tree crops such as coconut, cacao, coffee, and rubber with other tree crops, shade trees, and/or herbaceous crops
Protein banks (a form of silvopasture)	Production of protein-rich tree fodder on farms/rangelands for cut-and-carry fodder production
Shelterbelts and windbreaks	Use of trees to protect fields from wind damage, sea encroachment, and floods
Temperate-zone agroforestry	
Alley cropping	Trees planted in single or grouped rows with herbaceous (agricultural or horticultural) crops in the wide alleys between the tree rows
Forest farming	Utilizing forested areas for producing specialty crops that are sold for medicinal, ornamental, or culinary uses
Riparian buffer strips	Strips of perennial vegetation (trees/shrubs/grass) planted between croplands/pastures and streams, lakes, wetlands, and ponds
Silvopasture	Combining trees with forage (pasture or hay) and livestock production
Windbreaks	Row trees around farms and fields, planted and managed as part of crop or livestock operation to protect crops, animals, and soil from wind hazards

Source: Nair PKR (1993) *An Introduction to Agroforestry*. Dordrecht, The Netherlands: Kluwer; and Garrett HE, Rietveld WJ, and Fisher RF (eds) (2000) *North American Agroforestry: An Integrated Science and Practice*. Madison, WI: American Society of Agronomy.

coffee (*Coffea* spp.) under the shade of the timber tree, *Cordia alliodora*, in Costa Rica and growing cacao (*Theobroma cacao*) under the shade of coconuts (*Cocos nucifera*) in Kerala, India, are two examples of the same (or similar) practice, but they represent different systems. In spite of these differences, in common usage, the words 'systems' and 'practices' are used synonymously in agroforestry just as in other land-use forms. To make matters worse semantically, the same 'base' word is retained even after a traditional system or practice may have been improved through scientific intervention. For example, there are traditional fallows, enriched fallows, and improved fallows, and traditional intercropping and hedgerow intercropping. Another anomaly in the use of these terms is the variants or 'forms' of the same practice in different contexts. For example, the term 'alley cropping' as used in the tropics is different from temperate-zone alley cropping (Table 1), although conceptually they are similar.

Within the tropics, the complexity of agroforestry systems is greatest in the lowland humid and subhumid tropics, where climatic conditions generally favor rapid growth of a large number of plant species. Homegardens, shaded perennial systems (or plantation-crop combinations), and multilayer tree gardens are common in such regions with high human populations, and less intensive systems such as taungya and shifting cultivation are common in areas with less population density. In the semiarid tropics, the nature of agroforestry systems is also influenced by population pressure: homegardens and multilayer tree gardens are found in the relatively wetter areas; windbreaks and shelterbelts, and multipurpose trees on croplands are found in the drier regions. In the highland tropics (with favorable rainfall regimes), sloping lands and rolling topography make soil erosion an issue of major concern; consequently, soil conservation is one of the main objectives of agroforestry in these regions. Shaded perennial systems, use

of woody perennials in soil conservation, improved fallows, and silvopastoral systems are the major forms of agroforestry in these tropical highlands. Several other specific systems also exist in the tropics: for example, apiculture with trees, aquaculture involving trees and shrubs, and woodlots of multipurpose trees.

Alley cropping, forest farming, riparian buffer strips, silvopasture, and windbreaks are the five major agroforestry practices recognized in North America (Table 1). Other temperate-zone agroforestry systems include ancient tree-based agriculture involving a large number of multipurpose trees such as chestnuts (*Castanea* spp.), oaks (*Quercus* spp.), carob (*Ceratonia siliqua*), olive (*Olea europa*), and figs (*Ficus* spp.) in the Mediterranean region. The 'dehesa' system of land use, involving grazing under oak trees with strong linkages to recurrent cereal cropping in rangelands, is also a very old system in this region.

The nature, complexity, and objectives of agroforestry systems vary considerably between the tropics and the temperate region. The climatic conditions in many parts of the tropics favor longer production cycles within a year and a large diversity of species, facilitating the existence of more numerous and diverse agroforestry systems in the tropics and subtropics than in the industrialized countries. Besides, socioeconomic factors such as human population pressure, availability of labor, land-holding size, land tenure, and proximity to markets have a major influence on the nature and form of agroforestry systems. As a consequence, considerable variations are found among systems existing in similar agroclimatic conditions. In general, small family farms, subsistence food crops, and emphasis on the role of trees in improving soil quality of agricultural lands are characteristic of tropical agroforestry systems. On the other hand, the driving force of agroforestry in the industrialized nations is environmental protection. The emphasis on monocultural production systems of agriculture and forestry in these countries has led to reduced biodiversity and loss of forest resources and wildlife habitat, increased erosion, nonpoint-source pollution of groundwater and rivers, greenhouse gas emissions, and social changes such as deterioration of family farms.

In summary, the primary objectives of tropical agroforestry systems are to exploit the role of trees on farms to provide products (such as fuelwood, poles, timber, animal fodder, food, fruits, and medicines) and ecosystem services (such as nutrient input and cycling, and soil erosion control). In the temperate zone, environmental amelioration and enhanced economic returns from tree-production systems are the key motivations.

Examples of Common Agroforestry Systems

Innumerable examples of agroforestry systems that exist in different parts of the world have been documented at various levels of detail. This section gives some examples of systems that have received research and development attention during the 1980s and 1990s.

Improved Fallow

The term implies the deliberate planting of species, usually legumes, with the primary purpose of improving soil fertility, mainly through nitrogen fixation and nutrient cycling, in a crop–fallow rotation. Growing herbaceous green-manure species in rotation with economic crops, usually called green manuring, is an age-old agricultural practice. The use of woody species as short-rotation fallows in the context of reduced fallow length in shifting cultivation cycles is a relatively new practice, and that is what is implied by the term 'improved fallow' in the context of agroforestry. The term 'managed fallow' is sometimes used to refer to such improved fallows in order to distinguish them from natural fallows characterized by colonization by natural vegetation, as in traditional shifting cultivation. Research on improved fallow has increased substantially during the 1990s, spear-headed by ICRAF, primarily in Africa. The main woody species used are leguminous, belonging to the genera *Sesbania*, *Tephrosia*, *Crotalaria*, *Mimosa*, and *Cajanus*.

Alley Cropping

In the tropics, alley cropping was developed in the 1970s and 1980s in an effort to find alternatives to long-fallow shifting cultivation. In this practice, arable crops are grown between hedgerows of planted shrubs and trees, preferably leguminous species that are periodically pruned to prevent shading to crops (Figure 1). The biomass, sometimes called prunings, obtained by pruning the trees and shrubs, is a source of mulch and green manure. Besides, leguminous woody species add nitrogen to the system through biological nitrogen fixation. The main motivation and rationale for alley cropping were the perceived soil-improving potential of fast-growing, leguminous woody species and their ability to withstand repeated cutting. It was hypothesized that by integrating such woody species into food crop production systems in simultaneous combinations, some of the benefits of natural fallows of shifting cultivation could be realized. Following that, many studies on alley cropping were undertaken in various parts of the tropics. Indeed, alley cropping was the

Figure 1 Hedgerow intercropping, commonly called alley cropping, is a much-researched tropical agroforestry technology. A large number of tree/shrub species have been tried as hedgerow species in association with a number of agricultural crops in different situations. This picture shows an alley-cropping field with *Leucaena leucocephala* in contour hedgerows, with cow pea (*Vigna unguiculata*) as the crop in the alleys, in Ibadan, Nigeria. Courtesy of B.T. Kang.

Figure 2 In the temperate region, alley cropping refers to growing agricultural (herbaceous) crops in the wide alleys between rows of trees; each tree row bordering the alley could be single-, double-, or multiple-row thick. Black walnut, *Juglans nigra*, is the most widely used tree for alley cropping in North America. The photo shows an alley-cropping plot of black walnut and maize (*Zea mays*) in Indiana, USA. Courtesy of S. Jose.

most widely researched topic in tropical agroforestry during the 1980s and much of the 1990s. In interpreting the results of these studies, some experts have used the data to defend alley cropping, others to denigrate it. A key issue concerning alley cropping is its ecological adaptability. The provision of nutrients through decomposing mulch, a basic feature of alley cropping, depends on the quantity of the mulch as well as on its quality and time of application. If the ecological conditions do not favor the production of sufficient quantities of mulch (as is the case in the dry tropics), there is no perceptible advantage in using alley cropping.

In temperate regions, alley cropping refers to growing herbaceous (agricultural) crops in the wide alleys between rows of trees that are planted in single or grouped rows (Table 1). Black walnut (*Juglans nigra*), a valuable timber- and nut-producing tree, is the species most widely used in alley cropping in North America (Figure 2). The common spacing adopted is 12.5 m between tree rows and 3 m between trees within a row (270 trees ha^{-1}). The driving force behind most temperate-zone alley cropping is economic benefits from the intercrop. Soil conservation in gently sloping lands could be an additional advantage. Quite a large amount of research has been conducted on various aspects of this system, especially in the eastern and Mid-western USA and Canada. Other similar systems in the temperate regions include the traditional intercropping practice used in the establishment of fruit and nut trees, including olive (*Olea* spp.) and grapes (*Vitis* spp.) in Europe, with pecan (*Carya illinoensis*) trees in south-eastern USA, and with paulownia (*Paulownia* spp.) trees in China.

Homegarden, Shaded-Perennial, and Multistrata Systems

These terms are used to describe a set of intimate plant associations in the highly populated parts of tropics. The word 'homegarden' is used rather loosely to denote diverse practices from growing vegetables in backyards, to complex multistrata systems (Figure 3). Homegardens can be found in almost all tropical and subtropical ecozones where subsistence land-use systems predominate. The choice of plant species in homegardens is determined by several factors: in addition to environmental and socioeconomic factors, dietary habits of people and market demands of the locality are important. The nature of species is remarkably similar among different homegardens in various places: fruit trees and other food-producing trees are the dominant components in most. The combination of crops with different production cycles provides an uninterrupted supply of food products throughout the year. These products are usually used for home consumption, with very little reaching the markets.

Growing shade-tolerant commercial species under the canopy of shade trees, commonly called shaded-perennial systems or plantation crop combinations, is another common agroforestry practice that is structurally similar to the homegarden. Traditionally, most of the so-called tropical plantation crops such as oil-palm (*Elaeis guineensis*), rubber (*Hevea brasiliensis*), coconut, cacao, cashew (*Anacardium occidentale*), tea (*Camellia sinensis*), and black pepper (*Piper*

Figure 3 Multistrata systems involving commercial crops such as cacao (*Theobroma cacao*) and coconut (*Cocos nucifera*) are common in many parts of the tropics. The photo shows such a system including young cacao plants in between rows of peach palm (*Bactris gasipaees*) in association with black pepper (*Piper nigrum*) in south-eastern Bahia, Brazil.

nigrum) were developed as monocultural production enterprises. Contrary to popular belief, however, substantial areas under these crops are grown by smallholder farmers who cultivate them in association with a number of other crops. Crop combinations with coconut and coffee and cacao under shade trees are two systems on which considerable research has been done in South Asia and Central America respectively (Figure 4). These plantation crop mixtures ensure efficient use of available resources such as water, nutrients, and sunlight, as well as production of different products throughout the year. Moreover, predominance of perennial rather than annual crops in the system results in a relatively high ratio of nutrients stored in the vegetation to those stored in the soil; this ensures an effective nutrient cycle and relatively small hazard for leaching and erosion.

Silvopastoral Systems

This term is used to denote animal production systems in association with trees; these represent a major form of agroforestry in both tropical and temperate environments. Both fodder-producing and commercial timber species of trees are used in these systems. The systems in many temperate regions such as Australia and New Zealand, southern Europe, and south and south-eastern USA involve mostly pasture and animals under commercial tree plantations (Figure 5). In these cases, where the trees do not have any fodder value, the main objective is to enhance the economic return from the enterprise during the early growth of plantation when there is no economic return from timber sales. Tropical silvopastoral systems include both grazing systems of pasture under commercial tree stands as in the temperate regions and growing and managing fodder-producing trees on farmlands or as community woodlots. In the latter case, trees and shrubs

Figure 4 Shaded-perennial crop systems involve growing shade-tolerant commercial species such as coffee (*Coffea* spp.) and cacao (*Theobroma cacao*) under overstorey shade trees. The photo shows coffee under *Erythrina poeppigiana*, a leguminous tree in Costa Rica. The tree is pollarded (pruned) two to three times a year to regulate shade received by the coffee bush. Pollarded tree trunks are in the foreground, while a nonpollarded tree is in the rear. Courtesy of R.G. Muschler.

Figure 5 Various types of silvopastoral systems exist in different parts of the world. This photo, from Florida, USA, shows a stand of widely spaced rows of slash pine (*Pinus elliottii*) with understorey growth of forage grass; the system allows cattle grazing for many years after plantation establishment.

that are known for their excellent fodder value are grown in intensive or extensive systems. Intensive systems include cut-and-carry systems or protein banks (Table 1) where the trees and shrubs are grown in block configurations or along field and plot boundaries or other designated places, and the foliage is lopped periodically and carried for stall-feeding of animals. Sometimes such fodder species are also grown in alley-cropping combinations. The fodder trees grown on field and plot boundaries may be used as live-fence posts, which are then managed as in cut-and-carry systems. Because of the importance of livestock in both subsistence and commercial production systems, tree-fodder and silvopastoral systems are a major area of research in agroforestry.

Agroforestry Applications for Soil Conservation and Reclamation

Direct or supplementary use of trees and shrubs to control soil erosion is a widespread agroforestry practice in both tropical and temperate regions. Direct use encompasses increased soil cover by trees and shrubs, permeable hedgerow barriers, natural terrace formation by soil accumulation up-slope of hedgerows, and increased soil resistance to erosion by maintenance of organic matter. In supplementary use, the trees and shrubs are not the primary means of checking erosion, but they support other soil conservation structures through the stabilizing effect of the tree root system, while yielding valuable products such as fruits, animal fodder, and firewood. Use of trees and shrubs as a windbreak is a major form of agroforestry practice for soil conservation in drier and wind-prone areas. Windbreaks are narrow strips of trees, shrubs, and/or grasses planted to protect fields, homes, canals, and other areas from the wind and blowing sand. Shelterbelts are a type of windbreak consisting of long, multiple rows of trees and shrubs, usually along the seacoast, to protect agricultural fields from inundation by tidal waves. Riparian buffer strips formed by planting strips of perennial vegetation (of trees, shrubs, and grasses) in between crop fields or pastures and water bodies such as streams, ponds, lakes, and wetlands, are used increasingly in North America to rehabilitate degraded waterways that drain such heavily fertilized lands. Investigations on multiple benefits of stream rehabilitation and bioremediation using multispecies riparian buffer strips in the Midwestern USA have shown the substantial buffering capacity of such vegetated riparian buffers to remedy nonpoint-source contamination of waterways.

Other Agroforestry Applications

In addition to the common types of practices and systems described above, there are many others that have been relatively little studied. These include extensive silvopastoral systems, fuelwood lots, scattered multipurpose trees on farmlands, and tree planting for reclamation of wastelands and problem soils. For example, establishing woodlots of multipurpose (fodder- and fuelwood-yielding), salt-tolerant trees is widely practiced for reclaiming extensive tracts of salt-affected soils in north-western India. The parkland system is an extensive tree-intercropping system in sub-Saharan Africa that originated from the traditional shifting cultivation. During clearance of forest for agricultural production, farmers conserve trees that grow on crop fields for shading, fodder, fruits, and medicines. The trees that are most common include *Faidherbia albida*, *Parkia biglobosa*, and *Vitelleria paradoxa*. Other agroforestry systems involving a whole host of trees and production of a variety of nontimber forest products such as fruits, food, tannins, medicines, resins, and honey are present in several tropical and some temperate regions. The role of agroforestry in the buffer zones around protected parks and bioreserves is another aspect that is often mentioned, but not studied in deserving seriousness.

Forest farming is another practice, recognized especially in temperate-zone agroforestry (Table 1). It refers to growing specialty crops and products in forests and other wooded areas on a regular annual basis. Most notable among them in North America include cultivation of ginseng (*Panax* spp.) for medicinal purposes and the pharmaceutical industry, ferns and other ornamental species, mushrooms and other specialty food products, and pine straw for mulching. The concept of forest farming is not unique, however, to the temperate regions. As mentioned earlier, agroforestry is also used as an approach to exploiting a large number of nontimber forest products around the tropics.

Several indigenous agroforestry systems involve a multitude of such lesser-known species, especially woody species. They have come to be known as 'multipurpose trees' or 'multipurpose trees and shrubs.' The exploitation of these species and the agroforestry practices involving them have wide implications in food security and environmental protection, as well as conservation and use of genetic resources to meet current and future needs. The main difficulty in basing large-scale development plans on these practices is that the available qualitative descriptions by themselves often do not provide the necessary

technical basis for preparing sound projects that involve large-scale investments.

Agroforestry and Ecosystem Services

Soil Productivity and Protection

One of the main conceptual foundations of tropical agroforestry is that trees and other vegetation improve the soil beneath them. Observations of interactions in natural ecosystems and subsequent scientific studies have identified a number of facts that support this concept. Research results during the past two decades show that three main tree-mediated processes determine the extent and rate of soil improvement in agroforestry systems. These are: (1) increased N input through biological nitrogen fixation by nitrogen-fixing trees; (2) enhanced availability of nutrients resulting from production and decomposition of substantial quantities of tree biomass; and (3) greater uptake and utilization of nutrients from deeper layers of soils by deep-rooted trees.

Nitrogen-fixing trees (NFTs) are a valuable resource in agroforestry systems. Some of the widely held assumptions about their benefits could, however, be wrong or incomplete. Because of methodological difficulties in quantifying N_2 fixation, especially in older trees, our understanding of the extent of N_2 fixation, and, therefore, the benefit that is actually realized by using NFTs in agroforestry systems is unsatisfactory. Furthermore, it is not clearly understood how much of the N_2 that is fixed by an NFT is actually utilized or potentially made available to an associated crop during its growth cycle, and how much goes into the soil's N store for eventual use by subsequent crops.

Biomass decomposition patterns vary greatly among agroforestry tree species. Several biomass (litter)-quality parameters, based on the chemical composition of plant tissues, have been developed to interpret these patterns: ratios of C to N, polyphenols to N, lignin to N, and (polyphenols + lignin) to N. Using this information, we can develop management strategies to manipulate the decomposition of plant biomass in agroforestry systems, thereby regulating the rates of nutrient release in the short term, and improving soil fertility, via improved soil organic-matter status, in the long term.

Compared to our knowledge of aboveground biomass additions in agroforestry systems, much less is known about the dynamics of belowground biomass. Experimental evidence shows that deep roots of trees can take up subsoil nitrate from beyond the rooting depth of crops. There seems to be little potential, however, for tree roots to take up immobile nutrients such as P from below or beyond the root zone of interplanted crops. Furthermore, the role of trees in the uptake of nutrients other than N and P is little studied in agroforestry systems.

The other major avenue of soil improvement through agroforestry is through soil conservation. When properly designed and managed, agroforestry techniques can contribute to ecosystem protection and restoration functions by reducing water and wind erosion and enhancing soil productivity.

Carbon Sequestration

Considerable interest has been evinced lately in the scientific community about the carbon sequestration potential of agroforestry. Agroforestry systems such as fuelwood plantations, shelterbelt/windbreak systems and woodlots may have the potential to sequester C, or offset fossil fuel emissions by substituting sustainably produced fuelwood and fodder. Based on a preliminary assessment of national and global terrestrial C sinks, two primary beneficial attributes of agroforestry systems have been identified: (1) direct near-term C storage (decades to centuries) in trees and soils; and (2) the potential to offset immediate greenhouse gas emissions associated with deforestation and subsequent shifting cultivation. A projection of carbon stocks for smallholder agroforestry systems indicated C sequestration rates ranging from 1.5 to 3.5 Mg C ha^{-1} $year^{-1}$ and a tripling of C stocks in a 20-year period, to 70 Mg C ha^{-1}. According to one estimate, median carbon storage by tropical agroforestry practices is around 9, 21, and 50 Mg C ha^{-1} in semiarid, subhumid, and humid ecozones, respectively. The total carbon emission from global deforestation at the currently estimated rate of 17 million ha per year is 1.6 Pg. Assuming that 1 ha of agroforestry could save 5 ha from deforestation and that agroforestry systems could be established in up to 2 million ha in the low-latitude (tropical) regions annually, a significant portion of carbon emission caused by deforestation could be reduced by establishing agroforestry systems. These are rough estimates however. C sequestered in agroforestry systems varies with a number of site- and system-specific characteristics, including climate, soil type, tree-planting densities, and tree management. The key point is that agroforestry could have a major role to play in the global terrestrial C budget; this is an area that needs to be researched in more depth.

Other ecosystem benefits such as water-quality improvement through agroforestry have received little attention in tropical agroforestry research, compared with the level of interest and research in this area in the temperate zone.

Agroforestry Research

The establishment of ICRAF in 1977 is considered to be the starting point of agroforestry research on a global scale. Since it happened in the wake of frustrations arising from the Green Revolution's failure to benefit poor farmers and escalating land-management problems such as tropical deforestation, fuelwood shortage, and soil degradation, the expectation was that investments in agroforestry research would contribute substantially to addressing these problems. Research results in agroforestry were therefore expected to be of a problem-solving and application-oriented nature.

Although the scientific gains in agroforestry during recent decades have been impressive, doubts and concerns have been expressed that the original expectations of agroforestry research have not been fulfilled. Some of the reasons for these concerns can be related to the prevailing myths about the science of agroforestry, and misconceptions about the practice of conducting agroforestry research. A major myth centers around the popular perception that agroforestry involves 'miracle' tree species and their 'magical' ability to improve soil productivity in agroforestry systems. Our understanding of the main tree-mediated processes that determine the extent and rate of nutrient cycling and soil improvement in agroforestry systems is far from satisfactory, so some of the widely held assumptions about their benefits could be wrong or incomplete.

Misconceptions also abound in agroforestry research. A major one is about research methodologies. In agroforestry research, be they of biophysical or socioeconomic nature, we rely on methodologies that have been developed in specific disciplines, often for conditions that are simpler than (or different from) those of agroforestry. Such methodologies may not encompass the special conditions and requirements of agroforestry. Agroforestry systems are often a puzzle to economists too. Since the basic-needs approach of classical political economics has been replaced by the market-superiority premise, economic efficiencies of farm enterprises are calculated based on their profit generation. Although tropical agroforestry systems that are primarily of subsistence nature fulfill the basic needs of farm families, they rank very low in the value premises and theoretical assumptions that underline neoclassical analysis. Yet, agroforestry systems have flourished for a long time. Thus, agroforestry systems pose challenges to ecologists and economists alike. Many such misconceptions about research in agroforestry are, however, not unique to agroforestry, but are characteristic of most such application-oriented research on low-input, integrated land-use systems to support the evasive and ill-defined goals of rural development.

Future Directions

The gains and developments of more than two decades of agroforestry research and development are certainly impressive. In particular, the second half of the 1990s has witnessed a large number of top-quality research and specialized publications. Undoubtedly agroforestry is now on a firm scientific footing. Today agroforestry is no longer a mysterious enigma that defies science and scientific principles, as it was perceived two decades ago. Agroforestry is well on its way to becoming a specialized science at a level similar to those of crop science and forestry science.

These hard-earned scientific gains need to be put to practice for solving the problems we had set out to address when we started research. In order to accomplish that, a two-pronged approach seems to be the best strategy. First, process-oriented research should be extended to hitherto underresearched aspects of agroforestry, and second, a rethinking is needed on overall research and development efforts at all levels (local, regional, and global). Relatively underresearched aspects include the 'agro' (i.e., crop) component of agroforestry. All crops and their cultivars/varieties that are used in agroforestry research today are those that have been developed for sole-crop conditions, and they may not be the best for mixed stands, as in agroforestry systems. Domestication and improvement of indigenous and underexploited trees is an area that can yield immediate dividends. Furthermore, research endeavors have so far been limited mostly to plot and field levels; they need to be scaled up to landscape, watershed, and ecosystem levels. Another research priority is policy and market issues pertaining to valuation of nontimber forest products. Enhancing the adoption potential of agroforestry technologies through improved extension techniques and methodologies for impact assessment are two other areas that need immediate research attention in the near future.

Agroforestry systems, be they practiced in the tropics or temperate regions, provide many basic needs and ecosystem services, and thus contribute to many regional developmental goals. These underexploited systems have the potential to develop into a set of major land-use options in the twenty-first century.

Acknowledgment

This chapter is adapted from the author's work: Nair PKR (2001) Agroforestry. In: Tolba MK (ed.) *Our Fragile World: Challenges and Opportunities for*

Sustainable Development. Forerunner to *The Encyclopedia of Life Support Systems*, vol. I, pp. 375–393. Paris, France: UNESCO.

List of Technical Nomenclature

Agroforestry	Purposeful growing of trees, crops, and sometimes animals in interacting combinations for a variety of objectives. Agrisilviculture = trees + crops; silvopasture = trees + pasture/animals; agrosilvopasture = trees + crops + animals/pasture
Alley cropping	Growing crops in the interspaces or alleys between planted trees or shrubs. In the tropics, alley cropping usually takes the form of hedgerow intercropping, where crops are grown in the alleys between regularly pruned hedgerows of planted shrubs or trees
Fallow	Land resting from cropping, which may be grazed or left unused, often colonized by natural vegetation. An improved fallow refers to deliberate planting of fast-growing species for rapid replenishment of soil fertility
Homegarden	A subsistence farming system consisting of integrated mixtures of multipurpose trees and shrubs in association with crops and sometimes livestock around homes, the whole unit managed intensively by family labor
ICRAF	International Centre for Research in Agroforestry, now World Agroforestry Centre (Nairobi, Kenya)
Multipurpose tree (and shrub)	A tree/shrub that is grown for multiple products and/or services
Multistoried or multistrata system	An arrangement of plants forming distinct layers from the lower (usually herbaceous) layer to the uppermost tree canopy

See also: **Windbreaks and Shelterbelts**

Further Reading

Agroforestry Systems. The Netherlands: Kluwer Academic. Available online at: www.kluweronline.nl.

Buck LE, Lassoie JP, and Fernandes ECM (eds) (1999) *Agroforestry in Sustainable Agricultural Systems*. Boca Raton, FL: Lewis/CRC.

Buresh RJ and Cooper PJM (eds) (1999) The science and practice of short-term improved fallows. *Agroforestry Systems* 47: special issue.

CABI International (1998) *The Forestry Compendium*. Wallingford, Oxon, UK: CABI. CD-ROM.

Garrett HE, Rietveld WJ, and Fisher RF (eds) (2000) *North American Agroforestry: An Integrated Science and Practice*. Madison, WI: American Society of Agronomy.

Leakey RRB, Temu AB, Melynk M, and Vantomme P (eds) (1996) *Domestication and Commercialization of Non-Timber Forest Products in Agroforestry Systems*. Rome, Italy: FAO.

Nair PKR (ed.) (1989) *Agroforestry Systems in the Tropics*. Dordrecht, The Netherlands: Kluwer.

Nair PKR (1993) *An Introduction to Agroforestry*. Dordrecht, The Netherlands: Kluwer.

Nair PKR and Latt CR (eds) (1998) *Directions in Tropical Agroforestry Research*. Dordrecht, The Netherlands: Kluwer.

Nair PKR, Rao MR, and Buck LE (2004) *New Vistas in Agroforestry: A Compendium for the First World Congress of Agroforestry 2004*. Dordrecht, The Netherlands: Kluwer.

The Overstory. Available online at: www.agroforester.com/overstory/ovs.html.

Young A (1997) *Agroforestry for Soil Management*, 2nd edn. Wallingford, Oxon, UK: CAB International.

Air Phase *See* **Aeration**; **Diffusion**

Albedo *See* **Energy Balance**; **Radiation Balance**

Allophane and Imogolite *See* **Amorphous Materials**

ALLUVIUM AND ALLUVIAL SOILS

J L Boettinger, Utah State University, Logan, UT, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Alluvial soils are some of the world's most useful and productive soil resources. Alluvium, the parent material of alluvial soils, is the sediment deposited by fluvial systems such as rivers and streams. Alluvium includes a wide variety of compositions and textures, depending on the source of geologic materials and the depositional environment. Alluvium occurs in all climate regimes and underlies geomorphic surfaces ranging in age from zero to millions of years. Therefore, alluvial soils are also extremely diverse.

Alluvium

Alluvium is ultimately dervied from the weathering and erosion of bedrock such as basalt or granite, or other unconsolidated sedimentary deposits, including colluvium (gravity-deposited), loess or eolian sand (wind-deposited), or alluvium. Important characteristics of alluvium, which ultimately influence the properties of alluvial soils, include composition, texture, and landform.

Composition

The chemical and mineralogical composition of alluvium depends on the lithology, or type of rock materials, from which the sediments are derived: alluvium derived from calcareous sedimentary rocks such as limestone and dolomite will be calcareous; alluvium derived from gypsum-bearing shales will be gypsiferous; alluvium derived from siliceous crystalline igneous rocks will be rich in quartz and feldspar, and lack carbonates and gypsum; alluvium derived from several lithologic sources will have a mixed composition; and so on. For example, the composition of alluvium in the vast San Joaquin Valley of central California differs from east to west. The eastern side of the valley comprises alluvium derived from granitic rocks of the Sierra Nevada mountain range and therefore is rich in quartz and feldspar, with minor amounts of mafic minerals such as hornblende and biotite. The western side of the valley comprises alluvium from the dominantly marine sedimentary rocks of the southern Coastal Ranges. The west-side alluvium contains varying amounts of sulfates, carbonates, and a diversity of silicate minerals, reflecting the range in composition of the marine shales, siltstones, sandstones, and limestones.

Texture

The texture of alluvium, or relative distribution of particle size, depends primarily on the energy of the fluvial environment in which the sediments were deposited. High-energy fluvial environments such as channels of braided streams can carry and therefore ultimately deposit relatively large particles. For example, glacial outwash deposited by high-discharge streams carrying glacial meltwater is typically cobbly to gravelly. Lower-energy fluvial environments deposit finer-textured alluvium. For example, sediments deposited by overbank flooding are typically clays, silts, and fine sands. An example of gravelly channel deposits and silty overbank flood and distal fan deposits is shown in Figure 1.

Different components of the stream channel vary in their ability to carry and deposit sediment, and the stream channel migrates across the landscape in response to changing discharge and bed loads. As a result, alluvium is typically stratified, i.e., composed of alternating layers of different textures (Figure 2b). Increasingly finer textured deposits, or a fining-upward sequence, often overlie coarser-textured alluvium.

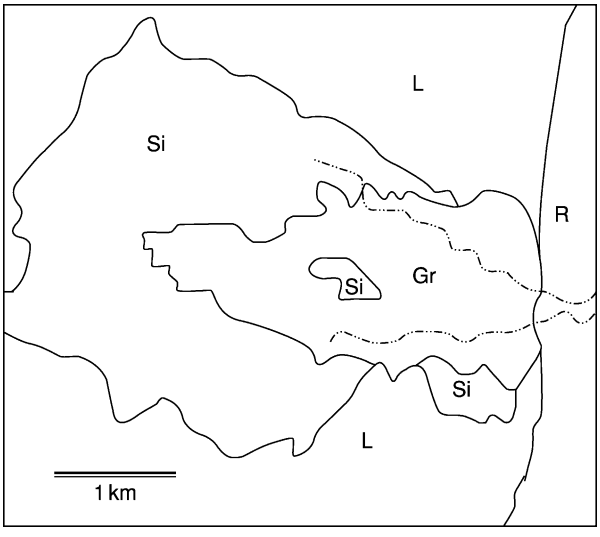


Figure 1 Schematic representation of the Holocene Green Canyon alluvial fan cut into and deposited on lacustrine deposits (L) of Pleistocene Lake Bonneville, Cache Valley, northern Utah, USA. Streams draining rocky limestone and dolomite uplands (R) deposited gravelly channel alluvium (Gr) and silty overbank and distal fan alluvium (Si). Holocene soils are the Green Canyon series (loamy–skeletal Typic Haploxerolls) formed in gravelly alluvium, and the Millville series (coarse–silty Typic Haploxerolls) formed in silty alluvium. (Information derived from the Soil Survey of Cache Valley Area, Utah.)

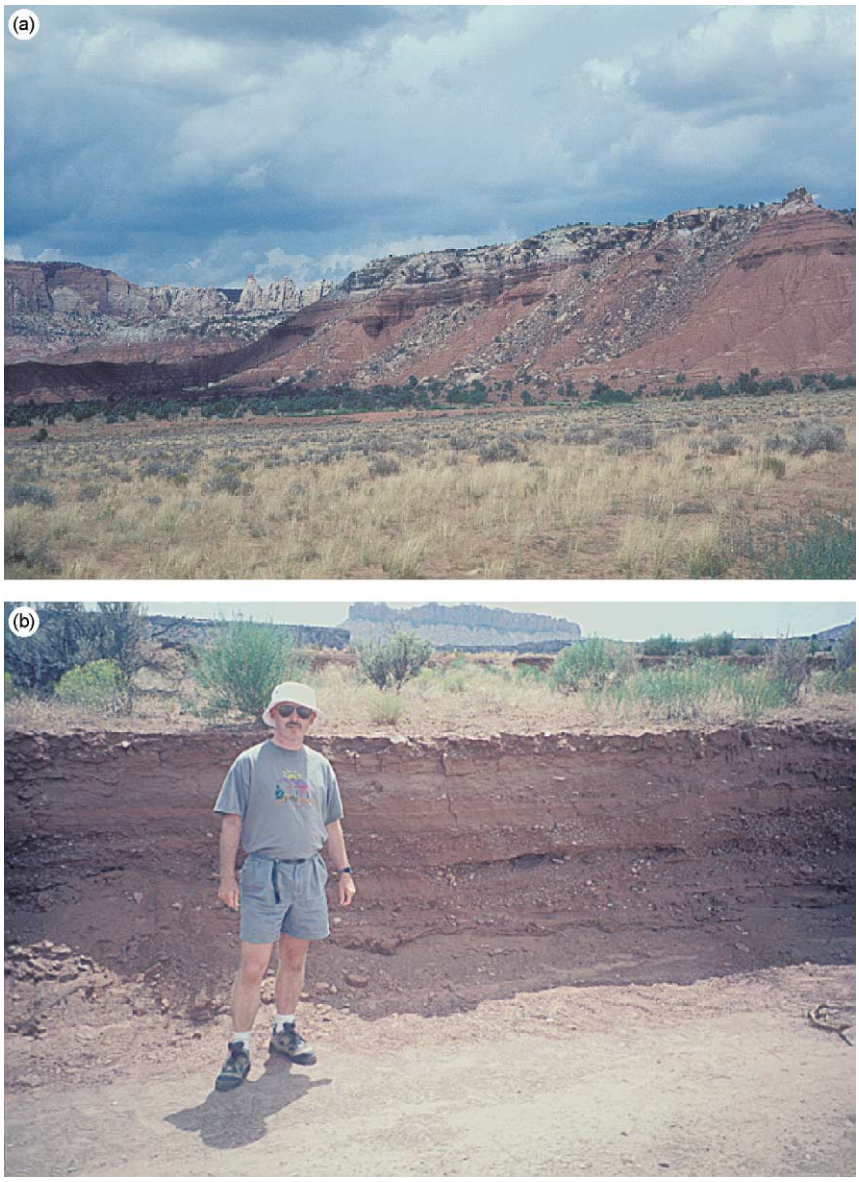


Figure 2 (a) Alluvium from the Mesozoic sedimentary rocks of the Circle Cliffs area of southern Utah, USA, deposited in a small, low-gradient alluvial fan that slopes away from the sediment source to the right of the frame; (b) recent alluvial soil (Ustic Torrifluvent) in the Circle Cliffs, USA, showing typical stratification of alluvial sediments. The man is approx. 1.7 m tall.

Alluvial deposits that fine upward reflect the migration of the stream channel away from its initial position on the landscape. The stratigraphy of alluvial deposits can provide clues to the history of landscape evolution.

The composition of the geologic source material also influences the texture of alluvium. Clearly, alluvium derived from sandstone is typically sandy, whereas alluvium derived from loess is silty. The physical weathering characteristics of the rock source will also influence texture of the alluvium. For example, alluvium derived from granitic rocks is typically sandy to fine gravelly – the size range of the individual mineral grains – because the rock fabric typically shatters into individual grains.

Landforms

Alluvium occurs in deposits that compose a variety of different landforms. Recent alluvium occurs on nearly level floodplains adjacent to streams and rivers. Older alluvium occurs on alluvial terraces, which are generally higher than associated floodplains and are not subject to frequent flooding. Higher terraces subject to upland erosion are often dissected. Alluvial terraces are also known as stream

or river terraces. Pleistocene glacial outwash occurs on outwash terraces.

Streams draining mountainous terrain that discharge into wide valleys or basins, such as in the Basin and Range Province of western North America, deposit alluvial fans. Components of the alluvial fans include floodplains, fan terraces, and dissected fan remnants. Alluvial fans emanating from adjacent streams can coalesce and form bajadas. Finer-textured distal fan deposits can form alluvial plains or basins on the valley floor, and overbank deposits can form interfan basins between low-gradient alluvial fans. Alluvial fans are generally sloped away from the sediment source (Figure 2a), with steeper slopes near the source and lower gradient slopes far from the source.

Alluvial Soils

The morphological, physical, chemical, and mineralogical properties of alluvial soils depend greatly on the characteristics of the alluvial parent material in which the soils formed, especially when the soils are young. As alluvial soils develop with time, the other soil-forming factors influence the resulting soil properties.

Recent Alluvial Soils

Recent alluvial soils are often highly stratified, containing layers of alluvium that were deposited successively and/or in fining-upward sequences (Figure 2b). Soils on active floodplains receive deposits of new alluvium with each flooding episode. The amount of alluvium deposited during each event will vary. Small amounts of material deposited on the soil can be barely perceptible and incorporated into the underlying surface horizon rapidly, the rate of which depends on the climate and biota. Larger amounts of new alluvium can completely bury underlying soils.

Because of periodic disturbance by flooding, soils on recent floodplains often develop only A or O horizons, resulting from the near-surface deposition and decomposition of plant material. Subsequent deposition of new alluvium and reinitiation of landform stability and soil formation results in soils containing one or more buried A or O horizons.

Recent alluvial soils typically can have somewhat elevated concentrations of organic carbon at depth. New alluvium is often derived from the eroded A or O horizons of upland and/or upstream soils. In addition, soils with buried A or O horizons clearly demonstrate an irregular decrease in organic carbon with increasing depth (Figure 3).

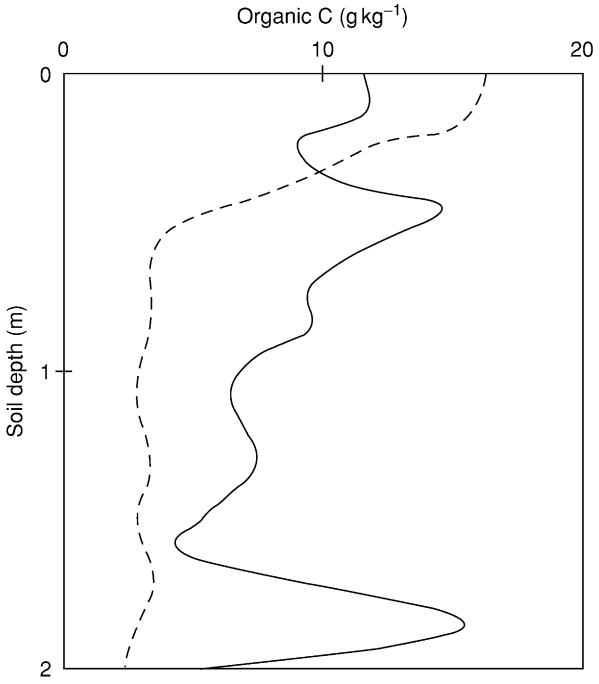


Figure 3 Idealized representation of a recent alluvial soil with an irregular decrease in organic carbon with depth (solid line) compared with an older alluvial soil with a regular decrease in organic carbon (dashed line). The recent alluvial soil probably has a buried A horizon (Ab) at approx. 1.8 m. Organic carbon distributions are typical of alluvial soils in the semiarid, temperate climate of southeastern Utah, USA.

Climate and associated biota further influence the properties of recent alluvial soils. These soils in humid climates generally support dense vegetation, thus developing A and/or O horizons more rapidly than in arid, semiarid, or subhumid climates. Highly soluble minerals such as gypsum and salts, if present in the parent alluvium, will be rapidly dissolved and leached in humid climates, whereas these constituents are often retained in arid climates.

Older Alluvial Soils

In general, older alluvial soils develop when they are no longer subject to periodic flooding events. Surfaces are more stable and thus able to support a stable vegetation cover. Organic carbon in the subsoil is eventually decomposed and the soil develops a regular distribution of organic carbon with increasing depth (Figure 3).

Climate, which influences vegetation and associated biota, further modifies the properties of alluvial soils. In cold climates with permafrost, cryoturbation can disrupt stratification of alluvial layers; in warmer climates, the available precipitation influences the resulting soil properties; in humid climates, alluvial

soils are commonly leached. Depending on the mineralogical composition and texture of the parent material, B horizons develop in the subsoil and accumulate constituents such as silicate clay, free iron oxides, and metal humus complexes. In subhumid, semiarid, and arid climates, alluvial soils are incompletely leached. Depending on alluvium composition and texture, B horizons can accumulate carbonates (nearly ubiquitous), gypsum, soluble salts, etc., with or without silicate clay.

Old alluvial soils have often been subject to changes in climate during their development. This is particularly the case in areas that have existed as alluvial valleys or basins for hundreds of thousands to millions of years, such as the valleys and basins of the Basin and Range physiographic province of western North America. Very old alluvial soils can show the imprinting of several climate regimes, e.g., well developed, clay-rich B horizons that have been engulfed by calcium carbonate accumulations such that clay skins are no longer visible in the field.

Alluvial deposits, landforms, and the associated soils can be subject to active fluvial processes at one time or another. Therefore all or a portion of an alluvial soil profile is subject to erosion. If only a portion of the soil is removed by erosion, the soil has been truncated. Often, soils are eroded down to the more resistant clay- or carbonate-rich B horizon; older, truncated alluvial soils can have a B horizon exposed at the soil surface. Recent alluvial deposits may bury a truncated soil. As soil development begins again, soil-forming processes influence the recent overlying alluvium and the underlying older alluvial soil if shallow enough, resulting in a soil with a welded profile.

Alluvial soils often retain at least a partial record of the history of alluvial deposition. As soils develop on stable or truncated surfaces, they also record the passage of time since deposition, landform stability, and/or truncation. Therefore, alluvial soils are useful for soil-geomorphic studies. Soil development on alluvial surfaces has been used to correlate surfaces of similar ages, decipher paleoclimate, and estimate activity of faults that cut alluvial surfaces.

Classification

The diversity of alluvial soils results in a complex array of potential soil classifications. Recent alluvial soils that lack significant development of surface or subsurface diagnostic features are classified into the order of Entisols (US Department of Agriculture Soil Taxonomy System) or Fluvisols (Food and Agriculture Organization World Reference Base for Soil Resources).

In the Soil Taxonomy System, the formative element 'fluv' is used to connote the alluvial origin and stratified nature of recent alluvial soils. Better-drained Entisols that have high organic carbon levels (more than 0.2%) at depth (125 cm) or an irregular decrease in organic carbon with depth (see **Figure 3**) are in the suborder of Fluvents. Fluvents are further subdivided into 'great groups' by soil moisture regimes (e.g., Udifluvents, Torrifluvents, etc.). Alluvial Entisols that are saturated for prolonged periods during a normal year are Fluvaquents. In arid climates, the organic carbon level in the soil is often low; thus, many youthful alluvial soils in arid climates are classified as Torriorthents.

With greater landform stability and soil development, alluvial soils can develop into a myriad of different soils and thus can occur in all of the other soil orders of the Soil Taxonomy System. Alluvial soils with mollic epipedons and high base status throughout are Mollisols. Alluvial soils in arid climates that develop subsurface diagnostic horizons (e.g., argillic, calcic, gypsic) are Aridisols. Other alluvial soils with ochric or umbric epipedons are Inceptisols, Alfisols, Ultisols, Oxisols, and Spodosols, depending on the diagnostic subsurface horizons and properties present and the degree of leaching. Alluvial soils subject to cryoturbation are Gelisols; alluvial soils that accumulate deep organic soil materials are Histosols. Vertisols and Andisols are also possible, depending on the mineralogical composition, texture, and the degree of soil development.

Human Use of Alluvial Soils

For millennia, humans have used alluvial soils, especially those that are young and less developed, for the production of food. New alluvium rich in organic matter and nutrients provided fertile soils for agriculture.

Alluvial soils have been the sites of the earliest agriculture. This is illustrated by the concentration of early civilization advancement along river corridors in the Mediterranean Basin. For example, early agricultural societies developed along the Tigres and Euphrates rivers in the Middle East, the Nile in Egypt, and the Ebro in the Iberian Peninsula.

Indigenous cultures throughout the world still rely on the productivity and ease of cultivation of recent alluvial soils. Periodic flooding results in the rejuvenation of soil fertility by depositing organic-matter rich sediment on soil surfaces. Zuni and other indigenous peoples of the arid and semiarid southwestern US rely also on the periodic deposition of sand, which acts as mulch to increase infiltration rates and decrease soil moisture loss via evaporation.

Even today, alluvial soils underlie the most productive agricultural regions of the world. For example, the Central Valley of California supports a vibrant, multimillion-dollar agricultural industry. However, many of the soil and landscape properties that make alluvial soils attractive for agriculture are also attractive for urban development. Alluvial soils often have low slopes and occur in wide valleys or plains and are easy to excavate. Therefore, thousands of hectares of productive agricultural soils are being lost each year to urbanization and industrialization, making the preservation of prime agricultural land a world-wide concern.

List of Technical Nomenclature

Alluvial	Relating to alluvium
Alluvial fans	Fan-shaped landforms deposited by streams usually emanating from mountains and emerging on to lower gradient valleys
Alluvium	Sediment deposited by fluvial systems such as rivers and streams
Bajadas	Continuous skirt of sediment adjacent to a mountain range formed by coalescent alluvial fans
Bed loads	Debris moved along the bottom of the stream
Braided streams	Stream that flows in a network of dividing and rejoining channels because flow is diverted around sediment deposited by the stream
Buried soil	Older, developed soil overlain with more recently deposited sediment
Channel	Deepest part of a stream where the main flow of water occurs
Discharge	Rate of flow (volume per unit time)
Dissected	Incised by erosion
Distal fan deposits	Sediments deposited at the far end of the alluvial fan where the stream runs out of water and energy
Floodplains	Landform adjacent to the stream constructed by materials deposited by present flooding activities, and still covered with water during floods
Glacial outwash	Glacier-derived sediments moved by high-discharge steams carrying glacial meltwater
Landscape evolution	Development of deposits and landforms through time
Overbank flooding	Flooding where water rises and flow over the tops of the banks of the stream
Paleoclimate	Past climate(s)
Stratified	Composed of alternating layers
Terrace	Relatively flat to gently sloped surface generally bounded by short, steep slopes
Truncated soil	Soil from which the upper part was removed by erosion

Further Reading

Birkeland PW (1999) *Soils and Geomorphology*, 2nd edn. New York: Oxford University Press.

Busacca AJ, Singer MJ, and Verosub KL (1989) *Late Cenozoic Stratigraphy of the Feather and Yuba Rivers Area, California, with a Section on Soil Development in Mixed Alluvium at Honcut Creek*. In: US Geological Survey Bulletin 1590-G, pp. G1–132. Washington, DC: Government Publishing Office.

Eghbal MK and Southard RJ (1993) Stratigraphy and genesis of Durorthids and Haplargids on dissected alluvial fans, western Mojave Desert. *Soil Science Society of America Journal* 59: 151–174.

Gile LH and Grossman RB (1979) *The Desert Project Soil Monograph: Soils and the Landscapes of a Desert Region Astride the Rio Grande Valley, Near Las Cruces, New Mexico*. Washington. DC: USDA Soil Conservation Service.

Gile LH and Hawley JW (1966) Periodic sedimentation and soil formation on an alluvial-fan piedmont in southern New Mexico. *Soil Science Society of America Journal* 30: 261–268.

Harden JW (1987) *Soils Developed in Granitic Alluvium Near Merced, California*. In: US Geological Survey Bulletin 1590-G, pp. A1–65. Washington, DC: Government Printing Office.

Hillel D (1991) *Out of the Earth: Civilization and the Life of the Soil*. New York: Free Press.

Norton JB (2003) Hillslope soils and organic matter dynamics within a Native American agroecosystem on the Colorado Plateau. *Soil Science Society of America Journal* 67: 225–234.

ALUMINUM SPECIATION

D R Parker, University of California, Riverside, CA, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Aluminum (Al) is the third most abundant element in the Earth's crust, and thus plays a pivotal role in soil science, geochemistry, ecology, and numerous related disciplines. Although it has only one stable oxidation number in nature (+III), the chemistry of Al is extremely complex, and has been studied intensively for more than a century. Although only sparingly soluble at environmental pH values, Al is demonstrably toxic to a variety of organisms, including higher plants, soil microorganisms, phytoplankton, fish, and other aquatic animals.

During the last two decades, considerable emphasis has been placed on the aqueous speciation of Al, specifically its tendency to form stable complexes with an assortment of inorganic and organic ligands. Aluminum is classified as an 'A-type' or 'hard-sphere' metal, and thus has considerable affinity for oxygen-containing ligands (including carboxyl groups on organic molecules) and fluoride (F^-). Understanding the speciation of Al is essential to explaining and predicting its toxicity, as only certain aqueous species seem to elicit any toxic response. An accurate description of Al speciation is also necessary for modeling its geochemical behavior. The solubility of Al-bearing minerals, as an example, is described thermodynamically in terms of the activity of the free, trivalent metal ion, and is independent of the 'side reactions' involving other ligands.

Inorganic Aluminum Species in Soil Solution

The most notable feature of the chemistry of Al(+III) in water is its tendency to hydrolyze. In simple, dilute solutions of low pH, the dissolved Al is primarily the hexaaquo ion, $Al(H_2O)_6^{3+}$. As pH is increased, however, Al reacts with water according to the general scheme:

$$x\mathrm{Al}^{3+} + y\mathrm{H_2O} \rightleftarrows \mathrm{Al_x(OH)_y^{(3x-y)+}} + y\mathrm{H}^+ \qquad [1]$$

with overall formation constants of the form:

$$\beta_{x,y} = \frac{\left(\mathrm{Al_x(OH)_y^{(3x-y)+}}\right)\left(\mathrm{H}^+\right)^y}{\left(\mathrm{Al}^{3+}\right)^x} \qquad [2]$$

where the waters of hydration have been omitted for simplicity, and parentheses indicate ionic activities. The four mononuclear hydrolysis products (i.e., the coefficient $x = 1$) are reasonably well-understood and agreed-upon, and formation constants (log β) are provided in Table 1.

Polynuclear hydroxy complexes of Al (where $x > 1$) can be readily demonstrated in the laboratory, most often by adding a strong base (e.g., NaOH) to a solution of a soluble salt such as $AlCl_3$. As long as the ratio of added OH^- to total Al is <3, stable and soluble polynuclear complexes are formed that can persist for months or even years. The exact structure of these complexes has been the subject of fierce debate, but the most convincing evidence is for a tridecameric, 'Keggin' structure formulated as $AlO_4Al_{12}(OH)_{24}(H_2O)_{12}^{7+}$, and often denoted simply as 'Al_{13}.' This tridecamer consists of a highly symmetrical tetrahedrally coordinated Al centralized in a cage-like structure composed of 12 octahedrally coordinated Al atoms. At present, however, there is little evidence that this polynuclear Al complex exists outside the laboratory, and its existence in soil solution is especially uncertain.

The third hydrolytic reaction of Al involves its precipitation as oxyhydroxide minerals such as gibbsite and bayerite. The solubility of these minerals shows an extremely strong dependence on solution pH, as predicted by the stoichiometry of reaction:

$$\mathrm{Al}^{3+} + 3\mathrm{H_2O} \rightleftarrows \mathrm{Al(OH)_{3(s)}} + 3\mathrm{H}^+ \qquad [3]$$

The mineral gibbsite is often assumed to control the solubility of Al in soils, and measured solubility

Table 1 Aluminum hydrolysis and complexation constants at infinite dilution and 25°C (waters of hydration are omitted for simplicity)

Reaction	*log* β
$Al^{3+} + H_2O \rightleftarrows Al(OH)^{2+} + H^+$	−5.0
$Al^{3+} + 2H_2O \rightleftarrows Al(OH)_2^+ + 2H^+$	−10.1
$Al^{3+} + 3H_2O \rightleftarrows Al(OH)_3^0 + 3H^+$	−16.8
$Al^{3+} + 4H_2O \rightleftarrows Al(OH)_4^- + 4H^+$	−23.0
$Al^{3+} + SO_4^{2-} \rightleftarrows AlSO_4^+$	3.5
$Al^{3+} + 2SO_4^{2-} \rightleftarrows Al(SO_4)_2^-$	5.0
$Al^{3+} + F^- \rightleftarrows AlF^{2+}$	7.0
$Al^{3+} + 2F^- \rightleftarrows AlF_2^+$	12.7
$Al^{3+} + 3F^- \rightleftarrows AlF_3^0$	16.8
$Al^{3+} + 4F^- \rightleftarrows AlF_4^-$	19.4
$Al^{3+} + 5F^- \rightleftarrows AlF_5^{2-}$	20.6
$Al^{3+} + 6F^- \rightleftarrows AlF_6^{3-}$	20.6
$Al^{3+} + H_2PO_4^- \rightleftarrows AlH_2PO_4^+$	~3
$Al^{3+} + HPO_4^{2-} \rightleftarrows AlHPO_4^-$	~7

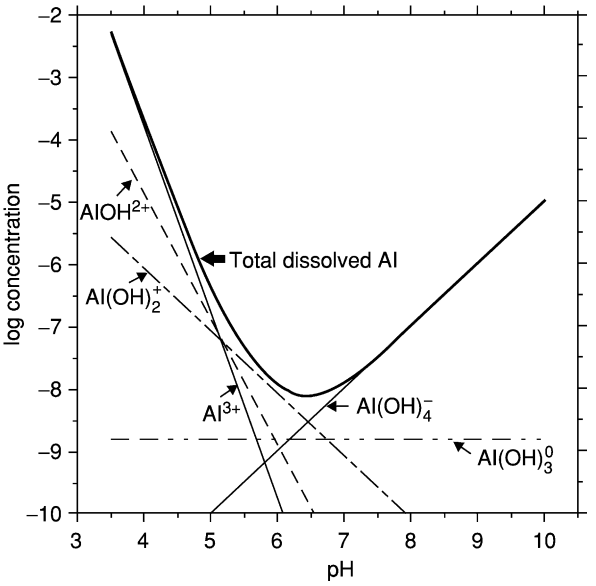


Figure 1 Distribution of mononuclear Al species in a simple system contains no ligands other than the hydroxyl ion that arises from hydrolysis. The total Al in solution is limited, especially at near-neutral pH values, by sparingly soluble oxyhydroxide minerals such as gibbsite, modeled here as $Al^{3+} + 3H_2O \rightleftarrows Al(OH)_{3\,(s)} + 3H^+$, with $\log \beta = -8.0$.

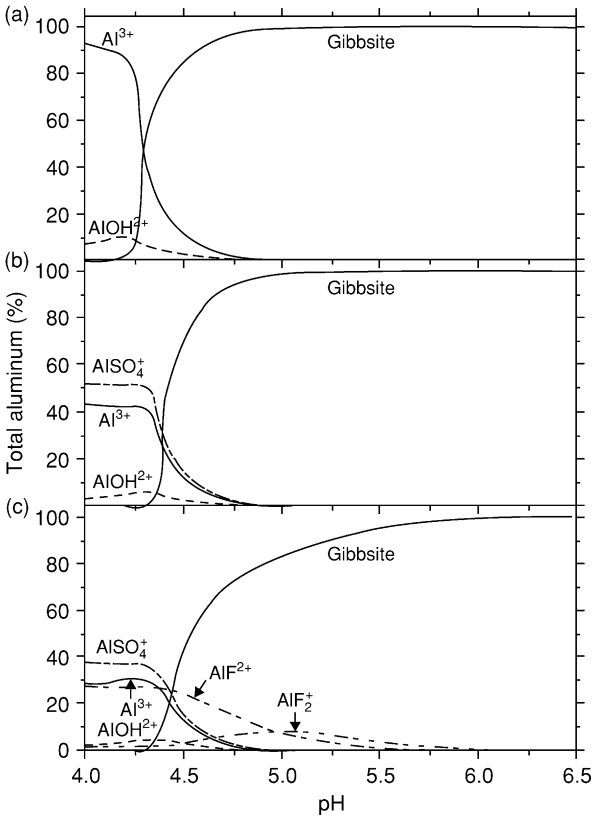


Figure 2 Distribution of Al as a function of pH for a simple solution initially containing 50 μmol l^{-1} total Al. Species comprising $\leq$1% of the total Al are omitted for clarity. (a) Ionic medium is 1 mmol l^{-1} $Ca(NO_3)_2$; (b) ionic medium is 1 mmol l^{-1} $CaSO_4$; (c) ionic medium is 1 mmol l^{-1} $CaSO_4$, and includes 15 μmol l^{-1} total F in solution. In all three cases, the overall solubility of Al is limited above pH 4.5. Speciation was calculated using the formation constants in **Table 1** and the assumed gibbsite solubility depicted in **Figure 1**.

products have varied significantly. However, a log β of -8.0 for the foregoing reaction would be a reasonable value, and the resulting solubility of Al as a function of pH is depicted in **Figure 1**. Note the logarithmic nature of the y-axis; at pH $\sim$6.5, the expected concentration of dissolved Al is only about 10^{-8} mol l^{-1}. In acidic soils with pH $<$5.0, the free Al^{3+} ion dominates, and toxicity is often best correlated with its concentration (or activity). Interestingly, Al solubility also increases at alkaline pH: the solubility of Al at pH 8.7 is about what it is at pH 5.0 (**Figure 1**). Yet, no toxic effects of Al have been demonstrated at alkaline pH, presumably because of the dominance of an anionic species, $Al(OH)_4^-$.

Other inorganic ligands that are important to the aqueous speciation of Al in soils include sulfate, fluoride, and perhaps phosphate; formation constants for the relevant complexes are given in **Table 1**. Interactions between Al and anions such as chloride, nitrate, and (bi)carbonate are too weak to be of any general significance. In light of the stability of the Al-O-Si bonds in aluminosilicate minerals, one might expect significant complexation of Al by aqueous Si, but this does not seem to be the case, presumably due in part to the low solubility of silicic acid in water.

In **Figure 2**, the distribution of Al as a function of pH in acidic soils (e.g., 4.0–6.5) is presented using a linear rather than logarithmic y-axis. In a simple matrix of 1 mmol l^{-1} $Ca(NO_3)_2$, the only significant solution species are Al^{3+} and $AlOH^{2+}$; precipitation of gibbsite becomes dominant at about pH 4.5 (**Figure 2a**). If the nitrate is replaced by sulfate, then a codominant complex – $AlSO_4^+$ – prevails until gibbsite's insolubility again dominates (**Figure 2b**). When a small amount of the stronger ligand F is also included, the precipitation of gibbsite is partially suppressed, and somewhat more Al is maintained in the solution phase out to a pH of about 5.5 (**Figure 2c**). Note the complexity of the Al speciation in **Figure 3** over the pH range 4.0–5.0, even in the absence of any organic ligands – a complexity with significant implications for direct experimental measurement of Al speciation (see below).

Organic Complexes of Aluminum

All soil solutions naturally contain dissolved organic carbon (DOC), although the quantities vary

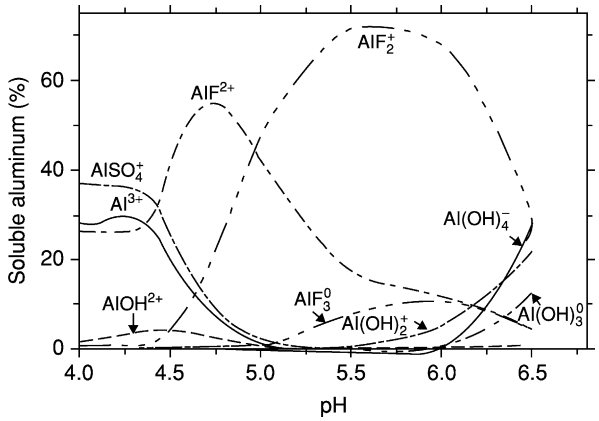


Figure 3 Distribution of Al species in **Figure 2c**, but expressed here as a percentage of the Al that remains in solution after gibbsite precipitates. Species comprising $\leq$1% of the total dissolved Al are omitted for clarity. The absolute concentration of these species is quite low above pH ~5.0, and the distribution of the numerous complexes present shows a marked dependence on pH.

enormously, and may be vanishingly small in some cases. Because of the affinity of carboxyl groups (and, to a lesser degree, acidic hydroxyls and amines) for Al, DOC is potentially a potent chelator of Al. The most ubiquitous form of DOC in soils is 'fulvic acid' (FA), a term used to describe a group of relatively high-molecular-weight compounds that are water-soluble at near-neutral pH values. One of the stable end-products of microbial processing of soil carbon, the FAs have highly variable and complex macromolecular structures, and contain a number of functional groups of varying affinity for Al. Attempts to model predictively the binding between Al and FA have ranged from rather simple to quite elaborate conceptual and mathematical representations. For the former, conditional stability constants (i.e., those that are specific to a given pH and ionic strength) for the formation of a '1:1' complex (here, the '1' for the ligand refers to a single binding site on the macromolecule) have generally been estimated to be in range of 10^3–10^6. Thus, the FAs may have a binding affinity approaching that of F^-, but may also be as weak a chelator of Al as is SO_4 (**Table 1**). The range in reported binding affinities is clearly quite wide, and generalizations about the overall importance of DOC are therefore difficult to make. In certain soils, the FAs may dominate the speciation of Al, while in others they be of little or no significance.

The other notable class of organic ligands for Al are the organic acid anions such as citrate, oxalate, malate, succinate, and propionate, among others. These are common by-products of both plant and microbial metabolism, and are sometimes present at measurable levels in soil solution. They are particularly likely to be abundant in the immediate vicinity of the root surface (i.e., the 'rhizosphere'), because roots often exude organic acid anions and because they are heavily colonized by microflora. The appearance of these ligands may be extremely transient, however, as they are readily metabolized as a carbon and energy source by heterotrophic microorganisms.

Root exudation of organic acid anions has recently been implicated as an important mechanism of the resistance to Al toxicity exhibited by certain plant taxa, consistent with the notion that complexed and chelated forms of Al are relatively harmless, and that it is the free Al^{3+} ion that elicits the toxic response. The stability constants for binding of Al by these ligands are not always precisely known, as different researchers have utilized quite disparate reaction schemes to fit similar potentiometric titration data (which are often highly problematic due to the uncontrolled formation of polynuclear hydroxy species, discussed above). But it seems clear that the tricarboxylate citrate has the highest binding affinity for Al, with 1:1 stability constant of about 10^8. Oxalate and perhaps malate are also sufficiently strong chelators of Al to reduce its toxicity in the rhizosphere, but the other common dicarboxylic acids are probably not under most conditions; monocarboxylic acids such as lactate and acetate are almost certainly of no general significance to Al speciation in soils.

Phenolic groups are important components of humic substances, and various phenolic acids, such as salicylic acid, caffeic acid, and gallic acid, have at times been implicated in the solubilization and mobility of Al in soils. It has been tempting to assign significance to these compounds because their reported log β values for Al binding are often quite high (>10). However, it is important to recognize that these binding constants refer to the fully deprotonated ligand. With the phenolic acids, the ring hydroxyl does not deprotonate except at extremely alkaline pH values; thus the conditional constants that describe their binding affinity in the acidic pH range of relevance are orders of magnitude lower. In comparison to the aliphatic organic acids described above, the phenolics are probably of minor significance to the speciation of Al in most soils.

The speciation of Al is of particular significance to its ecotoxicological effects. With both terrestrial and aquatic organisms, toxicity seems to correlate best with the concentration of free Al^{3+}, an affirmation of the 'free ion model' that is observed with a number of other trace metals. In general, none of the other inorganic or organic mononuclear species have been shown to contribute to the toxic response, including the aluminate anion that prevails at alkaline pH. The Al_{13} polynuclear complex has been shown to be uniquely and acutely toxic to both plants and algae reared in laboratory

solutions, but the environmental significance of this finding remains unknown.

Determining Aluminum Speciation in Experimental Samples

Computational Methods

If all of the Al-complexing ligands present in a soil-solution sample could be accurately quantified, and if all of the pertinent binding constants were exactly known, then the speciation of Al could be readily computed using any one of a number of 'off-the-shelf' chemical equilibrium models (GEOCHEM-PC, MINEQL+, MINTEQA2, etc.). However, several obstacles to this procedure usually persist. Not all of the ligands can be readily identified and quantified at the levels that typically prevail in soil solution. Fluoride is problematic in that few analytical methods exist that are accurate in the needed micromolar range, and F has some affinity for metals other than Al and perhaps for colloidal forms of Al. Polynuclear forms of Al, if present, cannot be predicted using the thermodynamically based ion-association models that underlie all chemical speciation programs. Other problems may include difficulties in the accurate separation of dissolved Al from microcolloidal forms that might pass a submicron filter prior to analysis for total Al.

The highly heterogeneous nature of DOC is particularly problematic, and makes it difficult or impossible routinely to quantify and accurately model the Al-binding characteristics of organic ligands such as FA. The interactions between FA and Al have been modeled by treating the individual binding sites as a hypothetical soluble ligand, and binding stoichiometries include simple AlL complexes, as well those of the form AlL_n and Al_nL. Sometimes a single metal–ligand complex is employed, but often two or more ligands or binding stoichiometries are invoked to simulate the variation in binding affinity among the many sites on natural FA. The most elaborate computational models take into account electrostatic effects and make use of polyelectrolyte theory. To date, none of these approaches has shown any general predictive capability, and the models have usually required site- or even sample-specific empirical recalibration.

These limitations have led researchers to develop and employ a host of empirical methods for the speciation of Al, some of which rely partly on the computational approach described above. It should be noted, however, that chemical equilibrium modeling is often the method of choice for 'clean' solutions prepared from known components in the laboratory, and is invaluable as a heuristic tool for probing the complexities of aqueous Al chemistry in hypothetical solutions such as those depicted in Figures 1–3.

Fractionation Based on Size, Charge, and Reactivity

Much of the purported 'speciation' of Al in the literature would not meet a strict definition of the term, as the chemical entities in solution were not directly probed and quantified. More commonly, the dissolved Al is analytically allocated to several operationally defined categories based on physical separation due to differences in size and/or charge, or based on chemical reactivity with an Al-complexing agent.

Because much of the natural DOC in soil solutions or extracts consists of high-molecular-weight compounds, the DOC-complexed Al should be separable from the smaller species using dialysis or ultrafiltration membranes. Nominal pore diameters as small as 1 nm are available, and the method has been employed with some success. Care must be taken to avoid adsorptive losses on to the membrane, as well as contamination with reactive metals and ligands. The largest difficulty with such methods is achieving an unambiguous cutoff with respect to molecular size. Transport efficiency across the membrane can become very low as the molecular diameter approaches the nominal pore size, and there may be a range of molecular weights present with natural DOC.

A number of chromogenic Al-complexing reagents have been employed over the years to fractionate dissolved Al, including ferron, 8-hydroxyquinoline, chrome azurol S (CAS), and pyrocatechol videt (PCV), among others. When bound to Al, these ligands form a distinctly colored complex that can be quantified using ultraviolet-visible spectrometry (UV-Vis) or, occasionally, fluorimetry. Most typically, these are timed analyses in that the reading must be exactly made after a fixed period of reaction (e.g., 60 s), and this has been achieved in some cases by automating the method using flow-injection analysis (FIA). Alternatively, the reaction is similarly stopped by extraction of the Al complex into an organic solvent. Less commonly, the formation of the colored complex is monitored continuously, and the absorbance versus time data are then fitted to a kinetic model. In all cases, the method depends critically on the assumption that certain Al species are measured quantitatively during the brief reaction with the analytical ligand, and that other, less labile species are completely excluded from the analysis.

When examined critically, this assumption has seldom been found to be completely valid. The mononuclear hydrolysis products are, through proton exchange, in rapid equilibrium with the free Al^{3+} ion, so all should be quantified together. The sulfate

complexes are weak, and seem to be readily 'stripped' by the Al-complexing reagent. With stronger Al complexes, however, such as those with F or citrate, the Al may or may not be exchanged on to the analytical reagent and, if it is, the rate may be both sluggish and unknown. The reactivity of Al–DOC complexes (e.g., the FAs) would be even more difficult to predict.

In a number of cases, this 'rapidly reacting' Al has been assumed to represent the sum of all of the inorganic, mononuclear Al species. Regardless of the validity of this assumption, it is important to note that, in order to construct a detailed picture of the Al speciation (for example, to know the concentration of free Al^{3+}), it is necessary to model computationally the distribution of these species as described previously. The measured, reactive Al is used as the 'total' Al concentration, and the concentrations of F and SO_4 must also be determined and used as model inputs. Thus, many of the empirical fractionation procedures in use are actually combination or 'hybrid' methods, in that part of the 'speciation' is done empirically, and part is done computationally.

Cation exchange resins of both strong- and weak-acid character have been used by a number of researchers to fractionate Al, often in combination with other methods. For example, a column of strong-acid resin may be used to separate inorganic and organic forms of mononuclear Al. The underlying assumption is that the inorganic complexes are positively charged and/or sufficiently labile to be retained by the resin, while organic complexes are neutral or negatively charged, and are sufficiently strong so as not to dissociate while passing through the column. Thus the inorganic species are retained, while the organic complexes are collected in the eluent where they are quantified by analyzing for total Al. Figure 4 presents a typical fractionation scheme wherein total mononuclear Al is quantified by one of the timed spectrophotometric methods, the organic subfraction by ion exchange, and the inorganic species by difference. The method also allows differentiation of any very nonreactive Al such as small colloidal solids, and of dissolved forms that are particularly refractory (e.g., large polynuclear complexes, unusually strong organic chelates).

Note that all of these empirical approaches contain some critical underlying assumptions. Foremost is that, during the analysis or separation of Al, the equilibrium distribution of species is not perturbed. Of course this cannot be strictly true as the separation of some of the Al into different phases, or its selective complexation by a reagent ligand, must inevitably shift the speciation. However, if kinetic hindrances sufficiently retard this reequilibration, then the method may not be severely compromised. In addition, many of the methods

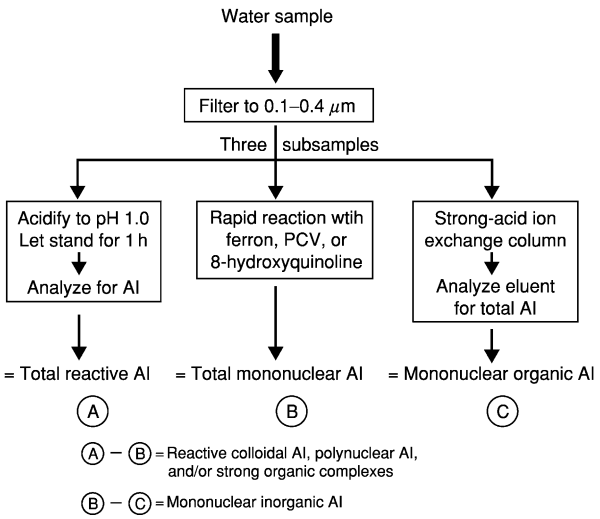


Figure 4 Schematic diagram of a typical fractionation procedure where 'dissolved' Al is allocated into five operationally defined chemical fractions. PCV, pyrocatechol violet.

cause shifts in pH and ionic strength as compared to the original sample due to, for example, the inclusion of concentrated buffers in many of the timed spectrophotometric methods. It must be assumed that these do not cause major shifts in the distribution of Al species, although the pH-sensitivity that is well-illustrated in Figures 1–3 should instill extra caution in this regard.

Speciation Using More Direct Analytical Methods

Aluminum has a single natural isotope, ^{27}Al, with a small quadrupole moment and high resonance frequency, making it an ideal candidate for analysis using nuclear magnetic resonance (NMR). This is an appealing tool for aqueous speciation, because samples can be probed nondestructively at ambient temperatures, and no reagents or resins need be added to the sample matrix; original chemical conditions are thus readily preserved. Indeed, ^{27}Al-NMR has been widely used to study the hydrolysis behavior of Al and, increasingly, its complexation and chelation by various ligands in laboratory-synthesized solutions. An NMR spectrum reflects differences in coordination chemistry in two ways, via chemical shifts and via changes in line width. The former are most profound when the coordination number changes, for example from octahedral to tetrahedral coordination. Line broadening results from asymmetry in the ligand field which, due to ^{27}Al's quadrupole moment, is a general problem that often limits the resolution of the technique.

To date, the greatest limitation to the application of ^{27}Al-NMR to 'real-world' speciation has been its lack of sensitivity: total Al concentrations of

$\sim 10^{-4}$ mol l^{-1}, or, more typically, $\geq 10^{-3}$ mol l^{-1}, have been required using conventional instrumentation. Line broadening is partially to blame for this poor sensitivity, but the presence of Al as an ubiquitous impurity in NMR probes and glass tubes is also responsible. Recently, utilization of specially constructed probes, coils, and sample tubes using low-Al materials has enhanced sensitivity such that concentrations approaching 10^{-6} mol l^{-1} can be addressed. It remains to be seen what the utility of such enhanced-sensitivity ^{27}Al-NMR might be for complex environmental samples containing multiple ligands. In acidic samples, virtually all of the Al is in octahedral coordination, and line-broadening can make it difficult or impossible to resolve the overlapping resonances arising from multiple Al species.

Various forms of liquid chromatography (e.g., IEC, HPLC, FPLC, SEC: see Table 2 for definitions) have been utilized for some 20 years in attempts to separate analytically and quantify various complexes of Al. The number of published methods is very large, and the last decade has seen a tremendous surge in interest in these techniques. The earlier papers tended to use postcolumn reaction with one of the aforementioned chromogenic reagents, with subsequent detection using UV-Vis or fluorimetry. More recently, there has been increased emphasis in 'hyphenated' techniques, where the chromatographic separation is coupled with an atomic spectroscopy instrument such as ICP-MS, ICP-AES, or ET-AAS (see Table 2 for definitions). These element-specific methods of detection have greatly improved sensitivity; detection limits are now $\sim 10^{-7}$ mol l^{-1} and even lower. The chromatographic separations have been based on differences in chemical properties such as charge, size (e.g., SEC), and hydrophobicity (in the case of organic complexes using RPC).

To date, most of the published studies have focused on synthetic laboratory solutions and biological fluids (e.g., human serum). Among the organic ligands, the emphasis as been on mammalian proteins and the aliphatic organic acids; comparatively little work has been done on environmental samples containing heterogeneous DOC. Thus, the utility of these chromatographic methods for soil solutions remains unknown, although the emergence of techniques such LC-ES-MS will undoubtedly help us to understand better the structure of Al-complexing humic substances. As with the fractionation methods described previously, a general caution applies to these chromatographic techniques: because of the eluents used, there can be marked changes in pH, ionic strength, or other chemical properties during the analysis. In order for the results to reflect the *in situ* speciation of the sample, assurance must be gained that the complexes of interest do not shift due to these changes. Such stability during analysis might arise thermodynamically, but more likely would be a result of kinetic stability that requires careful evaluation and verification.

Finally, mention can be made of two indirect approaches to the speciation of soil solution Al. The first utilizes an F-selective electrode to measure directly the activity of the free F^- ion in the sample. A separate measurement is made after addition of a metal-complexing buffer to liberate all of the F, thus generating a value for total F in solution. These two values are combined with the measured prevailing pH and, utilizing the equilibrium relations shown in Table 1, the free Al^{3+} concentration (or activity) can be calculated. A similar approach involves the addition of a trace quantity of the reagent morin to the sample, which forms a fluorescent complex with Al that can be quantified at very low concentrations. Again, by knowing the concentrations of both total morin and the Al–morin complex, the free Al^{3+} concentration can be 'backed out' computationally. There are several limitations to these methods. With the F-electrode method, there must be sufficient free F^- in the sample to obtain accurate readings. When the pH exceeds ~ 5 and the hydrolysis products are more prevalent, small errors in measured F^- can lead to relatively large errors in computed speciation. With the morin method, it is necessary that the quantity of morin added is small enough to cause only minimal shifts in the original speciation, requiring some foreknowledge of the prevailing Al chemistry; fluorescence of natural DOC in the sample can also cause interferences. Finally, the information gained from these two methods is limited

Table 2 Abbreviations used in the description of aluminum speciation methods

Abbreviation	*Definition*
CAS	Chrome azurol S
Ferron	8-Hydroxy-7-iodo-5 quinoline-sulfonic acid
FIA	Flow-injection analysis
FPLC	Fast protein liquid chromatography
HPLC	High-performance liquid chromatography
ICP-MS	Inductively coupled plasma mass spectrometry
ICP-AES	Inductively coupled plasma atomic emission spectrometry
ET-AAS	Electrothermal atomic absorption spectrometry (graphite furnace)
IEC	Ion-exchange chromatography
LC-ES-MS	Liquid chromatography electrospray mass spectrometry
morin	2,3,4,5,7-Pentahydroxy-flavone
NMR	Nuclear magnetic resonance
PCV	Pyrocatechol violet
RPC	Reversed-phase chromatography
SEC	Size-exclusion chromatography
UV-Vis	Ultraviolet-visible spectrometry

to an estimate of the concentrations of free Al^{3+} ion and the mononuclear hydrolysis products (which are linked by pH); little is learned about the relative importance of the various Al complexes that cause total Al and free Al^{3+} to differ widely in many samples. Neither method has enjoyed widespread utilization at this time.

Further Reading

Bertsch PM and Parker DR (1996) Aqueous polynuclear aluminium species. In: Sposito G (ed.) *The Environmental Chemistry of Aluminum*, 2nd edn, pp. 117–168. Boca Raton, FL: CRC Press.

Bi S-P, Yang Y-D, Zhang F-P, Wang X-L, and Zou G-W (2001) Analytical methodologies for aluminium speciation in environmental and biological samples – a review. *Fresenius Journal of Analytical Chemistry* 370: 984–996.

Bloom PR and Erich MS (1996) The quantitation of aqueous aluminium. In: Sposito G (ed.) *The Environmental Chemistry of Aluminum*, 2nd edn, pp. 1–38. Boca Raton, FL: CRC Press.

Campbell PGC (1995) Interactions between trace metals and aquatic organisms: a critique of the free-ion activity model. In: Tessier A and Turner DR (eds) *Metal Speciation and Bioavailability in Aquatic Systems*, pp. 45–102. New York: John Wiley.

Gensemer RW and Playle RC (1999) The bioavailability and toxicity of aluminum in aquatic environments. *Critical Reviews in Environmental Science and Technology* 29: 315–450.

Kinraide TB (1991) Identity of the rhizotoxic aluminum species. *Plant and Soil* 134: 167–178.

LaZerte BD, van Loon G, and Anderson B (1997) Aluminum in water. In: Yokel RA and Golub MS (eds) *Research Issues in Aluminum Toxicity*, pp. 17–45. Washington, DC: Taylor & Francis.

Ryan PR, Delhaize E, and Jones DL (2001) Function and mechanism of organic anion exudation from plant roots. *Annual Review of Plant Physiology and Plant Molecular Biology* 52: 527–560.

Stumm W and Morgan JJ (1996) *Aquatic Chemistry*, 3rd edn. New York: John Wiley.

Tipping E (1998) Humic ion binding model VI: an improved description of the interactions of protons and metal ions with humic substances. *Aquatic Geochemistry* 4: 3–48.

AMMONIA

D E Kissel and M L Cabrera, University of Georgia, Athens, GA, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

The primary forms of nitrogen (N) taken up from soil by plants are ammonium (NH_4^+) and nitrate (NO_3^-). Soils used for the production of food and fiber often store large quantities of N in soil organic matter; for example, a highly fertile prairie soil in the central USA may typically contain 10 000 kg N per hectare in the top meter of soil, with most in organic compounds in the soil's organic matter and not available to plants. A small proportion of this N, around 1%, undergoes biological reactions each year to form ammoniacal nitrogen, i.e., either ammonia (NH_3) or NH_4^+. A natural equilibrium exists between NH_3 and NH_4^+ that depends largely on pH and temperature. The physical properties of these two forms of ammoniacal N are totally different. Both forms are water-soluble, but whereas NH_4^+ can be easily metabolized by plants and other organisms, NH_3 is highly toxic to living organisms. Because the pH of most soils is in the range of 5–8, most ammoniacal N exists in the soil as NH_4^+. However, due to the volatility and toxicity of NH_3, its concentration in soil may become sufficiently high above pH 7 to be toxic to organisms or to be lost to the atmosphere.

Over the past 50 years, there has been an approximately linear increase in the world production of industrially fixed N fertilizers, from around 10 million metric tons in 1960 to just over 80 million metric tons by 2000. Over 75% of this N is in the form of ammoniacal N, or in forms of N such as urea (about 40%) that react in the soil to produce ammoniacal N. Because N fertilizers are often added to soils in concentrated zones by surface or band applications, subsequent chemical reactions are often more pronounced than fertilizers that are uniformly distributed throughout a larger soil volume. The resulting chemical and physical reactions are of considerable practical importance because N fertilizer is sometimes lost into the atmosphere. Nitrogen applications with less loss result in more efficient and economical use by crops, and they minimize off-site losses to the ecosystem.

Chemical Reactions of Ammoniacal N

Ammonia–Ammonium Equilibrium

Ammonium in the soil solution can be considered as a weak acid in equilibrium with its dissociation products NH_3 and H^+ as follows:

$$NH_4^+ = NH_3 + H^+ \qquad pKa = 9.25 \qquad [1]$$

The pKa value indicates the pH at which half of the ammoniacal N is NH_4^+ and half is NH_3. Together with pH, the pKa can be used to estimate the proportion of ammoniacal N as NH_3 using the equation:

$$pH = pKa - \log[NH_3/NH_4^+] \qquad [2]$$

The value of the pKa at 25°C is 9.25; however, this dissociation is temperature-sensitive. The effect of temperature on the proportion of NH_3 in the mixture is substantial. For example, at a pH of 8, the proportion of ammoniacal N as NH_3 is about 2% at 10°C, and 13% at 40°C (Figure 1). Another physical constant that affects the behavior of ammoniacal N is the Henry's constant that describes the equilibrium of partitioning of NH_3 between a solution and the gaseous phase in contact with the solution. It too is affected by temperature, with an approximately threefold increase in the proportion of NH_3 that partitions into the air when the temperature is raised from 10 to 40°C. Taken together, higher temperatures enhance ammonia loss from ammoniacal N due to its effect on both of these physical constants.

Retention of Ammonia by Soils

In some agricultural production areas, anhydrous NH_3 is applied directly to soils as a source of N fertilizer for crop production. It is typically released into the soil from application points attached to tools that are pulled through the soil, resulting in line (banded) applications. The applied NH_3 reacts first with soil nearest the line of release, and then diffuses outward, resulting in cylindrical NH_3 retention zones just below the soil surface. The properties of the NH_3 retention zones are shown in Figure 2. Because NH_3 is a base when dissolved in water, it raises soil pH following its application to soil. The measured value of pH will vary with distance from the injection point or line, and it will also depend on the amount of NH_3 added and the H^+-buffering properties of the soil. As noted in Figure 2, the soil pH at the center of the retention zone is typically about 9 by 1 day after application. The outward expansion of the retention zone is complete by 1 day following application. The spatial distribution of nitrate concentrations at 1, 2, and 4 weeks following application is also shown in Figure 2. It indicates that nitrification proceeds most rapidly at an intermediate radial distance from the retention zone center, due to more favorable conditions for nitrification. Nitrate concentrations are lowest in the retention zone center where high concentrations of NH_3 are toxic to nitrifying organisms, thereby slowing the rate of nitrification. With more time and continued nitrification, soil pH decreases, and nitrification rates increase in the retention zone center.

The concentration of NH_3 retained in the center of a retention zone by 1 day after application can be predicted from the soil's titratable acidity to pH 9. Titratable acidity to pH 9 is a measure of the quantity of H^+ in soil that will potentially react with the applied NH_3 at the retention zone center, as well as the soil's pH-dependent negative charge that will result from raising soil pH to 9 that will in turn adsorb NH_4^+. This reaction can be written as:

$$NH_3 + H - Soil = NH_4 - Soil \qquad [3]$$

'H – Soil' in this case represents titratable acidity that is reactive with a base but not cation-exchangeable with an unbuffered salt. Because titratable acidity to pH 9 is a good predictor of ammoniacal N concentration in the center of the retention zone, it follows that initial soil pH should also affect the ammoniacal N concentration in the center of the retention zone. As the pH of a soil becomes more acid, the reduction in the soil's effective cation exchange capacity is balanced by an equivalent increase in the soils unionized pH-dependent charge that is measured as titratable acidity. If NH_3 is applied to two soils equal in all respects except for their pH, the differences in the ammoniacal N concentrations in the center of their retention zones will be equal to the difference

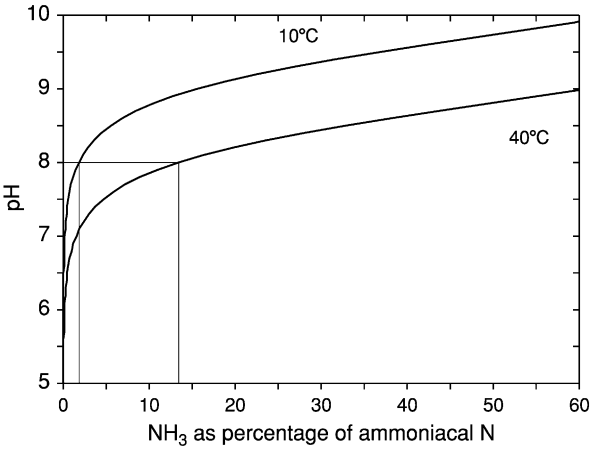


Figure 1 The percentage ammonia (NH_3) of a dilute ammoniacal N solution as affected by solution pH and temperature.

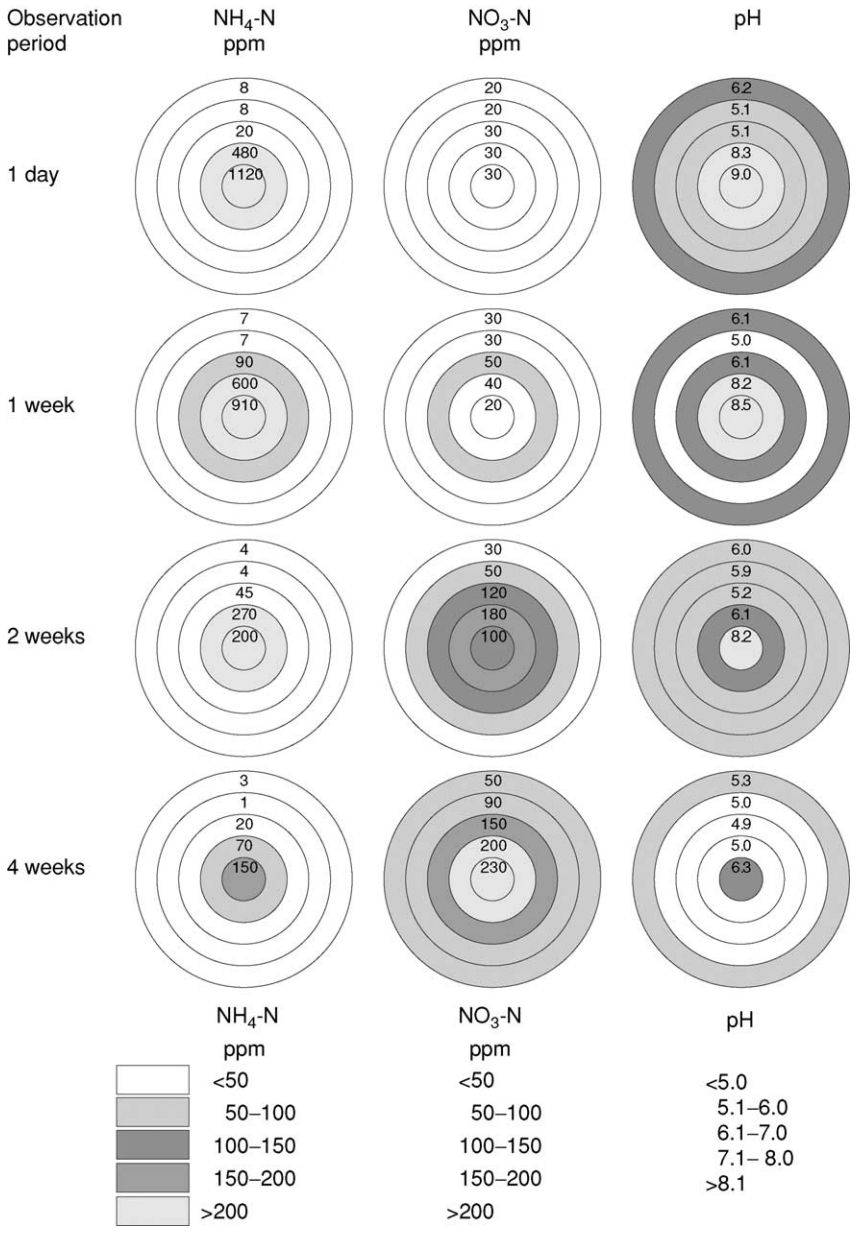


Figure 2 Distribution of NH_4^+, NO_3^-, and pH in soil with distance of 0–1.5, 1.5–3, 3–5, 5–7, and 7–9 cm from the injection line. Ammonia applied at 116 kg N ha^{-1} at 25 cm spacing. (Reproduced with permission from Nommik H and Vahtras K (1982) Retention and fixation of ammonium and ammonia in soils. *Agronomy* 22: 123–171.)

in titratable acidity to pH 9, being greater, of course, in the soil of lower pH. Likewise, as noted above, two soils with identical pH but with different unionized pH-dependent charge (titratable acidity) to pH 9.0 will also differ in NH_3 retention, with greater retention in the soil with greater titratable acidity.

Because the application of NH_3 varies greatly in practice by factors such as the spacing between adjacent retention zones, as well as the application rate per hectare, it follows that these too must be considered when estimating the properties of an NH_3 retention zone. The factors that will finally determine the properties of an NH_3 retention zone will be the soil's capacity of unionized pH-dependent charge per volume of soil (determined by its pH buffering capacity, its pH, and its bulk density), plus the retention zone spacing and the amount of NH_3 applied per hectare. A computer simulation model has been developed that describes the distribution of ammoniacal N and pH, as well as ammonia loss based on soil bulk density and its titratable acidity to pH 9, depth and spacing of application bands, and the amount of ammonia applied per hectare.

Retention and Loss of Surface-Applied Ammoniacal N in Soils

Nitrogen fertilizer is sometimes uniformly spread on the soil surface and not incorporated into the soil by tillage. This application method is often employed with established crops for which tillage will be detrimental, such as the topdressing of small grain, forages, or crops produced under no-tillage culture. In these cases, the chemical, physical, and biological reactions that control the formation and adsorption of NH_3 at the soil surface are of special interest. These reactions determine if NH_3 losses are significant, and the subsequent availability of the N fertilizer to crops. Although other biological reactions may be important in affecting the availability of the N fertilizer to crops, only those affecting the loss and retention of NH_3 will be considered here. Two different cases of N fertilizer reactions will be considered: (1) those involving urea fertilizers applied to all soils, and (2) those involving the reaction of NH_4^+ fertilizers with calcareous soils. Each will be considered separately.

Reactions of Urea in Soils

When added to soil, urea fertilizer granules will typically dissolve in the soil water after a short time (minutes to hours), depending on the soil water content and temperature. Even applications at the soil surface will dissolve, typically within a few days, from water near the soil surface and from dews. The dissolved urea then reacts (catalyzed by soil urease enzyme) to form NH_4^+ and HCO_3^-, as described by eqn [4]. Concurrently, bicarbonate produced in eqn [4] forms carbon dioxide in eqn [5]. Below pH 8.2, eqn [5] will continue to consume all of the bicarbonate, provided that CO_2 diffuses freely from the soil, which would be the case for surface applications of urea. In the case that eqns [4] and [5] go to completion, two H^+ are consumed and two NH_4^+ ions released for each urea hydrolyzed (or one H^+ consumed for each NH_4^+ released).

$$CO(NH_2)_2 + 2H_2O + H^+ = 2NH_4^+ + HCO_3^- \quad [4]$$

$$HCO_3^- + H^+ = CO_2 + H_2O \quad [5]$$

The ratio of one H^+ consumed for each NH_4^+ released results in an effective cation exchange site being formed and occupied by an NH_4^+ ion, as shown in eqn [6]. The H^+ on the right side of eqn [6] is consumed in eqns [4] or [5]. The term 'H^+ – Soil' in eqn [6] refers to the pH-dependent charge in the soil, which can be measured as titratable acidity.

$$H^+ - Soil + NH_4^+ = NH_4 - Soil + H^+ \quad [6]$$

Some of the soil cations are not adsorbed on the cation exchange sites. These cations (typically Ca^{2+}, Mg^{2+}, K^+, NH_4^+, and Na^+) exist in the soil solution and exchange freely with the same cation species on the exchange sites. Since the equilibria described by eqn [1] only apply to the ions in the soil solution, the cation exchange equilibria, shown in eqn [7] for the case of Ca^{2+}, affects the amount of ammoniacal N in the soil solution, and therefore the proportion that will exist in solution as NH_3.

$$1/2Ca^{2+} + NH_4 - Soil = NH_4^+ + Ca_{1/2} - Soil \quad [7]$$

The consumption of H^+ causes soil pH near the dissolved urea to rise. The amount of pH rise depends on the soil's H^+ buffering capacity, which will be discussed below. If the soil pH rises above 7, a significant amount of NH_3 can form, which depends primarily on the soil pH, temperature, and the concentration of NH_4^+ in the soil solution, as described by the equilibrium in eqns [7] and [1]. Based on equilibrium in eqn [1], any movement of NH_3 away from the reaction site will release H^+, thereby lowering pH.

Eqns [1], [4], [5], [6], and [7] are the predominant ones below a soil pH of 8.2. The process of nitrification, in which NH_4^+ is oxidized to NO_3^-, will add two H^+ to the soil for each NH_4^+ oxidized. This process will not be discussed any further here. Its effect on NH_3 volatilization is not very pronounced in the first week following application, since there is often a lag time of a few days in the buildup of a nitrifier population following N application.

Above pH 8.2, eqns [8] and [9] become important, and CO_2 formation from reaction [5] stops, as shown graphically in **Figure 3**. The predominant species is HCO_3^- (99.2% HCO_3 and 0.8% CO_3^{2-}) around pH 8.2. Any HCO_3^- from eqn [4] remains in the soil solution or is changed to CO_3^{2-} as pH is raised further, as shown by eqn [8]. If there are sufficient Ca^{2+} ions in the soil solution to exceed the solubility product of $CaCO_3$, then solid $CaCO_3$ is formed in the soil, as shown in eqn [9].

$$HCO_3^- = CO_3^{2-} + H^+ \quad [8]$$

$$Ca^{2+} + CO_3^{2-} = CaCO_{3(solid)} \quad [9]$$

The formation of solid $CaCO_3$ removes CO_3^{2-} from the soil solution, causing a new equilibrium to be established in eqn [8], releasing H^+ that slows the rise in pH. This effect is shown in **Figure 4**. In this study, soils were titrated with either NH_4OH or urea + urease enzyme to raise soil pH. Up to pH 8.2, soil pH was raised identically by NH_4OH and urea that had completely hydrolyzed. Above pH 8.2, urea was

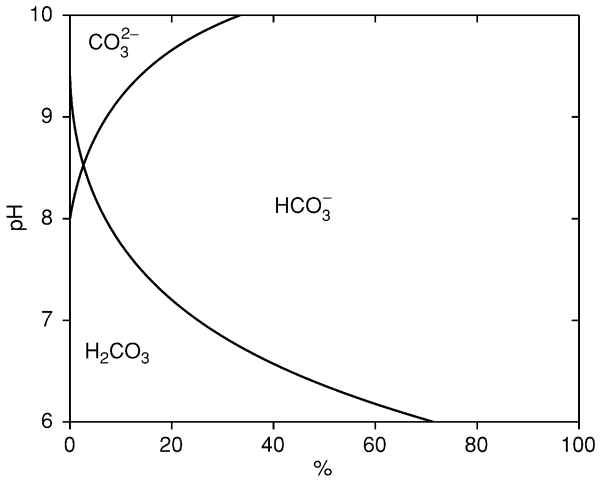


Figure 3 The distribution of inorganic C species in a dilute solution as affected by pH. (Reproduced with permission from Koelliker JK and Kissel DE (1988) Chemical equlibria affecting ammonia volatilization. In: Bock BR and Kissel DE (eds) *Ammonia Volatilization from Urea Fertilizers*, pp. 37–52. Bulletin Y-206. National Fertilizer Development Center. Muscle Shoals, AL: Tennessee Valley Authority.)

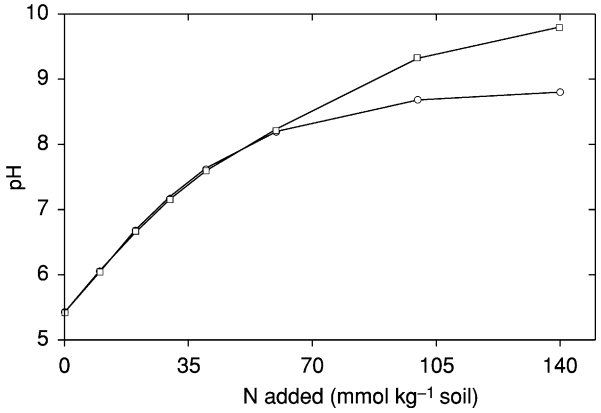


Figure 4 The change in soil pH of Kahola silt loam as affected by the addition of NH_4OH (squares) or urea (circles) allowed to hydrolyze completely. (Reproduced with permission from Kissel DE, Cabrera ML, and Ferguson RB (1988) Reactions of ammonia and urea hydrolysis reaction products with soil. *Soil Science Society of America Journal* 52: 1793–1796.)

less effective than NH_4OH in raising soil pH. The lower effectiveness of urea was due in part to the urea-C not forming CO_2 but remaining as HCO_3^- and CO_3^{2-} in the soil solution or by precipitating as $CaCO_3$.

Environmental Conditions and Ammonia Formation

When urea is surface-applied, the formation of NH_3 at or near the soil surface may allow some NH_3 volatilization, and when urea is banded near germinating seeds or seedlings, some NH_3 toxicity may result, causing plant damage or loss. The severity of both depends primarily on the concentration of NH_3 formed. The concentration of NH_3 depends on the amount and method of urea application, soil properties, and soil environmental conditions for several days after application. Specifically, the most important factors affecting NH_3 concentrations are:

1. The concentration of urea in soil immediately following application directly affects the rise in soil pH. If urea is more concentrated, eqns [4–6] take place in a smaller mass of soil, consuming H^+ from a smaller volume, which will raise soil pH to higher levels. More concentrated urea applications will also result in more concentrated levels of ammoniacal N.
2. The soil pH for a 3- to 5-day period following urea application affects NH_3 concentrations because pH determines the proportion of the two forms of ammoniacal N, as described in eqn [1].
3. The rate of urea hydrolysis affects NH_3 concentration because the reactants and reaction products all tend to become less concentrated with time due to diffusion from the site of the reactions. Therefore, faster urea hydrolysis rates will increase the concentration of NH_3 at the site of the reactions.

Each of these factors will be discussed further in the following sections.

Urea concentration The concentration of urea in soil depends largely on two factors: (1) the amount applied per hectare; and (2) the method of application. Methods such as band application or surface application cause higher urea concentrations than urea that is incorporated and mixed with soil by tillage. Rainfall or irrigation immediately after surface application can also move urea into soil and cause it to disperse and to be less concentrated. Molecular diffusion of urea in soil water also disperses urea and makes it more dilute, allowing it to interact with more urease. Hydrogen ions from the soil are consumed when urea is hydrolyzed, so it follows that the H^+ will be consumed from less soil where urea is more concentrated, provided there is enough urease enzyme and adequate temperature and water to sustain a fast reaction. Rapid hydrolysis and consumption of H^+ from a small volume of soil will cause the soil pH at the site of application to increase further and cause more NH_3 to form. When urea hydrolysis is complete, the rise in pH will be directly proportional to the initial urea concentration, and it will be highly soil-dependent, as shown in **Table 1** for two soils from Kansas. In this comparison, the pH was measured after all urea had hydrolyzed due to the addition of urease enzyme. For comparison, the soil

Table 1 Soil pH of two Kansas soils 13 h after adding urea and urease

Urea-N concentration (mg N kg^{-1})	*Haynie (soil pH)*	*Kahola (soil pH)*
0	5.74	5.42
140	7.17	6.05
280	8.13	6.67

Source: Kissel DE, Cabrera ML, and Ferguson RB (1988) Reactions of ammonia and urea hydrolysis reaction products with soil. *Soil Science Society of America Journal* 52: 1793–1796.

pH after the hydrolysis of 280 mg urea-N kg^{-1} was 6.67 for the silt loam soil (Kahola) and 8.13 for the sandy loam (Haynie). The differences in the rise in pH were due to differences in soil H^+ buffering capacities of the two soils, which can be measured from a pH titration of the soils, to be discussed further in the next section.

Soil pH and H^+ buffering capacity The pH rise that occurs when a given concentration of urea hydrolyzes depends largely on the H^+ buffering capacity of the soil, which is a measure of the quantity of H^+ that will be adsorbed or released by the soil per unit pH change. The H^+ buffering capacity arises primarily from soil organic matter and clays, and is sometimes referred to as pH-dependent charge in soils. The type of clay mineral greatly affects buffering. Kaolinite clays have very little H^+ buffering capacity, whereas 2:1 layer clays, especially hydroxy-interlayered vermiculites, have significant amounts. Sandy soils are often low in both clay and organic matter, and therefore have less H^+ buffering capacity. The effect of soil buffering on soil pH change for a uniformly mixed sample is shown in Figure 5. The effect of H^+ buffering capacity on loss from surface-applied urea will depend in part on how deep and how quickly the applied urea and urea reaction products diffuse into the soil. Urea and its reaction products move by molecular diffusion to depths of approximately 25 mm within 24 h after the urea dissolves at the soil surface. With movement this deep into the soil, it is apparent that differences in H^+ buffering capacity will affect the rate and amount of loss, as shown in Figure 5. In this study, soil mixes were prepared that differed in their H^+ buffering capacities, but had the same initial pH and cation exchange capacities. Urea fertilizer was then applied to the surface of each soil mix and NH_3 loss and surface pH were measured for 19 days. As would be expected, the NH_3 loss was greatest from the soil mix that had the least H^+ buffering capacity (soil mix 1). Because this soil mix had the fewest H^+ ions, it allowed the soil surface pH to be higher than the other soil mixes 3–4 days after urea was surface-applied. The higher pH allowed a greater proportion of the ammoniacal N to be in the gaseous NH_3 form, which resulted in more loss. In contrast, the soil mix with the most H^+ buffering capacity (soil mix 3) had the lowest surface soil pH, and the least NH_3 loss. In summary, it can be generalized that soils with more H^+ buffering capacity will have a lower pH increase following surface application of urea and will therefore lose less NH_3.

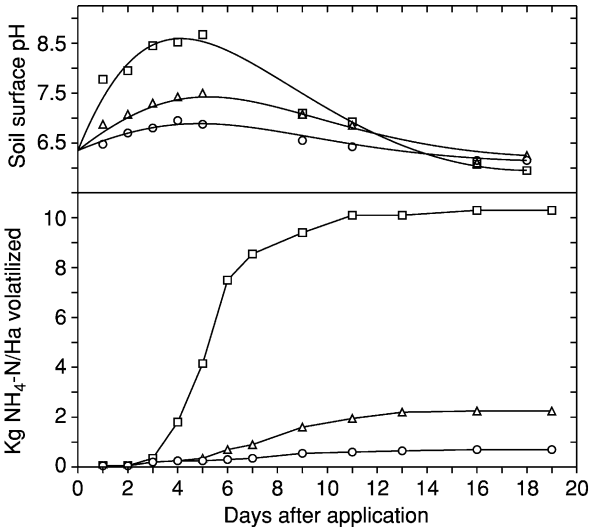


Figure 5 Soil surface pH and total NH_3 loss with time after surface applications of urea to soil mixes with different H^+ buffering capacity. Squares, soil mix 1; triangles, soil mix 2; circles, soil mix 3. (Reproduced with permission from Ferguson RB, Kissel DE, Koelliker JK, and Basel W (1984) Ammonia volatilization from surface-applied urea: effect of hydrogen ion buffering capacity. *Soil Science Society of America Journal* 48: 578–582.)

Factors affecting the rate of urea hydrolysis The factors that are most important in affecting the rate of urea hydrolysis are the concentration of urease enzyme in soil, urea concentration, soil temperature, soil water content, and soil pH. Each will be discussed separately.

Soil urease concentration The number of active urease molecules in soils cannot be measured directly. However, urease activity (the rate of urea hydrolysis) can be measured under standard conditions of temperature, water content, pH, and urea concentration. The measured value of a soil's urease activity under the standard conditions (including a defined urea concentration) then represents the number of urease molecules. Several investigators have attempted to find the soil property that correlates best with urease activity in cultivated soils. There was general agreement among these studies that soil organic C was a relatively good predictor of soil urease activity. In the

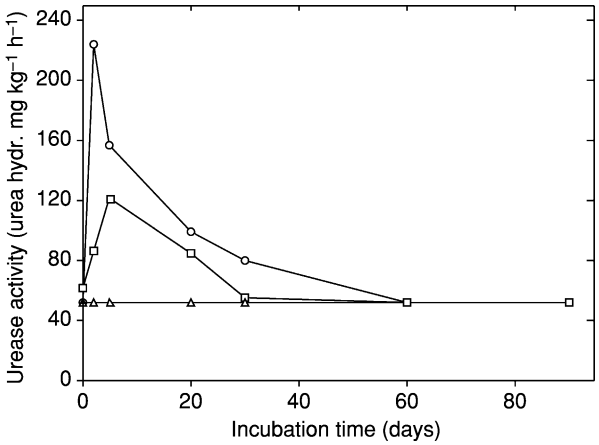


Figure 6 The effect of incubation time on the urease activity of unamended soil (triangles) and soil amended with 5000 mg of corn residue (squares) or glucose (circles) per kg of soil. (Reproduced with permission from Zantua MI and Bremner JM (1976) Production and persistence of urease activity in soils. *Soil Biology and Biochemistry* 8: 369–374.)

absence of a measurement of a soil's urease activity, it can be estimated from the organic C content of cultivated soils. However, this estimation is complicated in soils that contain a significant amount of decomposing residues, as shown in Figure 6. In this study, soil samples were treated with glucose or crop residues that enhanced microbial activity, which increased the urease activity of the soil. The persistence of the urease produced depended on the soil type and the amendments added, but eventually the level of urease activity returned to that found in unamended soil. Apparently, this is due to a fixed number of protective sites for urease molecules, which varies with soil type.

In the practice of surface application of urea, the urea hydrolysis rates observed during the first 5–20 h after surface application of urea are sometimes slower than subsequent hydrolysis rates. These observations are not due to an increase in urease activity. Studies have shown that urea uniformly mixed with soil results in linear increases in NH_4^+ over the first 6 or 7 h after application, indicating no change in urease activity. A more appropriate explanation for the lag in hydrolysis rates from surface application of urea is that time is required for urea to diffuse down into the soil and to interact with more urease molecules as time progresses. This increased contact of urea with urease as time progresses explains the lag in hydrolysis rates over the first couple of days. Of course, NH_3 loss lags even more because of the time required to increase soil pH and levels of ammoniacal N.

Urea concentration When urea fertilizer granules dissolve in soil, the concentration of urea in the soil solution varies from a nearly saturated solution near the granule to very low concentrations at the outer reaches of the area affected. In laboratory studies of urea mixed uniformly with soil, the effects of urea concentration across the full range of concentrations have been shown to be well described by the sum of two Michaelis–Menten reactions, but with some slowing of hydrolysis rates at concentrations near saturation, apparently due to inhibition of the enzyme.

Soil temperature The effect of soil temperature on urea hydrolysis has been shown to follow a Q_{10} of 2, i.e., a doubling of reaction rate with each 10°C rise in temperature. For example, a rise in soil temperature from 5 to 25°C will cause the rate of urea hydrolysis to be approximately four times faster. Urea hydrolysis is a relatively fast process, even at cool temperatures that would typically occur when urea fertilizers are applied to cool-season crops. For example, for a typical silt loam surface soil with about 2% organic matter, a water potential of −0.5 MPa or wetter, and a neutral pH, one would expect urea hydrolysis at 27°C to be 90% complete by 3 days after application, whereas at 2°C, urea hydrolysis would still be 90% complete by 9 days after application.

Soil water content Urea hydrolysis proceeds at optimum rates when soil water contents are in the range that is readily available to plants. In the water content range from wilting point (−1.5 MPa) to air-dry, the rate slows greatly, and essentially stops as the soil approaches air dryness.

Soil pH The effect of soil pH on urease activity has been described in the scientific literature primarily as a single effect, regardless of the urea concentration. The optimum pH has been reported to vary from around 7 to as high as 8.8–9.0. More recent research indicates that the optimum pH is different for the two reactions described briefly in the section on urea concentration.

Reactions of NH_4^+-N Applied to Calcareous Soils

Based on experimental evidence with surface-applied urea fertilizers, NH_3 loss is not very significant from most soils unless the pH at the soil surface reaches a value of 6.5 or higher (Figure 5). Because of this, the loss of NH_3 from most NH_4^+-N fertilizers applied to the surface of neutral to acid pH soils is very low or insignificant. This may also be partly true because the most common NH_4^+-N sources such as NH_4NO_3, $(NH_4)_2SO_4$, and NH_4Cl are also slightly acid, which may decrease the pH of the soil surface even further. An exception is $(NH_4)_2HPO_4$, which has a pH of about 8.5.

When NH_4^+-N sources are applied to calcareous soils, NH_3 loss can occur, not only because the pH of calcareous soils is above 7 (typically in the range of 7.5–8.2), but also because some of the NH_4^+ sources will react with $CaCO_3$ to form reaction products that increase soil pH to even higher values, as shown in Figure 7. Ammonium sources that form relatively insoluble calcium reaction products with the anion of the NH_4^+ fertilizer will have higher pH for the first 5–20 h after application than those sources that do not. For the sources shown in Figure 7, the solubilities were 0.0016, 0.20, and 134 g 100 ml^{-1} H_2O for CaF_2, $CaSO_4$, and $Ca(NO_3)_2$, respectively. Losses of NH_3 by 24 h after surface application were 60, 40, and 8% respectively for NH_4F, $(NH_4)_2SO_4$, and NH_4NO_3 respectively. These losses are consistent with the relative soil pHs for the three sources for the first day following application. Although pH following NH_4NO_3 application was by far the lowest, around pH 7, losses were still significant, due to the high concentration of NH_4^+ in the soil solution near the soil surface. The reactions responsible for the high losses from these sources are the following:

$$CaCO_{3(solid)} = Ca^{2+} + CO_3^{2-} \qquad [10]$$

Carbonate ions dissolved from the solid phase at pH of around 8 will immediately revert to HCO_3^-, as shown by eqn [6] and Figure 3. As Ca^{2+} is removed from solution by precipitation with the anion of the NH_4^+ N source (for those sources with a relatively insoluble reaction product), more $CaCO_3$ will be dissolved and the CO_3^{2-} will react via eqn [8] to remove H^+ from the soil, which will tend to raise pH, as will loss of CO_2 because eqn [5] will reestablish equilibrium. Based on these reactions, $CaCO_3$ will be consumed in an amount chemically equivalent to the NH_3 lost from the soil. This consumption of $CaCO_3$ can also be understood from an H^+ balance, since the fertilizers are applied to the calcareous soils as NH_4^+, and volatilize as NH_3 with a difference of one H^+ added to the soil for each NH_3 molecule volatilized.

Since $CaCO_3$ will be consumed from the calcareous soil due to these NH_3 loss reactions, and since the reactions occur within a time period of hours, the $CaCO_3$ must be of sufficiently small particle size to dissolve quickly to sustain the loss reactions. A concentration of about 10% by weight of clay size $CaCO_3$ sustains the reaction at rates determined by other factors, such as temperature and rate of N application.

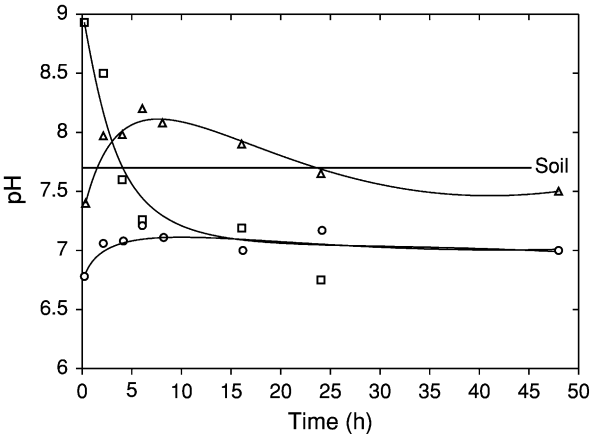


Figure 7 Change in soil pH with time of the surface 6.4 mm of Houston Black clay soil at 33°C for three ammonium compounds surface applied at 550 kg N ha^{-1}. Squares, NH_4F; triangles, $(NH_4)_2SO_4$; circles, NH_4NO_3. (Reproduced with permission from Fenn LB and Kissel DE (1973) Ammonia volatilization from surface applications of ammonium compounds on calcareous soils. I. General theory. *Soil Science Society of America Proceedings* 37: 855–859.)

List of Technical Nomenclature

(MPa)	Water potential
(mg kg^{-1})	Concentration
(multiplicative reaction rate (10°C)$^{-1}$: Q_{10})	Temperature quotient

See also: **Nitrogen in Soils:** Cycle

Further Reading

Bates RG and Pinching GD (1949) Acidic dissociation constant of ammonium ion at 0° to 50°C, and the base strength of ammonia. *Journal of Research of the National Bureau of Standards* 42: 419–430.

Fenn LB and Kissel DE (1973) Ammonia volatilization from surface applications of ammonium compounds on calcareous soils. I. General theory. *Soil Science Society of America Proceedings* 37: 855–859.

Ferguson RB, Kissel DE, Koelliker JK, and Basel W (1984) Ammonia volatilization from surface-applied urea: effect of hydrogen ion buffering capacity. *Soil Science Society of America Journal* 48: 578–582.

Izaurralde RC, Kissel DE, and Cabrera ML (1987) Titratable acidity to estimate ammonia retention. *Soil Science Society of America Journal* 51: 1050–1054.

Izaurralde RC, Kissel DE, and Cabrera ML (1990) Simulation model of banded ammonia in soils. *Soil Science Society of America Journal* 54: 917–922.

Kissel DE and Cabrera ML (1988) Factors affecting urea hydrolysis. In: Bock BR and Kissel DE (eds) *Ammonia Volatilization from Urea Fertilizers*, pp. 53–66. Bulletin Y-206, National Fertilizer Development Center. Muscle Shoals, AL: Tennessee Valley Authority.

Kissel DE, Cabrera ML, and Ferguson RB (1988) Reactions of ammonia and urea hydrolysis reaction products

with soil. *Soil Science Society of America Journal* 52: 1793–1796.

Koelliker JK and Kissel DE (1988) Chemical equilibria affecting ammonia volatilization. In: Bock BR and Kissel DE (eds) *Ammonia Volatilization from Urea Fertilizers*, pp. 37–52. Bulletin Y-206. National Fertilizer Development Center. Muscle Shoals, AL: Tennessee Valley Authority.

Moyo CC, Kissel DE, and Cabrera ML (1989) Temperature effects on soil urease activity. *Soil Biology and Biochemistry* 21: 935–938.

Nommik H and Vahtras K (1982) *Retention and Fixation of Ammonium and Ammonia in Soils*. Agronomy. 22, pp. 123–171. Madison, WI: American Society of Agronomy.

Rachhpal-Singh and Nye PH (1984) The effect of soil pH and high urea concentrations on urease activity in soil. *Journal of Soil Science* 35: 519–527.

Zantua MI and Bremner JM (1976) Production and persistence of urease activity in soils. *Soil Biology and Biochemistry* 8: 369–374.

AMORPHOUS MATERIALS

J Harsh, Washington State University, Pullman, WA, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Poorly crystalline aluminosilicates ('amorphous materials') in soils consist primarily of the mineral imogolite and allophane, a term that refers to all remaining short range-ordered (SRO) aluminosilicate clays. Although long thought to derive only from volcanic ash parent material, these materials are ubiquitous in soils under a variety of vegetation, parent materials, and climate. They are characterized not only by their poor crystallinity, but by their small particle size, imparting high specific surface areas, and by their variable and permanent surface charge properties. These characteristics make them important ion exchange materials even when present at relatively low concentration. SRO materials also play roles in Al solubility, organic matter stabilization, and soil shrink–swell properties. As a result, they affect a number of environmental processes, including solute transport, mineral weathering, formation of humic materials, and soil stability.

Occurrence in Soils

Allophane and imogolite result from the rapid precipitation of soluble Al and Si that are released either by labile parent materials, such as volcanic ash, or by intense weathering of any parent material ranging from sandstone to granite. They are metastable products that form in favor of more stable crystalline minerals such as kaolinite, because SRO materials have lower surface tension and, thus, nucleate more rapidly in aqueous solutions. Eventually they are replaced by the more crystalline aluminosilicate clay minerals and, as Si is weathered and leached from the soil, by Fe and Al (hydr)oxides. Allophane and imogolite may begin to form immediately in an ash deposit but increase in concentration over tens of thousands of years before declining in favor of the more crystalline minerals. Therefore, their metastable nature does not diminish their importance in soils.

Poorly crystalline aluminosilicates can form in any environment where weathering leads to sufficient Al and Si in solution. The association of allophane and imogolite with soils of volcanic origin arises from the fact that ash, tephra, and other pyroclastic materials contain amorphous volcanic ash that can rapidly release Al and Si. Allophane formation is favored by Si concentrations between 0.1 and 4 mmol l^{-1}. The type of allophane that forms will depend on the Si-to-Al ratio of the solution, the pH, and soluble organic matter. Soils derived from volcanic debris and dominated by SRO materials (or by Al–humus complexes) often fall into the Andisol soil order under the US classification system. It would be a mistake to assume that allophane and imogolite are limited to soils derived from volcanic materials and formed in humid environments. They have been found in soils that have formed from gneiss, sandstone, igneous and sedimentary rock, and loess. Allophane has been found in soils of six of the 12 orders: Entisols, Inceptisols, Spodosols, Alfisols, Aridisols, and Ultisols, as well as in humic, xeric, and arid moisture regimes, including the deserts of Iceland. The nearly ubiquitous nature of SRO materials underscores the fact that it is the presence of sufficient Al and Si in solution that determines their formation, not any particular environment or parent material.

The environment and parent material largely determine the nature of the materials formed. The Al/Si ratio in solution largely governs the type of allophane that results. Allophanes range from Al-rich, which have an Al/Si molar ratio of approximately 2, to

Si-rich, which have Al/Si of approximately 1. At high soluble Si concentration in the soil solution, there is a tendency for halloysite, a 1:1 phyllosilicate clay mineral, to form. Halloysite is favored at soluble Si concentrations greater than $10^{-3.45}$ mol l^{-1}, whereas Al-rich allophane and imogolite are favored at lower concentrations. When the parent material contains volcanic glass, an equilibrium concentration of $10^{-2.7}$ mol l^{-1} is expected, and the formation of these minerals will depend upon the rate at which Si is leached from the system. Illustrative of this climatic influence is a climatosequence of a 170-ky soil in Hawaii where 'noncrystalline' materials (primarily allophane and imogolite) increase and halloysite decreases with increasing rainfall (Figure 1). Increasing rainfall lowers readily soluble Si, favoring allophane and imogolite.

Identification

The poorly crystalline nature of allophane and imogolite defies their easy identification in soils by X-ray diffraction, the method of choice for the phyllosilicate clay minerals and many oxides. Instead, a combination of methods is used to identify and characterize these materials, including electron microscopy, selective chemical dissolution, Fourier transform infrared (FTIR) spectroscopy, and differential thermal analysis. Synthetic allophane and imogolite or purified samples from soils can be further characterized by nuclear magnetic resonance (NMR) and X-ray absorption spectroscopy (XAS). Electron microscopy is useful for defining the morphology of individual particles and aggregates of SRO materials, whereas the other methods give information on the structural and chemical bonding properties.

Allophane and imogolite are not truly noncrystalline or amorphous materials and do exhibit X-ray diffraction patterns, albeit with diffuse and often weak diffraction peaks. These patterns can be used to identify imogolite and distinguish it from Al-rich allophane (protoimogolite) (Figure 2), something not readily accomplished with FTIR. All allophanes, regardless of Al/Si ratio, display the two broad peaks

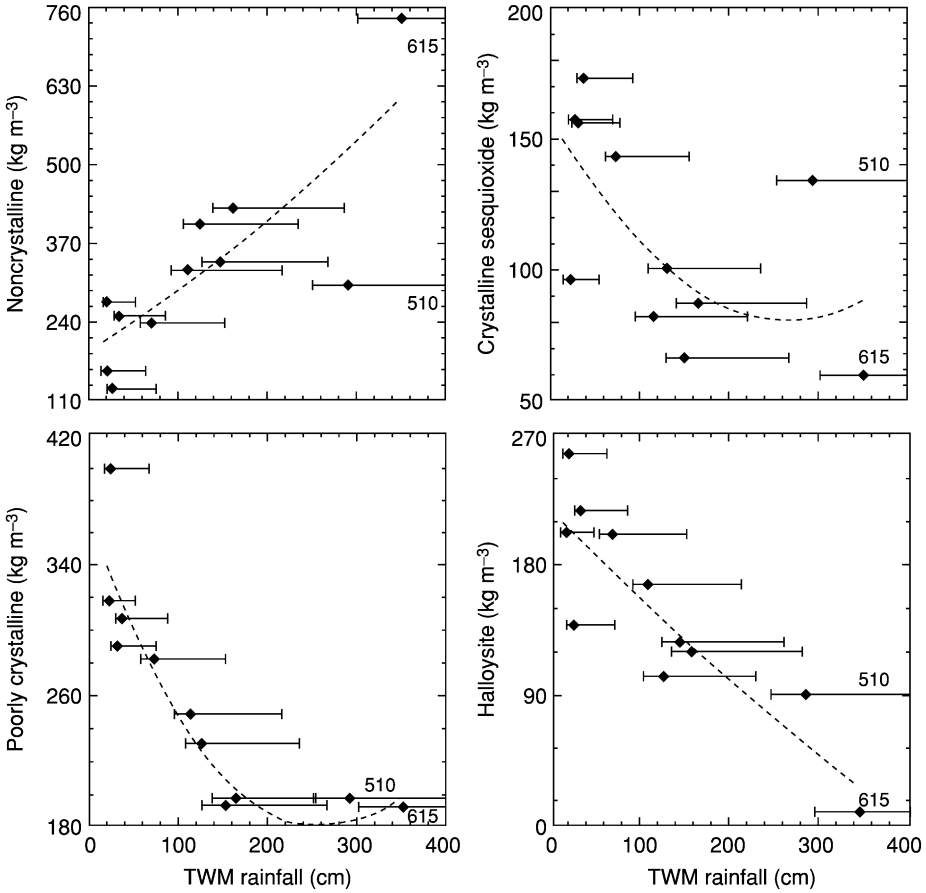


Figure 1 Soil components extracted from a climatosequence of soils in Hawaii. The 'noncrystalline' components are acid oxalate-extracted aluminosilicates – primarily allophane and imogolite. TWM, total weighted mean. (Source: Chadwick OA, Olson CG, Hendricks DM, Kelly EF, and Gavenda RT (1994) Quantifying climatic effects on mineral weathering and neoformation in Hawaii. *Transactions of the 15th World Congress of Soil Science,* International Soil Science Society and the Mexican Society of Soil Science.)

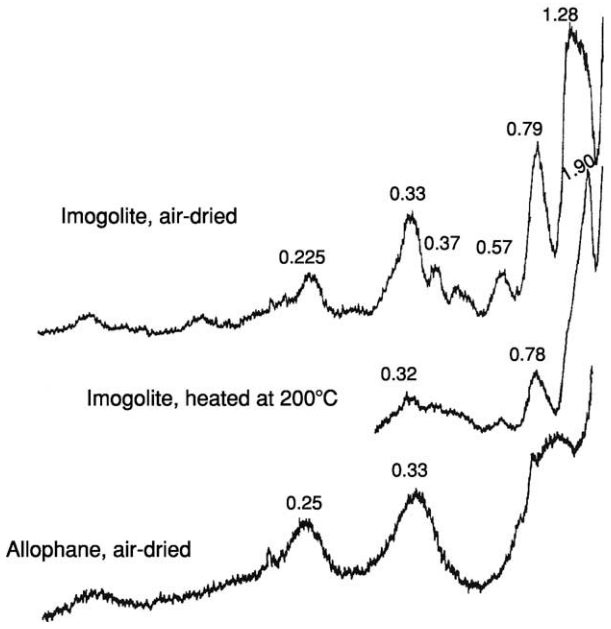


Figure 2 X-ray diffraction patterns of naturally occurring imogolite and allophane. (Reproduced from Wilson MJ (1987) *A Handbook of Determinative Methods in Clay Mineralogy*. Glasgow, UK: Blackie & Son.)

centered around 0.33 and 0.25 nm. The imogolite peaks result both from the long-range order along the longitudinal direction of the imogolite fiber and from the close packing structure of fiber bundles.

The clearest way to distinguish between allophane and imogolite is with electron microscopy. Transmission electron micrographs reveal the fibrous bundles of imogolite (Figure 3) that are readily identifiable even in complex soil mixtures. The imogolite fibers have cylindrical morphology, consisting of a gibbsite-like sheet of aluminum hydroxide octahedra pulled into a cylinder by individual silica tetrahedra. A two-dimensional (2D) ball-and-stick model in Figure 4 represents a cross-section of the imogolite tube. Allophane appears amorphous under the electron microscope, appearing gel-like (Figure 5) or as aggregates of small spherules. The structure, composition, and morphology of allophane depend largely on its Al/Si ratio.

Estimates of the quantity of SRO materials can be made by selective dissolution of these materials from soil. Extraction with acid ammonium oxalate removes Al and Si from poorly crystalline aluminosilicates and iron oxides (i.e., ferrihydrite) and Al from complexes with organic matter. Iron oxide- and organic-associated Al and Si can be removed with dithionite-citrate bicarbonate (DCB) and sodium pyrophosphate, respectively, before extraction with oxalate to quantify the amount coming from sources other than allophane and imogolite.

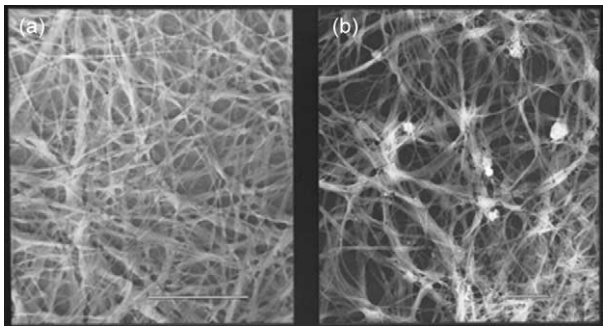


Figure 3 Transmission electron micrographs and electron diffraction patterns of (a) synthetic imogolite and (b) natural imogolite.

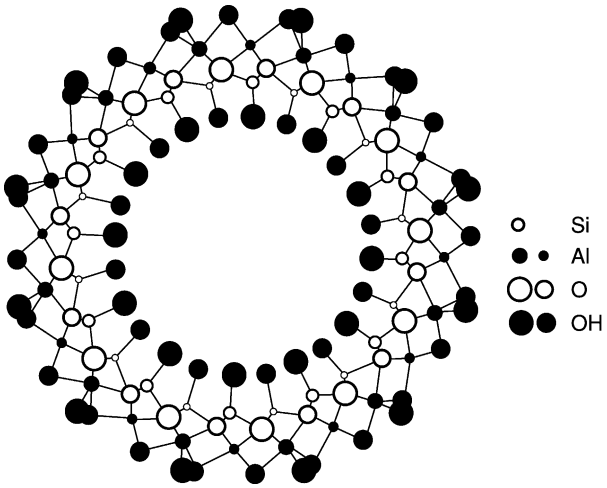


Figure 4 Crystallographic representation of the cross-section of an imogolite tube. (Reproduced from Cradwick PDG, (Farmer VC, Russell JD, Masson CR, Wada K, and Yoshinaga N (1972) Imogolite, a hydrated aluminum silicate of tubular structure. *Nature Physical Science* 240:187–189.)

Neither DCB nor pyrophosphate dissolves a significant amount of aluminosilicate material. Once the amount of Al and Si coming solely from SRO materials has been calculated, the percentage of allophane and imogolite in the soil can be estimated from:

$$\%\ \text{allophane in soil} = \text{Si}/(-0.067\,\text{Al/Si} + 0.27) \quad [1]$$

where Si is the percentage Si from SRO aluminosilicates. This formula is based on an average Al/Si ratio for allophanes found in soils.

Imogolite can be distinguished from allophane and its concentration in soil estimated by both differential thermal analysis (DTA) and FTIR. Imogolite gives rise to a unique endothermic peak at 400°C due to structural water loss (Figure 6). The area of this peak is proportional to the quantity of imogolite in the soil. An FTIR band at 348 cm^{-1} is indicative of

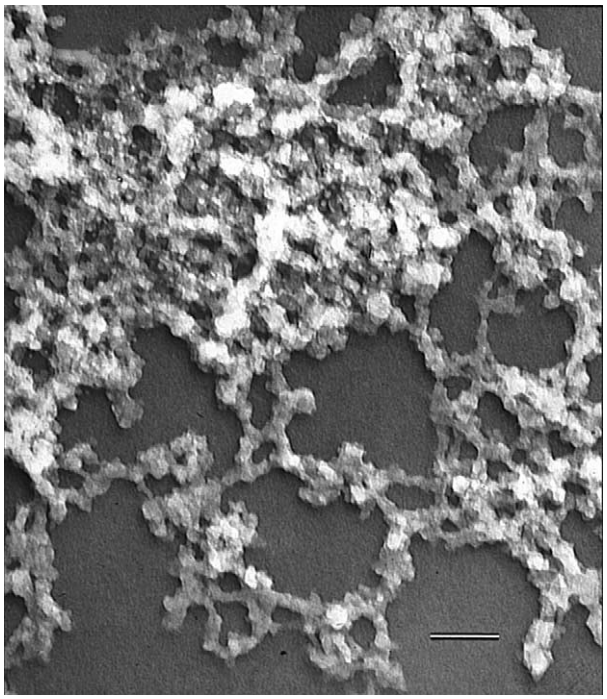


Figure 5 Transmission electron micrographs and diffraction pattern of synthetic allophane with a 1.12 Al/Si ratio.

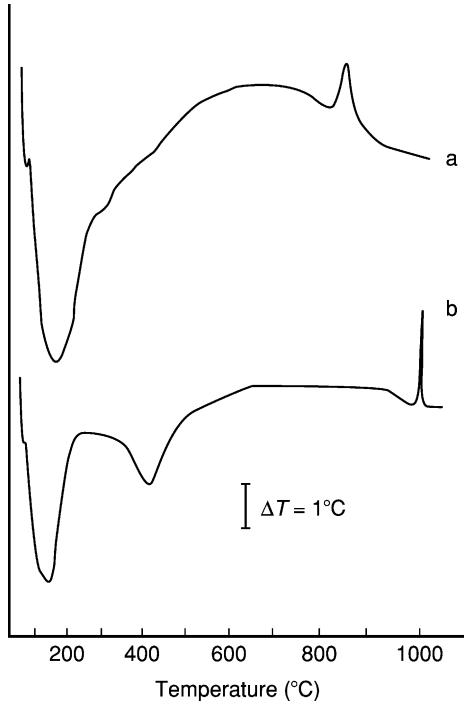


Figure 6 Differential thermal analysis curves of (a) Al-rich allophane and (b) imogolite.

the structure of both imogolite and protoimogolite allophane. This peak can be normalized to provide the quantity of 'imogolite-like structures' in a given soil.

Structure and Charge

Infrared and NMR spectra have been instrumental in helping to infer the structure of both imogolite and allophane. The IR band at 940 cm^{-1} is assigned to the unshared OH groups of the orthosilicate groups in imogolite and protoimogolite allophane. For all other allophanes, the Si-O stretching band shifts from 975 to 1020 cm^{-1} as the Al/Si ratio decreases (Figure 7). This indicates that polymerization of the silicate groups increases with increasing Si concentration. The −79 ppm NMR shift is also indicative of Si coordinated to three Al-O groups and one OH (Figure 8). The negative shift in the NMR spectrum as Al/Si decreases supports increased O sharing among Si atoms.

Polymerization is accompanied by an increase in tetrahedrally coordinated Al, as indicated by ^{27}Al-NMR and XAS spectra, inferring that the Si-rich allophanes have a feldspathoid-like structure. The feldspathoids are network silicates that, like zeolites, are characterized by high structural negative charge. Such charge also increases in allophanes as Al/Si decreases.

In Al-rich allophane and imogolite, the surface charge arises almost exclusively from variably charged aluminol (Al-OH) and silanol (Si-OH) groups. Chemically, the surface charging reactions can be represented as protonation–deprotonation reactions on these groups:

$$\text{Al-O}^- + \text{H}^+ \leftrightarrow \text{Al-OH} + \text{H}^+ \leftrightarrow \text{Al-OH}_2^+ \qquad [2]$$

$$\text{Si-O}^- + \text{H}^+ \leftrightarrow \text{Si-OH} + \text{H}^+ \leftrightarrow \text{Si-OH}^+ \qquad [3]$$

Aluminol groups in gibbsite have no net charge around pH 9, whereas silanol groups in quartz have a net neutral charge near pH 2. As a result, it is assumed that the acidic silanol groups contribute negative charge to the SRO aluminosilicates, and aluminol groups primarily bear positive charge, because, typically, the pH range of soils falls between 3 and 9.

Figure 9 shows the variable-charge behavior of an Al-rich allophane with the same Al/Si ratio as imogolite (i.e., protoimogolite). The data points represent the amount of Na and Cl adsorption determined as a function of pH and ionic strength. The magnitude of the adsorption increases with increasing ionic strength, whereas the sign of the charge depends on the pH. The pH at which Na and Cl adsorption is equal (i.e., no net surface charge) is known as the point of zero net charge (PZNC). The dotted line in Figure 9 represents the surface charge determined by adsorption of H^+ and OH^- (σ_H). The pH at which σ_H is zero is known as the point of zero proton charge (PZNPC). The PZNC and PZNPC are equivalent

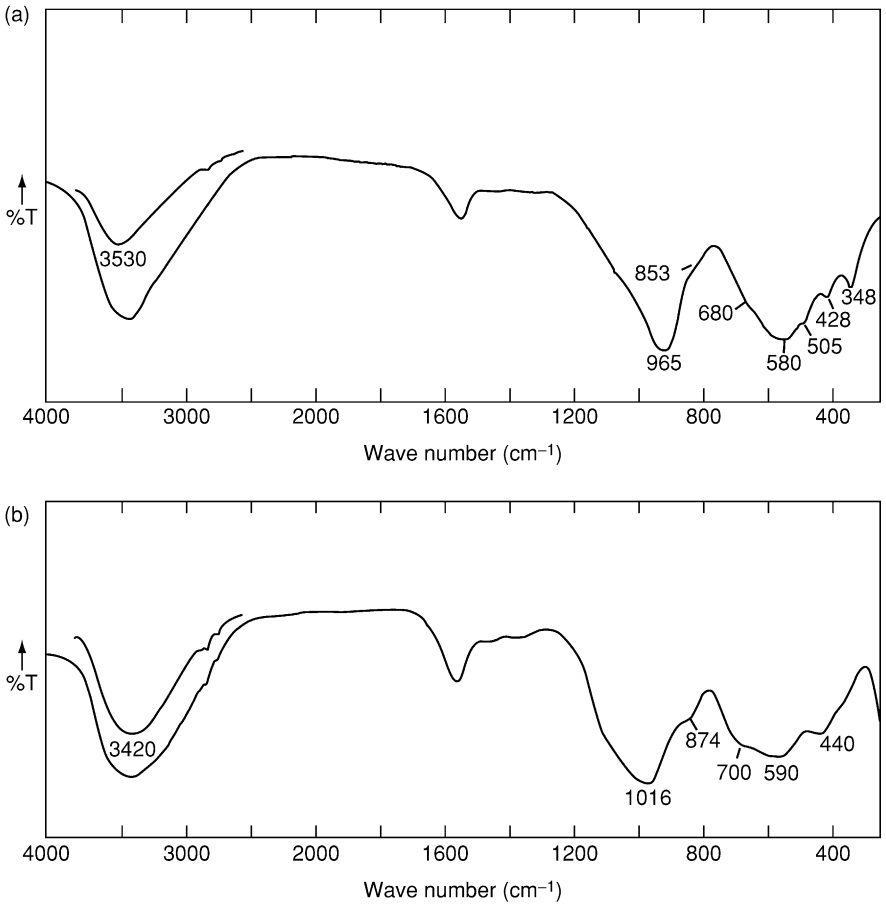


Figure 7 Infrared spectra of (a) Al-rich (protoimogolite) and (b) Si-rich allophanes. The *y*-axis is %T, which represents percentage of transmittance.

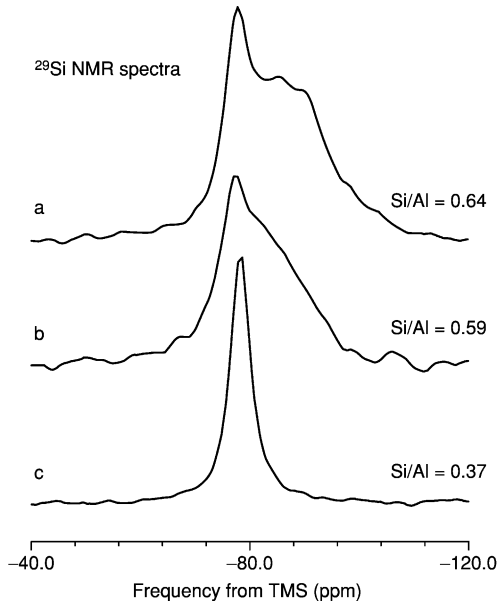


Figure 8 The ^{29}Si-NMR spectra of three allophanes with decreasing Si/Al molar ratio: A, 0.64; B, 0.59; C, 0.37.

when there is no permanent structural charge in the allophane. The fact that these two are nearly equivalent in Figure 9 also shows that Al is not substituting for Si in tetrahedral sites, giving rise to structural charge.

It is evident from Figure 9 that protoimogolite allophane is positively charged in all but calcareous and other alkaline soils. The amount of charge is significant, rivaling smectites in magnitude of charge in both acid and highly alkaline soils. Indeed, allophane and imogolite may be among the most important anion exchangers in soils. The use of soils containing allophane as filters for anions such as arsenate, iodide, and technetium has been considered. Silica-rich allophanes, on the other hand, bear significant negative structural charge and will function as important cation exchangers.

Chemisorption and Ligand Exchange

In addition to the nonspecific cation and anion exchange reactions that occur on positively and negatively charged sites, respectively, metals and ligands

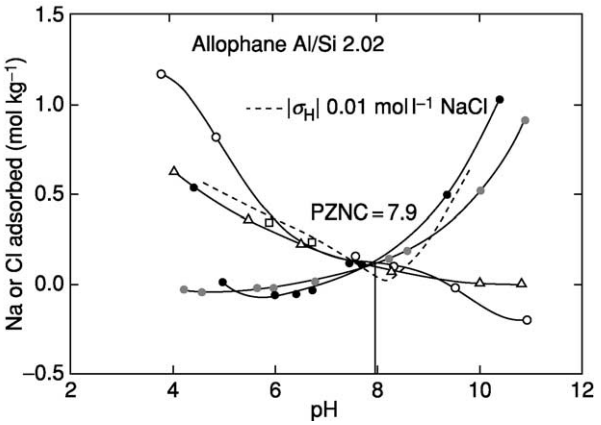


Figure 9 Ion adsorption and proton charge (σ_H) of synthetic Al-rich allophane in NaCl. (Reproduced with permission from Su C, Harsh JB, and Bertsch PM (1992) Sodium and chloride sorption by imogolite and allophanes. *Clays Clay Mineral* 40: 280–286.)

can chemisorb to the reactive aluminol and silanol groups. The nonspecific interactions on charged surfaces are characterized by the fact that the ions can be easily exchanged with other ions of like charge from the soil solution. The adsorbed ions remain hydrated (outer-sphere complexes), are kinetically labile, and equivalently balance the surface charge. Chemisorbed ions, on the other hand, bond directly to the aluminol and silanol groups (inner-sphere complexes), are not easily or rapidly exchanged, and do not require an oppositely charged surface group to adsorb. These interactions have the ability to govern the transport, stability, and bioavailability of solutes including toxic metals, essential plant nutrients, and soil organic matter.

Chemisorption of heavy metals such as Zn, Cu, Pb, Co, and Cd generally involves exchange with a proton from a surface silanol or aluminol group. An inner-sphere complex is formed as the metal bonds directly with the oxygen of the surface group. For example, on an aluminol group on allophane:

$$\text{Al-OH} + \text{M}^{2+}(\text{aq}) \leftrightarrow \text{Al-OM}^{+} + \text{H}^{+}(\text{aq}) \qquad [4]$$

represents the chemisorption of a divalent metal cation (M^{2+}) where one proton is exchanged. The result, in this example, is a more positively charged surface, a reduction in the mobility of the metal cation, and a decrease in soil pH. Given the high specific surface areas of SRO aluminosilicates and the relatively high number of reactive surface groups, allophane and imogolite can play an important role in metal behavior in soils, even when these materials are present in low concentration.

The cation exchange behavior of metals depends primarily on their charge and ionic radius. The selectivity for a negatively charged site tends to increase with both charge and size as long as the cation remains hydrated. In the case of the chemisorbed metals, however, the selectivity depends on the nature of the reaction with the aluminol or silanol group. The selectivity series for metals on aluminol groups follows the order:

$$\text{Cu}^{2+} > \text{Pb}^{2+} > \text{Zn}^{2+} > \text{Ni}^{2+} > \text{Co}^{2+} > \text{Cd}^{2+} > \text{Mg}^{2+} > \text{Sr}^{2+}$$

This series represents the position of the adsorption edge when the amount of metal adsorbed is plotted against pH. It implies that the metals such as Cu and Pb will be chemisorbed to aluminol groups even in acid soils, whereas the alkaline metals like Mg and Sr are not chemisorbed in slightly alkaline to acid soils, but only undergo cation exchange. Chemisorption to silanol groups is similar.

Just as allophane and imogolite can be important anion exchangers in soil, they are also capable of chemisorbing anions and neutral solutes through ligand exchange:

$$\text{Al-OH}_2^{+} + \text{A}^{-} \leftrightarrow \text{Al-A}^{-} + \text{H}_2\text{O} \qquad [5]$$

In the above reaction, an anion (A^-) exchanges with a water molecule, increasing the negative charge on the surface. The anion bonds directly with the surface Al or Si ion, forming an inner-sphere complex. Several solutes are known to chemisorb to allophane and imogolite through ligand exchange, including phosphate, fluoride, selenite, and organic acids.

The sorption of phosphate to SRO materials has a major effect on the availability of phosphate to crops grown in allophanic soils. Although it was long assumed that this reaction was governed strictly by ligand exchange, it is more likely that phosphate reacts with allophane to form insoluble Al-phosphate phases similar to variscite ($AlPO_4 \cdot 2H_2O$). The high retention of phosphate may limit fertility of both agricultural and forested soils. Solutions to this potential problem have included not only high application of inorganic fertilizers, but addition of silica and organic matter to compete with phosphate for reactive sites.

Fluoride is strongly adsorbed to both silanol and aluminol groups and results in the release of OH^- by the reaction:

$$\text{Si-OH} + \text{F}^{-} \leftrightarrow \text{Si-F} + \text{OH}^{-} \qquad [6]$$

This reaction has been used as a test for the presence of SRO aluminosilicates in soil. The high concentration of reactive groups can easily result in a pH >10

when a few grams of an allophanic soil is reacted with $0.1\,\text{mol}\,\text{l}^{-1}$ NaF, initially at neutral pH. Other SRO materials, such as ferrihydrite, and Al-humic complexes also contribute to this reaction. As in the reaction with phosphate, insoluble metal fluorides can replace allophane.

Reaction with Soil Organic Matter

Soils that are derived from volcanic ash or contain high amounts of SRO materials often fall into the Andisol order. The term 'andisol' originated in Japan and refers to 'dark soils.' These nearly black soils resulted from relatively high quantities of organic matter in their surface horizons. Organic matter contents as high as 20% can be found in Andisols that occur in warm and humid climates, where degradation should be rapid. The association between allophanic soils and high organic matter has long been known, but is not well understood.

The high amount of organic matter in allophanic soils has been attributed to several possible factors. One is a high concentration of Al-organic complexes, which inhibit microbial activity. These complexes not only slow organic matter decomposition, but may preclude the formation of SRO aluminosilicates. In this case, it is the Al complexes themselves that impart allophanic properties to the soil. Chemisorption of humic materials to the high-surface-area SRO materials is assumed to protect them from degradation through steric factors or by limiting microbe accessibility via occlusion. It is also possible that reaction between allophane and organic matter alters the humification process, resulting in more recalcitrant, polymethylene-type molecules.

Aluminum Solubility

The activity of Al in soils is an important consideration from an environmental standpoint, because Al is toxic to various plant species, especially agricultural crops, and to aquatic organisms. Aluminum activity increases 1000-fold for every unit decrease in pH when Al solubility is controlled by solid phases such as gibbsite, kaolinite, and imogolite. Among the soils for which a solid-phase dissolution mechanism is likely to control Al activity in the soil solution are those containing SRO aluminosilicates. The B horizons of Spodosols, in particular, near equilibrium with an imogolite-like phase as determined in both field and laboratory studies. As mentioned above, the SRO aluminosilicates are metastable with respect to more crystalline minerals such as kaolinite; however, they are persistent in soils and approach equilibrium solubility fairly rapidly in soils.

The dissolution reaction for imogolite can be written as follows:

$$0.5\text{Al}_2\text{SiO}_3(\text{OH})_4(\text{s}) + 3\text{H}^+(\text{aq}) \leftrightarrow \text{Al}^{3+}(\text{aq}) + 0.5\text{Si}(\text{OH})_4(\text{aq}) + 1.5\text{H}_2\text{O}(\text{l}) \quad [7]$$

and the log solubility product ($\log K_{sp}$) is:

$$\log K_{sp} = \log(\text{Al}^{3+}(\text{aq})) + 0.5\log(\text{Si}(\text{OH})_4(\text{aq})) + 3\text{pH} \quad [8]$$

where parentheses denote activities.

Aluminum solubility can then be obtained by rearranging:

$$\log(\text{Al}^{3+}) = \log K_{sp} - 0.5\log(\text{Si}(\text{OH})_4(\text{aq})) - 3\text{pH} \quad [9]$$

Thus, the activity of Al^{3+} in a soil solution controlled by imogolite depends negatively on $\text{Si}(\text{OH})_4(\text{aq})$ activity and pH.

A stability diagram based on some of the published values for imogolite and protoimogolite allophane in relation to more crystalline minerals is shown in Figure 10. The two imogolite lines are from two different investigations. The more stable of the two synthetic imogolites crosses the halloysite stability line at an aqueous silica activity near $10^{-3.5}$, consistent with the formation of halloysite in soils with higher values. However, the formation of halloysite in preference to imogolite may be kinetic, requiring significant supersaturation. Although the solubility products of these minerals are not yet well defined, Figure 10 still shows the metastability of SRO

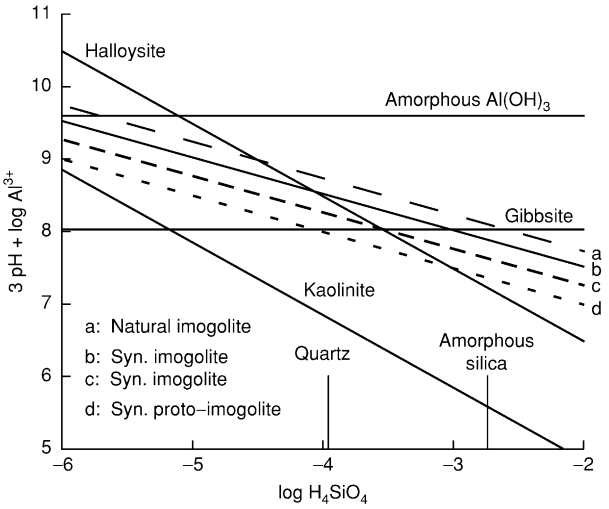


Figure 10 Stability diagram for selected Al hydrous oxide and aluminosilicate solid phases. Syn, synthetic.

materials with respect to kaolinite at nearly all soluble silica activities: the relative stability of gibbsite being higher at low ($Si(OH)_4(aq)$) and of halloysite being higher ($Si(OH)_4(aq)$).

Physical Properties

Soils high in SRO materials generally have very low bulk densities, typically less than $0.85\,Mg\,m^{-3}$. Indeed, this is a defining characteristic of Andisols or soils classified as having 'andic' properties. The low bulk densities result from high microporosity brought about by particle–particle associations and from the strong association between allophanes and soil organic matter, discussed above. Allophane particles form linear associations that flocculate into microporous aggregates (**Figure 2**) when near the point of zero charge or the point where the particles have zero mobility in an electric field. Individual allophane and imogolite particles have densities of 2.7–2.8 $Mg\,m^{-3}$ and 2.6–2.75 $Mg\,m^{-3}$, respectively, which are higher than for crystalline clay minerals.

The surfaces of allophane and imogolite can be assumed to be largely accessible to the soil solution. Specific surface areas determined by a polar molecule such as ethylene glycol monoethyl ether are around $10^6\,m^2\,kg^{-1}$, rivaling the specific surface of smectites. Specific surface areas of allophanes determined with N_2 are significantly lower, ranging from 0.3 to $0.7 \times 10^6\,m^2\,g^{-1}$.

The physical characteristics of SRO materials and associated organic matter have a profound effect on soils containing these materials. In addition to the low bulk densities, such soils have unusually high water retention at both 33 kPa ('field capacity') and 1500 kPa ('permanent wilting point') suction. Saturated water retention has been reported as high as $1.8\,kg\,kg^{-1}$. This can be attributed to the high microporosity and specific surface and leads to favorable conditions for plant growth, particularly in forested soils.

Meso- and macroporosity are also high in andic soils as a result of granular structure brought about by the aggregation of particles and organic matter. Consequently, saturated hydraulic conductivity is generally high in allophanic soils.

Both physical and chemical alteration of these soils can change hydraulic conductivity and other physical properties. When soils containing allophane and imogolite are air-dried, there is an irreversible increase in void volume, which results in higher saturated hydraulic conductivity and lower water retention. It is evidently difficult to rehydrate all surfaces and micropores after drying. In the field, mechanical disturbance that enhances drying or disrupts through shearing can lead to significant lowering of soil plasticity, whereas moist andic soils are considered highly plastic. Finally, adjusting soil pH to values higher or lower than the near-neutral point of zero charge of Al-rich allophanes disturbs particle–particle associations and can reduce hydraulic conductivity as much as 75% or more.

In general, the physical properties of soils containing allophanic materials are conducive to soil productivity. High water retention and hydraulic conductivity not only make water available for plant growth, but result in soils resistant to erosion. Under the wetting and drying conditions common in cropping systems, these soils tend to resist compaction by machinery and mechanical disruption by rainfall. This environment is also conducive to the physical aspects of plant growth, including seedling emergence and root penetration.

See also: **Clay Minerals**

Further Reading

Dahlgren RA (1994) Quantification of allophane and imogolite. In: Amonette J and Zelazny LW (eds) *Quantitative Methods in Soil Mineralogy*, pp. 430–451. Madison, WI: Soil Science Society of America.

Farmer VC and Russell JD (1990) The structure and genesis of allophanes and imogolite: their distribution in non-volcanic soils. In: De Boodt MF, Hayes MHB, and Herbillon A (eds) *Soil Colloids and Their Association in Aggregates*, pp. 165–178. New York: Plenum Press.

Harsh J, Chorover J, and Nizeyimana E (2002) Allophane and imogolite in soil. In: Dixon JB and Schulze DE (eds) *Soil Mineralogy with Environmental Applications*. Madison, WI: Soil Science Society of America.

Huang PM (1995) The role of short-range ordered mineral colloids in abiotic transformations of organic components in the environment. In: Huang PM, Berthelin J, Bollag J-M, McGill WB, and Page AL (eds) *Environmental Impact of Soil Component Interactions,* vol. 1, pp. 135–167. Boca Raton, FL: CRC Press.

Maeda T, Takenaka H, and Warkentin BP (1977) Physical properties of allophane soils. *Advances in Agronomy* 29: 229–269.

Shoji S, Nanzo M, and Dahlgren RA (1993) Volcanic ash soils: genesis, properties and utilization. *Developments in Soil Science* 21: 145–187.

Su C and Harsh JB (1994) Gibbs free energies of formation at 298 K for imogolite and gibbsite from solubility measurements. *Geochimica Cosmochimica Acta* 58: 1667–1677.

Wada K (1989) Allophane and imogolite. In: Dixon JB and Weed SB (eds) *Minerals in Soil Environments*, 2nd edn, pp. 1051–1087. SSSA Book Series. Madison, WI: Soil Science Society of America.

ANAEROBIC SOILS

P W Inglett, K R Reddy, and R Corstanje, University of Florida, Gainesville, FL, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Anaerobic soils occur in areas where oxygen consumption by soil biota exceeds the diffusion of oxygen into the soil profile. This condition is also termed 'soil anaerobiosis' and results in a predominantly oxygen-free environment in the soil profile. Anaerobic soils occur in a number of environments in the landscape, including: wetlands; paddy soils; organic soils; poorly drained, heavy textured soils; areas with a high water table; soils amended with heavy rates of organic materials such as animal wastes, biosolids, and composts; and soils treated with ammoniacal fertilizers.

In upland environments, anaerobic soil conditions may be temporary and may not last more than a few days; while, in wetland environments, soil anaerobic conditions last for several months. Thus anaerobic soils include: all types of wetland soils (swamps, marshes, floodplains, coastal wetlands, and bottomland hardwood forests), hydric soils, paddy soils, organic soils, and any other waterlogged or flooded soils. Because much of our knowledge of anaerobic soils has been gained through research in wetlands and rice paddies, the discussion in this paper will largely focus on the morphological and biogeochemical features of these types of soils.

Wetland soils are widely distributed throughout the world and can be found in all climates, ranging from the tropics to tundra, with the exception of Antarctica. Approximately 6% of the Earth's land surface, which equals approximately 800 million hectares (approx. 2 billion acres) is covered by wetlands. The USA alone contains approximately 12% of the world's wetlands, or approximately 111 million hectares (274 million acres). In any given landscape, wetlands are located in areas with a low elevation and a high water table. Wetlands can be broadly defined as marshes, swamps, bogs, and similar areas. These areas are poorly drained and retain water during rainy periods. Thus, the physical, chemical, and biological characteristics of anaerobic soils are important in determining the properties and functioning of wetlands.

The creation of anaerobic soil conditions is predicated in the situation where demand exceeds the supply of oxygen. Once a soil becomes saturated, the supply of oxygen is immediately reduced owing to the displacement of oxygen contained in the available pore space. Following consumption of the relatively small amount of available oxygen in the pore water, oxygen can only be supplied to respiring organisms through the process of diffusion from the nearest aerobic zone. This process is comparatively slow under saturated soil conditions as oxygen diffusion in water is approximately 10 000 times slower than through air. Under these conditions, even moderate rates of soil or root respiration can quickly deplete available oxygen and result in anaerobic soil conditions.

Depending on hydrologic conditions, wetland soils can be present: (1) flooded, with defined water depth above the soil surface; (2) under saturated soil conditions, with no excess floodwater; and (3) when the water table lies below the soil surface at a certain depth, depending on soil characteristics. Under the first two conditions, wetland soils can be classified as hydric soils, while the third group can mimic the characteristics of both wet- and upland soils, depending on soil type and hydrologic conditions. Soil taxonomy classifies soils with these characteristics into a suborder 'aquic,' which implies that soil pores are filled with water (from soil surface to a depth of 2 m), and many of the oxidized compounds are enzymatically reduced, with end products of these reductive processes accumulating in the soil. Soil taxonomists classify aquic soils according to soil color and not the accumulation of reduced products. Gray colors or low chroma (2) are used generally as indicators of soil anaerobiosis.

Physical Characteristics

Soil volume primarily comprises solid matter, water, and air. When soils are flooded, most of their pore volume is occupied by water. Upland mineral soils generally consist of about 50% by volume of solids, 25% of water, and 25% of air. In wet mineral soils, approximately 50% of the soil volume is solids, while the remaining 50% is occupied by water. In wetland organic soils, a large proportion (up to 80%) of soil volume is occupied by water, with soil organic matter and mineral matter occupying less than 20%.

Generally, reduced compounds are not found in upland soils. Gaseous exchange is not restricted because of continuation of air spaces in upland soils, and oxygen dominates the respiratory and chemical environment. Gaseous composition of soil pores is approximately 10–21% O_2, 0.03–1% CO_2, and trace amounts of N_2O and NH_3. In wetland soils, there is less oxygen, because soil pores are filled with water. In the absence of oxygen, reduced

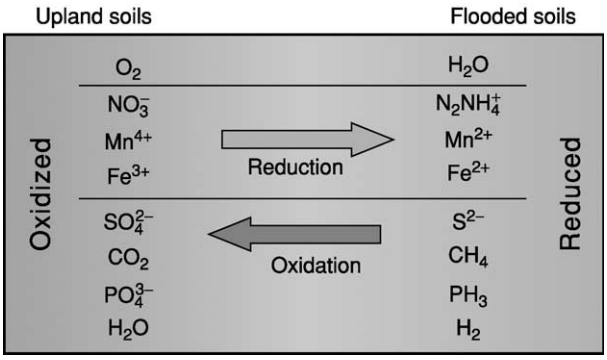


Figure 1 Various inorganic oxidized and reduced compounds in upland and wetland soils.

chemical forms predominate and are regulated by associated biogeochemical processes (Figure 1). In recently flooded soils, N_2O can be present as a result of denitrification of nitrate nitrogen (NO_3-N). In moderately reduced soils, H_2S can be observed, followed by CH_4 under more reducing conditions. In highly reduced soils, C_2H_4 and PH_3 (phosphine) can be observed. However, the presence and accumulation of these gases depend on respective oxidants available for reduction.

Biological Characteristics

Saturated soil conditions support microbial populations adapted to anaerobic environments. Aerobic microbial populations are restricted to zones where O_2 is available. Most of the aerobic organisms become quiescent or die, and new inhabitants, largely facultative (organisms which can function under both aerobic and anaerobic environments) and obligate anaerobic bacteria, take over.

Fungi, which are active in upland environment, are inhibited in the anaerobic wetland soil environment. This is primarily due to absence of O_2 and alteration in soil pH (acid to neutral) under anaerobic conditions. Similarly, microbial biomass decreases under saturated soil conditions. This decrease in microbial activity is primarily due to the shift from aerobic to anaerobic respiration. Thus, under wetland soil conditions, rates of many microbially mediated reactions decline, and some reactions may be eliminated and replaced by new ones.

Saturated soil conditions can support the growth of microphytic communities, including a variety of planktonic, epiphytic, and benthic algae at the soil–floodwater interface. The species composition varies with physicochemical conditions within the wetland. Many of the species in these microbial assemblages have the capacity to carry out photosynthesis. Diel fluctuations in dissolved O_2 produced as a consequence of photosynthesis often increase the O_2 levels in the floodwater beyond saturation levels during daytime and to low levels during nighttime. These large fluctuations in available oxygen have special significance in wetlands for regulating biogeochemical cycles of nutrients.

Wetland or anaerobic soil conditions also support the presence of hydrophytic vegetation, or plants that are adapted to the reducing wetland environment. These plants have unique characteristics to adapt to oxygen-deficient conditions, including physiological adaptations (such as capability to respire anaerobically), anatomical adaptations (such as development of intercellular air spaces), and morphological adaptations (such as water roots and adventitious roots). With these adaptations, hydrophytic plants are able to survive under reducing conditions considered toxic to other macrophytes. In many cases, adapted plant communities become the dominant source of organic matter in wetland systems.

Chemical Characteristics

When oxygen availability becomes limited, bacteria must utilize other compounds as electron acceptors to maintain their metabolism. These compounds, many of which are nutrient elements, can exist in both dissolved and solid phases, and include oxidized forms of elements such as N, Fe, Mn, and S. As they are utilized during respiration, these elements gain electrons and thus become chemically reduced. The result of microbial metabolism therefore is the conversion of oxidized elements to the corresponding reduced form under anaerobic conditions. When wetland soils are drained, many of the reduced compounds are oxidized either by chemical or biochemical reactions. Therefore, in upland and/or drained soils, oxidized forms of chemical species dominate the system, while reduced forms dominate the wet soil system (Figure 1).

Reduction–oxidation, or redox potential (E_h), reflects the intensity of reduction *or* a measure of electron (e^-) activity analogous to pH (which measures H^+ activity). Depending on soil characteristics, E_h generally decreases with time and approaches a steady value after flooding. Redox potential is the most common parameter used to measure degree of soil wetness or intensity of soil anaerobic conditions. The range of E_h values observed in wetland soils is from +700 to −300 mV (Figure 2). Negative values represent high electron activity and intense anaerobic conditions typical of permanently waterlogged soils. Positive values represent low electron activity and aerobic to moderately anaerobic conditions typical of wetlands in transition zone. Soils with

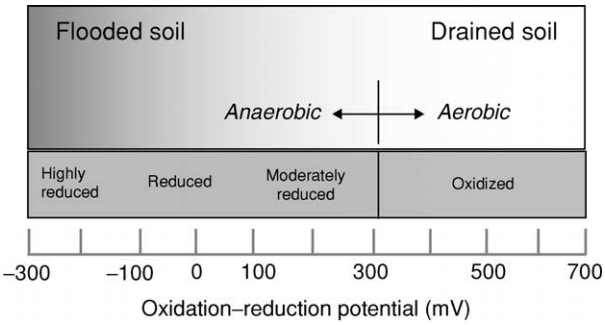


Figure 2 Schematic showing relationship between oxidation–reduction potential and oxidized and reduced conditions in wetland soils.

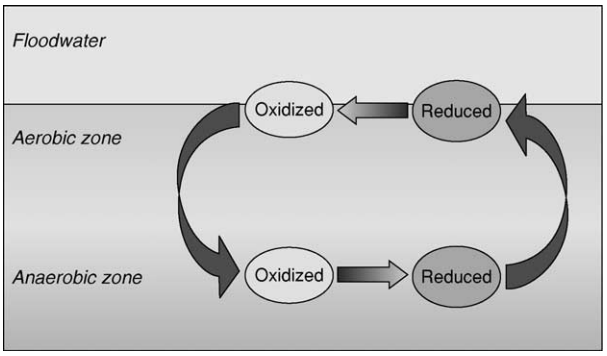


Figure 3 Diffusional patterns of reduced and oxidized compounds in response to anaerobic gradients in flooded soils.

$E_h > 300$ mV are considered aerobic or upland. Under these conditions, oxygen is used as the dominant electron acceptor. Soils with $E_h < 300$ mV are considered anaerobic or wetland.

The chemical nature of the reduction process also affects soil pH, electrical conductivity, cation exchange capacity, and sorption and desorption processes. In general, saturated soil conditions result in an increase in pH, electrical conductivity, and ionic strength, but a decrease in soil redox potential. The pH of most soils tends to approach the neutral point under flooded conditions, with acid soils increasing and alkaline soils decreasing in pH. Increase in pH of acid soils depends on the activities of oxidants (such as NO_3^-, Fe^{3+}, Mn^{2+}, and SO_4^{2-}) and proton consumption during reduction of these oxidants under flooded conditions. In alkaline soils, pH is controlled (and generally lowered) by the accumulation of dissolved CO_2 and organic acids.

Accumulation of reduced compounds in the anaerobic soil layer results in the establishment of concentration gradients across the aerobic–anaerobic interface. The concentration of reduced compounds is usually higher in the anaerobic layer, which results in upward diffusion into aerobic soil or floodwater, where they are oxidized (Figure 3). Similarly, some of the dissolved, oxidized compounds diffuse downward, i.e., from floodwater or aerobic soil layer into underlying anaerobic soil layer, where they will be reduced. The exchange rates between soil and overlying water determine whether the wetlands soils or sediments are functioning as a sink or source for nutrients. The rate of exchange of dissolved species depends upon: (1) concentration of dissolved species in soil pore water; (2) soil type and other related physicochemical properties (pH, cation exchange capacity, organic matter content, and bulk density); (3) concentration of dissolved species in the floodwater; and (4) kinetics of related biogeochemical processes in soil and floodwater.

Morphological Characteristics of Wetland Soils

Wetland protection now requires identification of the boundaries between uplands and wetlands. Criteria based on hydrology, vegetation, and soils are individually or together used to determine these boundaries. Among these three components, soils assessment is particularly critical because, while vegetation and hydrology are temporally affected by climatic fluctuations, soils are the most stable and respond only to long-term inundation. The term 'hydric soils' is now commonly used in jurisdictional language synonymous with wetland soils. Hydric soils are defined as soils formed under conditions of saturation, flooding, or ponding long enough during the growing season to develop anaerobic conditions in the upper part of the soil profile.

Saturated soils develop several unique morphological characteristics as a result of several oxidation–reduction reactions. These features are now used as soil indicators to evaluate independently wetland boundaries. Some of the key hydric soil indicators are formed by the accumulation or loss of iron and manganese, hydrogen sulfide, or accumulation of organic matter. In many cases, soil color is used to assess both the accumulation of organic matter (dark horizons) and the reduction of iron species (formation of gray or gley colors).

Biogeochemical Characteristics

Microbial communities in anaerobic soils play a key role in regulating a number of essential biogeochemical cycles such as carbon, nitrogen, phosphorus, and sulfur. Organic matter released by the primary producers is degraded by microbial communities, releasing nutrients back into the environment. Degradation of organic materials allows heterotrophic microbial groups to obtain energy and nutrients

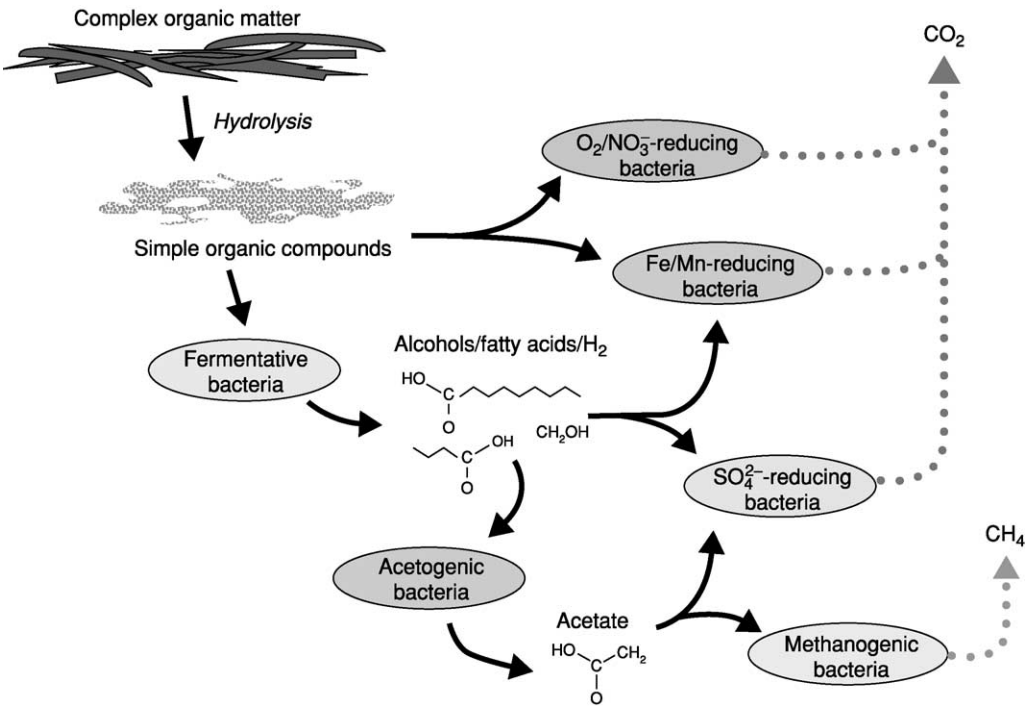


Figure 4 Pathways of organic matter decomposition in wetland soils.

for growth. Enzymatic hydrolysis of organic matter produces several monomers, including glucose, xylose, fatty acids, and amino acids. These simple reduced compounds then serve as substrates for microbial metabolism as chemical energy is released during their oxidation.

Anaerobic decomposition of organic matter is more complex and less energetically favorable than aerobic decomposition. It is mediated by a complex group of physiologically different microorganisms which participate in the decomposition pathways (Figure 4). Often, the end product of one metabolism is substrate for the next until the decomposition is complete. Initially, fermenting bacteria mediate the extracellular hydrolysis of high-molecular-weight polymers (i.e., proteins, polysaccharides, lipids, and nucleic acids) and ferment their respective monomers (e.g., amino acids, sugars, fatty acids, and nucleotides) to CO_2, H_2, acetate, propionate, butyrate, and other fatty acids and alcohols.

Microbial Respiration

Oxygen, if present in wetland soils, provides an electron acceptor (oxidant) for supporting microbial oxidation of reduced carbon compounds (respiration). Thermodynamically, oxygen is the most preferred electron acceptor by microorganisms (Table 1), and as a result, aerobic microorganisms maintain a competitive advantage while oxygen is present. Oxygen is used preferentially because it receives electrons from the reductant material (organic matter) more readily than do other oxidants. The greater energy yield during the aerobic process is due to: (1) complete oxidation of carbon atoms in the organic substrates to CO_2, and (2) the oxygen redox couple having a relatively high, positive reduction potential. This leads to a large net difference in electrical potentials between electron donor (organic substrate) and terminal electron acceptor (oxygen).

Once oxygen is depleted in the soil, bacteria must respire anaerobically. Under these conditions, bacteria capable of utilizing the electron acceptor with the next-highest thermodynamic potential dominate. Thus, the microbial use of electron acceptors proceeds in a sequential manner dependent on their electron affinity, energy yield, and related enzyme systems in the bacteria. In this manner, the order of electron acceptor use following oxygen depletion is: NO_3^-, Mn^{4+}, Fe^{3+}, SO_4^{2-} and finally CO_2 (Figure 5, Table 2). Anaerobic respiration is typically reflected in vertical profiles of these electron acceptors, with soil and/or sediment depth, as the most favorable electron acceptors are utilized first. However, the exclusion of less-favorable respiration pathways is not complete, resulting in considerable overlap between the different forms of organic carbon mineralization. The rate at which electron acceptors are consumed in soil systems depends on their concentration, the availability of organic compounds, and the activity of the microbial population involved in the reductive processes.

Table 1 Microbial groups involved in various redox reactions in wetland soils

Redox potential (mV)	*Electron acceptor*	*Decomposition end products*	*Microbial groups*
Aerobic			
>300	O_2	CO_2, H_2O	Aerobic fungi and bacteria
Fermenting			
less than −100 to +300	Organics	Organic acids, CO_2, H_2, alcohols, amino acids	Fermenting bacteria
Facultative anaerobic			
100–300	NO_3^-	N_2O, N_2, CO_2, H_2O	Denitrifying bacteria
	Mn^{4+}	Mn^{2+}, CO_2, H_2O	Mn(IV) reducers
	Fe^{3+}	Fe^{2+}, CO_2, H_2O	Fe(III) reducers
Obligate anaerobic			
less than −100	SO_4^{2-}	HS^-, CO_2, H_2O	Sulfate reducers
	CO_2 and acetate	CH_4, CO_2, H_2O	Methanogens
	Organic acids	Acetate, CO_2, H_2	H_2-producing bacteria

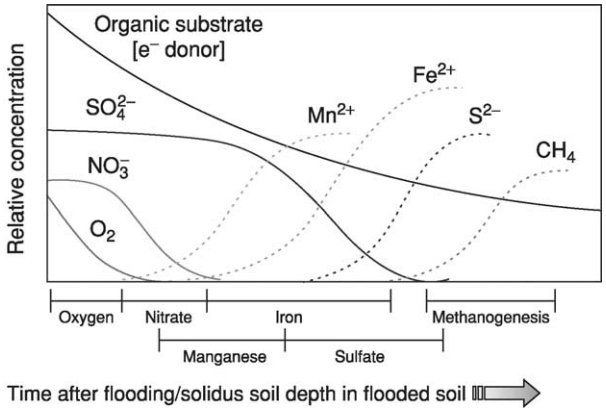


Figure 5 Sequential reduction of oxidants (oxidized compounds) and accumulation of reductants (reduced compounds) in wetland soils.

Nitrate reduction can occur in wetlands according to two major pathways:

- Dissimilatory nitrate reduction to ammonia (DNRA):

$$NO_3^- \rightarrow NO_2^- \rightarrow NH_4^+$$

- Denitrification:

$$NO_3^- \rightarrow NO_2^- \rightarrow NO \rightarrow N_2O \rightarrow N_2$$

Dissimilatory reduction of NO_3^- is performed by a variety of facultative anaerobic bacteria. During this process, NO_3^- is first converted to nitrite NO_2^-, which may be further reduced to NH_4^+. Reduced NH_4^+ produced through dissimilatory NO_3^- reduction results in high NH_4^+ levels characteristic of wetland soils and sediments. Denitrifiers are heterotrophic bacteria (most of them facultative anaerobic) that couple the oxidation of organic substrates to the reduction of NO_3^- to either N_2O or N_2. This reaction occurs in moderately reduced conditions in the absence of oxygen and is one of the dominant mechanisms for removal of nitrogen from aquatic systems.

The oxidation and reduction of iron in many soils is possibly one of the main components of soil formation. Its relatively ubiquitous presence in soils and sediments makes this respiratory pathway a major contributor of organic-matter mineralization. A large variety of microorganisms are capable of reducing iron, including fungi. When Fe(III) is reduced, Fe(II) is the reduced end product. For Mn, Mn(II) is generally accepted as the end product of Mn(IV) reduction; however, Mn(III) may also be encountered as an intermediate species.

In the general Fe(III) and Mn(IV) reduction model, complex organic matter is hydrolyzed to smaller components (i.e., sugars, amino acids, fatty acids). The sugars and amino acids are metabolized by fermentative microorganisms, which may reduce a small amount of Fe(III) or Mn(IV) in the process. The majority of the primary products from this first stage of the metabolism of sugars and amino acids are short-chain fatty acids and possibly hydrogen. This hydrogen can then be oxidized by Fe(III) and Mn(IV) reducers (e.g., *Pseudomonas* sp.), while other fermentation products are oxidized through Fe(III) or Mn(IV) reduction by species such as *Shewanella putrefaciens*. Alternatively, *Thiobacillus thiooxidans* or *T. ferrooxidans* can reduce Fe(III) or Mn(IV), with elemental sulfur S^0 as the electron donor.

Sulfate reducers are obligate anaerobes that couple oxidation of organic substrates to CO_2 with the reduction of terminal electron acceptor SO_4^{2-} to sulfides ($-S^{2-}$). Gram-negative bacteria such as *Desulfobacterium*, *Desulfobulbus*, and *Desulfotomaculum* are the most common types of sulfate-reducing bacteria in freshwater sediments. Sulfate-reducing bacteria cannot

Table 2 Summary reactions for microbial respiration pathways in wetland soils

Electron acceptor	*Reaction coupled to glucose oxidation*	ΔG_f^o *(kJ mol^{-1})*
O_2	$C_6H_{12}O_6 + 6O_2 \rightarrow 6CO_2 + 6H_2O$	−2879
NO_3^-	$5C_6H_{12}O_6 + 24NO_3^- + 24H^+ \rightarrow 30CO_2 + 12N_2 + 42H_2O$	−2713
MnO_2	$C_6H_{12}O_6 + 12MnO_2 = 6CO_2 + 12M_n^{2+} + 18H_2O$	−1916
$Fe(OH)_3$	$C_6H_{12}O_6 + 24Fe(OH)_3 + 48H^+ = 6CO_2 + 24Fe^{2+} + 66H_2O$	−418
SO_4^{2-}	$C_6H_{12}O_6 + 4H_2O \rightarrow 2CH_3COO^- + 2HCO_3^- + 4H_2 + 4H^+$	−207
	$CH_3COO^- + SO_4^{2-} + 3H^+ = 2CO_2 + H_2S + 2H_2O$	−63
CO_2	$C_6H_{12}O_6 + 4H_2O \rightarrow 2CH_3COO^- + 2HCO_3^- + 4H_2 + 4H^+$	−207
	$2CH_3COO^- + 2H_2O \rightarrow 2CH_4 + 2HCO_3^-$	−31

synthesize enzymes to hydrolyze polymers such as polysaccharides. Also, many groups of sulfate-reducing bacteria cannot use monomers such as monosaccharides (e.g., glucose) as substrates for energy, and thus sulfate reducers are dependent on fermenting bacteria to produce simple organic compounds (e.g., acetate, propionate).

Sulfate reducers are widely studied groups of microorganisms with special significance in coastal wetland ecosystems due to the high concentration of sulfate in seawater. Sulfate reduction can occur over a wide range of pH, temperature, and salinity. One product of sulfate reduction, hydrogen sulfide, is extremely toxic to aerobic organisms, because it reacts with the heavy metal groups of the cytochrome system. Hydrogen sulfide is very reactive with metals and usually results in the precipitation of metallic sulfides (e.g., FeS).

The terminal step in the anaerobic degradation of organic macromolecules, in the absence of all other electron acceptors, is the conversion of acetate and H_2/CO_2 to methane. This is an intricate process involving a net of interactions, possibly encompassing the largest set of microbial dependencies. Methanogens are obligate anaerobes that grow autotrophically (they use CO_2 as C source and as electron acceptor) and heterotrophically (they use organic substrates as energy source).

Like sulfate reducers, methanogens cannot directly utilize large-molecular-weight polymers; so methanogens must depend on at least three groups of microbes, including hydrolytic, fermentative, and H_2-producing acetogenic bacteria. Methanogens are typically found in the archaeal families of Methanobacteriaceae, Methanomicrobiaceae, Methanosaeteaceae, and Methanosarcinaceae.

Fermentation pathways vary depending on the original substrate, and quantity and presence of alternate electron acceptors. Denitrification, and Fe(III) and Mn(IV) reduction may utilize any of these fermentation products as the final step in respiration. Acetate and H_2, and other small organic acids are utilized directly by sulfate-reducing bacteria, while methanogens can only use acetate and H_2. Acetogenic bacteria cleave organic acids and alcohols into acetate, H_2, and CO_2. This conversion is only possible in the presence of sulfate-reducing bacteria or methanogens that consume H_2, resulting in low hydrogen concentrations, ensuring that acetogenesis is thermodynamically favorable.

Agronomic, Ecologic, and Environmental Significance

Anaerobic soils occupy an important niche in the biosphere, and their importance in wetlands and paddy soils is widely recognized by scientists, environmental managers, and policy-makers. Agronomically, anaerobic soils commonly known as paddy soils are widely used throughout the world for rice production. Anaerobic soils in wetlands are primary drivers of natural ecosystem function, as many of the biogeochemical processes have important feedback to ecosystem productivity and function.

Anaerobic soils are primary nutrient sources to plants grown in the paddy soils or wetlands. The decomposition process described here results in production of bioavailable nitrogen and phosphorus, which supports the productivity of plants. Furthermore, the extent of Fe(III) and/or Mn(IV) reduction can strongly influence the distribution of toxic trace metals and availability of P.

Environmentally, anaerobic soils may have both positive and negative attributes. One negative aspect is that wetlands are one of the primary sources of methane, a potent greenhouse gas. Approximately 25% of methane emitted to the atmosphere is derived from wetlands. Alternatively, anaerobic soils in wetlands also function as sinks, sources, transformers of nutrients and contaminants, and their role in improving water quality is widely recognized. This function of anaerobic soils has resulted in developing low-cost constructed wetland technology for water treatment. At present several thousands of such wetlands are in operation throughout the world.

See also: **Carbon Cycle in Soils:** Dynamics and Management; **Hydric Soils**; **Microbial Processes:** Environmental Factors; **Nitrogen in Soils:** Cycle; **Organic Soils**; **Paddy Soils**; **Sulfur in Soils:** Overview; **Wetlands, Naturally Occurring**

Further Reading

Cowardin LM, Carter V, Golet FC, and LaRoe ET (1979) *Classification of Wetlands and Deepwater Habitats of the United States*. Fish and Wildlife Service, US Department of Interior. Washington, DC: US Government Printing Office.

Federal Register, Notice (1994) *Changes in Hydric Soils of the United States*. Federal Register, vol. 59. Washington, DC: US Government Printing Office.

Federal Register, Notice (1995) *Hydric Soils of the United States*. Federal Register, vol. 60. Washington, DC: US Government Printing Office.

Holmer M and Storkholm P (2001) Sulphate reduction and sulphur cycling in lake sediments: a review. *Freshwater Biology* 46: 431–455.

Lovley DR (1991) Dissimilatory Fe(III) and Mn(IV) reduction. *Microbiological Reviews* 55(2): 259–287.

National Technical Committee for Hydric Soils (1991) *Hydric Soils of the United States*. USDA–SCS Miscellaneous Publication 1491. Washington DC: USDA–SCS.

Ponnamperuma FN (1972) The chemistry of submerged soils. *Advances in Agronomy* 24: 29–96.

Reddy KR and D'Angelo EM (1994) Soil processes regulating water quality in wetlands. In: Mitsch W (ed.) *Global Wetlands – Old World and New*, pp. 309–324. New York: Elsevier.

Reddy KR, D'Angelo EM, and Harris WG (2000) Biogeochemistry of wetlands. In: Sumner ME (ed.) *Handbook of Soil Science*, pp. G89–G119. Boca Raton, FL: CRC Press.

Schwertmann U (1993) Relations between iron oxides, soil color, and soil formation. In: Bigham JW and Crolkosz EJ (eds) *Soil Color*, pp. 51–70. Madison, WI: Soil Science Society of America.

US Department of Agriculture, Natural Resource Conservation Service (1998) *Field Indicators of Hydric Soils in the United States*. version 4.0. Ft. Worth, TX: USDA–NRCS.

Vepraskas MJ (1992) *Redoximorphic Features for Identifying Aquic Conditions*. Technical Bulletin 301, North Carolina Agricultural Research Service. Raleigh, NC: North Carolina State University.

Anion Exchange *See* **Cation Exchange**

APPLICATIONS OF SOILS DATA

P J Lawrence, USDA Natural Resources Conservation Service, Washington, DC, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Soil survey, classification, and interpretation efforts are well established in countries around the world, and play significant roles in land use and natural resource decision-making. Scientific inquiry into soils began in the nineteenth century, with centers in Western Europe, Russia, and the USA. The early US pioneers in soil science and survey emphasized the soil genesis concepts developed in Russia in the 1870s and worked to validate and apply these landmark principles. Soils interpretations bring information on tangible and measur- able soil properties together with desired uses to make predictions about site suitability or limitations. Applications of soils data and survey interpretations have steadily broadened in range and precision, from the earliest interpretations that identified potential for salt accumulation in soils to today's sophisticated uses for precision agriculture and evaluating nonpoint-source pollution risks. This article examines the development and uses of soils interpretations and data in the USA.

Development of Interpretations in the USA

Interpretations were emphasized to varying degrees throughout the history of the US soil survey. Milton Whitney, the first Chief of the Soil Survey Division, US Department of Agriculture (USDA), started the systematic survey of soils in four areas of the country in 1899. These early surveys concentrated on measurable soil characteristics that correlated with the

properties Whitney thought important: soil texture, soil moisture, soil temperature, and concentration of soluble salts in soils.

The most accurate and economically valuable interpretation early soil surveyors could make was to identify the potential for 'alkali' soils – reflecting the presence of soluble salts in soil and water. In 1901 and 1903, interpretations from work in the Imperial Valley of Southern California accurately identified large areas of heavy-textured impervious soils of high salt content that would pose substantial limitations for drainage and reclamation. This report coincided with a large-scale irrigation project promoted by an influential company. A subsequent survey for the Modesto-Turlock area of the San Joaquin Valley in California (1909) also reported 'alkali' presence. Predictably, such information was not always popular with landowners, developers, and others who perceived that the information diminished land values and development opportunities.

Whitney's successor, Curtis Marbut, generally deemphasized the federal role in interpretations, but worked assiduously at establishing the concepts of soil genesis first introduced in Russia by Dokuchaiev, developing a soil classification system, and establishing the scientific reputation of the published soil survey. State governments, through cooperating soil surveyors and other agricultural scientists, took the lead in preparing and publishing interpretations of completed surveys. A common theme was to identify soils particularly suited to certain crops, for example, apples or tobacco. Other federal agencies, such as the Bureau of Reclamation and the Bureau of Public Roads, established cooperative studies with the Bureau of Soils in acquiring soils information for their particular uses.

Despite Marbut's view on interpretations, some effort was made to develop soil erosion interpretations. In the late 1920s, as Hugh Hammond Bennett was escalating his campaign for research on erosion and soil conservation measures, H.E. Middleton of the Bureau of Chemistry and Soils was making substantial progress toward understanding the complex processes related to erodibility. Bennett identified areas where the combination of geology and agricultural practices combined to produce serious soil erosion and he took this discovery to both the public and politicians (Figure 1). Subsequently, Congress authorized a series of soil erosion experiment stations where interdisciplinary teams of researchers measured erosion conditions under different crops, soils, rotations, and a variety of structural and management practices. These pioneering studies added to interpretations and led to national-level soil erodibility data – the origins of the erodibility data that support current conservation planning tools such as the Universal (and the Revised) Soil Loss Equation.

When Charles E. Kellogg became Chief of the Soil Survey Division in 1935, the period of scant federal interest in soil interpretation ended. Kellogg brought new vigor to developing soil interpretations, which coincided with Bennett's success in influencing Congress to create the Soil Conservation Service (SCS). (The Soil Conservation Service was renamed the Natural Resources Conservation Service (NRCS) in 1994.) Kellogg felt strongly that pedologists had a

Figure 1 The combination of severe drought, geology, and agricultural practices contributed to severe erosion problems in the US Dust Bowl days. (Photo: Location unknown, 1930s). Courtesy of the Natural Resources Conservation Service USDA.

responsibility to summarize and make public the results of their work. As a student at Michigan State University, he had worked on the 1922 Michigan Land Economic Survey, a landmark in multidisciplinary research and inventorying for land classification. Kellogg also established a policy that soil surveys would include productivity ratings. Earlier attempts to develop natural or inherent productivity ratings had met with criticism as interpreters wrestled with how to rate soils that were productive under certain management but nonarable otherwise. Under Kellogg, productivity as an interpretation was to be relative to the technology and management applied. Kellogg's re-energized emphasis on interpretations was timely, corresponding with the emergence of new agricultural technologies that when matched to refined differentiation of soil types could generate substantial productivity increases.

The 1940s found the USDA in the position of having two soil survey efforts: the federal component of the National Cooperative Soil Survey (NCSS), which was aimed at producing a national inventory of soils, and the more localized farm planning surveys of soils that were conducted by SCS for use in assisting farmers and ranchers to apply soil-conserving practices. While the general soil classification framework was similar, scale, legends, and interpretations differed considerably. In 1945, USDA made clear that its requirements for legend preparation, field reviews, and soil correlation applied only to the NCSS. The SCS soil maps produced for farm planning would not be integrated into the NCSS national portrait of soil resources. Later, however, the two survey efforts were united. In 1952, the Federal NCSS component (the Soil Survey Division) was integrated into SCS; mapping was accelerated, and interpretations benefited from the expertise of a large cadre of conservationists with training in range science, agronomy, forestry, biology, engineering, and other disciplines. The merger also linked the soil survey to a major user group – the agricultural landowners with whom the SCS conservationists worked directly (Figure 2).

During this period, a new soil classification system was being developed under the leadership of Guy D. Smith. The new system, *Soil Taxonomy*, retained the deep-rooted concept of the soil series, but made classification more quantitative, and provided a more accurate, clearly defined basis for making interpretations.

Figure 2 Hugh Hammond Bennett, 1946. In 1946, over a decade after the establishment of the Soil Conservation Service (SCS), Hugh Hammond Bennett (left) visits a farm in Coon Valley, Wisconsin. Coon Valley was the first SCS field demonstration of the benefits of conservation systems to agricultural productivity and sustainability. Photograph by M.F. Schweers. Courtesy of the Natural Resources Conservation Service USDA.

Interpretations Beyond Agriculture

Before the 1950s, the primary applications of soil surveys were farming, ranching, and forestry, although some states recognized applications for highway planning as early as the late 1920s. In the post-World War II environment, however, nonagricultural interpretations flourished. Rapid urbanization brought forth example after example of problems stemming from development on poorly suited sites. Soil scientists made the point that soils information could be used to avoid some problems, but more information was needed about the response of soils to developmental uses. In the 1950s, soil scientists, engineers, and others worked to develop interpretations of soils for building sites, sewage-disposal systems, highways, pipelines, and recreation. By the late 1950s, some states and counties had begun to integrate soil-survey data and interpretations into their land-use planning functions. Fairfax County, Virginia, a rapidly urbanizing suburb of Washington, DC, is thought to be the first county in the USA to hire a full-time soil scientist.

A special symposium at the 1965 annual meeting of the Soil Science Society of America highlighted the rapid evolution of soil interpretations in the postwar period, especially those related to land-use planning. The resulting publication, *Soil Surveys and Land Use Planning*, represented widely ranging disciplines – soil science, civil engineering, architecture, and city and regional planning, among others. The symposium revealed the cost-effectiveness of integrating soil-survey interpretations into planning efforts. By passing the Soil Information Assistance for Community Planning and Resource Development Act of 1966, Congress clarified the legal basis for conducting soil surveys on nonfarm

areas in order to assist states and other geopolitical units in planning and resource development.

The decades of Kellogg's leadership were active for interpretations. The land capability classification system was refined and updated to help identify appropriate land uses and needed conservation practices for agricultural sustainability (Figure 3). The classification system, however, did not adequately address range or forestland capability, thus prompting the first forest soil-site correlations in the 1960s. The work on range and forestland quickly became firmly integrated into the soil survey. Soon biologists were developing criteria and ratings for soils related to food and cover for wildlife. The soil survey was well along in its transition from a purely agricultural productivity perspective to a much broader view.

Today's Interpretive Categories

Today, soil-survey information helps to answer a wide range of soil-related questions. The National Soil Survey Manual of the Natural Resources Conservation Service (NRCS) identifies 12 standard interpretive categories for soil-survey data:

National Inventory Groupings

echnical soil groupings present soil features or attributes of specific interest on a national scale. Such groupings have been developed as criteria for national-level programs, for example, prime farmlands, unique farmlands, hydric soils, and highly erodible lands.

Land-Use Planning

Interpretations provide information that allows evaluation of the potential environmental and economic effects of proposed or competing land uses. Interpretive maps at different scales and with different taxonomic levels are used as appropriate to the planning area. For example, regional planning may employ maps on entire associations of soil series and at higher taxonomic levels.

Farmland

Farmland interpretations traditionally place soils in management groups, identifying the soil properties important to crop production, and conservation needs, among other aspects of agriculture. Interpretations may be made to determine areas suitable for specialty crops, match crops with appropriate soils, delineate fields for greater soil homogeneity, and identify needs for specific management practices.

Productivity interpretations are one of the most important farmland interpretations and are described in terms of the output of product per unit land area (or as carrying capacity or liveweight gain for pastureland) under defined management. Interpretations may also present potential productivity ratings based on best practices or generalized soil productivity based on a number of different crops in the survey area.

Resiliency interpretations provide information on the ability of a soil to rebound from depletion or degradation. Resiliency ratings are important in evaluating alternative management systems and long-term effects of management systems on soils.

Rangeland

Soil-range site correlation identifies the suitability of the soil to produce various kinds, proportions, and amounts of plants, which is important in developing management alternatives to maintain site productivity. Range site descriptions provide information on the landscape, climate, soil, and vegetation factors, and the typical species (plant and animal) and their dynamics. The site interpretation describes the potential importance of the site for each of its uses and the feasibility of reclaiming degraded or depleted areas of the site.

Forest Land

Soil-forest site correlation describes the suitability of the soil to produce wood products and provides information on erosion and windthrow hazards, equipment limitations, seedling mortality, plant competition, and recommended trees for reforestation. Estimated productivity of the common trees may be given for each soil in the survey area if it is an important component of the overall survey. Surveys provide information on understory vegetation, potential problems associated with typical forestry operations, and effects of potential hazards such as burning, soil-borne pests, and diseases.

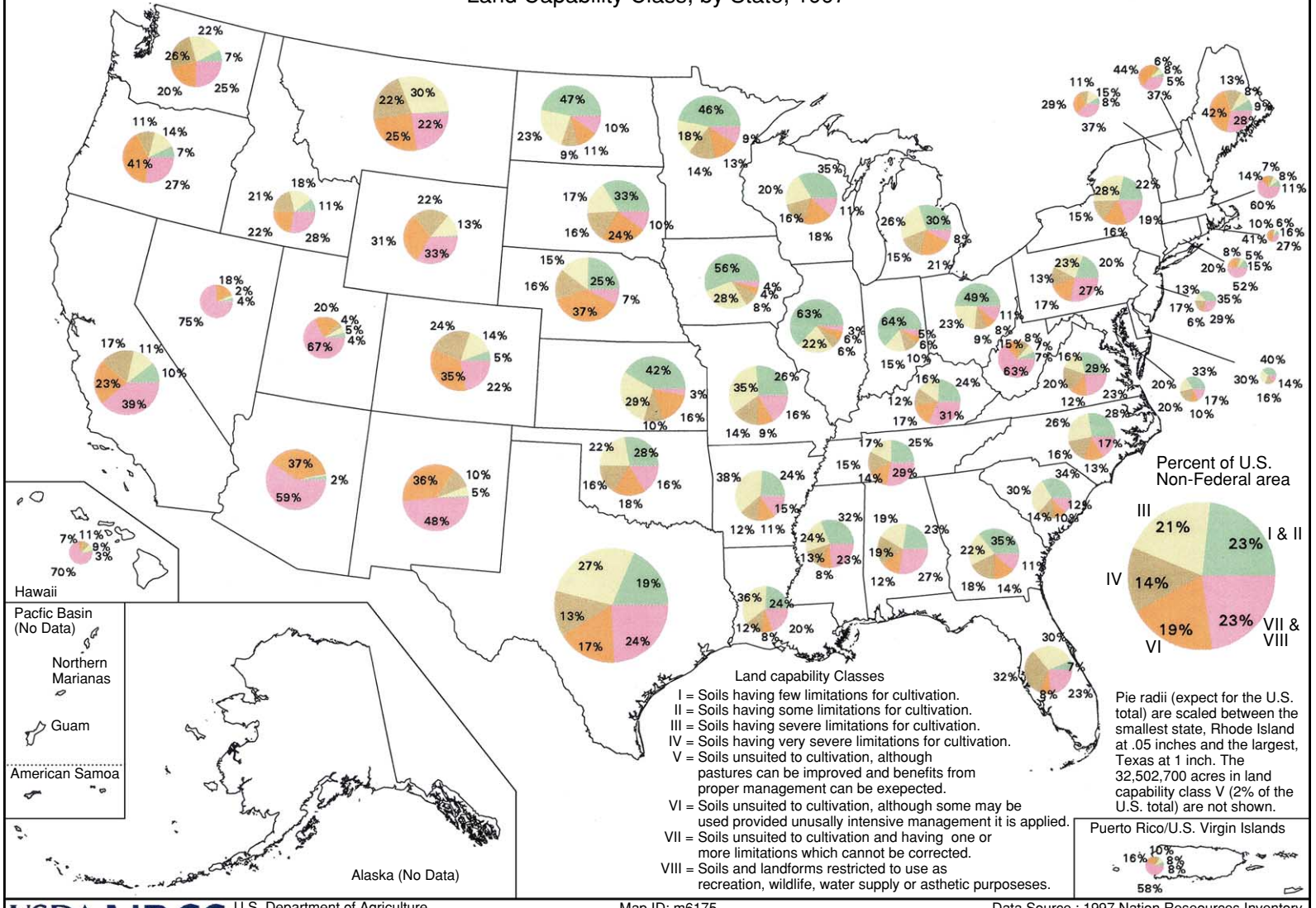
Windbreaks

Windbreaks of suitable species are used to protect soil resources, conserve moisture and energy, provide wildlife habitat, and protect homes, among other purposes. Correlation of soil properties and adaptable windbreak species helps in the selection of appropriate species to achieve the intended objective.

Recreation

Interpretations are made for a variety of recreational purposes such as golf courses, picnic sites, playgrounds, hiking paths, ski areas, snowmobile trails, and campsites. Restrictive soil-interpretative properties such as slope and texture of surface horizons tend

Land Capability Class, by State, 1997



Land capability Classes

I = Soils having few limitations for cultivation.
II = Soils having some limitations for cultivation.
III = Soils having severe limitations for cultivation.
IV = Soils having very severe limitations for cultivation.
V = Soils unsuited to cultivation, although pastures can be improved and benefits from proper management can be exepected.
VI = Soils unsuited to cultivation, although some may be used provided unusally intensive management it is applied.
VII = Soils unsuited to cultivation and having one or more limitations which cannot be corrected.
VIII = Soils and landforms restricted to use as recreation, wildlife, water supply or asthetic purposeses.

Pie radii (expect for the U.S. total) are scaled between the smallest state, Rhode Island at .05 inches and the largest, Texas at 1 inch. The 32,502,700 acres in land capability class V (2% of the U.S. total) are not shown.

USDA NRCS
U.S. Department of Agriculture
Natural Resources Conservation Service
Resource Assessment Division
Washington DC, December 2001

Map ID: m6175
For proper interprtation, see Explanation of Analysis for this map at our web site. Search for "USDASOTL" to locate our map index.

Data Source : 1997 Nation Reseources Inventory
Revised December 2000

to form the basis for ratings. Other factors such as location, accessibility, and infrastructure also tend to be important; however, they are not evaluated in map units for the survey.

Wildlife Habitat

Interpretations for wildlife habitat describe the suitability for different vegetation groups or habitat elements (e.g., hardwood trees, shallow-water areas), and a rating for types of wildlife supported (for example species adapted to open spaces, woodland, or wetlands). Current land use and existing vegetation and wildlife populations are not considered because they are subject to change; they nevertheless may express a significant influence on the potential for an area to support wildlife.

Construction Materials

Interpretations describe the suitability of the soil as a construction material and locations for obtaining materials such as gravel, sand, or low-shrink–swell potential soils. Survey information may also rate soils as potential sources of materials for other purposes, such as organic or mineral soil material that may be used as soil amendments or in horticultural uses. Soils can be rated as possible sources of these materials, but the quality of the site generally cannot be defined.

Building Sites

Interpretations describe the suitability of sites for small-building construction, as well as for road, street, and utility installation, among other construction purposes. Building-site interpretations provide information that can be used to compare suitability of alternative sites, but onsite evaluation is necessary for site selection.

Waste Disposal

Interpretations rate soils for their capacity to handle waste in a relatively small area, such as a septic-tank absorption field, or as distributed at low rates over a larger area. Typically, the evaluation begins with determining how disposal systems have performed on specific kinds of soil in the area, either from experience or related research. Soil scientists and specialists in other disciplines determine what properties are critical and how to appraise the effects of the properties.

Water Management

Interpretations for water management focus on the construction of small to medium impoundments, waterway control, installation of drainage and irrigation systems, and control of surface runoff. Detailed onsite evaluations are required to design engineered projects; however, interpretations assist in the evaluation of alternative sites.

Uses of Soils Data and Interpretations

While soils information is sometimes used alone, it is also used as one layer of information in integrated systems that consider other natural resources, demographics, climate, and ecological and environmental factors. Soil-survey data are used in models that deal with regional planning, erosion prediction, crop yields, and even global change. The following examines a few important or unique ways in which soil survey interpretations have been and are used today.

Federal and State Programs

While all USDA conservation programs and activities rely to some degree on soils information and interpretations, certain efforts and programs have more prominent links. The Farmland Protection Policy Act (FPPA; PL 97–98, 7 USC 4201), administered by USDA, was enacted by Congress in 1981 after finding that federally funded projects were playing a role in the loss of important agricultural lands. The FPPA mandates federal entities to conduct a land evaluation and site assessment to determine the potential for proposed projects to result in conversion of important agricultural lands to nonagricultural uses – soils information is the basis for identifying the presence of those important agricultural lands. Soils information is also a basic element in USDA farm program eligibility determinations. The Food Security Act of 1985 (1985 Farm Bill, PL 99–198, 7 USC 1631)

Figure 3 Land capability class, by state, 1997. The land capability classification system refined by Kellogg is still used today. This map allows a general comparison of land capability among states. Courtesy of the Natural Resources Conservation Service USDA.

included landmark provisions linking eligibility for USDA farm programs to compliance with the highly erodible land (HEL) and wetlands requirements in the legislation (Figure 4). Soils information forms the basis for determining the presence of HEL or wetlands. Soils information is also used to determine if lands are eligible for farmland protection programs. The presence of prime and important soils is a requirement for participation in USDA's Farm and Ranch Lands Protection Program, which works in conjunction with state, local, or nongovernmental organization programs to purchase conservation easements to protect important agricultural lands.

State and Local Planning

Interpretations have formed the foundation for land-use planning ordinances across the USA (e.g., floodplain setbacks; slope protection; agricultural zoning). One of the earliest uses of the soil-survey information for planning was in Fairfax, VA in the 1950s. So useful were the survey and interpretations that the county retained a soil scientist to provide consultation to planners. Within 4 years the county had a soil scientist on permanent staff to assess: (1) septic tank disposal systems; (2) rezoning cases; (3) floodplain extent; (4) determining best trees/shrubs for planting; (5) type of and depth to bedrock; (6) slope stabilization and soil slippage problems; and (7) new school sites. Today, soil survey data are used to develop suitability models that evaluate soils for their capacity for specific uses (for example, as water recharge or environmentally sensitive areas), and thus support local land-use zoning decisions.

Natural Resource Management

Early interpretations focused on the capabilities of land for agricultural production including productivity ratings and soil suitability for specific crops. One of the earliest agronomic interpretations was identifying 'tobacco soils' in the southeast for expansion of production to other suitable areas. Over time, interpretations have become much broader, covering a wide array of environmental effects.

With the emergence of precision farming, soils information is increasingly being used in geographic information systems (GIS) to develop variable-rate application plans tailored to specific field conditions. While the resolution of data needed in precision agriculture may exceed that generally provided by soil surveys, the development of improved dissemination and expert systems will likely benefit some kinds of precision agriculture.

Soil-landscape information provided by soils interpretations is an important tool for managing forestland, rangeland, and wildlife. Range and forest site correlations are being enhanced to distinguish sites by their ability to produce a characteristic natural plant and animal community. This ecological site data for forestland and rangeland will be available through the NRCS Ecological Site Information System (ESIS) and will provide data on the ecological site, plant and animal composition, history, and condition of the community over time. The National Park Service uses soil-survey information to support vegetation management in the park system. These soil–ecological site interpretations may also be used to identify forest ecosystems with high potential for containing specialty forest products such as truffles or ginseng.

Interpretations can have a large role in the conservation of wildlife populations. Soils maps help botanists identify areas with high probabilities of having rare or endemic species – often species that are endangered or threatened as they are restricted to unusual soils (e.g., soils produced from substrates like gypsum, marine clays, or from the interactions of substrates and dynamic geomorphic processes). Some such searches have led to the discovery of additional populations. This method has proved successful for locating new populations of green pitcher plant (*Sarracenia oreophilia*), Mohr's Barbara button (*Marshallia mohri*), and geocarpon (*Geocarpon minimum*).

Land Appraisal and Assessment

Soil interpretations historically have had significant influence in property sales. The most vociferous opposition to early soil surveys arose in response to interpretations that identified limitations on land potentials, primarily of an agricultural nature. While early uses of soils information included land valuation for tax assessment, the same information has been used to reverse assessments. In one case, a community won a reduced tax assessment as they convincingly demonstrated for the court that land values were lower than assessed because of the soils' natural unsuitability to support building foundations.

Soils data and interpretations can be used to evaluate potential risks and insurance needs. Insurance underwriters commonly use soil-survey information and interpretations that identify high flood potentials, shrink–swell soils unsuitable for construction, and other important characteristics upon which development and insurance decisions depend. On the St Regis Mohawk reservation in New York, soil composition was used to identify homes that should be tested for radon levels. Twenty-five buildings on soil with a high correlation to radon emissions were targeted for testing; of these, four were identified as exceeding the standard.

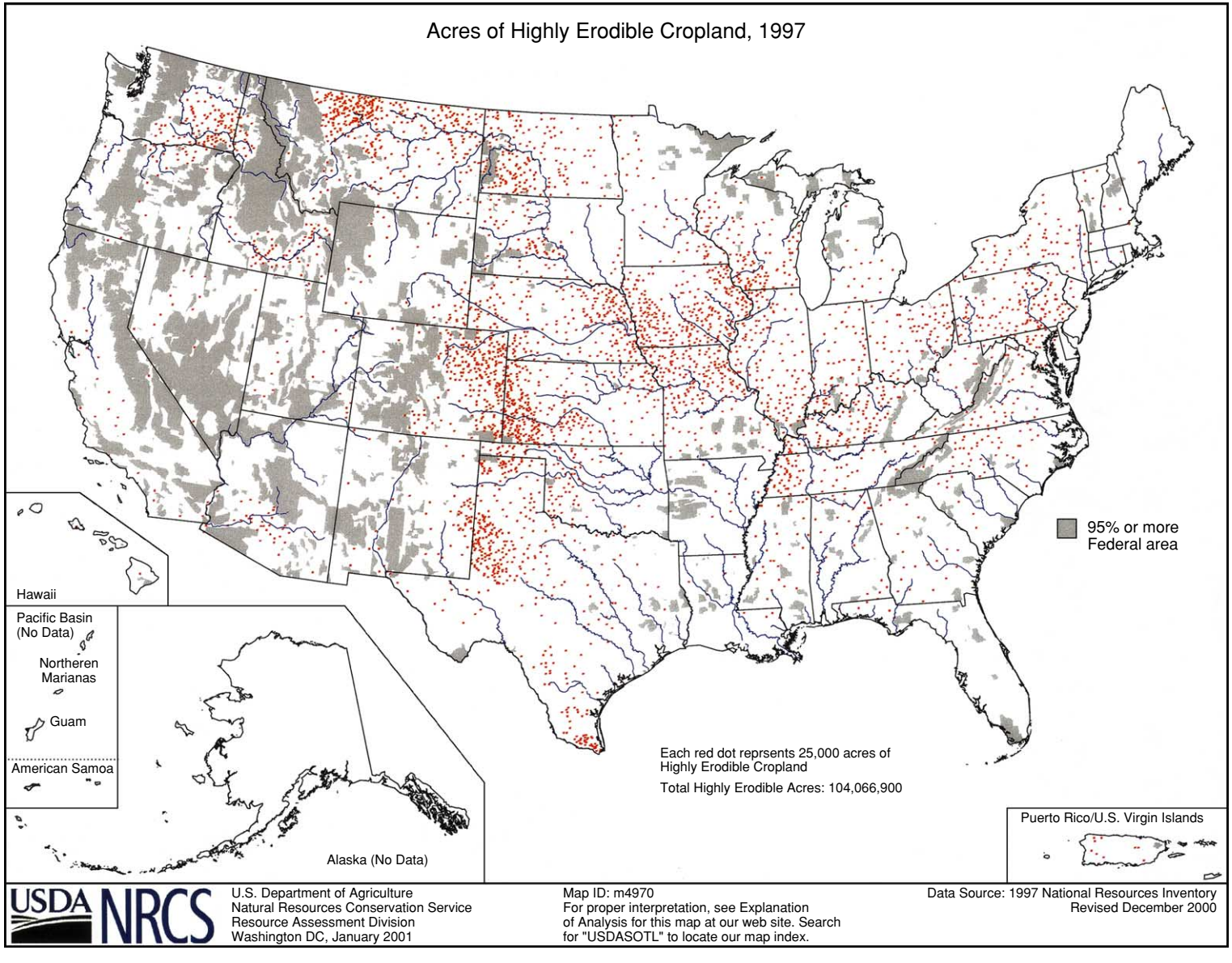


Figure 4 Acres of highly erodible cropland in the USA, 1997. A focus on the conservation needs of highly erodible land led to the origins of the Soil Conservation Service (SCS) in 1935 and a half-century later it was also a focal point of the 1985 Food Security Act. This map shows the total number and distribution of highly erodible acres as defined by the 1985 Act. This presentation is useful for national-scale evaluation of areas potentially needing conservation treatment, however, a site-specific evaluation is needed to establish the erodibility index for a specific land area. Courtesy of the Natural Resources Conservation Service USDA.

Engineering and Construction

Soils maps and interpretations are used widely in siting development such as highways, buildings, and other construction. Selecting sites with soils with sufficient load-bearing quality, not subject to flooding, and suitable for on-site sewage disposal result in substantial costs-savings, and reduce the potential for unforeseen failures. Soils data used by officials in one Illinois county helped decision-makers select a highway site with the fewest limitations, which resulted in savings of several thousand dollars per acre because of reduced excavation and construction costs.

Interpretations can also be used in estimating costs. The cost of laying buried pipelines, for example, is affected by a variety of soil factors. Soils interpretations can inform decisions on special material needs (e.g., pipe composition and rigidity), mitigation required (e.g., extremely wet or differentially draining soils creating needs for protective coatings or cathodic protection), and costs of excavation (e.g., short depth to bedrock or presence of hardpan can increase excavation costs).

Hazardous Waste, Brownfields, and Remediation

Evaluation for toxic substances, such as heavy metals, is an emerging area for soil surveys, particularly in urban or developed settings. The survey for LaTourette Park on Staten Island, New York, includes an evaluation of heavy metals in soils in addition to the more traditional interpretations for playgrounds and picnic areas. A soil survey completed before the Plattsburgh, New York, Air Force Base closure in 1995 is now a resource for redevelopment. The survey and interpretations are providing the information needed to undertake environmental-impact statements, planning, and remediation for reclamation of contaminated sites (e.g., dumps, fuel spills, etc.).

Research and Analyses

Soil survey data and interpretations are a substantial resource for research efforts. Interpretations have been vital to the development of improved management practices that optimize productivity. Interpretations are also used to identify soils with desired characteristics for specific experiments (e.g., soils of varying acidity) and to evaluate site suitability and limitations for agricultural research. This use of interpretations was first seen in 1911 when the state of Alabama requested a detailed survey and interpretation of its agricultural experiment stations. During World War II, interpretations helped to identify areas of high potential for producing goods in short supply. Interpretations identified areas suitable for the production of guayule as a substitute for rubber; oil-producing alternatives such as castor bean; and fiber crops such as American hemp.

Interpretations have proven useful in historic and archeological research. Soil maps and interpretations provide a resource to identify features associated with early settlements (e.g., access to rivers or tributaries, productive soils) as well as physiographic changes over time. Comparison of older surveys with current aerial photos can identify shifts that help to target potential sites. Some uses have included verifying property boundaries, which were commonly associated with stream channels. Early survey maps also included farmstead boundaries, rural cemeteries, and other settlement features (e.g., churches, towns, city buildings, etc.) – now of historic interest.

The array of analytical uses of soils data continues to expand. Detailed soil survey data are available through the Soil Survey Geographic (SSURGO) database to support GIS uses (mapping scales from 1:12 000 to 1:63 360). Field-mapping methods using national standards are used to construct SSURGO soil maps and SSURGO digitizing duplicates the original soil-survey maps. This level of mapping is designed for use by landowners, townships, and county natural-resource planning and management. State Soil Geographic (STATSGO) database maps present smaller-scale generalized soil-survey data designed for broad planning and management uses covering state, regional, and multistate areas (Figure 5) (mapping scales of 1:250 000, with the exception of Alaska, which is 1:1 000 000).

Soils data are an essential component of natural-resource modeling. For example, a soils data layer is a fundamental source of information for modelers simulating nonpoint-source pollution potentials from agricultural areas. Some productivity models, such as the Environmental Policy Integrated Cimate model (EPIC), integrate soil, climatic, economic, management, and other variables to simulate impacts of cropping systems on the environment and productivity. EPIC uses up to 11 variables to describe the soil and up to 20 variables to describe physical and chemical characteristics of each identifiable soil layer in the profile.

Today, soil survey data are being used to evaluate potential for carbon sequestration and other soil conditions that relate to global change. The role that soils play in mediating the effects of agriculture and forestry on the global atmospheric composition of greenhouse gases is a major component of the USDA Global Change research and development program.

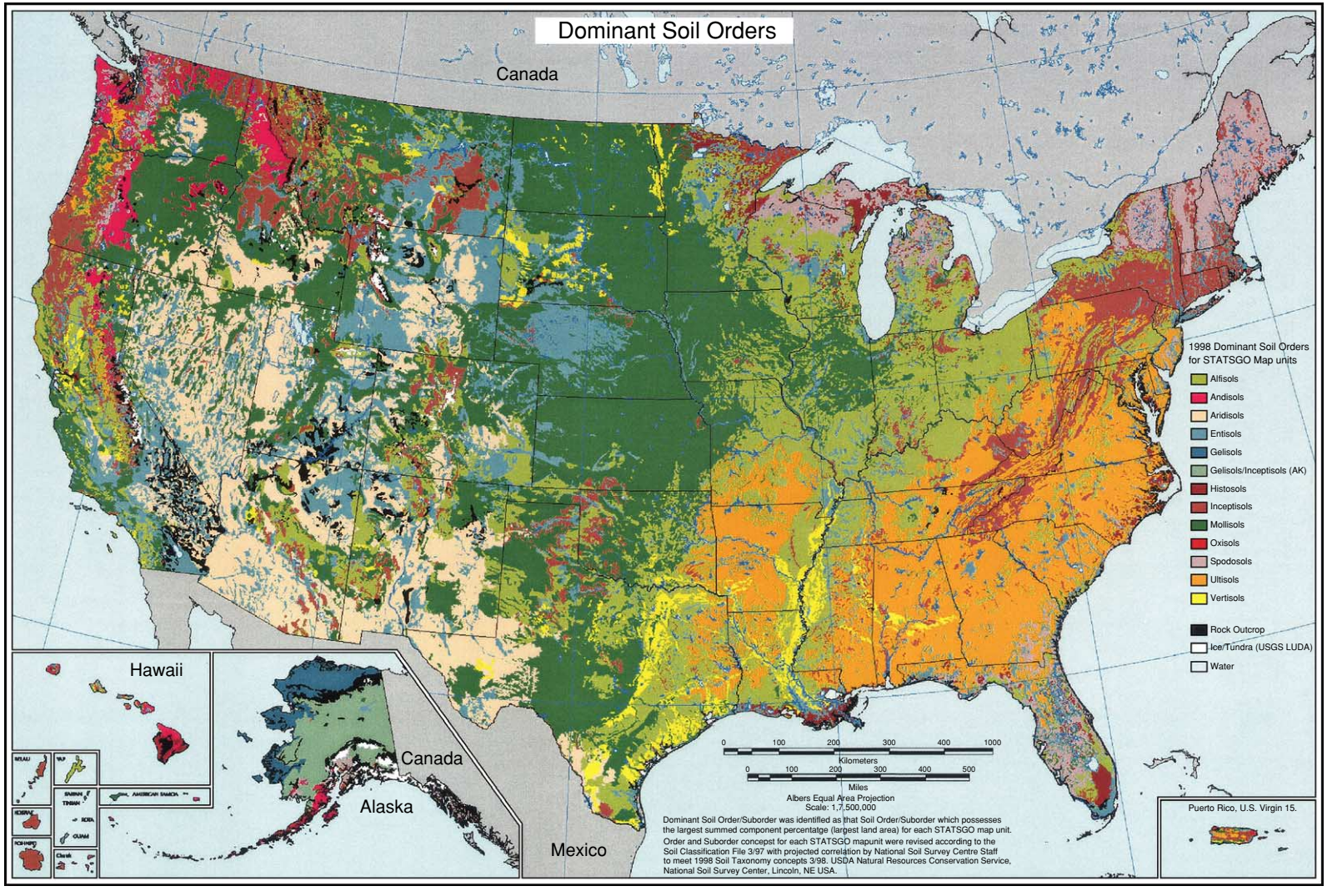


Figure 5 Dominant soil orders in the USA. This polygon map shows the dominant soil order present in each State Soil Geographic (STATSGO) database map unit. Dominant soil orders represent the largest land area for each map unit. This type of geographical information systems analysis allows users to compare the general distribution of soil orders. STATSGO-level mapping is designed to be used for broad planning and management uses covering regional, state, and multistate areas. Courtesy of the Natural Resources Conservation Service USDA.

Conclusion

Applications of soils data and survey interpretations have steadily broadened in range and precision in response to economic, environmental, social, and political influences. Today, interpretations are becoming increasingly dynamic with the application of information technology to support decision-making. As soils data are integrated into analytical and expert systems, access to digitized soils data is essential. Advances in digital orthophotography (digital imagery that has been rectified to remove distortions resulting from topography and the camera angle, thus equalizing distances represented on the image), maps and digital data for computer manipulation and retrieval enhance the delivery efficiency of soil-survey information. Detailed resource information on specific land areas can now be provided quickly and interactively to help landowners, communities, and others in land-use decision-making.

Further Reading

Bartelli LJ, Klingebiel AA, Baird JV, and Heddleson MR (eds) (1966) *Soil Surveys and Land Use Planning*. Madison, Wisconsin: Soil Science Society of America and American Society of Agronomy.

Gardner DR (1998) *The National Cooperative Soil Survey of the United States*. Historical Notes Number 7. Washington, DC: Natural Resources Conservation Service.

Helms D, Effland A, and Durana P (eds) (2001) *Profiles in the History of the U.S. Soil Survey*. Ames, IA: Iowa State Press.

Jenny H (1961) *E. W. Hilgard and The Birth Of Modern Soil Science*. Pisa, Italy: Collana Della Rivista 'Agrochimica.'

Kellogg CE (1943) *The Soils That Support Us*. New York: Macmillan.

Klingebiel AA (1991) Development of soil survey interpretations. *Soil Survey Horizons* 32(3): 53–65.

Olson GW (1981) *Soils and the Environment, A Guide to Soil Surveys and Their Applications*. New York: Chapman & Hall.

Simonson RW (1987) Historical aspects of soil survey and soil classification. *Soil Survey Horizons*. Madison, Wisconsin: Soil Science Society of America.

Smith GD (1983) Historical development of soil taxonomy – background. In: Wilding LP, Smeck NE, and Hall GF (eds) *Pedogenesis and Soil Taxonomy*, pp. 23–49. New York: Elsevier.

US Department of Agriculture, Natural Resources Conservation Service, Soil Survey Staff (1999) *Soil Taxonomy, A Basic System of Soil Classification for Making and Interpreting Soil Surveys*, 2nd edn. Agriculture Handbook No. 436. Washington, DC: US Department of Agriculture.

US Department of Agriculture, Natural Resources Conservation Service, Soil Survey Division Staff (1993) *Soil Survey Manual*. USDA Handbook No. 18. Washington, DC: US Department of Agriculture.

ARCHAEA

J E T McLain, USDA Agricultural Research Service, Tucson, AZ, USA

Published by Elsevier Ltd.

Introduction

In 1977, Carl Woese and his colleagues announced the discovery of 'a new form of life.' The immediate reaction of biologists was largely skeptical. Prior to this, the existing paradigm was that all organisms, except viruses, could be assigned to one of two primary groups, prokaryotes and eukaryotes. The rRNA work of Woese and others confirmed the existence of the archaeal domain, and currently recognized biotic diversity now consists of three groups, two of which are exclusively microbial (Archaea and Bacteria), while the third (Eukarya) contains both microbial life (as unicellular protists) and multicellular organisms.

Although they are metabolically diverse, a property common to the majority of archaeal organisms identified to date is the ability to exist in extreme habitats, including environments of high salt, high temperature, low pH, and acute anoxia. Three general ecological categories represent the overall patterns of archaeal adaptations to extreme environments: thermophilic (heat-loving), methanogenic (methane-producing), and halophilic (salt-loving). Most Archaea belong to at least one of these categories, and a number belong to two. This does not mean, however, that Archaea are limited to extreme environments. Archaea also thrive in freshwater sediments, temperate soils, and other less extreme conditions, confirming that microbes of this domain are ubiquitous.

The harsh environments in which many Archaea flourish have intrigued scientists interested in strategies of coping with life at the extremes. The recent

sequencing of the entire genomes of several Archaea have provided a wealth of knowledge, including the fact that some archaeal genes, including those encoding major metabolic pathway enzymes, are similar to those of Bacteria, while others, such as those for RNA polymerase subunits, are more similar to eukaryal genes, while still others appear to be Archaea-specific.

Archaeal Evolution

Phylogenetic relationships derived on the basis of 16S and 23S rRNA indicate that the domain Archaea consists of three major kingdoms, the Crenarchaeota, Euryarchaeota, and Korarchaeota (Figure 1). One of the two main branches of this phylogenetic tree contains, in large part, the thermophilic sulfur-dependent Archaea, while the other branch contains the methanogens, the extreme halophiles, and a few thermophilic organisms.

Current phylogenetic evidence deduced from comparison of 16S rRNA sequences suggests that Archaea have evolved more slowly than either Bacteria or the Eukaryotes. This is especially true of hyperthermophilic Archaea. It is not known why Archaea are the slowest-evolving of the three domains, but it may be related to their habitation of extreme environments. For example, organisms living in hyperthermal environments must maintain those genes that specify phenotypic characteristics critical to life at high temperatures, as evolutionary alteration of gene expression could impair organism survival.

Because thermophilic Archaea may have evolved very slowly, it has been proposed that these organisms are likely to have been among the earliest life forms on Earth. The phenotypic properties of thermophiles, including the ability to withstand high temperatures and the use of anaerobic chemoorganotrophic (the use of organic chemicals as electron donors) or chemolithotrophic (the use of organic chemicals as electron donors) metabolism, agree well with the phenotype of primitive organisms one would predict, given the geochemical conditions present on the Earth three billion years ago or more. If indeed life first occurred in the form of thermophiles in boiling hot springs deep on the ocean floor, this may explain the mystery surrounding the importance of phosphate in information storage and energy transfer in living cells. In other sea and freshwater environments, phosphate is present in very low concentrations and is often the limiting nutrient for organisms that live there, but water exiting hydrothermal systems often percolates through phosphate-rich minerals, assuring a rich phosphate supply to organisms inhabiting this environment. If thermophiles were present as an early life form, then life on Earth may have evolved in a phosphate-rich environment, thus developing a necessity for this mineral for cellular function.

The Archaeal Cell

Archaeal cells display a wide variety of morphological types. Some, similar to species of the bacterial domain, are strictly rod-shaped or spherical. Others

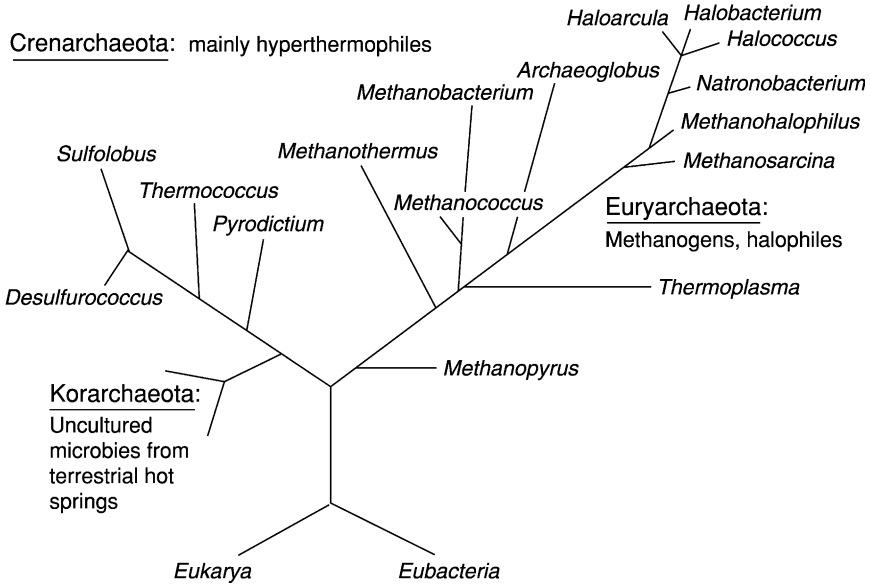


Figure 1 The phylogenetic tree of the Archaea. (Reproduced with permission from Howland JL (2000)) *The Surprising Archaea: Discovering Another Domain of Life*. New York: Oxford University Press.

Table 1 Some cellular features of the Archaea

Feature of cell	*Most similar to domain*
Morphology	Some unique
Cell wall with pseudopeptidoglycan	Resembles bacteria
Genome consisting of single circular piece of DNA	Bacteria
DNA polymerase, DNA helicase, DNA ligase	Bacteria
Protein trafficking for sugars and inorganic ions	Bacteria
Multisubunit RNA polymerase	Eukarya
Ether-linked lipids	Unique
Promoter site for transcription	Eukarya
Lack of nuclear membrane	Eukarya
Flagellar composition and assembly	Unique
Chaperonin heat-shock proteins	Bacteria

are disks, spirals, or filaments, or exhibit amebal irregularity, with variable protuberances. Still others have a mineral-like geometry, with shapes similar to cubes or triangles. Other irregular cells are bumpy spheres (e.g., *Thermococcus*) or are vaguely rod-shaped, but of highly variable diameter (e.g., *Pyrodictium*).

Archaeal cells possess many characteristics similar to those of the Bacteria and Eukarya, and others that are unique to the archaeal domain (Table 1). With one exception, all Archaea contain a cell wall that, like the cell wall of bacteria, functions to prevent osmotic lysis and to define cell shape. The ability of Archaea to adapt to extreme environments is assisted by the possession of unique cell wall types, which vary from those containing molecules composed of pseudopeptidoglycan, closely resembling bacterial peptidoglycan, to cell walls completely lacking a polysaccharide component. A common wall type is a paracrystalline surface layer (S-layer), consisting of protein or glycoprotein, generally of hexagonal symmetry. Although S-layers are common to all groups of Archaea, the biochemical makeup of S-layers among species is very diverse and, in some cases, this layer is too supple to contribute to the stability of the cell. In these cases, the function of the S-layer is unknown, but it has been proposed that the space between the cytoplasmic membrane and the outer surface of the S-layer may fulfill the role of a periplasmic space reminiscent of Gram-negative bacteria.

Archaeal RNA polymerases are complex, consisting of up to 14 subunits (compared to 4 in the bacterium *Escherichia coli*). Like Bacteria, Archaea possess a single RNA polymerase, but the archaeal RNA polymerase resembles those of eukaryotes in multisubunit complexity and sequence homology. In addition, archaeal RNA polymerases are unable to initiate transcription *in vitro*, a feature also seen in eukaryotes where general transcription factors are required for initiation.

Three Archaeal Groups

Extreme Halophiles

Extremely halophilic Archaea inhabit environments with salt concentrations high enough to kill most organisms. They occur in hypersaline bodies of water such as the Dead Sea or the Great Salt Lake, saline soils, and have also been isolated from dry deposits where salt is mined commercially. Halobacteria can also be found on salted fish, salted hides, bacon, and sausage, and these microorganisms can often be attributed to the spoilage of these foods.

The formation of saline waters throughout the world can result from the geologic separation of seawater from the open ocean and concentration of salts by evaporation, or may result from the dissolution of salts from rocks into bodies of water. The ionic composition of hypersaline water bodies varies widely, from high Na^+ ($\sim 100\,g\,l^{-1}$) and Cl^- ($\sim 200\,g\,l^{-1}$) in the Great Salt Lake, to high Cl^- ($>200\,g\,l^{-1}$) in the Dead Sea, to high SO_4^{2-} ($>200\,g\,l^{-1}$) in Hot Lake in Washington state, USA. The varied chemical and physical properties of saline habitats result in colonization by a widely diverse group of archaeal prokaryotes, which possess some common features that allow them to thrive in these high-solute environments.

Research has shown that extreme halophiles not only tolerate salty conditions, but they also require high concentrations of salt. A generally accepted definition of an extreme halophile is that the organism requires at least $1.5\,mol\,l^{-1}$ NaCl for growth, but most species require 2–4 $mol\,l^{-1}$ for optimal growth. Virtually all extreme halophiles can grow at the limit of saturation for NaCl ($5.5\,mol\,l^{-1}$), although some species grow only very slowly at this salinity. The plasma membrane of halophiles does not exclude salts, but can select for certain ions so that internal concentrations of ions can be controlled and cellular functioning can be maintained. Cells of *Halobacterium*, for example, pump large amounts of potassium from the environment into the cell such that the concentration of K^+ inside the cell is higher than the Na^+ concentration outside. Thus, the total ionic strength remains the same on both sides of the plasma membrane but potassium, required for many cellular functions, is the prevailing cation on the inside. In this manner, *Halobacterium* employs an inorganic ion as its compatible solute and remains in positive water balance

The cell wall of *Halobacterium* is stabilized by sodium ions and in low-Na^+ environments the

cell wall breaks down, resulting in cell lysis. The halobacterial cell wall is composed of a glycoprotein with an exceptionally high content of the acidic amino acids aspartate and glutamate, and the sodium ions shield the negative charges contributed by the carboxyl groups of these amino acids. When sodium is diluted, the negatively charged parts of the proteins actively repel each other, leading to cell lysis. Thus, the cellular components of halophiles exposed to the external environment require high sodium for stability, whereas internal components require high potassium. This requirement for specific cations in such high amounts is a feature unique to halophilic bacteria. Another unique quality of Halobacteria are cytoplasmic proteins with very low levels of hydrophobic amino acids, perhaps representing an evolutionary adaptation to the highly ionic cytoplasm of extreme halophiles. In environments of high ionic strength, polar proteins would tend to remain in solution, whereas nonpolar molecules would tend to cluster and perhaps lose activity.

Some extreme halophiles possess a unique light-mediated synthesis of adenosine triphosphate (ATP) that does not involve chlorophyll pigments. When light is available and their ability to obtain energy through respiration is compromised, *Halobacterium salinarum* and certain other extreme halophiles synthesize and insert a protein, bacteriorhodopsin, into their membranes. Conjugated to this is a molecule of retinal, a carotenoid-like molecule that can absorb light and catalyze the transfer of protons across the cytoplasmic membrane. Light-mediated ATP production in *H. salinarum* has been shown to support slow growth of this organism anaerobically in the absence of organic energy sources and under conditions in which other energy-generating reactions do not occur.

Methanogens

Two features common to all methanogenic Archaea are the inability to tolerate oxygen or reactive oxygen species and the ability to produce methane gas. Methane production in soils characteristically occurs under anaerobic, highly reducing conditions in the absence of nitrate, sulfate, or ferric iron, including the mud of swamps and marshes, the beds of fresh and marine bodies of water, and mud originating from sewage plants and rice paddies. In some instances, methanogens live in small anoxic pockets in soils in an otherwise oxygen-rich area. Such regions are often formed by the action of microorganisms that locally consume all of the available oxygen, and several studies have confirmed that methanogens can be present throughout a macroscopically oxic soil and that methane-producing activity can respond rapidly to the establishment of appropriate anoxic conditions.

Methanogens also flourish in the digestive systems of ruminants. Production of methane gas worldwide from cattle, goats, sheep, and camels is greater than methane production from paddy fields and swamps. The rumen utilizes a mixed population of microorganisms, including Eubacteria, methanogenic Archaea, and anaerobic protozoa to carry out the digestion of cellulose and other polymeric sugars. Methanogens are responsible for regulating the overall fermentation in the rumen by removing hydrogen gas during methane production. This action keeps the hydrogen concentration in the rumen low, encouraging the activity of hydrogen-producing species and altering their metabolism towards higher-yielding pathways. These pathways result in the synthesis of more microbial cells, increasing the available protein to the ruminant.

Methane is the final product of a complex community of organisms breaking organic materials down into the simple methanogenic substrates and thus, methanogens complete the last step in the anaerobic decomposition of organic matter. Because they are dependent on other organisms for provision of substrate and the establishment of reducing conditions, methanogens can only function as members of microbial communities. Methanogenesis is frequently rate-limited by the activities of the other members of the community, and particularly by how rapidly hydrogen or acetate is made available. At least 10 substrates can be converted to methane by methanogens (Table 2) and three main classes of metabolic reactions can be used to create energy for ATP synthesis. The first class utilizes CO_2-type substrates (Table 2, lines 1–5), while the second class of reaction involves reduction of the methyl group of methyl-containing compounds to methane (Table 2, lines

Table 2 Reactions for methanogenesis by methanogenic bacteria

Substrates		*Products*
$4H_2 + CO_2$	→	$CH_4 + 2H_2O$
$4H_2 + HCO_3^- + H^+$	→	$CH_4 + 3H_2O$
4 Formate $+ 4H^+$	→	$CH_4 + 3CO_2 + 2H_2O$
4 (2-Proponal) $+ CO_2$	→	$CH_4 +$ 4 acetone $+ 2H_2O$
2 Ethanol $+ CO_2$	→	$CH_4 +$ 2 acetate $+ 2H^+$
4 Methanol	→	$3CH_4 + CO_2 + 2H_2O$
4 Methanol	→	$3CH_4 + HCO_3^- + H^+ + H_2O$
4 Methylamine $+ 2H_2O$	→	$3CH_4 + CO_2 + 4NH_4^+$
Methanol $+ H_2$	→	$CH_4 + H_2O$
Acetate $+ H^+$	→	$CH_4 + CO_2$
Acetate $+ H_2O$	→	$CH_4 + HCO_3^-$

6–9). The third class of methanogenic reactions is acetotrophic, the cleavage of acetate to CH_4 plus CO_2 (Table 2, lines 10–11). The conversion of acetate to methane appears to be a very significant ecological process, especially in sewage digesters and in freshwater anoxic environments where competition for acetate between sulfate-reducing bacteria and methanogenic bacteria is not extensive. The reduction of CO_2 to CH_4 is generally H_2-dependent, but formate, carbon monoxide, and even elemental iron can serve as electron donors for methanogenesis. A few methanogens can utilize simple organic compounds as electron supplies for CO_2 reduction.

Methanogenesis employs a unique collection of cofactors in its reactions. Methanogens often contain high concentrations of these compounds, whose presence can be used to determine the prevalence of methanogenic Archaea. One cofactor commonly used to identify methanogens is coenzyme 5-deazaflavin F420, an electron carrier involved in several reactions in the methanogenic pathway which also has a role in DNA photorepair. Another cofactor, coenzyme F430, contains nickel, which all methanogens require for growth. A third cofactor unique to methanogenesis, coenzyme M, acts as a methyl-carrying coenzyme in the last step of the methanogenic pathway and is thus involved in the final reduction of a methyl group to form methane.

Hyperthermophiles

The archaeal kingdom Crenarchaeota consists of a great diversity of organisms. Some form a stable part of the soil microbial community in boreal environments, while others are defined by their extremely thermophilic nature. This latter group, the hyperthermophiles, contains organisms that are the most heat-loving of all known prokaryotes. Several hyperthermophiles are capable of growth at temperatures above the normal boiling point of water, and all have temperature optima above 80°C.

Many hyperthermophilic Archaea have been isolated from geothermally heated soils or waters containing elemental sulfur and sulfides, and most hyperthermophilic species metabolize sulfur in some way. In fact, the first hyperthermophile discovered, *Sulfolobus*, grows in sulfur-rich hot acid springs. Sulfur-rich environments are found throughout the world, and extensive studies of hyperthermophilic Archaea have been made in Yellowstone National Park (USA), where the highest concentration of sulfur-rich thermal features in the world has been attracting researchers since the first scientific study of the region in the late nineteenth century.

Hyperthermophilic Archaea have also been found in artificial thermal habitats such as the boiling outflow of geothermal power plants. In addition, a phylogenetically distinct set of hyperthermophilic Archaea has been isolated from submarine volcanic habitats, where the pressure of even a few meters of seawater can raise the boiling point of water sufficiently to select for organisms capable of growth above 100°C. *Pyrolobus fumarii*, a submarine organism, has a growth optimum at 106°C but can grow at 113°C, and can even survive autoclaving at 121°C for 1 h.

Depending on the surrounding geology, geothermally heated environments may be either slightly alkaline to mildly acidic (pH 5–8) or extremely acidic, with pH values below 1 not uncommon. Such extreme acidity does not deter the hyperthermophiles: *Picrophilus oshimae* has a pH optimum for growth of 0.7 and can grow significantly at pH values approaching zero; however, the majority of thermophilic Archaea inhabit neutral or mildly acidic habitats.

With a few exceptions, hyperthermophiles are obligate anaerobes. Their sulfur requirement is based on the need for an electron acceptor to carry out anaerobic respiration or an electron donor for chemolithotrophic metabolism. Organisms of the genera *Thermococcus* and *Thermoproteus* oxidize a variety of organic compounds (e.g., small peptides, glucose, starch) anaerobically in the presence of S^0 as an electron acceptor. *Sulfolobus* is an obligate aerobe capable of oxidizing organic compounds, H_2S, or S^0 to H_2SO_4 and fixing CO_2 as a carbon source. Many hyperthermophilic Archaea can grow chemolithotrophically with H_2 as an energy source. *Acidianus*, a facultative aerobe resembling *Sulfolobus*, grows anaerobically using S^0 as an electron acceptor and H_2 as an electron donor, forming H_2S as the reduced product, and *Pyrodictium* can be cultured under strict anaerobic conditions in a mineral-salts medium supplemented with H_2 and S^0 at temperatures up to 110°C.

A unique property shared by many hyperthermophiles such as *Sulfolobus* and *Acidianus* is an unusually low guanine-cytosine (GC) base ratio. The DNA of *Sulfolobus* is ~38% GC, whereas that of *Acidianus* is ~31%. These low GC base ratios are intriguing: in a testtube, DNA of 30–40% GC content would melt almost instantly at 90°C. Research into how DNA of these organisms is prevented from melting is ongoing; however, it is hypothesized that the DNA may be protected, in part, by high cytoplasmic solute concentrations (the melting temperature of DNA increases as the solute concentration increases). Cytoplasmic concentrations of solutes such as cyclic 2,3-diphosphoglycerate in cells of *Methanopyrus fervidus* and other thermophilic methanogens are

closely correlated with the temperature at which the organisms are grown in laboratory incubations. In addition, it is thought that specific DNA binding proteins of hyperthermophiles somehow prevent DNA from melting, perhaps by folding the DNA into a conformation consistent with thermal stability. *Pyrodictium* cells grown at 110°C produce 80% of their protein biomass as a single protein that functions as a molecular chaperonin, stabilizing other cellular proteins by refolding them as they begin to denature near the upper temperature limits of growth. At 100°C (near the optimum for growth of *Pyrodictium*), very little of this chaperonin protein is made, suggesting that only at very extreme temperatures do the otherwise thermally stable proteins of this organism begin to denature. Additional protein adaptations of hyperthermophiles that allow these organisms to withstand extreme temperatures include sequence modifications, addition of salt bridges, increased hydrophobic interactions, additional ion pairing and hydrogen bonding, improved core packing, and shortening of loops. These strategies, used to differing extents by different thermophilic proteins, not only confer higher thermal stability but also enhance rigidity and resistance to chemical denaturation.

Within the last several years, molecular techniques have uncovered a unique lineage of the kingdom Crenarchaeota that is phylogenetically distinct from the hyperthermophiles. Nonthermophilic Crenarchaeota, which have been identified in marine picoplankton, freshwater sediments, soils, and in continental shelf anoxic sediments, have been shown to account for as much as 2% of microbial rRNA in soils analyzed. Phylogenetic analyses suggest that the nonthermophilic Crenarchaeota may have a common ancestor with the hyperthermophiles, but as yet the ecological significance of the nonthermophiles remains unknown.

Psychrophiles

Archaea have also been detected in ecosystems with characteristics in direct contrast to hyperthermophilic environments. Psychrophilic (cold-loving) Archaea account for over a third of the prokaryotic biomass in coastal Antarctic surface waters, and the hypersaline lakes of the Vestfold Hills lake system in Eastern Antarctica have been the subjects of a number of studies on microbial distribution. One of these lakes is Deep Lake, with a salinity of $320\,g\,l^{-1}$ and temperatures between -14 and -18°C. The biodiversity of Deep Lake is low, and is dominated by Archaea of the family Halobacteriaceae. To date, only three psychrophilic archaeal strains, all free-living and members of the subdomain Eukaryarchaeota, are available in pure culture. *Methanococcoides burtonii* (T_{min} -2.5°C) and *Methanogenium frigidum* (T_{min} -10°C) were originally isolated from the bottom of Ace Lake, Antarctica, where the methane-saturated waters remain between 1 and 2°C.

Thermoplasma

Thermoplasma acidophilum is a prokaryote that does not possess a cell wall and in this respect resembles the mycoplasma. Phylogenetically, however, *Thermoplasma* is a member of the Archaea. *Thermoplasma* is an acidophilic, aerobic, thermophilic chemoorganotroph, and with one exception, all strains of *Thermoplasma* have been obtained from self-heating coal refuse piles, which contain coal fragments, pyrite, and other organic materials extracted from coal. When this refuse is dumped into piles in coal-mining operations, it tends to self-heat by spontaneous combustion and creates conditions conducive to growth of *Thermoplasma*, which apparently metabolizes leached organic compounds. A chemically unique cell membrane allows *Thermoplasma* to survive the osmotic stresses of life without a cell wall and to withstand the dual environmental extremes of low pH and high temperature. This membrane contains a unique lipopolysaccharide that, together with other molecules, renders the *Thermoplasma* stable to hot acidic conditions.

Biotechnological Use of Extremoenzymes

Biotechnologically useful enzymes represent the main focus of industrial interest in the Archaea, as a result of the abilities of these microbes to function at the temperature, salinity, and pH limits of life. Heat-tolerant enzymes are currently the most investigated of all extremoenzymes because performing biotechnologically related processes at higher temperatures is often advantageous for many reasons. In chemical reactions involving organic solvents, decreased viscosity and increased diffusion at elevated temperatures result in higher reaction rates. In addition, performing reactions at higher temperatures reduces the possibility of complications resulting from contamination. One thermophilic compound of particular interest is DNA polymerase, an enzyme that is responsible for the elongation of the primer strand of a growing DNA molecule and is thus central to the polymerase chain reaction for DNA amplification. DNA polymerases from various hyperthermoarchaea (including *Pwo* from *Pyrococcus* woesei and *Pfu* from *P. furiosus*) are showing biotechnological promise, based on

their stringent proofreading abilities and suitability for the amplification of longer DNA fragments. These hyperthermophilic DNA polymerases possess error rates that are five- to 10-fold lower than that of the widely used thermobacterial *Taq* polymerase from *Thermus aquaticus*.

Applied uses exist or have been proposed for a variety of other archaeally derived materials. The extremely stable lipids of archaeal membranes may represent a novel drug delivery system because of their enhanced stability under temperature extremes. Archaeal components such as the S-layer glycoprotein have drawn interest for their use as possible vaccine carriers and other nanotechnological potentials, and it has been shown that much higher immune responses in mice are shown to protein antigens encapsulated in archaeosomes than in conventional liposomes. A thermostable ligase for the ligase chain reaction (an amplification method that involves the ligation of two sets of adjacent oligonucleotides) would be of obvious benefit because the ligation must be carried out near the melting temperature of the DNA, and the ligase enzymes must be stable during the dissociation step that follows. Currently, a ligase from *T. aquaticus* is used, but a more stable equivalent may be available from hyperthermophilic Archaea. Haloarchaeal polymers have been considered as a raw material for biodegradable plastics. Hydrolases from hyperthermophiles could be used in the food-processing industry to hydrolyze fats at high temperatures, reducing bacterial contamination problems. Addition of polymer-hydrolyzing extremoenzymes such as beta-glycanases from psychrophiles to detergents would allow for efficient washing in cold water. The food industry could exploit pectinases that act at lower temperatures in the processing of fruit juices or cheeses.

Often, the mere presence of archaeal communities carries considerable potential economic value. Methanogenic Archaea have proven to be quite valuable in their capacity as clean and inexpensive energy sources, and acidophilic Archaea have been identified at several acid mine drainage sites where their mineral-sulfide oxidizing abilities play an important role in the geochemical sulfur cycle.

The days of Archaea being considered as just 'odd bacteria' adapted to living in extreme environments appear to be over. In the past few years, information on the isolation, characterization, description, and applications of Archaea has mushroomed. Researchers will continue to study the adaptations that allow Archaea to grow at the extremes and to search for new species that will extend the boundaries that limit life of Earth. There is no doubt that many novel features remain to be discovered about these microorganisms, and their continued study will have a major impact on science in the decades to come.

See also: **Bacteria:** Soil

Further Reading

Bernander R (1998) Archaea and the cell cycle. *Molecular Microbiology* 29: 955–961.

Eichler J (2001) Biotechnological uses of archaeal extremozymes. *Biotechnology Advances* 19: 261–278.

Howland JL (2000) *The Surprising Archaea: Discovering Another Domain of Life*. New York: Oxford University Press.

Jarrell KF, Bayley DP, Correia JD, and Thomas NA (1999) Recent excitement about the Archaea (Archaebacteria). *BioScience* 49: 530–542.

Kates M, Kushner DJ, and Matheson AT (eds) (1993) *The Biochemistry of Archaea (Archaebacteria)*. New Comprehensive Biochemistry Series, vol. 26. Amsterdam, Netherlands: Elsevier Science.

Madigan MM, Martinko J, and Parker J (2000) *Brock Biology of Microorganisms*. Upper Saddle River, NJ: Prentice-Hall.

Oren A (2002) Molecular ecology of extremely halophilic Archaea and Bacteria. *FEMS Microbiology Ecology* 39: 1–7.

Reysenbach A-L, Voytek M, and Mancinelli R (eds) (2001) *Thermophiles: Biodiversity, Ecology, and Evolution*. New York: Kluwer Academic Plenum.

Vreeland RH and Hochstein LI (eds) (1993) *The Biology of Halophilic Bacteria*. Boca Raton, FL: CRC Press.

Woese CR and Wolfe RS (eds) (1985) *The Bacteria: A Treatise on Structure and Function*, vol. VIII. *Archaebacteria*. Orlando, FL: Academic Press.

ARCHEOLOGY IN RELATION TO SOILS

J A Homburg, Statistical Research Inc., Tucson, AZ, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Archeology, soil science, and geology are intimately related disciplines because they are all historical sciences. Records of each of these scientific disciplines take time to develop, and indicators of each are encoded in the sedimentary and soil matrix in which the archeological record is embedded. Physical and chemical soil properties are commonly used to make interpretations of the archeological record, including how the soil was formed, how it has been altered, and how it is preserved. Soils are also used to draw archeological inferences about human behavior in the past. Examples include identifying chemical signatures of different human activities that may leave no physical traces, assessing the long-term effects of cultivation on soil productivity, and reconstructing how earthen material was used to make pottery and build many types of cultural features (e.g., adobe houses, earth ovens, clay-lined fire hearths, roasting pits, postmolds, etc.). An understanding of soil variability within a site is important in deciding where archeological test pits should be placed and how deep they should be excavated.

As an interdisciplinary science, archeology is unrivaled in terms of the number of fields on which it draws, except perhaps by forensic science. Today, archeological projects often include one or more soil scientists who work with experts in geomorphology, geochemistry, ethnobotany, pollen, phytoliths, paleontology, paleoclimatology, dating methods (e.g., radiocarbon, tree-ring, archeomagnetic, optical luminescence, and obsidian hydration dating), and many other disciplines. Soil science has major applications in archeological research. Archeology can also contribute significantly to pedology, especially with the use of cross-dating techniques, whereby temporally diagnostic artifacts such as pottery and projectile points from archeological deposits are used to determine the age of soil horizon development.

Pedostratigraphy and the Archeological Record

A basic understanding of soil science, especially pedology (soil morphology and pedogenesis) and soil chemistry, is essential for making meaningful interpretations of archeological context and site-formation processes that account for the contemporary archeological record. This record, as expressed in surficial, buried, and stratified cultural deposits, is an imperfect and biased record of past human activities. Many postoccupational disturbance processes (e.g., animal burrowing, displacement by root growth, and infilling of soil cracks caused by freeze–thaw and wetting–drying cycles) modify the archeological record by translocating and mixing artifacts. The degree of mixing can be assessed by observing and documenting soil profiles, focusing on the extent of krotovinas (that is, infilled animal burrows) and the level of soil horizonation. Preservation of materials in the archeological record is also highly variable. Durable materials such as stone and ceramic artifacts are usually preserved for very long periods, but perishable items made of wood and leather are rarely preserved except in dry caves and anaerobic wetland soils.

Calcareous materials such as bone and shell may or may not be preserved in the archeological record, depending on the climate, soil pH, and how long they were exposed to weathering before being buried. Bone, which consists mainly of the calcium phosphate mineral hydroxyapatite, is best preserved at pH 7.88 and it becomes increasingly soluble above and below this level. It is well known that bone is poorly preserved in acidic soils, but it is also poorly preserved in moderately to very strongly alkaline soils. It is interesting to note that human bone at the Port Hudson Confederate Cemetery in Louisiana was almost completely decomposed in acidic soils in approximately 120 years, with a soil pH of approximately 5.5. Calcareous remains such as shellfish in coastal shell middens have been added in such large quantities in some cases that soil pH has altered sufficiently to preserve bone.

Site-formation processes, including soil formation, if properly interpreted, are useful for helping to explain the archeological record. Archeological context refers to the relationship between cultural and natural deposits, as well as to the relationship between artifacts and cultural features within archeological sites. An understanding of soils and how they develop is crucial for explaining why archeological deposits are located or concentrated in particular places and for assessing the stratigraphic integrity of cultural deposits. This is because the interaction of many soil-forming factors, including climate, topography, vegetation, and parent material, were also important considerations to humans in deciding where

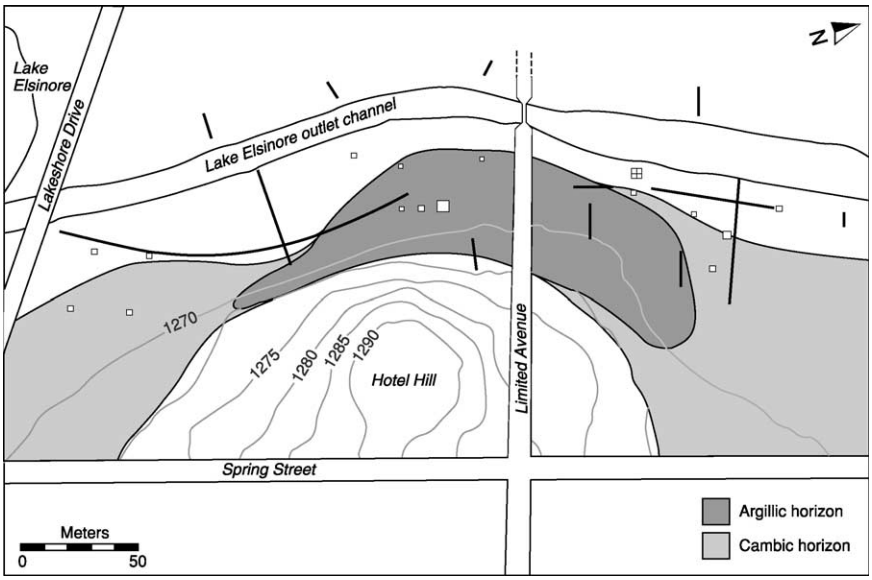


Figure 1 Map of the distribution of cambic and argillic horizons at the Elsinore Site. (Adapted from Grenda DR (1997) *Continuity and Change: 8500 Years of Lacustrine Adaptation on the Shores of Lake Elsinore*, Fig. 26. Technical Series No. 59. Tucson, AZ: Statistical Research/The University of Arizona Press, with permission.)

habitations and other site types were established. Soil maps are commonly used by archeologists in developing models to predict where sites of particular ages and functions are located, and how land-use patterns changed through time.

Soil development is helpful in assessing geomorphic stability, or the extent to which landforms have been preserved, buried, or modified. Soils showing the most development are usually the oldest in a given study area, and these soils are places where older archeological deposits may be preserved. Artifacts are commonly concentrated in well-developed A horizons, because these are surface horizons that served as occupation surfaces where artifacts and refuse were discarded and incorporated into the soil. Older archeological deposits commonly overlie or are found within diagnostic subsurface horizons such as argillic and calcic horizons. An example of this relationship has been evidenced from backhoe trenches and testpits at the Elsinore site, an 8500-year-old site located near the outlet channel of Lake Elsinore in southern California. Excavations have revealed that the richest and oldest cultural deposits are in and above sediments where an argillic horizon has developed. The youngest cultural remains, by contrast, are on the fringe of the site, where a weakly developed soil horizon, a cambic horizon, has formed (Figure 1).

Soil properties are useful for identifying places where past human activities have been concentrated and explaining why this is so. At the Admiralty site, a coastal shell midden in the Ballona Wetlands of Los Angeles, an E horizon has been identified in

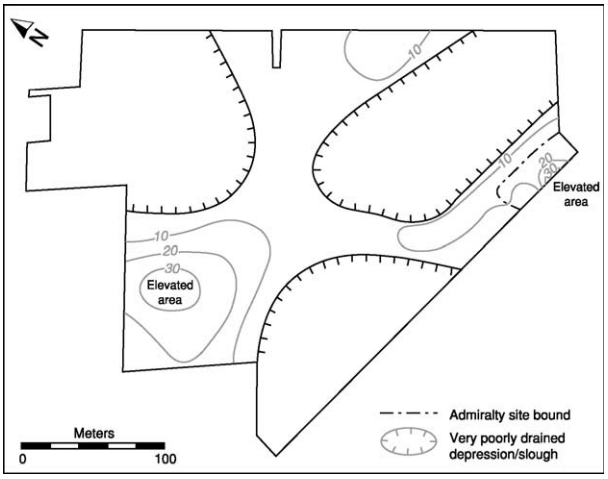


Figure 2 Map of the Channel Gateway project area in Los Angeles, showing E horizon thickness, depression, and the Admiralty Way site. (Adapted from Altschul JH, Homburg JA, and Ciolek-Torrello RS (1992) *Life in the Ballona: Archaeological Investigations at the Admiralty Site and the Channel Gateway Site*, Fig. 71. Technical Series No. 33. Tucson, AZ: Statistical Research/The University of Arizona Press, with permission.)

subsurface tests. Soils in this wetland setting, although seasonally moist or saturated, are dry enough most of the year for organic matter, clay, and bases to be translocated downward to form an eluvial horizon where these materials have been depleted. Thickness of a grayish eluvial horizon, the E horizon, marks the best-drained areas on and adjacent to the site (Figure 2). On- and off-site subsurface tests show that the site location coincides with an area with

the thickest E horizons. This use of soil morphology has helped to explain why the site is located where it is in the wetlands.

Soil morphology has also been used to study the Louisiana State University Campus Mounds, a 5000-year-old earthen-mound site built by Native Americans during a time that archeologists refer to as the Middle Archaic Period. Three soil cores were collected from each of the two 5-m-tall mounds, to determine their age and construction history. Three radiocarbon samples of humates in the A horizon buried under one of the mounds and in the lower mound fill have yielded dates of approximately 5000 years BP, which makes this mound site one of the earliest in North America. Although similar in appearance from the outside, the soil stratigraphy varies greatly between the two mounds, which suggests different construction histories (Figure 3). Many thin, dark-colored bands of A horizon material have been documented in Mound A. If these A horizons have formed *in situ*, it implies different construction stages, with enough time between stages for natural A horizons to form as organic matter was added to the soil from decomposing vegetation. The boundaries between the dark-brown bands are abrupt to very abrupt in Mound A, which contrasts sharply with the clear-to-gradual horizon boundaries in the off-mound natural soils. Consequently, the bands have been interpreted as basket-loaded A horizon material that has been placed there during mound construction, not *in situ* A horizons marking construction stages. Brief interruptions during the construction process have been identified, however, one of the indicators being a reddish oxidized zone that apparently marks a fire hearth, which may have served a ritual purpose during mound construction, and a laminated zone, resulting from slope-wash deposition during a storm. The fill of Mound B was quite different, consisting mainly of dark yellowish-brown subsurface soil derived from the argillic and fragic soil (a Btx horizon) from off-mound areas adjacent to the mounds. Thin grayish bands are also found in Mound B, and these are interpreted as a mix of material from the E horizons that overlie the Btx horizon off-mound and from infilled vertical cracks (or interprisms) of the fragipan. Soil chemistry supports the interpretation that Mound A was built mainly from near-surface soil and Mound B from subsurface material. The Mound A fill has a significantly higher pH, organic matter content, and phosphorus and calcium concentrations. Interestingly, a fragipan, a loamy soil horizon with a high bulk density that can restrict root and water penetration, has been found on the lower slopes of both mounds. The fragipan has formed *in situ*, thus indicating that this diagnostic subsurface horizon can develop within 5000 years. This is a good example of how archeology can help ascertain how long certain soil features take to develop.

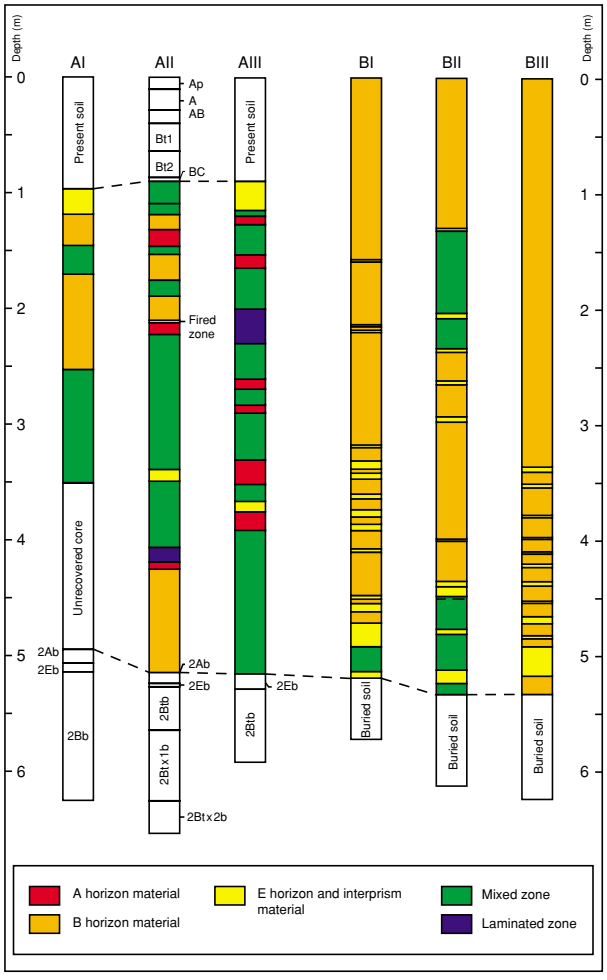


Figure 3 Stratigraphy of cores at the Campus Mounds of Louisiana State University. (Adapted from Homburg JA (1988) Archaeological investigations at the LSU Campus Mounds. *Louisiana Archaeology* 15: 31–204, Fig. 18, with permission.)

Archeological Soil Chemistry

Soil chemistry is being utilized increasingly in archeological interpretations. It is sometimes used for paleoenvironmental reconstruction, such as the use of carbon isotopes to reconstruct vegetation, such as prairie versus forest vegetation, in places where the boundary has shifted. More commonly, however, soil chemistry is used to reconstruct past human activities, identify activity areas within sites, and define site boundaries and stratigraphic relationships. Detecting traces of past human activities depends largely on the degree and kind of impact that humans had on soils. Two types of surface horizons (or epipedons) that are clearly the result of human activities are anthropic

horizons, which are characterized by elevated levels of phosphorus that result from fertilizer additions, and plaggens, which result from many years of manure additions. Other signatures of human activity are much more subtle and subject to interpretive difficulties. This is because humans impact soils in such a wide variety of ways and with differing intensities, and at all scales of time and space.

Elements commonly enriched in soil due to human activity include carbon, nitrogen, phosphorus, and calcium, and, to a lesser degree, potassium, magnesium, sulfur, copper, and zinc. The most common addition to soil in preindustrial societies that is often easy to recognize today is organic matter from plant and animal residues composed mainly of carbon, nitrogen, phosphorus, sulfur, and humus. This type of human effect typically results in more organic matter in the surface soil and a thickened A horizon that has a lower chroma value than unmodified soils; for example, the dark, fertile soils of the Amazon Basin, known as Terra Preta. Differentiating organic matter that is anthropogenic from natural soils often relies on other traces of human activity such as artifacts, remains of plant foods (e.g., pollen, phytoliths, charcoal, and ash), and chemicals such as phosphates and fatty acids. Traces of these materials can be identified in thin sections using soil micromorphology.

Because of its high stability and immobility in many soils, phosphorus has served as a key element for identifying areas of human activity. Phosphorus is concentrated in surface horizons owing to many kinds of human activity, especially disposal of rubbish and waste products from humans and livestock. Similarly, aluminum, calcium, manganese, and iron levels may be elevated in anthropogenic soils, because they are often bound in phosphate compounds. Calcium and calcium carbonate levels have been used to identify places where animal bone and limestone were processed. Potassium is enriched where animal remains decompose and where wood ash is added to the soil as a result of burning vegetation. Nitrogen, which is abundant in animal and human waste, may also become elevated in soils, but, because it rapidly decomposes, traces of this element rarely persist for long periods in the archeological record. Various heavy metals have also been useful for archeological studies, especially in reconstructing activities associated with metal-working and processing in cultures that used iron, copper, and other metals.

Since the 1990s, researchers have used a wide range of analytical techniques to characterize the elemental chemistry of anthropogenic soils relative to natural, unaltered soil. Two of the most sophisticated methods include inductively coupled plasma–atomic emission spectroscopy (ICP/AES) and inductively coupled plasma–mass spectroscopy (ICP/MS). Soil samples are typically collected using grid- or transect-sampling methods to cover both archeological sites and off-site areas. In some cases interpretations of ancient activity areas have been improved by studying activity areas in modern or recently abandoned preindustrial societies. This research is a type of ethnoarcheology, which involves studying existing cultures and how they use space for different activities to bolster interpretations of the archeological past. Examples of this research include studies of modern farming villages in the Mayan area and of pastoral societies in Africa such as the Masai. It is important to note that research on human impacts on soils in archeological sites is relevant to interpretations of impacts on soils today. The archeological record provides a long-term perspective on human impacts that is available from no other source.

Soil as a Resource

Soil has always been a basic resource for sustaining human populations. This section focuses on the agricultural use of soils, including archeological studies of the long-term effects of cultivation on soil productivity. Soils served as the planting medium for ancient agricultural systems, and also provided material for building earthen architectural structures and for making pottery.

Studies of ancient agricultural soils can contribute significantly to research on agricultural sustainability and soil quality in the context of modern and ancient farming systems. Ancient agricultural soils in arid and semiarid regions are particularly useful for these kinds of studies because: (1) soil-formation processes (e.g., weathering, leaching, and illuviation) proceeds slowly, so soil changes caused by ancient cultivation practices tend to persist and be detectable for 1000 years or longer; (2) many ancient fields have not been cultivated since they were abandoned, so historic farming practices such as plowing and artificial-fertilizer applications have not masked or erased soil properties that reflect ancient farming practices; and (3) the presence or absence of agricultural features for planting crops, such as rock alignments, rock piles, and terraces, provide important clues for discerning, sampling, and comparing cultivated and uncultivated soils.

Since the 1960s, several studies have been conducted in the American Southwest to assess the long-term, anthropogenic effects of cultivation. This work has been conducted in agricultural fields that are among the oldest identifiable fields in the New World. Ancient agricultural fields in the Sonoran desert are commonly associated with soil horizons

(e.g., argillic and petrocalcic horizons and duripans) or shallow bedrock that strongly impedes or blocks water infiltration, thereby holding moisture in the rooting zone for long periods after rain. Soil studies in Arizona and New Mexico indicate that the consequences of cultivation in terms of soil productivity and agricultural sustainability are highly variable due to many interacting factors (e.g., climate, topography, soil type, types of crops and native vegetation, agricultural technology, and the duration and intensity of cultivation). Cultivation effects are not easily predicted, so soil testing is essential for making informed evaluations of soil productivity and anthropogenic effects. Some investigations have found that ancient farming practices had long-lasting degradational effects on soils, whereas others have found that cultivation had little effect on soil quality or even that it was improved.

Degradational changes associated with organic matter depletion, compaction, erosion, and enrichment with salts are typically less severe for land where minimal or no-tillage farming has been practiced (as in the prehistoric American Southwest) than for intensively tilled or irrigated fields, but production levels can be reduced substantially nevertheless. Cultivation, especially of highly consumptive crops such as maize, a crop that was cultivated extensively in the New World, can rapidly deplete already low stores of N. Studies of Puebloan fields, at Mesa Verde and near Flagstaff and Santa Fe, and terraced Mimbres fields in southwest New Mexico have found that phosophorus reductions were so severe that agricultural soils became unproductive and were abandoned. By contrast, studies of rock mulch field systems in Arizona have found that soil productivity was not seriously degraded by cultivation. Rock mulch systems are ubiquitous in cobbly landscapes cultivated by the Hohokam, Anasazi, and other prehistoric cultures of the American Southwest. Similar rock mulch systems have also been documented in Israel, Italy, Peru, Argentina, New Zealand, China, the Canary Islands, and other places that have a soil moisture deficit during the growing season.

The Mimbres study has found that the primary anthropogenic soil changes were degradational, and that the effects persisted for more than 800 years after fields were abandoned. Rock alignments were built to function as dams to slow runoff, increase infiltration into agricultural soils, and thicken naturally thin A horizons in the terrace soils. Compared with uncultivated soils, cultivated terrace soils had lighter colors, thicker A horizons with blockier structures and higher bulk density, lower levels of organic carbon, nitrogen, and total and available phosphorus, and higher pH and manganese levels. Accompanying these soil changes were dramatic differences in vegetation patterns, with terrace soils having little to no grass cover compared with adjacent control. Chapalote maize grown in the terrace soils under controlled greenhouse conditions are significantly lower in weight than those grown in control soils. Different fertilizer treatments indicate that nitrogen is the most limiting nutrient for plant growth.

Studies of Classic Period (*ca.* AD 1150–1450) rock mulch field systems in the Gila River valley of southeast Arizona and in the Lower Verde River valley of central Arizona have found that agricultural soils were not degraded and that soil fertility was often improved. The lack of deleterious changes may reflect either direct cultivation effects or postcultivation vegetation associations with agricultural rock features, short-term use of fields, replenishment of nutrients by naturally deposited organic debris, cultivation of crops with low nutrient requirements, or recovery after abandonment. Use of gravel and cobble mulches on planting surfaces reduces soil erosion by wind and water, increases soil temperature to extend the growing season, increases water infiltration, and reduces evaporative water loss from the soil. Rock-pile soils in the Lower Verde have gravimetric water contents that average approximately 30% higher than adjacent controls (Figure 4). The Gila and Verde River studies have found that cultivated soils tend to have similar to significantly elevated levels of organic carbon (Figure 5), nitrogen, and total and available phophorus levels compared with uncultivated soils. Cultivated Gila soils had pH levels that were reduced from approximately 8.1–8.4 to 7.7–8.0, levels that would have been beneficial for crop production due to increased plant availability of many essential nutrients. Bulk density tests indicate that cultivation did not cause long-term soil compaction, which is a common degradational effect in many agricultural systems.

Determining what crops were or might have been cultivated in ancient fields can be very difficult, and

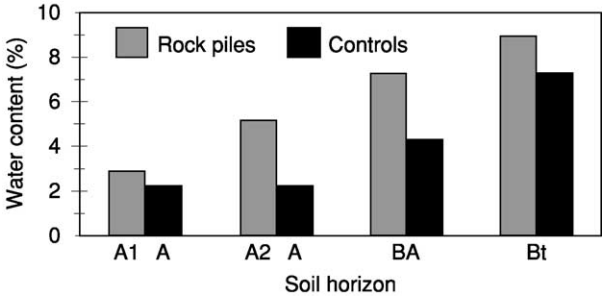


Figure 4 Gravimetric water content (grams of water per gram of soil, expressed as a percentage) of rock mulch soils in the Horseshoe Basin compared with adjacent controls. Reproduced from Homburg JA and Sandor JA (1997) Tucson, AZ: SRI Press.

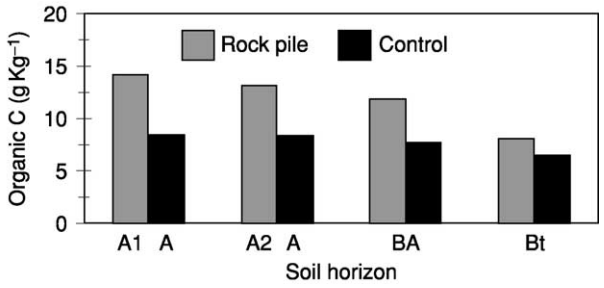


Figure 5 Organic C content of rock mulch soils in the Horseshoe Basin compared with adjacent soils. Reproduced from Homburg JA and Sandor JA (1997) Tucson, AZ: SRI Press.

soil data alone are insufficient to solve this problem. Where possible, evaluations of cultivation effects on soils need to make reference to the crops that were cultivated. Traces of plant remains (e.g., pollen, phytoliths, and carbonized plant remains) from soils in fields or in habitation features such as houses, agricultural processing areas, storage facilities, and waste deposits are used to reconstruct the range of crops that were cultivated. Soils can supply important clues about potential crops, however, especially when evaluated in the context of modern soil and native plant associations relative to prehistoric agricultural features. For example, agave thrives in cobbly, calcareous, droughty soils, even soils with low fertility in rugged terrain that support little other vegetation. Overall, soil nutrient levels in the Gila gridded-field complex are sufficient to have supported maize agriculture, but the thin soils, high temperatures, low rainfall, and low runoff throughout most landscape positions of the field suggest that crops such as agave or other drought-tolerant plants were the probable focus of agricultural production.

Studies of traditional Native American agriculture may supplement archeological studies of prehistoric agricultural systems in the Southwest. The rationale of these kinds of studies is that knowledge about prehistoric agriculture is inherently based on inference and reconstruction from sites abandoned for centuries. Historic and modern Native American agriculture may provide valuable sources of baseline information on identification of agricultural fields, placement of fields with respect to geomorphic and soil variables, crop and soil management, agricultural productivity, and the effects of agriculture on the environment. The relatively subtle nature of prehistoric agriculture, combined with subsequent overprinting by natural processes and cultural effects, underscores the need for such baseline studies. Although there are limitations in extrapolating from the past to the present, there are definite threads of continuity in agricultural strategies between prehistoric cultures and Native American groups that still practice similar forms of traditional agriculture.

One example is a study of the runoff agricultural system of the Zuni, a Puebloan tribe in west-central New Mexico. The Zuni and other Native American groups of the American Southwest have successfully cultivated maize and other crops for more than two millennia without using formal irrigation or artificial fertilizers. Zuni fields are among the oldest more or less continuously cultivated areas in the USA. Traditional Zuni agriculture is based on a runoff farming system that involves capturing storm water flow and organic-rich sediment from small watersheds and directing it on to agricultural fields. The Zuni and other Native American groups of the semiarid Southwest US have successfully cultivated maize and other crops for over three millennia without using artificial fertilizers. The Zuni soil study concludes that tillage in recent decades has altered some soil properties but there is no indication that agricultural soils have been degraded. Paired cultivated soils are 7.6% higher in bulk density on average and they have greater massive structure and reduced granularity. Compaction at this level, especially given the friability of the soils, suggests that root elongation is not restricted and there may even be advantages in moisture retention for this desert environment. No consistent differences in organic carbon, nitrogen, and available and total phosphorus have been identified in the cultivated, abandoned, and uncultivated soils, thus supporting the perception of Zuni farmers that long-term cultivation has not caused a decline in agricultural productivity.

List of Technical Nomenclature

Anthropogenic Relating to, or influenced by, human activity. Anthropogenic changes in soils can be degradational (e.g., accelerated erosion, organic matter and nutrient losses, salinization, and pollution) or improvements (e.g., elevated organic matter and nutrient status, and reduced erosion, salt, and pollution levels)

Argillic horizon A subsurface diagnostic horizon marked by significant accumulation of clay translocated from above. In general, a clay increase of at least 1.2 times that of an overlying horizon, and that is either 15 cm or more thick or at least one-tenth the thickness of all overlying horizons, is needed to qualify as an argillic horizon

Cambic horizon A subsurface diagnostic horizon marked by a pedogenic change from the soil parent material, such as reddening by

oxidation, development of soil structure, or accumulation of illuvial clay at levels that do not meet the definition of an argillic horizon

Geomorphic stability The degree or extent that a landform or landscape has been preserved or altered. Highly stable land surfaces have undergone little change for long periods of time, so associated soils are expected to be at or near equilibrium with major soil-forming factors, such as climate and vegetation cover. By contrast, unstable geomorphic surfaces are modified by erosion or buried by sediment, so the surface may change too fast for soil development to reach equilibrium with the environmental conditions

Illuviation A soil process whereby material (e.g., clay, iron oxides, calcium carbonate, or organic matter) accumulates in a soil due to translocation from overlying horizons. These materials are leached from above and translocated downward by wetting fronts in the soil

Midden A German term widely used in archeology that refers to anthropogenic accumulations of trash deposited either intentionally or unintentionally to form layers, mounds, or infilled pits. Midden soils are enriched with cultural materials, such as ash and carbonized plant debris, microfossil traces (pollen and phytoliths from cultivated and wild plant foods), butchered animal bones, shellfish, manure, and stone, ceramic, and metal artifacts are commonly found in middens, as well as human burials in some cultures. Midden soils are typically dark in color and have elevated organic matter and phosphorus levels relative to natural soils unaffected by human activities

Paleosol An ancient soil. Types of paleosols include: (1) buried soils – ones covered by sediment; (2) exhumed soils – ones formerly buried but now exposed by erosion; and (3) relict soils – ones formed under the influence of preexisting landscape or climatic regime that was never buried. This term is widely used in archeology and Quaternary geosciences, but there is no consensus as to a minimum age requirement or specific differences between soil formed in preexisting versus present landscapes and climates

Pedoarcheology A study that involves the application of soil science, especially studies in soil morphology and genesis, to answer archeological research questions, problems, or hypotheses

Pedostratigraphy Layering in soil material caused by soil formation processes. Soil horizons form through time in physically and chemically weathered sedimentary or residual geologic material, with differences in soil layers reflecting the integrated effects of climate, organisms, topography, and parent material. Pedostratigraphic units may or may not correlate to natural depositional layers, and they may result in layers that are well-preserved or mixed by burrowing soil fauna. Other kinds of stratigraphic layers include lithostratigraphic, chronostratigraphic, biostratigraphic, and ethno- or archeostratigraphic units

Soil micromorphology A study involving analysis of undisturbed soil at a microscopic scale. In archeological investigations, other kinds of earthen materials (e.g., unconsolidated sediments and cultural materials, like ceramics, adobe, bricks, and mortars) can be examined using micromorphology. Thin sections are 25–30-μm slices of soil or other earthen materials that are examined through a microscope using plane and/or crossed polarized light, or by other imaging techniques, such as ultramicroscopy (e.g., scanning or transmission electron microscopy). Thin sections are obtained by: (1) collecting and drying structurally intact samples; (2) impregnating them under vacuum with a solvent mixed with an unsaturated polyester or epoxy resin, with or without a stain or fluorescent dye; and (3) grinding, polishing, and mounting the thin section on glass slide, with or without a cover slip, depending on the kind of analysis

See also: **Applications of Soils Data**; **Civilization, Role of Soils**; **Factors of Soil Formation:** Biota; Climate; Human Impacts; Parent Material; Time; **Forensic Applications**; **Geographical Information Systems**

Further Reading

Brown AG (1997) Interpreting floodplain sediments and soils. In: Brothwell D, Market G, Dincauze D, and Renouf P (eds) *Alluvial Geoarchaeology: Floodplain Archaeology and Environmental Change*, pp. 63–103. Cambridge, UK: Cambridge University Press.

Collins ME, Carter BJ, Gladfelter BG, and Southard RJ (eds) (1995) *Pedological Perspectives in Achaeological Research*, SSSA Special Publication No. 44. Madison, WI: Soil Science Society of America.

Courty MA, Goldberg P, and MacPhail R (1989) *Soils and Micromorphology in Archaeology*. Cambridge, UK: Cambridge University Press.

Hill CL, Rapp GR, and Rapp G Jr (1998) Sediments and soils and the creation of the archaeological record. In: *Geoarchaeology: The Earth-Science Approach to Archaeological Interpretation*, pp. 18–49. New Haven, CN: Yale University Press.

Holliday VT (ed.) (1992) *Soils in Archaeology: Landscape Evolution and Human Occupation*. Washington, DC: Smithsonian Institution Press.

Holliday VT (1997) Stratigraphy, soils, and geochronology of Paleoindian sites. *Paleoindian Geoarchaeology of the Southern High Plains*, pp. 50–148. Austin, TX: University of Texas Press.

Holliday VT (2004) *Soils and Archaeological Research*. Oxford, UK: Oxford University Press.

Homburg JA (1988) Archaeological investigations at the LSU Campus Mounds. *Louisiana Archaeology* 15: 31–204.

Homburg JA and Sandor JA (1997) An agronomic study of two Classic Period agricultural fields in the Horseshoe Basin. In: Homburg JA and Ciolek-Torrello R (eds) *Vanishing River: Landscapes and Lives of the Lower Verde Valley: The Lower Verde Archaeological Project*, vol. 2, *Agricultural, Subsistence, and Environmental Studies*, pp. 127–148. CD-ROM. Tucson, AZ: SRI Press/The University of Arizona Press.

Limbrey S (1975) *Soil Science and Archaeology*. New York: Academic Press.

Mandel RD and Bettis EA III (2001) Use and analysis of soils by archaeologists: a North American perspective. In: Goldberg P, Holliday VT, and Ferring CR (eds) *Earth Sciences and Archaeology*, pp. 173–204. New York: Kluwer Academic/Plenum.

Sandor JA, Gersper PL, and Hawley JW (1986) Soils at prehistoric agricultural terracing sites in New Mexico. *Soil Science Society of America Journal* 50: 166–180.

Schiffer MB (1987) *Formation Processes of the Archaeological Record*. Albuquerque, NM: University of New Mexico Press.

Waters MR (1992) The geoarchaeological interpretation of sediments, soils, and stratigraphy. In: *Principles of Geoarchaeology: A North American Perspective*, pp. 88–113. Tucson, AZ: The University of Arizona Press.

Wood RW and Johnson DL (1978) A survey of disturbance processes in archaeological site formation. In: Schiffer MB (ed.) *Advances in Archaeological Method and Theory*, vol. 1, pp. 315–381. New York: Academic Press.

B

BACTERIA

Contents
Plant Growth-Promoting
Soil

Plant Growth-Promoting

Y Bashan and L E de-Bashan, Center for Biological Research of the Northwest (CIB), La Paz, Mexico

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Plant growth-promoting bacteria (PGPB) are defined as free-living soil, rhizosphere, rhizoplane, and phylosphere bacteria that, under some conditions, are beneficial for plants (Figure 1). Most of the activities of PGPB have been studied in the rhizosphere, and to lesser extent on the leaf surface; endophytic PGPB that reside inside the plant have also been found.

PGPB promote plant growth in two different ways: (1) They directly affect the metabolism of the plants by providing substances that are usually in short supply. These bacteria are capable of fixing atmospheric nitrogen, of solubilizing phosphorus and iron, and of producing plant hormones, such as auxins, gibberelins, cytokinins, and ethylene. Additionally, they improve a plant's tolerance to stresses, such as drought, high salinity, metal toxicity, and pesticide load. One or more of these mechanisms may contribute to the increases obtained in plant growth and development that are higher than normal for plants grown under standard cultivation conditions. However, these bacteria do not enhance the genetic capacity of the plant, as genetic material is not transferred. (2) A second group of PGPB, referred to as biocontrol-PGPB, indirectly promote plant growth by preventing the deleterious effects of phytopathogenic microorganisms (bacteria, fungi, and viruses). They produce substances that harm or inhibit other microbes, but not plants, by limiting the availability of iron to pathogens or by altering the metabolism of the host plant to increase its resistance to pathogen infection. Biocontrol-PGPB can also possess traits similar to PGPB; they may fix nitrogen or produce phytohormones, for example.

Plant Growth-Promoting Bacteria in Agriculture and the Environment

Many soil and especially rhizosphere bacteria can stimulate plant growth in the absence of a major pathogen by directly affecting plant metabolism. These bacteria belong to diverse genera, including *Acetobacter*, *Achromobacter*, *Anabaena*, *Arthrobacter*, *Azoarcos*, *Azospirillum*, *Azotobacter*, *Bacillus*, *Burkholderia*, *Clostridium*, *Enterobacter*, *Flavobacterium*, *Frankia*, *Hydrogenophaga*, *Kluyvera*, *Microcoleus*, *Phyllobacterium*, *Pseudomonas*, *Serratia*, *Staphylococcus*, *Streptomyces*, and *Vibrio* and including the legume-symbiotic genus *Rhizobium*.

Treatment of plants with agriculturally beneficial bacteria can be traced back for centuries. Inoculation of legumes with symbiotic *Rhizobium* has been practiced for almost 100 years and has had a major impact worldwide on crop yields. In Eastern Europe in the 1930s and 1940s, large-scale inoculation with associative nonsymbiotic bacteria such as *Azotobacter* and *Bacillus* failed. Two major breakthroughs in PGPB research that were largely responsible for the renewed interest in the field occurred in the late 1970s: in Brazil, the late Dr J Döbereiner and coworkers rediscovered that *Azospirillum* is capable of enhancing nonlegume plant growth, and in the USA, the work of JW Kloepper and MN Schroth and coworkers showed that biocontrol agents such as *Pseudomonas fluorescens* and *P. putida* can act as pesticides to control soilborne diseases.

The best-known among the nonsymbiotic PGPB are bacteria of the genus *Azospirillum*. These bacteria

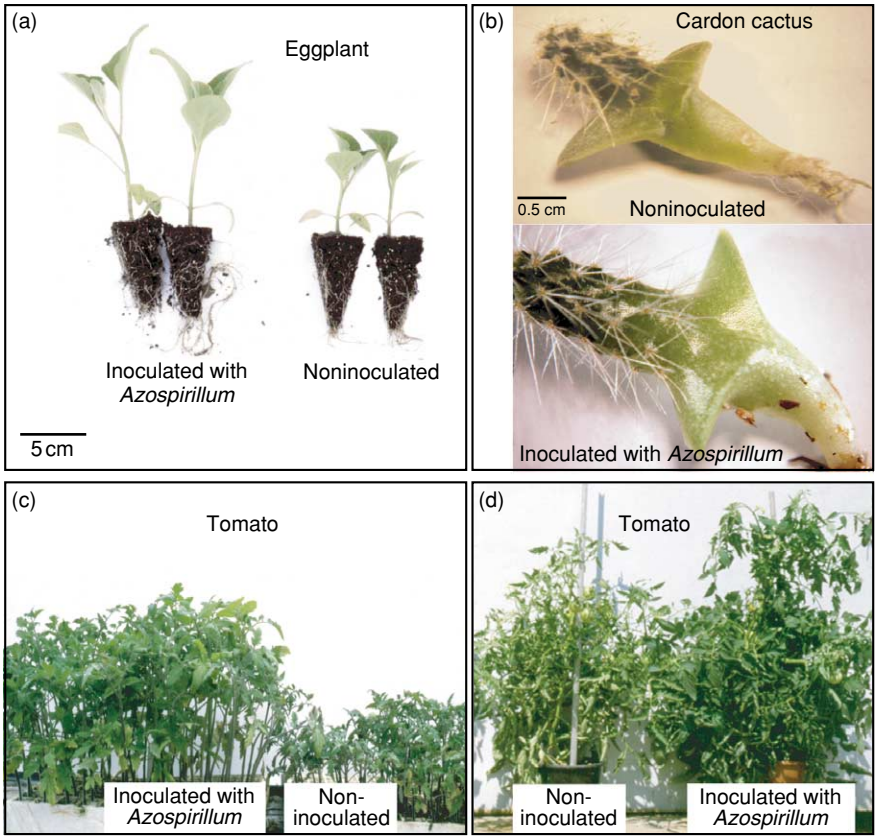


Figure 1 Enhanced seedling and plant growth after inoculation with *Azospirillum brasilense*. (a) Eggplant seedlings; (b) giant cardon cactus seedlings; (c) a nursery of tomatoes; (d) mature tomato plants. © Y Bashan.

enhance plant growth using a number of different mechanisms (Figure 2). Some are similar to those used by other PGPB (Table 1). Worldwide efforts to characterize this genus extensively have resulted in the availability of several commercial inoculants used for growth promotion of corn, wheat, rice, vegetables, and turf grass like Azogreen^{MR} (*Azospirillum lipoferum* on maize in France) and BioYield (*Paenobacillus macerans* and *Bacillus amyloliquefaciens* for biocontrol of tomato and pepper diseases in the USA). Although some strains of *Azospirillum* have an affinity for certain crops, the major advantage of this genus is that it is not plant-specific as it can enhance the growth of numerous plant species. Many field studies have shown that inoculation with *Azospirillum* increases crop yields by 5–30%; however, 30–40% of inoculations are unsuccessful. This inconsistency in yield stimulated experimentation with mixed inoculants, i.e., the combination of *Azospirillum* with other PGPB (Table 2). Enhanced plant growth following co-inoculation is due to the synergistic effect of both bacteria and *Azospirillum* functioning as a 'helper' bacterium to enhance the performance of other PGPB. Mixed inoculations have a higher success rate. It seems that in co-inoculated plants, nutrition is more balanced and the adsorption of nitrogen, phosphorus, and other mineral nutrients is significantly improved, yielding a better crop.

Although the PGPB described herein are associated with the plant, they are not symbiotic. Therefore, secure attachment of the bacteria to the root is essential for a long-term association for three main reasons: (1) If the bacteria are not attached to the root epidermal cells, plant growth substances excreted by the bacteria diffuse into the rhizosphere and are consumed by nutritionally versatile microorganisms before reaching the plant. (2) Without a secure attachment, water may wash the bacteria away from the rhizoplane to perish in the surrounding nutrient-deficient soil; many PGPB survive poorly in bulk soil. (3) Potential association sites for bacteria on roots, unoccupied by a PGPB, are vulnerable to colonization by aggressive, possibly nonbeneficial, root microbes. Thus, many PGPB have developed ways to remain attached to the roots, either temporarily or permanently. For example, *Azospirillum* has developed two modes of attachment (Figure 3). The first is a short-term attachment within hours after contact (after the bacteria migrate towards the roots by chemotaxis and aerotaxis, or the root reaches the

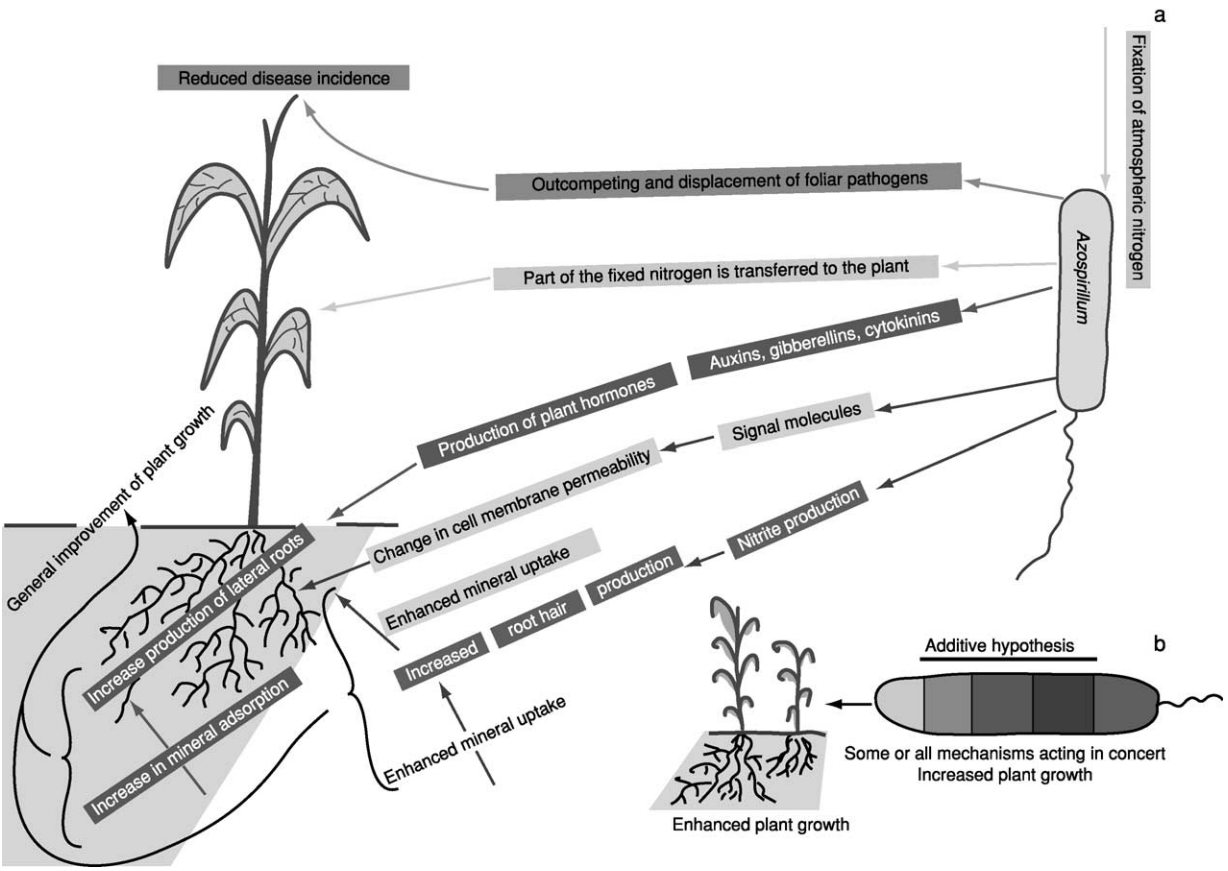


Figure 2 Mode of action of *Azospirillum* in promoting plant growth.

site of an applied inoculant). It involves hydrophobic interactions and lectin recognition between the bacteria and the plant cell wall. The second involves elaboration of a network of polysaccharide/protein fibrillar material, which anchors the bacteria permanently to the root surface. Eventually the bacterial cells multiply and form small aggregates (Figure 4) that provide an ecological advantage over the single cell state with respect to competition for nutrients that leak from the root. Similarly, the nitrogen-fixing cyanobacterium *Microcoleus chthonoplastes* enhances the production of thick mucigel layers on the roots of associated plants in which the bacteria are protected from the rhizosphere and from excessive oxygen which inhibits nitrogen fixation (Figure 5).

Because appropriate plants for colonization are not always available, Gram-negative PGPB (nonspore-forming bacteria) have developed mechanisms that allow them to survive in the absence of a host. Cells can form cysts and flocs (large, visible aggregates) that protect them from desiccation, produce melanin blocking ultraviolet irradiation, and reduce cell metabolism to the minimum required for survival. Furthermore, in times of plenty, many PGPB store large amounts of poly-β-hydroxybutyrate, that can sustain them for prolonged periods of nutrient scarcity.

In addition to their usefulness as a crop inoculant, the potential benefits of PGPB were extended to environmental applications in recent years. For example, *Azospirillum* species can enhance bioremediation of wastewater by microalgae by increasing microalgal proliferation and metabolism, allowing the microalgae to clean the water better than when used alone.

Mangrove ecosystems enhance fisheries along tropical coasts because they serve as breeding, refuge, and feeding grounds for many marine animals in the tropics during their younger and more vulnerable life stages. *Azospirillum* and cyanobacteria species may improve mangrove reforestation by increasing the rate of survival and development of seedlings in an otherwise unfavorable environment. Inoculation with several PGPB of the genera *Vibrio*, *Bacillus*, and *Azospirillum* improves domestication of the wild oilseed

Table 1 Mechanisms employed by plant growth-promoting bacteria

Mechanisms	*Effect on plant growth*	*Examples of bacterial species*
Root-associated nitrogen fixation	Increase nitrogen content and biomass	*Azospirillum, Acetobacter, Azotobacter,* Cyanobacteria, *Herbaspirillum*
Production of plant hormones (auxin, giberellin, cytokinin)	Stimulate root branching, increase shoot and root biomass, and induce the reproductive cycle	*Azospirillum*
Phosphate solubilization	Increase biomass and P content	*Bacillus lichiniformis, Vibrio*
Inhibition of plant ethylene synthesis	Increase root length	*Pseudomonas putida*
Sulfur oxidation	Increase biomass and foliar nutrient content	Undefined
Production of signal molecules and enhanced proton extrusion	Change in plant metabolism related to mineral absorption	*Azospirillum, Achromobacter*
Increase root permeability	Increase biomass and nutrient uptake	*Azospirillum*
Enhance general mineral uptake	Increase biomass and nutrient uptake	*Azospirillum*
Increase nitrite production	Increase formation of lateral roots	*Azospirillum*
Increase nitrate accumulation	Increase biomass and nitrate content	*Azospirillum*
Reduce heavy-metal toxicity	Protection against nickel toxicity	*Kluyvera*
Increase legume nodules or size	Increase biomass, N content, and reproductive yield	*Azospirillum*
Increase alder root nodules or size	Increase biomass and N content	*Frankia*
Increase frequency of infection by endomycorrhizal fungi	Increase biomass	*Pseudomonas*
Increase number of ectomycorrhizal root tips	Increase biomass	*Pseudomonas*
Increase temporary 'rain root' production in cacti	Improve survival of seedlings during drought	*Azospirillum*
Increase resistance to adverse conditions (drought, salinity, compost toxicity)	Improve survival of seedlings and increase biomass	*Azospirillum*
Additive hypothesis	Increase biomass as a result of several small-magnitude mechanisms working in concert or at the same time	*Azospirillum*

Table 2 Mixed inoculation of *Azospirillum* with other microorganisms and plants – a sampler

Mixed inoculation with species	*Plant*	*Effect on plants*
Rhizobium	Soybean, French bean, lentil, chickpea, alfalfa, bean, peanuts	Increase in nodule stimulation and function, total number and weight of nodules, epidermal cell differentiation in root hairs, straw and grain yield, root surface area, yield
Bacillus polymyxa	Sorghum	Increase in N and P uptake, grain and dry matter
Agrobacterium radiobacter or *Arthrobacter mysoreus*	Barley	Increase in grain yield, N_2 fixation, N accumulation in plant
Azotobacter chroococcum or *Streptomyces mutabilis*	Wheat, sugarcane	Increase in plant growth, indole acetic acid, P, Mg, N, and total soluble sugars in shoots, soil N content
Mycorrhizal fungi	Sorghum, wheat, jute, strawberry	Increase in dry weight of roots and shoots, grain and straw yield, mycorrhizal infection, P content, and uptake of N, Zn, Cu, and Fe

plant *Salicornia*, normally grown in mangrove ecosystems, which could be used in a seawater-irrigated agriculture system.

Treatment of cacti with *Azospirillum* enhances seedling establishment and survival in eroded desert areas. Re-vegetation of eroded and disturbed desert areas, aided by PGPB and vesicular-arbuscular mycorrhizal (VAM) fungi invigoration of desert plants responsible for soil stabilization, prevents soil erosion and promotes abatement of dust pollution (Table 3 and Figure 6). Finally, plants inoculated with *Kluyvera* have reduced nickel toxicity and can therefore grow in and rehabilitate nickel-contaminated wastelands.

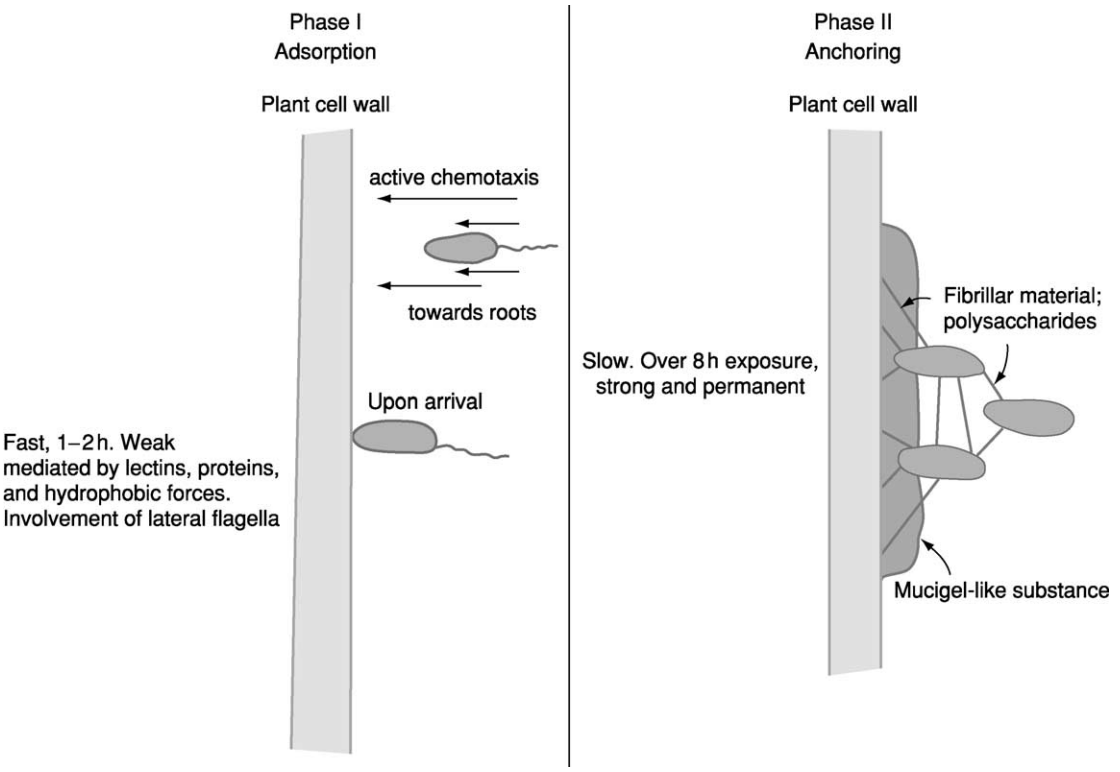


Figure 3 Mechanisms of attachment of *Azospirillum* to roots.

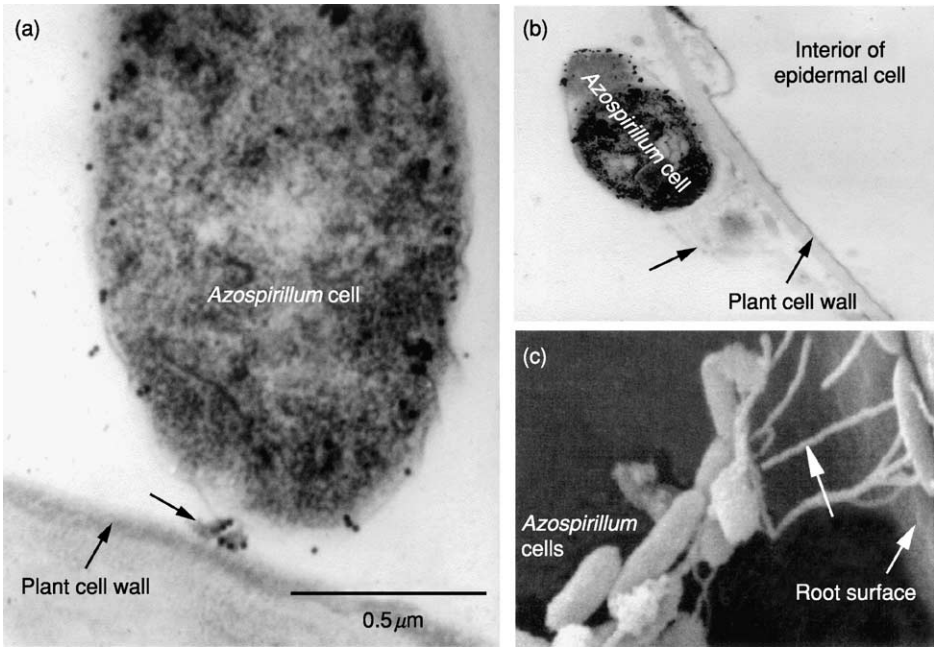


Figure 4 Transmission (a, b) and scanning (c) electron microscopy of attachment of *Azospirillum brasilense* to the root surface of wheat and cotton by fibrillar material. (a) polar attachment of bacterium to wheat plant cell wall by short fibrils (arrow); (b) nonpolar attachment of bacterium to wheat cell wall by unidentified electron-dense material (arrow); (c) permanent attachment of bacteria to cotton root surfaces through formation of long fibrils (arrow). © Y Bashan.

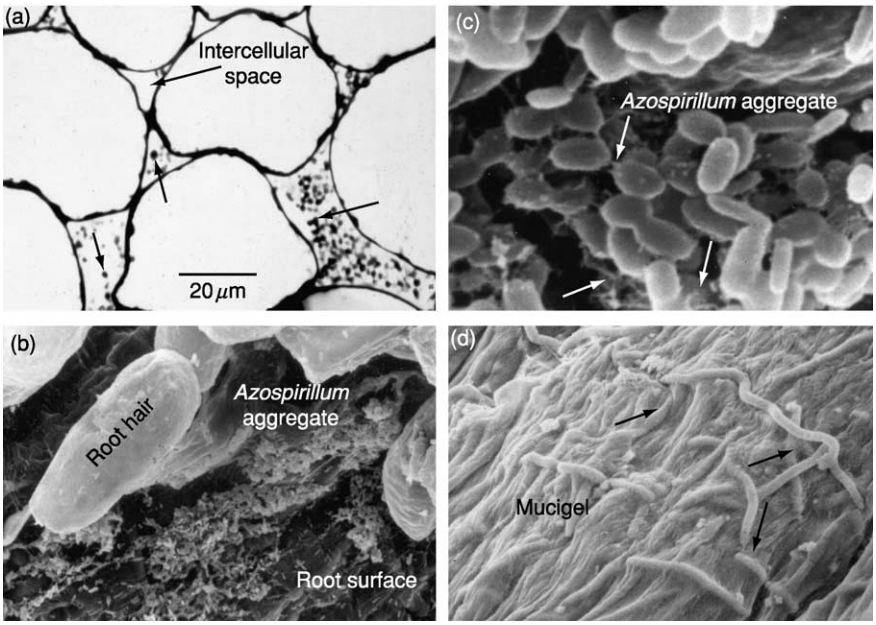


Figure 5 Light and scanning electron microscopy of interior and surface root colonization by plant growth-promoting bacteria (PGPB). (a) Light micrograph of thick cross-section of wheat roots showing the localization of *Azospirillum brasilense* within the roots. Bacteria (arrows) are located in the intercellular spaces of inner layers of cortical cells in the root elongation zone. (b, c) Root surface colonization of the elongation zone of wheat roots by inoculated *A. brasilense*. Note aggregation type of colonization. (c) is enlargement of small section of (b). Note fibril connections (arrows) between cells within the aggregate. (d) Production of mucigel layer on black mangrove roots in response to inoculation with the filamentous diazotroph cyanobacterium *Microcoleus chthonoplastes* in which the PGPB (arrows) is embedded. © Y Bashan.

Table 3 Plant growth-promoting bacteria usefulness in the environment

Bacterial species	*What it does*
Azospirillum	Helps microalgae *Chlorella* spp. to clean wastewater
Azospirillum + cyanobacteria *Microcoleus*	Improve reforestation of mangrove plants
Azospirillum + mycorrhizal fungi	Enhance seedling establishment and promote cactus growth to reduce soil erosion and dust pollution
Azospirillum, Bacillus, and *Vibrio*	Increase growth of wild oilseed plants destined for domestication
Kluyvera	Reduce nickel toxicity in polluted soil, which allows plant growth
Bacillus, Pseudomonas, and *Frankia*	Improve germination and increase growth of forest trees

Endophytic PGPB

Many bacterial species can live harmlessly as endophytes within plant tissues. They can reside latently or actively, and can colonize the plant within an organ or in the vascular system. Although most are saprophytes, some species are considered PGPB and biocontrol-PGPB because they increase plant growth and resistance to phytopathogens. For example, endophytic *Acetobacter diazotrophycus* can fix all the nitrogen required for cultivation of sugarcane and can promote pineapple growth. Some endophytic bacteria can promote growth of forest trees, from temperate pines to tree-shaped cacti. Seed endophytic PGPB from desert plants can weather rocks, unbind minerals essential for plant growth, and allow cactus seedlings to establish in barren areas as primary colonizers. These microbial activities are producing soil in bare rock areas, and consequently allow other plant species to grow. Endophytes, such as *Pseudomonas fluorescens* and *Bacillus* spp., can serve as biocontrol-PGPB controlling the soil pathogens *Fusarium* in cotton and *Rhizoctonia solani* and *Sclerotium rolfsii*.

Many endophytic bacteria invade plant tissues using mechanisms similar to pathogens, i.e., using hydrolytic enzymes, or natural (e.g., stomata) or artificial (wound) openings; however, their population density is generally lower than pathogens. Most are not recognized by the plants as potential pathogens. Endophytic bacteria rely heavily on nutrients supplied by the host plant; thus, variables affecting plant nutrition also affect the endophytic communities.

Biocontrol of Phytopathogens

Phytopathogenic microbes have an immense impact on agricultural productivity, greatly reducing crop yields and sometimes causing total crop loss. Usually,

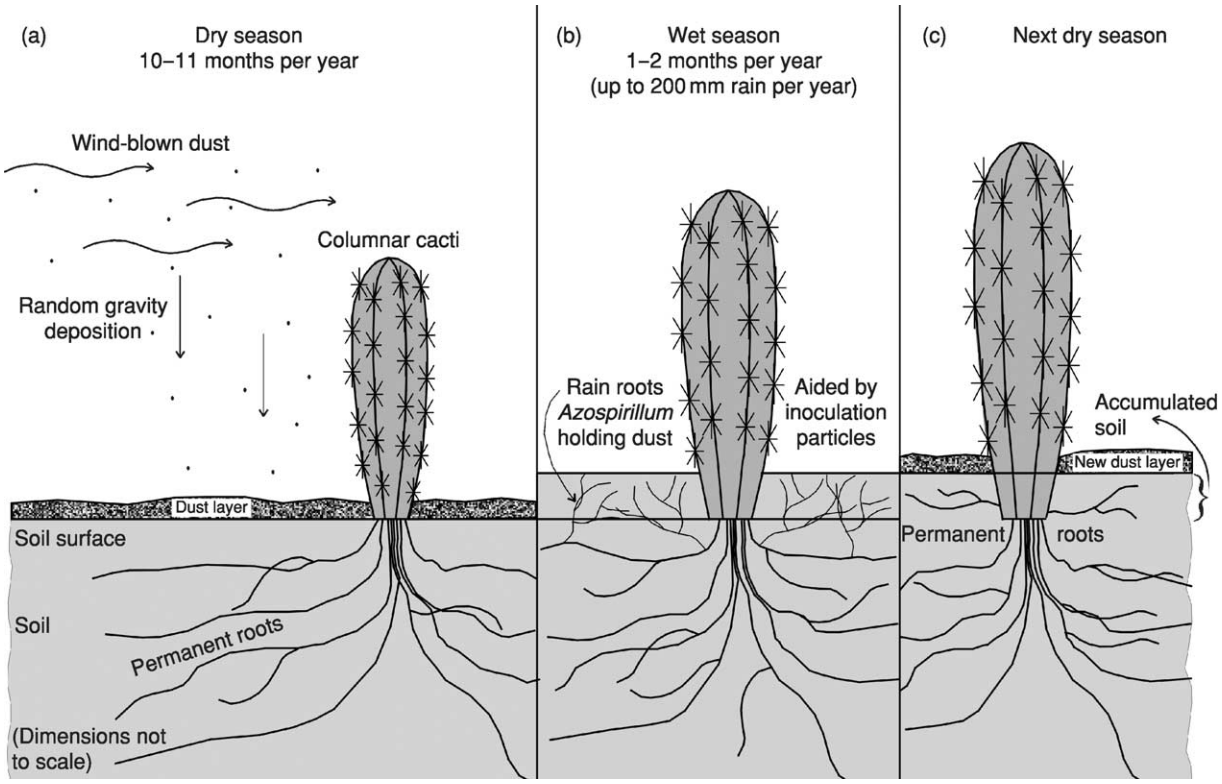


Figure 6 Mechanism of dust abatement and soil accumulation by cacti inoculated with *Azospirillum brasilense*.

growers manage phytopathogens by employing chemical pesticides and, to a lesser extent, expensive steam sterilization and 'soil solarization.' The main drawback of the chemical management strategy is that the target plants often remain infected but non-symptomatic for prolonged periods, thus, untreated. Small environmental shifts can produce uncontrollable epidemics. Additionally, pesticides are expensive, hazardous, affecting human and animal health when they accumulate in the plants and soil, and eliminate beneficial soil and biocontrol organisms. A better strategy to avert the development of epidemics is to treat the pathogen when its levels in the field are low, to prevent further increases over the growing season. Effective options include employing the pathogen's natural enemies as biological control agents or developing transgenic plants that are resistant to the pathogen. Both strategies are considered less destructive or more 'environmentally friendly' than chemical treatments. Several biocontrol-PGPB are commercially available.

Mechanisms Employed by Biocontrol-PGPB to Control Phytopathogens

The mechanisms employed by biocontrol-PGPB to deter phytopathogens can be chemical, environmental (outcompetition and displacement of pathogens), or metabolic (induction of acquired or induced systemic resistance and modification of hormonal levels in plants) (Table 4). A large array of microbial substances is involved in the suppression of pathogenic growth and subsequent reduction in damage to plants. These substances include antibiotics such as Agrocin 84, Agrocin 434, 2,4-diacetylphloroglucinol, herbicolin, phenazine, oomycin, pyoluteorin, and pyrrolnitrin, siderophores, small molecules such as hydrogen cyanide (HCN), and hydrolytic enzymes such as chitinase, laminarinase, β-1,3-glucanase, protease, and lipase. Most studies of biocontrol mechanisms were conducted under laboratory and greenhouse conditions; however, ultimately the efficacy of biocontrol-PGPB must be tested under field conditions. Thus, the importance of the below mechanisms in controlling pathogens should be considered conditional.

Production of Antibiotics

Antibiotic production by biocontrol-PGPB is perhaps the most powerful mechanism against phytopathogens. Many different types of antibiotics are produced and have been shown to be effective under laboratory conditions, although not necessarily under field conditions. Because the genes involved in the production of some antibiotics are known, it is possible to enhance antibiotic activity, hence enhance suppression

Table 4 Mechanisms employed by biocontrol plant growth-promoting bacteria against phytopathogens

Mechanisms	*Effect on plant growth*
Competition for Fe^{3+} ions through siderophore production	Reduced disease incidence and severity and increase biomass of plants
Antibiotic production	
Production of small toxic molecules	
Production of hydrolytic enzymes	
Competition for nutrients, colonization sites, and displacement of pathogens	
Induced and acquired systemic resistance	Make plant more resistant to infection by pathogens and increase biomass of plants
Change ethylene levels in plants	Reduce noxious effect of excess ethylene production in plants
Suppression of deleterious rhizobacteria	Increase biomass of plants

of phytopathogens, at least theoretically. Regardless of the progress under laboratory conditions, only one antibiotic-producing bacterium, *Agrobacterium radiobacter*, which produces the antibiotic Agrocin 84, is commercially available. This biocontrol-PGPB (which was later genetically modified to prevent the target pathogen from easily acquiring resistance) currently controls the pathogen *A. tumefaciens*, the causative agent of crown gall of stone fruit trees.

Production of Siderophores

Iron, an element essential for microbial growth, is mostly unavailable because it is mainly present in soil in a hard-to-solubilize mineral form (Fe^{3+}). To sequester iron from the environment, numerous soil microorganisms secrete low-molecular-weight, iron-binding molecules, called siderophores, which have a high capacity for binding Fe^{3+}. The now-soluble, bound iron is transported back to the microbial cell and is available for growth.

Siderophores produced by biocontrol-PGPB have a higher affinity for iron than the siderophores produced by fungal pathogens, allowing the former microbes to scavenge most of the available iron, and thereby prevent proliferation of fungal pathogens. Depletion of iron from the rhizosphere does not affect plant growth as plants can thrive on less iron than can microorganisms. Moreover, some plants can bind and release iron from bacterial iron–siderophore complexes, and use the iron for growth. Thus, the plant benefits in two ways: from the suppression of pathogens and from enhanced iron nutrition, resulting in increased plant growth (Figure 7).

Examples of the involvement of siderophores in disease suppression are many. A mutant strain of *P. putida* that overproduces siderophores was more effective than the wild bacterium in controlling the pathogenic fungus *Fusarium oxysporum* in tomato. Many wild strains that lose siderophore activity also lose biological control activity. The extent of disease suppression as a consequence of bacterial siderophore production is affected by several factors, including the specific pathogen, the species of biocontrol-PGPB, the soil type, the crop, and the affinity of the siderophore for iron. Thus, disease suppression under controlled laboratory conditions is only an indication of the efficacy of the biocontrol agent in the field.

Production of Small Molecules

Some biocontrol-PGPB produce a wide range of low-molecular-weight metabolites with antifungal potential. The best known is hydrogen cyanide (HCN), to which the producing bacterium, usually a pseudomonad, is resistant. HCN produced by bacteria can inhibit the black root rot pathogens of tobacco, *Thielabiopsis basicola*. In soil, a biocontrol pseudomonad was capable of using seed exudates of sugar beet to produce substances inhibitory to the pathogen *Pythium ultimum*, even though the pathogen was not inhibited when the two organisms were grown together in culture medium.

Production of Enzymes

Hydrolytic enzymes produced by some biocontrol-PGPB can lyse fungal cell walls, but not plant cell walls, and thereby prevent phytopathogens from proliferating to the extent that the plant is endangered. For example, *Pseudomonas stutzeri* produces extracellular chitinase and laminarinase that lyse the pathogen *Fusarium solani*. Similarly, *Burkholderia cepacia* produces β-1,3-glucanase and reduces disease caused by the fungi *Rhizoctonia solani, Scelrotium rolfsii*, and *Phytium ultimum*. Another strategy used by biocontrol-PGPBs to reduce disease severity in plants is the hydrolysis of fungal products that are harmful to the plant. *Cladosporium werneckii* and *B. cepacia* can hydrolyze fusaric acid (produced by the fungus *Fusarium*) that causes severe damage to plants.

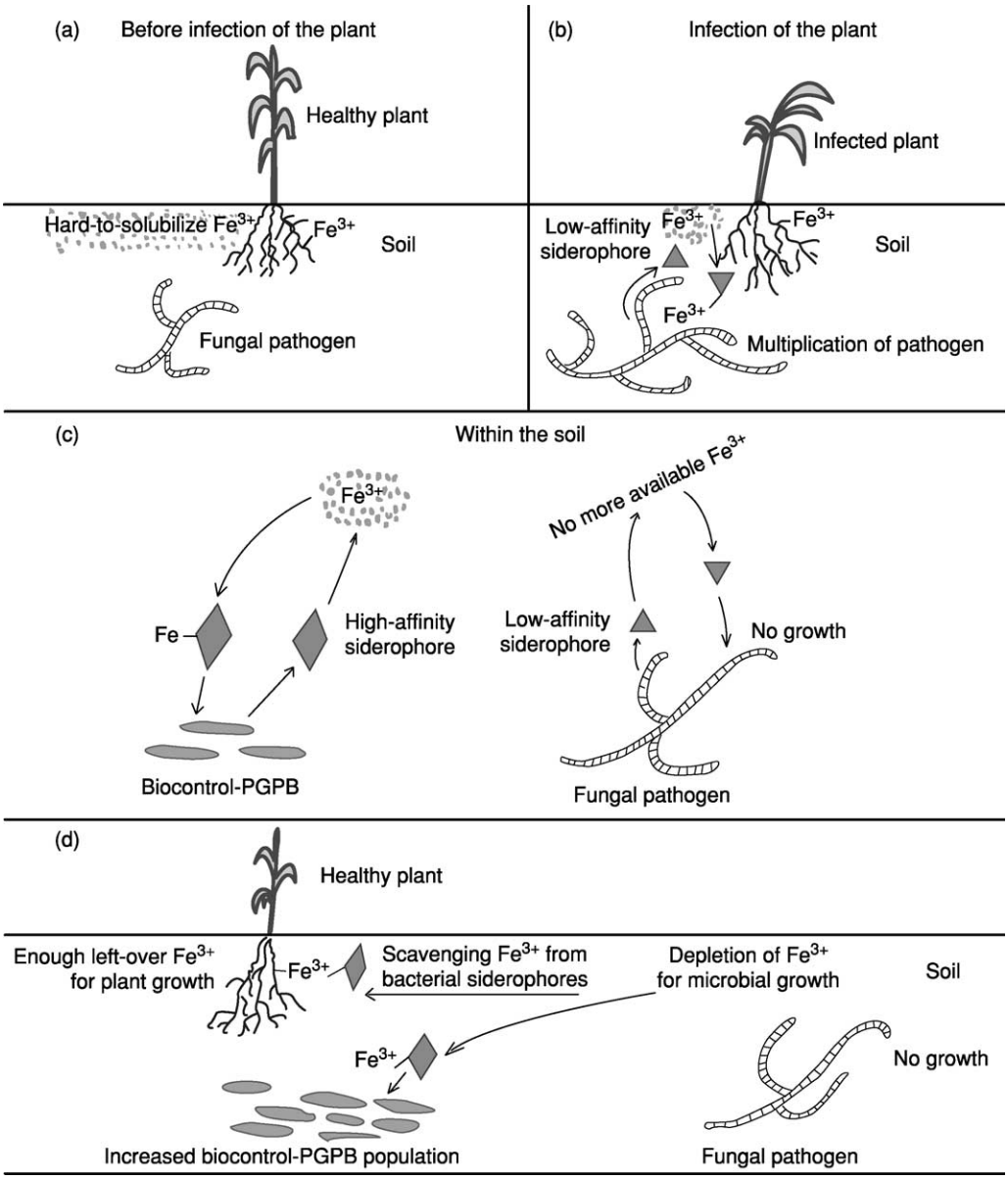


Figure 7 Biological control of fungal pathogens by biocontrol plant growth-promoting bacteria (PGPB) using siderophores.

Competition and Displacement of Pathogens

Competition for nutrients and suitable niches among pathogens and biocontrol-PGPB is another mechanism of biocontrol of some plant diseases. For example, high inoculum levels of a saprophytic *Pseudomonas syringae* protected pears against *Botrytis cinerea* (gray mold) and *Penicillium expansum* (blue mold).

On leaves there are a limited number of sites where a pathogen can attack the plant. Bacteria capable of multiplying on the leaf surface to form a large population can compete successfully with pathogens for these sites and often reduce disease. These agents can be saprophytic strains, PBPB, or nonvirulent strains of the pathogen. For example, the PGPB *A. brasilense* was able to displace the causal agent of bacterial speck disease of tomato, *P. syringae* pv. *tomato*, on tomato leaves, and consequently decreased disease development. Similarly, when a nonpathogenic strain of *P. syringae* pv. *tomato* was co-inoculated on to leaves with a pathogenic strain, disease incidence was significantly reduced. An ice-nucleation-deficient mutant of *P. syringae* displaced pathogenic *P. syringae*, and protected tomato and soybean against early frost induced by the pathogen.

Modification of Plant Metabolism

Induced and Acquired Systemic Resistance

Plants can be protected against pathogens for long periods and across a broad spectrum of disease-causing microbes by making them more resistant to infection. Exposure to pathogens, nonpathogens, PGPB, and microbial metabolites stimulates a plant's natural self-defense mechanisms before a pathogenic infection can be established, effectively 'immunizing' the plant against fungal, viral, and bacterial infections. Protection occurs by accumulation of compounds such as salicylic acid, which plays a central protective role in acquired systemic resistance, or by enhancement of the oxidative enzymes of the plant. While acquired systemic resistance is induced upon pathogen infection, induced systemic resistance can be stimulated by other agents, such as PGPB inoculants. The feasibility of protecting plants by induced systemic resistance has been demonstrated for several plant diseases. Plants inoculated with the biocontrol-PGPB, *P. putida* and *Serratia marcescens* were protected against the cucumber pathogen *P. syringae* pv. *lachrymans*.

Modification of Plant Ethylene Levels

In response to stress or pathogenic attack, plants commonly synthesize higher than normal amounts of the hormone ethylene. Ethylene stimulates senescence and leaf and fruit abscission, inhibits plant growth, and triggers cell death near infection sites. A biocontrol-PGPB that can lower plant ethylene levels after infection might be beneficial for the plant. Some PGPB synthesize the enzyme ACC deaminase, which has no known role in bacterial metabolism, but can lower the plant's ethylene levels and thereby stimulate plant growth when the plant is inoculated with the PGPB. This effect was demonstrated with the PGPB *P. putida* and with a strain of *A. brasilense* (which naturally lacks the enzyme) genetically manipulated to carry the gene for ACC deaminase.

Prospects for Improving PGPB by Genetic Manipulation

As our understanding of the mechanisms used by PGPB advances, it becomes feasible to enhance their capacity to stimulate plant growth by modifying promising traits. The activity and utility of a biocontrol-PGPB may be enhanced by supplanting it with genes responsible for the biosynthesis of antibiotics, extending the range of pathogens against which a single biocontrol-PGPB can be used, or by genetically manipulating the bacterium to increase production of the antibiotic. Similarly, it is possible to extend the range of iron–siderophore complexes that a single strain can use, allowing a biocontrol-PGPB strain to use siderophores synthesized by other soil microorganisms, hence giving it a competitive advantage. Because many of the enzymes that hydrolyze fungal cell walls are encoded by a single gene, it would be relatively easy to isolate these genes, transfer them to other biocontrol-PGPB, and thus construct new biocontrol-PGPB armed with antibiotics, siderophores, and hydrolytic enzymes. When developing an attenuated pathogenic strain to displace a pathogen in the environment, it is not only necessary to delete the virulence genes from the bacteria but also to insert copper-resistant genes in the chromosome, since copper is a major bactericide used in agriculture. It is possible to isolate bacterial genes for ACC deaminase and transfer them to biocontrol-PGPB that employ other plant growth-promoting mechanisms, allowing them to modulate ethylene levels in the host plant, and reduce disease severity.

Potential growth-promoting traits can be transferred from any bacteria to a PGPB. For example, the transfer of the gene for acid phosphatase from the saprophytic soil bacterium *Morganella morganii* to *B. cepacia* and *Azospirillum* strains would create a biocontrol-PGPB and a nitrogen-fixing bacterium with phosphate solubilization activity and therefore enhanced phosphate uptake.

Regardless of the type of genetic insertion aimed at improving the PGPB, as a general rule, wild strains (nontransformed) are likely to persist in the environment longer than their transformed relatives. However, a transformed PGPB with a short survival capacity (but long enough to last a growing season) is a bonus for commercial suppliers, who can then provide fresh inoculant to the grower on a regular basis. In addition, a short-lived genetically engineered PGPB may be more acceptable from an environmental standpoint.

PGPB Inoculants

A 'carrier' is a vehicle for delivery of live PGPB from the factory to the field. Without a suitable formulation, many promising PGPB will never reach the marketplace. A universal carrier or formulation is presently unavailable (**Table 5**). A good carrier has one essential characteristic: the capacity to deliver the right number of viable cells in good physiological condition at the right time (**Tables 6** and **7**). Peat is the most common carrier for rhizobia and many PGPB (**Figure 8**); however, more advanced formulations based on polymers such as alginate and liquid formulations in water and oils are constantly being

Table 5 Inoculant type carrier for plant growth-promoting bacteria

Category	*Materials*
Soils	Peat, coal, clays, and inorganic soil
Plant waste materials	Composts, farmyard manure, soybean meal, soybean and peanut oils, wheat bran, sugar industry waste, agricultural waste material, spent mushroom composts, and plant debris
Inert materials	Vermiculite, perlite, ground rock phosphate, calcium sulfate, polyacrylamide, polysaccharide-like alginate, and carraginan
Plain lyophilized microbial cultures	Culture media and cryoprotectants

Table 6 Characteristics of inoculant for plant growth-promoting bacteria (PGPB)

Characteristic	*Degree of importance*
Deliver right number of viable cells in good physiological condition at the right time	****
Bacteria released from the inoculant can inoculate the plant efficiently	****
Inexpensive raw material	****
Provide slow release of bacteria for long periods (or short periods in case of some biocontrol-PGPB)	****
Uniform, with consistent quality	***
Biodegradable by soil microorganisms	***
Contains large and uniform bacterial population	***
Ease of handling by the farmer	***
Nontoxic in nature, causes no ecological pollution (like air dispersion or entering the ground water)	**
Relatively small in volume and in nonrefrigerated conditions	**
Ease to manipulate its chemical properties in relation to the biological requirement of the PGPB	**
Applied with standard agrochemical field machinery	**
Sufficient shelf-life (1–2 years at room temperature)	**
Dry and synthetic	*
Easy quality control by the industry	*
Nearly sterile or easily sterilized	*
High water-holding capacity (for wet inoculants)	*
Suitable for many bacterial species or strains	*
Easily manufactured and mixed by existing industry	*
Allows the addition of nutrients and has an easily adjustable pH	*

Note: No single carrier can have all these qualities, but a good one should have as many as possible. Degrees of importance: *important to ****essential.

Table 7 Formulations of inoculants for plant growth-promoting bacteria

Dispersal form	*Main characteristics*	*Popularity of use*
Powders	Used as seed coating before planting. Sizes vary from 0.075 to 0.25 mm	Most common
Slurries	Powder-type inoculant suspended in liquid (usually water). Suspension drips into the furrow or seeds are dipped just prior to sowing	Less popular
Granular	Applied directly to the furrow together with the seeds. Size ranges from 0.35 to 1.18 mm	Popular
Liquids	Seeds are dipped into inoculant before sowing, or an applicator evenly sprays liquid inoculant on the seeds. After drying, the seeds are sown	Popular

evaluated. Since peat has reached its maximum development potential and still presents bacterial delivery difficulties, it appears that the future lies in the production of synthetic inoculant carriers in forms such as macro- and microbeads or powders on seed coatings (Figure 9).

Conclusions and Prospects

Application of PGPB in the field (apart from rhizobia) has yielded satisfactory results in controlled experiments, although results are less promising under agricultural conditions. Compared with chemical applications, their presence in the agricultural market is small, and nonexistent for environmental applications, where they are only experimental. However, the public, and therefore, the agrochemical industry, are now more sympathetic to the concept of PGPB inoculants. The notion prevailing today is that PGPB inoculants will complement the chemicals already on the market. It is relatively easy to isolate a new biocontrol-PGPB or to find a bacterium that will increase root development. Yet, identification of the best bacterium for the task is still difficult as the

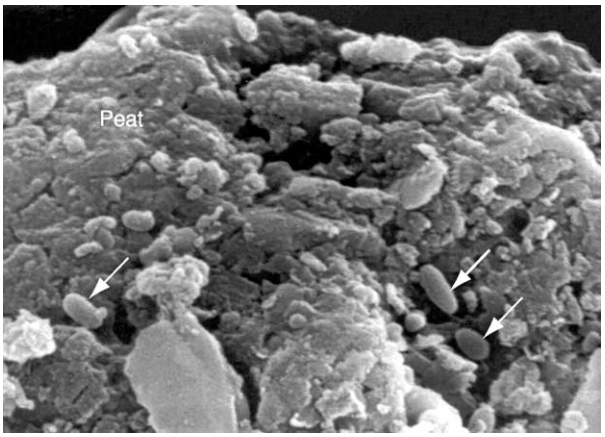


Figure 8 Scanning electron microscopy of peat inoculant (commonly used for *Rhizobium*) having a population of *Azospirillum brasilense* (arrows). © Y Bashan.

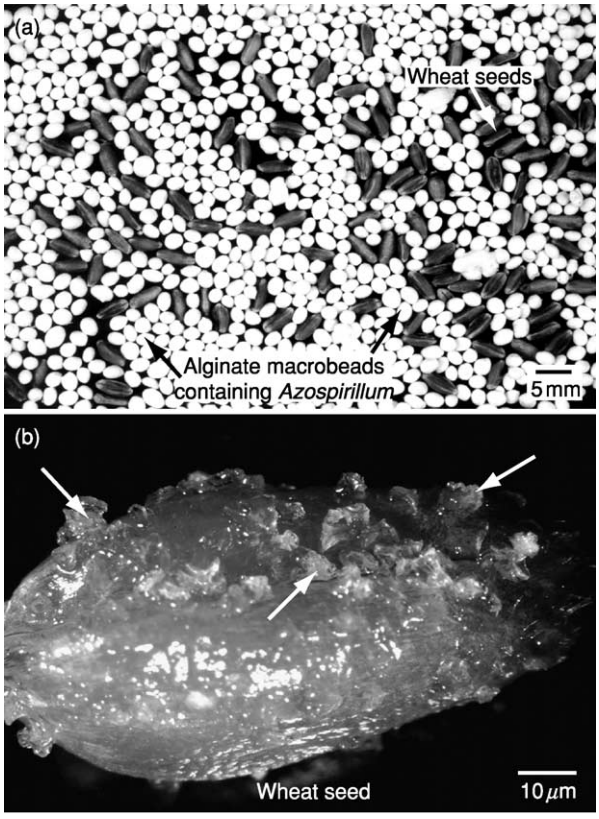


Figure 9 Synthetic inoculant carriers. (a) Macrobead inoculant made of alginate and *Azospirillum brasilense* mixed with wheat seeds before sowing. (b) Microbeads of alginate and *A. brasilense* (arrows) attached to the surface of a wheat seed. © Y Bashan.

characteristics necessary for such a PGPB are still poorly understood. Little is known about the features required for rhizo-competence and for survival and function in the environment after application.

While it is feasible to extract traits from PGPB controlled by a single gene and to use these genes to create transgenic plants that obviate the use of PGPB, PGPB use multiple mechanisms to promote plant growth, or mechanisms such as nitrogen-fixation that are impossible, as yet, to transfer. Thus, transfer of a mechanism encoded by a single gene from PGPB to plants may not provide significant benefits, although the engineering of plants with traits of PGPB has a significant presence in PGPB research.

Realistically, chemical fertilizers and pesticides will continue to dominate the marketplace in the near future. PGPB inoculants will only gradually and modestly displace chemical solutions to agricultural and environmental problems.

See also: **Mineral–Organic–Microbial Interactions**; **Mycorrhizal Fungi**; **Nitrogen in Soils:** Symbiotic Fixation

Further Reading

Bashan Y (1998) Inoculants of plant growth-promoting bacteria for use in agriculture. *Biotechnology Advances* 16: 729–770.

Bashan Y and Holguin G (2002) Plant growth-promoting bacteria: a potential tool for arid mangrove reforestation. *Trees Structure and Function* 16: 159–166.

Bashan Y, Bashan LE, and Moreno M (2001) Environmental applications of plant growth-promoting bacteria of the genus *Azospirillum*. In: De Boer SH (ed.) *Plant Pathogenic Bacteria*, pp. 68–74. Dordrecht, The Netherlands: Kluwer Academic Publishers.

Bashan Y, Holguin G, and de-Bashan LE (2004) *Azospirillum*–plant relationships: agricultural, physiological, molecular and environmental advances (1997–2003). *Canadian Journal of Microbiology* (in press).

Benizri E, Baudoin E, and Guckert A (2001) Root colonization by inoculated plant growth-promoting rhizobacteria. *Biocontrol Science and Technology* 11: 557–574.

Chanway CP (1997) Inoculation of tree roots with plant growth promoting soil bacteria: an emerging technology for reforestation. *Forest Science* 43: 99–112.

Cook RJ (1993) Making greater use of introduced microorganisms for biological control of plant pathogens. *Annual Review of Phytopathology* 31: 53–80.

Dobbelaere S, Vanderleyden J, and Okon Y (2003) Plant growth-promoting effects of diazotrophs in the rhizosphere. *Critical Reviews in Plant Sciences* 22: 107–149.

Glick BR (1995) The enhancement of plant growth by free-living bacteria. *Canadian Journal of Microbiology* 41: 109–117.

Glick BR (2003) Phytoremediation: synergistic use of plants and bacteria to clean up the environment. *Biotechnology Advances* 21: 383–393.

Glick BR, Patten CL, Holguin G, and Penrose DM (1999) *Biochemical and Genetic Mechanisms Used by Plant Growth Promoting Bacteria*. London: Imperial College Press.

Hallmann J, Quadt-Hallmann A, Mahaffee WF, and Kloepper JW (1997) Bacterial endophytes in agricultural crops. *Canadian Journal of Microbiology* 43: 895–914.

Jetiyanon K and Kloepper JW (2002) Mixtures of plant growth-promoting rhizobacteria for induction of systemic resistance against multiple plant diseases. *Biological Control* 24: 285–291.

Pankhurst CE and Lynch JM (1995) The role of soil microbiology in sustainable intensive agriculture. *Advances in Plant Pathology* 11: 229–247.

Sturz AV, Christie B, and Nowak J (2000) Bacterial endophytes: potential role in developing sustainable systems of crop production. *Critical Reviews in Plant Science* 19: 1–30.

Sutton JC and Peng G (1993) Manipulation and vectoring of biocontrol organisms to manage foliage and fruit diseases in cropping systems. *Annual Reviews of Phytopathology* 31: 473–493.

van Loon LC, Bakker PAHM, and Pieterse CMJ (1998) Systemic resistance induced by rhizosphere bacteria. *Annual Reviews of Phytopathology* 36: 453–483.

Vessey JK (2003) Plant growth-promoting rhizobacteria as biofertilizers. *Plant and Soil* 255: 571–586.

Zehnder GW, Yao C, Murphy JF *et al.* (1999) Microbe-induced resistance against pathogens and herbivores: evidence of effectiveness in agriculture. In: Agrawal AA (ed.) *Induced Plant Defenses Against Pathogens and Herbivores: Biochemistry, Ecology and Agriculture*, pp. 335–355. St. Paul, MN: APS Press.

Soil

L J Halverson, Iowa State University, Ames, IA, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

At densities ranging from 10^7 to 10^{11} per gram of soil, bacteria are clearly common residents of soil. The ability of bacteria to thrive in soil is due, in part, to their unequaled metabolic versatility and phenotypic plasticity, which allows them to colonize the vastly different habitats soils are comprised of. Soil is a remarkably diverse, complex, heterogeneous habitat comprised of solid, liquid, and gaseous phases that vary in both space and time, and whose complexity is influenced by the development, movement, and metabolic activity of plant roots. Furthermore, environmental conditions such as, for example, soil temperature, moisture, pH, and water availability, can fluctuate rapidly or slowly, creating stresses that bacteria have to cope with. By necessity soil bacteria need to utilize a range of colonization and survival strategies.

Given the complexity of the soil habitat, the questions arise as to how soil bacteria are able successfully to colonize a particular habitat, and what spectrum of strategies they employ for growth and survival.

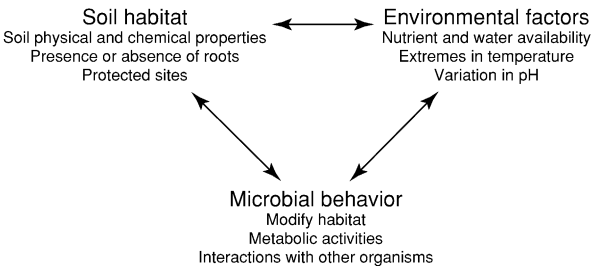


Figure 1 Model illustrating the interrelated attributes influencing bacterial fitness in soil.

To understand this requires developing an appreciation of what their habitat is like, what the landscape and environment is like, and how the residents, the neighbors, and the community behave, and how their combined behaviors influence growth and survival strategies. If the life history of each individual bacterium could be monitored we would know what adaptive strategies were employed, when they were employed, and the context in which they were employed. Instead, conceptual models are developed based on our knowledge of microbial physiology and ecology and of soils themselves to understand better the spectrum of strategies bacteria utilize to survive. One such model describes bacterial fitness as the ability to grow and/or survive in soil in a particular habitat under a given environmental regime. Fitness is therefore context-dependent, since it will depend on the habitat and the environment conditions in which it is being assessed (Figure 1). The ability of bacteria to deliberately alter their habitat in response to specific environmental signals to alleviate the effect of environmental stressors on cellular physiology or to increase their ability to successfully colonize a particular habitat is particularly noteworthy.

Bacterial Diversity

There are two types of microbial classification systems. Phenetic classification is where relationships among organisms are based on a maximum correlation of attributes when all properties are more-or-less weighed equally. In contrast, phylogenetic classification systems are based on evolutionary relatedness, which is usually based on the similarity of nucleic acids of highly conserved macromolecules such as ribosomal ribonucleic acids (rRNAs). From comparative analyses of the small subunit of rRNA, three phylogenetically distinct cellular lineages have been identified, two of which, called the Bacteria and Archaea, are prokaryotic in cell structure (Figure 2). Representatives of the Bacteria and Archaea domains are found in soil. Proteobacteria is comprised of five

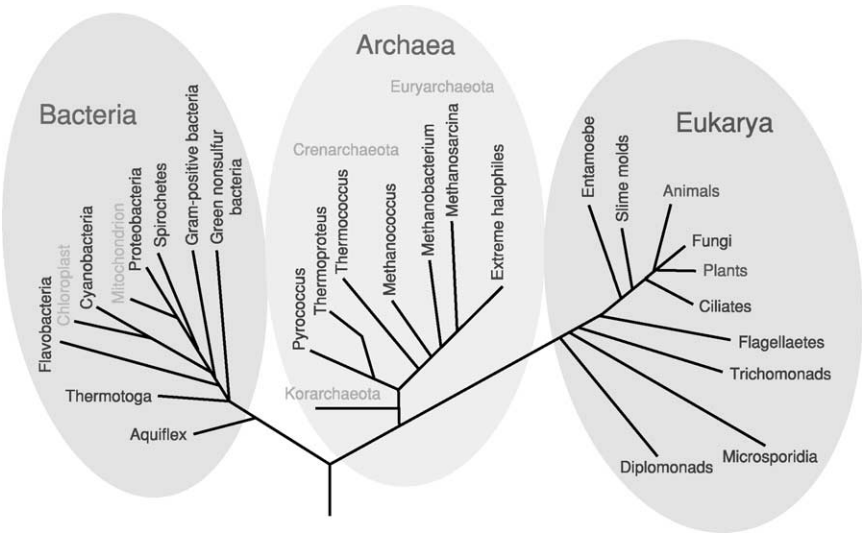


Figure 2 The three-domain classification of life based on comparative sequencing of small-subunit ribosomal ribonucleic acid molecules. This classification system separates prokaryotic organisms into two domains: Bacteria and Archaea. The third domain, Eukarya, contains all organisms composed of eukaryotic cells. Branch points and branch lengths do not reflect actual evolutionary position or distance. Figure based on data obtained from the Ribosomal Database Project II.

Table 1 Examples of physiological and ecologic classifications of soil bacteria

Category	*Classification*	*Description*
Nutritional	Heterotroph	Able to use organic compounds as energy and carbon source
	Phototroph	Able to use light as its energy source
	Autotroph	Able to use carbon dioxide as a sole carbon source
	Chemolithotroph	Able to use an inorganic substrate as an electron donor
	Saprophyte	Degrades nonliving organic material
Functional	Nitrifiers	Oxidation of ammonia to nitrite and nitrate
	Denitrifiers	Reduction of nitrate to dinitrogen gas
	Nitrogen-fixers	Reduction of dinitrogen gas to ammonia
	Sulfate-reducers	Reduction of sulfate to hydrogen sulfide
	Sulfate-oxidizers	Oxidation of elemental sulfur to sulfate
	Decomposers	Breakdown of complex compounds into simpler compounds
Interactions	Commensalism	One organism benefits while the other is neither harmed nor benefits
	Parasitism	One organism benefits while the other is harmed
	Mutualism	Both organisms benefit

subgroups and contains most of the Gram-negative bacteria found in soil. The ability to extract nucleic acids directly from soil has improved our understanding of the diversity of both cultivated and uncultivated bacteria, such as those in the division *Acidobacterium*, present in soil.

In addition to diversity at the organismal level (number of different taxonomic groups), soil microbiologists are also interested in genetic diversity (variation in genes caused by horizontal gene transfer and mutation). Currently we understand little about how genetic diversity is translated into organismal diversity and less about how genetic and organismal diversity influence ecologic and evolutionary processes. The metabolic activities and functions of soil bacteria are of especially strong interest, since they are vital to many ecologic processes.

Consequently, soil bacteria are frequently described by their physiological and ecologic properties rather than, or in addition to, their phylogenetic relationships (Table 1). This question remains as to how this vast organismal and functional diversity arise and how fitness traits, stress adaptation, and environmental parameters in soil influence this diversity.

The Microbe as the Scale of Study

It is important not to neglect the element of scale as we try to understand bacterial behavior. While macro- and mesoscale distributions of microbes

indicate the average abundance of microorganisms through volumes of soil, it is the patterns of distribution and metabolic activities at the scale of the microbe (the microscale) that will control many microbial behaviors. It is also the scale at which adaptive traits of individual cells will influence both short- and long-term survival strategies, which ultimately affect the outcomes of a population (diversification) or community. From the perspective of a bacterium, a volume of soil or a root several centimeters away can represent the distance needed to travel across a small state in the USA. The question arises as to how biological, chemical, and physical attributes that define a habitat and can vary along distances of less than 1 μm to 100 μm or more affect bacterial behavior. There is a limited understanding of how bacteria perceive their immediate environment or change to it on a scale that is most relevant to them.

Increasingly bioreporters utilizing reporter-gene technology are being used to provide information on the physiological activity of an individual microbe, usually at the level of transcription of a target gene that is responsive to a particular stimulus (Figure 3). Moreover, this information may help us gain an understanding of how bacteria perceive their habitat and the physical, chemical, and biological properties of that habitat. Reporter-gene technologies are also being developed that permit the identification of environmental conditions, such as bioavailability of a nutrient or pollutant, or patterns in expression of target genes even if the target compound is ephemeral or the gene is transiently expressed. The future of bioreporters in microbial ecology is great, since they provide a means to obtain a higher resolution of the bacterial habitat that is needed for understanding many bacterial processes.

Two Basic Survival Strategies

There are at least two strategies for bacteria to use to survive the large and often rapid fluctuations in environmental conditions that occur in a soil. One is a strategy of avoidance that requires the ability to seek and exploit sites that are protected from stresses. Alternatively, a strategy of tolerance requires the ability to tolerate environmental stresses (nutrient deprivation, low water availability, extremes in temperature). There is most likely a spectrum of strategies employed, from those which employ solely a tolerance strategy to those that employ both strategies to various extents. If a bacterium is to avoid a stress it must have active mechanisms by which to facilitate movement to sites that are amenable to growth and survival and, once there, it must be able to adapt to those specific conditions or to modify the habitat so that it is more suitable for survival.

Avoidance

To avoid a stress, bacteria need actively to position themselves to a more hospitable site. Avoidance can happen passively when bacteria are physically repositioned in the soil profile by root movement or through some other physical disturbance, but this is not an effective strategy for surviving stressful times. If bacteria grow or survive at particular sites in soils then the ability to move actively to such sites is an important arsenal in their survival strategies. Bacterial motility is likely to play a role in positioning them in a niche that

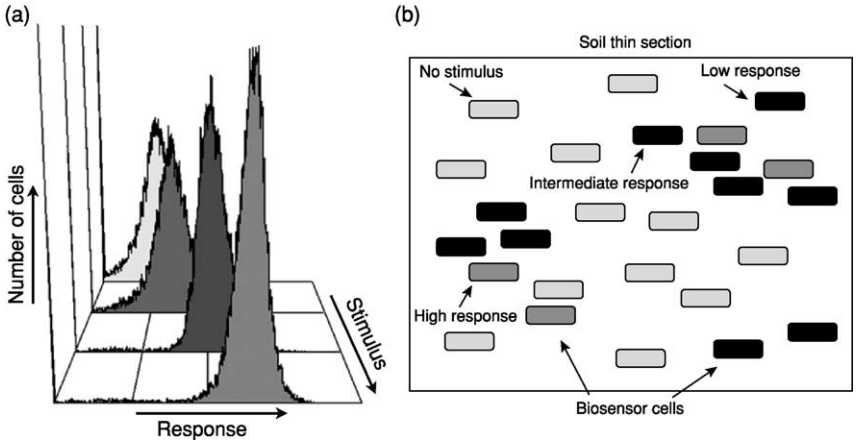


Figure 3 Use of living bacterial biosensors to describe properties of soil microhabitats: (a) response of a population of biosensor cells to a stimulus; (b) *in situ* observation of biosensor cells. For many bioreporters, reporter gene activity is proportional to the magnitude of the stimulus and is expressed as the response over the entire population (a) or on a cell-to-cell basis (b). As indicated by the different degrees of shading, there can be variation in a population response to a particular magnitude of a stimulus (a) or in the biosensors' ability to perceive a stimulus of a particular magnitude (b). Variation in stimulus exposure in a soil is inferred from the biosensors' behavior.

is abundant in desirable nutrients, low in toxic or undesirable compounds, or provides protective sites. Although flagellar motility requires a sufficiently thick water film, which is more common in wetter soils, it is possible that twitching or surface motility may play a role in positioning the cell in a favorable site, particularly when water films are too thin for flagellar motility, which more commonly occurs in unsaturated soils. It is uncertain what distances bacteria could move using this form of motility, but it would be shorter than that covered by flagellar motility.

Once a bacterium reaches a favorable site, the ability to resist removal may be advantageous. The strength of the selective pressure depends on the strength of the removal pressure, such as water percolation following rain, tillage, or movement of a root through soil. Once associated with a surface, bacteria can adhere to it by specific attachment structures such as cellulose fibrils, membrane proteins, pili, or extracellular polysaccharides (EPS). It has been hypothesized that bacteria employ specific adhesins in environments where there are strong physical forces. Due to the improbability of continually strong shear forces in soil, bacteria probably do not use adhesion extensively, except when they occupy sites that offer some protection, or are nutrient-rich, or on roots, because root movement could provide sufficiently strong physical forces to cause their removal.

Stress Tolerance

Bacteria in soil are exposed to a variety of environmental conditions that are going to affect them such as, for example, fluctuations in temperature, water availability, and nutrients, the presence of toxic environmental pollutants, or toxic metabolic wastes produced by themselves or by their neighbors. Survival in a continually changing environment requires a wide range of fast, adaptive responses, some of which do not require transcriptional activation of genes whose products facilitate coping with a given stress, since these responses take too long to complete. For example, modification of fatty acid composition in pseudomonads provides an extremely rapid response to maintain membrane fluidity when they are exposed to stresses such as desiccation or organic solvents that perturb membrane properties. Transcriptional activation of genes whose products facilitate coping with stress provide more long-term solutions. It is not feasible to review bacterial responses to every stress they will potentially encounter in soil. Consequently, reduced water availability is highlighted, and this is likely to be a dominant factor influencing bacteria in many soils; the mechanisms of adaptation to this stress are likely to be similar to cold or heat stress, which are two common stresses bacteria frequently experience in surface soils, since they all alter physical properties of bacterial membranes.

Water Availability

If solutes are abundant in free soil-water, they could become sufficiently concentrated upon drying to stress the resident bacteria. Osmotic stress is only one component of the total water stress that a bacterium may encounter in soil as it dries. The total soil-water potential is a quantitative term reflecting water availability. It is comprised primarily of the sum of the osmotic potential, which is due to the interaction of water with solutes, and the matric potential, which is due to the interaction of water with soil constituents that increases as the soil dries. From a bacterial perspective, the primary difference between these two stresses is that with an osmotic stress bacteria are bathed in water (albeit water with diminished activity), whereas with a matric stress bacteria become desiccated by removal of water from its environment, and the availability of the water that is remaining is reduced due to its sorptive interactions with soil constituents. There is ample evidence demonstrating that the stress imposed by a low matric potential has a stronger influence on a bacterial cell than an equivalent osmotic potential. This is due in part to the greater cellular dehydration that occurs under a matric than an osmotic stress. Dehydration produces deoxyribonucleic acid (DNA) damage, denatures proteins, and causes membranes to become less fluid and potentially damaged. Resistance to dehydration is crucial for microbial life in soil habitats. These include various strategies to counter the life-threatening, lipid-solidifying effects of desiccation, synthesizing solutes compatible with cellular physiology to create an intracellular water potential that is in equilibrium with the external environment, protecting and repairing DNA, and producing EPS.

Extracellular Polysaccharide Production

There is considerable evidence that soil bacteria are capable of producing EPS, and increased EPS production by soil bacteria has been shown to improve soil aggregation and aggregate stability, and water retention properties. These polysaccharides are likely to have a dominant role in the ecology of these organisms, since their presence defines the environment immediately surrounding the bacteria. As such, they probably mediate the interactions of the bacteria with the surface, other microorganisms, and the physical and chemical environment. The EPS matrix that surrounds them can have ion-exchange capabilities, which can concentrate nutrients from dilute sources

in the vicinity of the cell or provide protection from predators and shield cells from the action of lytic enzymes, antibiotics, and other inhibitory compounds. EPS production may be particularly important during desiccating conditions, because many EPS are hydroscopic and their production results in a more hydrated microenvironment in the immediate vicinity of the cell relative to the bulk environment. The greater water content in the immediate vicinity of cell membranes may reduce the physical damage to membrane properties that might occur otherwise. This is an excellent example of how habitat modification can benefit the resident bacterium. If individual cells are capable of producing EPS then the cooperative production of an EPS layer surrounding surface-dwelling bacteria to form a microcolony is likely to provide them with numerous benefits.

Biofilms

Many studies have revealed aggregates of bacteria in bulk soil or on root surfaces. These bacteria are closely packed and covered with an amorphous material that is presumably EPS. This association of aggregates of bacteria in a matrix is analogous to that of biofilms, which are defined as assemblages of bacteria on surfaces encased in EPS of their own making. The concept of biofilms in soil is essentially unused within the soil science and microbiology research community. One reason for this omission may be that biofilms are generally viewed as occurring in aquatic systems and experimentally examined in this context. Another reason is that biofilm growth is generally considered to be continuous (such as on the surface of submerged rocks in a stream) and not patchy like microbial growth in soils. However, by definition, bacteria in soil grow as assemblages on either biotic or abiotic surfaces and are enmeshed in a hydrated matrix of microbially synthesized, extracellular polymeric substances.

Biofilms are not simply organism-containing slime layers on surfaces, but instead represent communities with a high level of organization. A defining feature of a biofilm is the microcolony, which can exhibit a complex, three-dimensional architecture (Figure 4), and cells within the microcolony exhibit gene-expression patterns specific to their position in the microcolony. Biofilms are essentially a community of bacteria that live in a specialized habitat of their own making. Many species have shown distinct developmental steps in biofilm formation, including: (1) attachment to a surface, (2) formation of a microcolony, and (3) maturation of microcolonies into an EPS-encased structure. This implies that biofilm formation may require complex coordinated

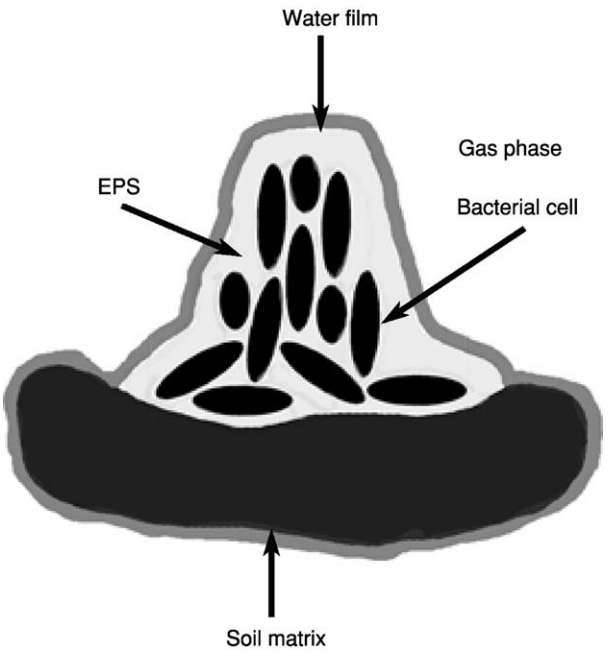


Figure 4 Ultrastructure of a bacterial aggregate or biofilm in an unsaturated soil. Cells are arranged in a structured fashion and are enmeshed in an extracellular polysaccharide (EPS) layer. Soil-water content controls water-film thickness surrounding the biofilm. Nutrients are provided by the underlying matrix that cells are adhered to, from those dissolved in soil water, or dead community members.

communication and behaviors between multiple bacterial cells and species. It is presently unclear whether biofilms or microcolonies in soil form the complex architecture that has been observed in aquatic systems. However, given the close physical proximity and density of bacteria in soil, and the fact that they frequently grow on surfaces, it is likely that they will assume some of the coordinated behaviors and interactions typically associated with biofilms in aquatic systems.

Competition for Resources

Bacteria of different species often occur together in soil and, if these species occupy the same microsites, the survival of any one depends on its ability to compete successfully for shared resources or to coexist by utilizing different resources. These resources can include the same physical site or be nutritionally based, and competition for nutrients is a driving factor in microbial ecology of bacteria in soil, particularly in the rhizosphere. Thus a broad nutrient utilization profile would increase the potential for an organism to survive in the presence of others and the ability to utilize an abundant resource should be highly advantageous as long as other resources are not limiting. Many soil bacteria, and in particular, rhizosphere (the

root surface and the region immediately surrounding a root) colonists have broad nutrient utilization profiles. Likewise, the ability of soil bacteria to tolerate prolonged nutrient deprivation would also be highly advantageous. If microbes must compete for limited resources, production of antimicrobial metabolites, such as an antibiotic, could improve their competitive ability. Many soil bacteria are capable of producing antibiotics that target dissimilar organisms or bacteriocins that target more closely related strains of the same species, and their production would conceivably provide a competitive advantage.

The rhizosphere constitutes an ecologic niche where nutrients are more readily available compared with the bulk soil. The ability to compete effectively for these nutrients is likely to be an important factor contributing to the ability of a bacterium to become a rhizosphere colonist. The deposition of organic material from the roots (rhizodeposition) enhances microbial growth, drives the structuring of microbial communities, and controls their metabolic activities in the rhizosphere. The interactions between the plant and the surrounding bacteria in the bulk soil select for the establishment of certain populations (rhizobacteria). While rhizodeposition strongly influences the size and activity of microbial populations at the soil–plant interface, the activity of these populations in turn affects plant health and thus influences both the quantity and quality of rhizodeposition. The potential for either an exudation response to bacteria or a response by bacteria to exudation suggests a certain degree of coevolution between plants and rhizobacteria.

Cooperation Among Bacteria

The ability to modify a habitat may be augmented by cooperative interactions among bacteria and these interactions may occur among both homogeneous and heterogeneous populations. The occurrence of aggregated populations of bacteria in soil or in association with plant roots provides an opportunity to maximize opportunities for interactions among community members. There have been several recent descriptions illustrating sophisticated cooperation among community members, even though at times they may appear not to be cooperative attributes.

Siderophores

Although iron is abundant, the extreme insolubility of ferric hydroxide limits the amount of free iron available in aerobic soil environments, since the iron exists in the form of insoluble oxides in neutral or alkaline pH soils. Many bacteria use siderophores and corresponding membrane receptors to acquire this essential nutrient. Siderophores, which are low-molecular-weight molecules produced under iron-limiting conditions, bind the ferric iron, the membrane receptor recognizes a particular siderophore–iron complex, and the iron is then transported into the cell. A particular bacterial species or strain generally produces one or more siderophores and the corresponding membrane receptors for transporting each siderophore–iron complex into the cell. The bacterium releases the siderophore into the environment and then the siderophore needs to chelate ferric iron and diffuse back to the cell for the cell to benefit from the energy spent to synthesize the siderophore. Yet many soil bacterial species have the capacity to utilize siderophores produced by other bacterial strains or species and possibly fungi. This cross-feeding clearly provides advantages to community members, since there would be a greater potential return on the metabolic investment of siderophore production by taking up ferric–siderophore complexes that it did not synthesize rather than depending solely on diffusion of the siderophore it produced back into the cell. This cooperativity is not surprising given that the entire microbial community is probably experiencing some level of iron deficiency, although there may be significant microsite heterogeneity in iron availability. This also illustrates that bacterial survival may not depend solely on the efforts of the individual but also, in part, on the efforts of the community.

Quorum-Sensing

Bacteria can communicate with one another using chemical signaling molecules as words. Specifically, they release, detect, and respond to the accumulation of molecules called autoinducers. Detection of autoinducers allows bacteria to distinguish between low and high cell population density, and to control gene expression in response to changes in cell number (Figure 5). In reality, this is not necessarily a function of high cell density, but is likely to be related more directly to high signal concentrations, which are likely to be achieved at lower population densities in soil due to their low water volumes that result in relatively high autoinducer concentrations. This process, termed 'quorum-sensing,' allows a population of bacteria to coordinately control gene expression of the entire community and allows the bacteria to behave as multicellular organisms and to reap the benefits that would be unattainable to them as individuals. Recent studies show that highly specific as well as universal quorum-sensing signaling systems exist which enable bacteria to communicate within and between species, and that more frequently this behavior is associated with bacteria that have been isolated from plant roots than from bulk soil. Many microbial

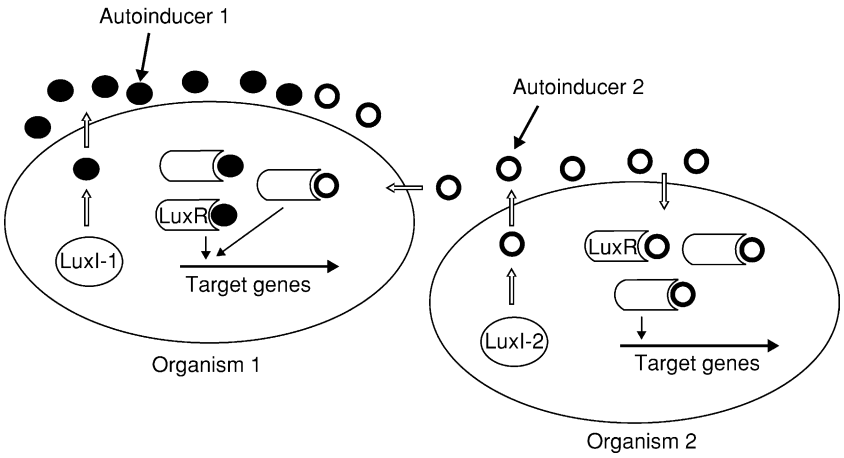


Figure 5 The LuxI/LuxR quorum-sensing signal transduction system. Two regulatory proteins control quorum-sensing. The LuxI-like proteins are the autoinducer synthases (*N*-acyl-homoserine lactones), which can freely diffuse through the cell membrane and accumulate in high density. At a particular intracellular concentration of the autoinducer, it will bind to the LuxR family of proteins that can then interact with target gene promoters to activate transcription. It is possible for a particular LuxR-family protein to bind to various autoinducers (produced by different species) that differ slightly chemically, although at different efficiencies, and still be able to activate transcription. The potential for cross-talk between two bacteria is demonstrated, because transcriptional activation of target genes in organism 1 can be mediated by autoinducer 1 and autoinducer 2 interacting with LuxR.

behaviors are regulated by quorum-sensing, including, for example, expression of nodulation genes in the *Rhizobium*–legume symbioses, virulence traits of bacterial pathogens of plants, conjugal transfer of plasmids between bacteria, antibiotic production, and biofilm formation. Bacterial biosensors using fluorescence-based reporter gene products have been particularly useful for visualizing cell–cell communication between different bacteria colonizing a plant rhizosphere. These studies illustrate the power and future of reporter-gene technology in microbial ecology research, and future studies, utilizing a spectrum of different-colored fluorescent proteins as bioreporters, will provide a more detailed picture of microbial behavior, habitat conditions, and interactions among community members.

Many bacteria have also developed mechanisms to interfere with bacterial quorum sensing, such as secreting enzymes that destroy autoinducers or producing autoinducer antagonists. Furthermore, some plants secrete substances that mimic or interfere with *N*-acyl-homoserine lactone autoinducer signal activities and affect population density-dependent behaviors in associated bacteria. These autoinducer signal-mimic compounds could prove to be important in determining the outcome of interactions between higher plants and a diversity of pathogenic, symbiotic, and saprophytic bacteria. Furthermore, some bacteria are capable of recognizing autoinducers produced by different bacterial species, indicating that interspecies quorum-sensing may play an important synergistic or competitive role in the dynamic behaviors of soil bacterial communities (Figure 5).

Death

One of the least understood aspects of bacterial life processes in soil is death. Clearly, the lack of sufficient types or amounts of nutrients, exposure to severe environmental stresses, or virus infections could lead to the death of individual cells, populations, or entire communities. The simplest examples of programmed cell death in soil bacteria are lysis of the mother cell during sporulation of bacilli and vegetative cells during myxobacterial fruiting-body formation. In these cases nutrient deprivation triggers developmental responses initiating spore formation, because spores are better suited than vegetative cells for surviving difficult times in soil. Furthermore, there is increasing evidence suggesting that bacteria have mechanisms to kill defective or damaged cells. From this perspective, programmed death of damaged cells may be beneficial to soil bacterial communities by releasing nutrients from dying cells to be used by neighbors, to provide a structural role during complex developmental cycles, such as during fruiting-body or biofilm formation, or to prevent spread of viral infections.

Summary

The ability of bacteria to survive in soil with its inherent complexity and heterogeneity under conditions that are often severe is a testament to their metabolic versatility and phenotypic plasticity (adaptability). Strategies for growth and survival in soil vary from transcriptional regulation of single genes to cooperative interactions among community members, and most likely a combination of both.

Many bacteria are able to modify their local environment, such as by producing EPS, to facilitate growth, and their increased numbers facilitate further habitat modification. Continued multiplication results in the formation of a microcolony (biofilm) that may be homogeneous if it contains progeny from only one cell or heterogeneous if multiple species are incorporated into it. Cooperative and coordinated behaviors are required for the development of such a community of bacteria as well as for maintaining the vitality and health of the community. Each individual member of the community, through its own actions, based in part on the signals it receives from other community members, initiates a spectrum of adaptive strategies that are appropriate for the particular habitat they reside in and the specific environmental challenges they are experiencing to optimize chances for survival.

See also: **Archaea**; **Bacteriophage**; **Fungi**; **Microbial Processes:** Environmental Factors; Community Analysis; Kinetics

Further Reading

Ashelford KE, Day MJ, and Fry JC (2003) Elevated abundance of bacteriophage infecting bacteria in soil. *Applied Environmental Microbiology* 69: 285–289.

Foster RC, Rovira AD, and Cock TW (1983) *Ultrastructure of the Root–Soil Interface*. St. Paul, MN: American Phytopathological Society.

Killham K (1994) *Soil Ecology*. Cambridge, UK: Cambridge University Press.

Koch B, Worm J, Jensen LE, Højberg O, and Nybroe O (2001) Carbon limitation induces σ^S-dependent gene expression in *Pseudomonas fluorescens* in soil. *Applied Environmental Microbiology* 67: 3363–3370.

Leveau JHJ and Lindow SE (2002) Bioreporters in microbial ecology. *Current Opinion in Microbiology* 5: 259–265.

O'Toole G, Kaplan GB, and Kolter R (2000) Biofilm formation as microbial development. *Annual Review of Microbiology* 54: 49–79.

Roberson EB and Firestone MK (1992) Relationship between desiccation and exopolysaccharide production in a soil *Pseudomonas* sp. *Applied Environmental Microbiology* 58: 1284–1291.

Torsvik V and Øvreås L (2002) Microbial diversity and function in soil: from genes to ecosystems. *Current Opinion in Microbiology* 5: 240–245.

Vogel J, Normand P, Thioulouse J, Nesme X, and Grundmann GL (2003) Relationship between spatial and genetic distance in *Agrobacterium* spp. in 1 cubic centimeter of soil. *Applied Environmental Microbiology* 69: 1482–1487.

Wisniewski-Dye F and Downie JA (2002) Quorum-sensing in *Rhizobium*. *Antonie van Leeuwenhoek* 81: 397–407.

BACTERIOPHAGE

M Radosevich, University of Tennessee, Knoxville, TN, USA
K E Williamson and K E Wommack, University of Delaware, Newark, DE, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Despite the importance of bacteriophage as model systems in genetics, until recently little attention was given to natural populations of bacteriophage. The observation that viruses commonly outnumber bacteria by a factor of 10 in aquatic environments revealed our limited understanding of the larger role bacteriophage play in the evolution, population biology, and ecology of bacteria. Even fewer studies have been conducted in soil ecosystems, and the overwhelming majority of these have focused on culturing bacteriophage of specific host bacteria.

Soil is the most biologically diverse ecosystem on Earth, with estimates of the number of bacterial species on the order of 5000–10 000 g^{-1} of surface soil. If we assume that the virus-to-bacteria ratio (VBR) of 10:1 that appears to be relatively constant across aquatic environments extends to the soil ecosystem, then there may be as many as 50 000–100 000 different viral taxa in 1 g of soil. The uncharacterized genetic diversity of soil bacteriophage communities is by extension simply staggering. Recent studies in aquatic environments and in rhizosphere soil suggest that viral infection may play a very critical role in shaping the structure and function of microbial communities. The literature is replete with investigations of the fate and transport of model bacteriophage of potential human pathogens and other nonindigenous viruses in porous media. There are relatively few culture-independent and model studies in bacteriophage ecology in soils, but parallels can be drawn from the fascinating findings that are more regularly emerging in aquatic ecosystems, and which raise major research questions.

Why Soil Bacteriophage?

Perhaps the most compelling reason to improve our understanding of soil bacteriophage ecology is that

infection of host bacterial communities by bacteriophages is a fundamental driver of host abundance, function, and diversity. Bacteriophage can affect host biology via three distinct mechanisms: (1) host mortality, (2) phage conversion, and (3) horizontal gene transfer by transduction. These three processes have serious implications on the natural ecologic function of microbial communities in native and managed ecosystems (e.g., nutrient cycling, control and/or abundance of plant pathogens, ecologic impact of genetically engineered organisms, and spread of antibiotic resistance genes) and bioremediation of contaminated soils (e.g., survivability of introduced microbial degraders and dissemination of catabolic genes). From a more applied perspective, a better understanding of bacteriophage–host interactions in natural ecosystems may provide an opportunity to develop bacteriophage as biocontrol agents of plant and animal pathogens.

Ecologic Role of Viruses in Microbial Communities

Bacteriophage Contain Genetic Elements that Can Increase Host Fitness: Bacteriophage Conversion

Through years of intensive study, it has been discovered that bacteriophage are an integral component in the survival and fitness of populations of pathogenic microorganisms. The wealth of evidence supporting this broad claim is too extensive to be provided here; however, a few examples are particularly illustrative of the possible role of viruses within soil microbial communities, as is the case for several pathogens, notably *Corynebacterium diphtheriae*, *Vibrio cholerae*, and enterohemorrhagic *Escherichia coli*. The critical element of the virulence of these pathogens is a toxic gene product carried on the genome of a temperate bacteriophage. Thus, only lysogenic strains of the pathogen gain the fitness-improving capability of invading and replicating within a host. It is possible that, like communities of pathogenic bacteria and their respective bacteriophage, intimate genetic relationships exist in which temperate bacteriophages confer advantageous phenotypes upon soil host populations. An example relationship between a bacteriophage and a phytopathogen is that of *Clavibacter toxicus*, the likely causative agent in ryegrass gumming disease. Strong circumstantial evidence indicates that bacteriophage infection of *C. toxicus* is important to the etiology of this disease. Acquisition of virulence determinants by disease-causing bacteria can be associated with transduction and lysogenic conversion and is a noteworthy reminder of the power of bacteriophage-mediated horizontal gene transfer. Additionally, bacteriophage and other genetic elements carry genes useful in the survival of host bacteria. Perhaps the best example is antibiotic and metal resistance. Given the collective example of pathogenic bacteria and the acknowledged role of bacteriophage in enhancing their survival and fitness, it seems appropriate to ask whether analogous systems exist between viruses and bacteria in soil.

The Role of Phage in Horizontal Gene Transfer and Bacterial Evolution

In the bacterial world, genetic recombination appears to be a relatively rare event in which a small amount of genetic material (one or a few alleles) is exchanged between two species. This occurs without the requirement that the two species be close genetic relatives. Because bacterial populations in active growth may exhibit short generation times and fast growth rates, even rare genetic recombination events can quickly alter (increase or decrease) the diversity of genotypes within a community. The mechanisms by which genetic exchange occurs in bacterial populations are well understood and have become valuable tools for microbial geneticists. However, these same processes that provide the technology selectively to engineer new bacterial phenotypes also provide the means by which genetic material can be introduced into the multitude of genetic backgrounds in natural environments. Concern about release of genetically engineered microorganisms (GEMs) continues to stimulate examination of genetic exchange processes. Whole-genome sequencing of bacteria has revealed that lateral transfer may also be an important factor in the evolution of prokaryotic genomes. Transduction (bacteriophage-mediated gene transfer) begins as a serendipitous event, in which the host deoxyribonucleic acid (DNA) (plasmid or chromosomal) is mistakenly packaged into the capsid during production of progeny bacteriophage particles. Two types of transduction are possible, generalized and specialized, according to the phenotype of the bacteriophage mediating the genetic transfer. Generalized transducing bacteriophage can carry, usually with equal frequency, any DNA contained within the host cell regardless of location, since integration of the bacteriophage genome into that of the host is entirely random. Both temperate and virulent bacteriophage can perform generalized transduction and these bacteriophage share common attributes with respect to bacteriophage genome replication and DNA packaging. Conversely, specialized transducing bacteriophage are solely temperate bacteriophage which integrate within the host chromosome and

are capable of transferring only specific chromosomal genes located close to the bacteriophage integration site. Because of its indiscriminant nature, generalized transduction provides the means through which large amounts of genetic material, i.e., approximately equal to the size of the transducing bacteriophage genome, can be passed between two bacterial cells.

Although transduction has generally shown the lowest frequencies of gene transfer, it has been speculated that the bacteriophage-as-carrier aspect of this form of horizontal gene transfer (HGT) provides specific advantages in that transducing bacteriophage particles can be long-lived in the environment, provide for both protection of DNA and dispersal of genetic material, and eliminate the requirement for cell-to-cell contact. The realization that bacteriophage are abundant within many environments and data showing high transduction frequency in marine systems indicate that transduction may be an underappreciated means of HGT in bacterial populations.

Some of the best evidence for the involvement of HGT in bacterial evolution has come from detailed examination of complete bacterial genomes. Properties such as the codon usage bias and G + C content of a gene serve as signatures identifying those genes that have been introduced into a new genetic background. Using such criteria it has been determined that more than 12% of the *E. coli* genome is probably transferred from other bacteria at the rate of 31 kbp per million years. The mechanisms responsible for the movement of genes between genomes, however, do not occur slowly over the course of millions of years, but are sometimes random and sometimes directed events which occur with frequencies on the level of one per 10^4–10^6 generations. As a testament to the occurrence of horizontal transfer, elements which probably cause the transposition, such as transposons, plasmid transfer origins, or bacteriophage attachment sites (*att* sites) are often found bordering presumed transferred genes. The process of HGT is probably a common event for most groups of microorganisms, but establishment of new genes within a genome is subject to the more discriminating and deliberate force of natural selection. None the less, all gene transfer must begin with a singular event of transfer (i.e., conjugation, transformation, or transduction). The collective rates of these processes ultimately determine the appearance of new genotypes on which natural selection will ultimately act.

Evidence for Natural Transduction in Microbial Communities

Several studies have shown the potential for transduction by inoculating soil microcosms with bacteriophage and an appropriate host. The transduction of chloramphenicol and mercury resistance genes into *E. coli* has been shown, using specialized transducing derivatives of bacteriophage P1. Transfer occurs in both sterile and nonsterile soil, and when lysogenic bacteria and recipient host cells are added to soil. Transduced *E. coli* are lysogenic for phage P1 by hybridization with a DNA probe derived from the bacteriophage. Although *E. coli* is a nonindigenous host bacterium, there is potential for transduction to play an important role in gene transfer in soil.

While the potential for transduction is clearly present in soil communities, it should be stressed that culture-independent assessment of transduction has not been demonstrated in the soil ecosystem. However, we draw on several elegant studies in aquatic ecosystems as evidence illustrating the importance of transduction as a viable mechanism of HGT and driving force of host diversity in natural microbial populations. Bacteriophage DNA probes derived from several bacteriophage of *Pseudomonas aeruginosa* are used to detect the presence of prophage (lysogenized bacteria) in natural samples of lake water, sediments, and sewage. Hybridization studies have shown that 40% of *P. aeruginosa* in lake water samples contain sequences homologous to bacteriophage DNA, thus indicating that bacteriophage are widely distributed in the environment and may play a significant role in natural gene transfer by transduction or lysogenic conversion.

Streptomycin resistance has been transferred by a generalized transducing phage (F116) to *P. aeruginosa* in lake-water microcosms incubated *in situ*. Transduction frequencies range from 1.4×10^{-5} to 8.3×10^{-2} transductants per recipient and are comparable with laboratory-determined transduction frequencies. A model describing how selection and transduction affect the establishment of a new phenotype has been developed and validated in a bacterial population. In nonsterile aquatic microcosms in which transduction is prevented, the density of inoculated mock transductants (cells of a strain which possess the transductant phenotype but are not lysogenic for F116) decline over time. When transduction is permitted, the number of transductants increases suggesting that transduction could, in a diverse microbial population, maintain a phenotype that would otherwise be lost. The results of this and other studies clearly show the potential for genes from lysogens to be maintained as part of the community gene pool through transduction even if specific strains with similar phenotypes do not survive and proliferate. Some studies have shown as much as a 100-fold-higher transduction frequency in the presence of solid particle surfaces, suggesting that aggregation to soil particles could enhance contact rates

between bacteriophage and host cells, thus increasing potential for infection.

If introducing bacteriophage DNA confers a selective advantage to the lysogen, such as antibiotic resistance, heavy metal resistance, ability to produce a toxin, or the ability to degrade a xenobiotic, then infection may result in enhanced host competitiveness. The long-term survival of a derivative of the temperate actinophage ØC31, called 'KC301,' which carries a thiostrepton resistance gene and confers resistance to the antibiotic to infected bacteria, has been investigated. KC301 was introduced as a lysogen of *Streptomyces lividans* TK24. The lysogen reached a population density slightly lower than the nonlysogenized counterpart. Addition of thiostrepton did not enhance survival or proliferation of KC301, but the antibiotic concentration was low and may not have provided a sufficient selective pressure to enhance the competitiveness of the lysogen. No comparable studies have been conducted in systems where xenobiotic degradation was used as the selective pressure for maintenance of lysogen-contaminated soils.

Cultivation-Independent Investigation of Soil Viral Communities

Microscopy of Soil Viruses

Bacteriophage can be found in any environment occupied by a suitable host; even in hosts that have been starved for long periods of time. Phage have been isolated from soil using specific enrichment with a variety of soil bacteria, including *Arthrobacter*, *Bacillus*, *Nocardia*, *Pseudomonas*, *Rhodococcus*, and *Streptomyces*, suggesting that phage are common inhabitants of soil. Early measurements of phage abundance in natural environments were quite low; however, these measurements were made using plaque formation in specific host bacteria and were not representative of the total bacteriophage populations. More recently, microscopic methods have been used to enumerate bacteriophage in environmental samples and the counts have increased by several orders of magnitude from previous determinations.

Major breakthroughs in our understanding of microbial ecology have come through technical improvements in direct enumeration of microorganisms. In the 1970s, application of epifluorescent light microscopy (EFM) to enumeration of bacteria revealed that natural abundance of this group in water and soil samples exceeds that obtained through cultivation-based techniques by 100- to 1000-fold. This realization has been pivotal in justifying the search for cultivation-independent approaches to the study of bacterial diversity such as the array of molecular genetic techniques based on small-subunit ribosomal ribonucleic acid (SSU rRNA). In addition, accurate estimation of bacterial abundance has led to dramatic improvements in understanding carbon and energy flow through natural ecosystems. The concept of the 'microbial loop,' in which a significant proportion of primary production, which would otherwise be lost to higher-order consumers, is incorporated into the bacterioplankton is supported by more recent knowledge of the high abundance of bacteria in many marine systems.

In comparison with the evolution of bacterial direct-counting approaches, the discovery of abundant viral populations within water samples was a dramatic, serendipitous event. Francisco Torrella and Richard Morita were the first directly to observe virus particles using transmission electron microscopy (TEM). In 0.2-μm filtered water samples, *c.* 10^4 virus-like particles (VLPs) per milliliter of seawater were observed; however, they felt natural abundance was probably higher, as prefiltration may have removed VLPs. A decade later, in a project originally designed to quantify the elemental composition of marine bacteria using TEM, extraordinarily high abundances of VLPs in water samples (*c.* 10^7–10^8 ml^{-1}) from a variety of marine environments were observed. In general, this and other early studies have demonstrated that viral abundance typically exceeds co-occurring host abundance by 10-fold; thus viruses are now widely acknowledged to be the most abundant members of marine microbial communities. The role of viruses and viral infection in influencing the composition, diversity, and productivity of marine microbial communities continues to be an open area of investigation.

TEM provided the earliest evidence of high viral abundance in water samples and continues to be the standard through which all succeeding EFM-based enumeration techniques are judged. The principal advantage of TEM-based enumeration is the provision of morphological data (e.g., capsid size and head–tail morphology) that confirms positive identification of a virus particle and enables additional characterization of a viral community. However, in comparison with epifluorescence techniques, TEM is difficult with samples that contain even a small proportion of particulate material and requires more time for sample preparation and observation. Several nucleic acid-binding fluorescent stains have been employed for direct EFM enumeration of viruses within water samples, including 4′-6′-diamidino-2-phenylindole (DAPI), YoPro, and the SYBR stains. Of these, the SYBR stains appear to be the most widely accepted due to ease of staining, exceptional brightness, and

low nonspecific binding and fluorescence. Because EFM provides no morphological detail, counts using these techniques are reported as VLPs. Numerous comparative studies have shown that TEM virus counts are similar, but generally lower than counts of VLPs using EFM. The possible interference of particulates in TEM counting and nonviral, stain-positive particles in EFM are hypothesized to result in under- and overestimates of these two approaches, respectively. Nevertheless, the ease and relative precision of EFM techniques have resulted in widespread application of these approaches to viral enumeration.

In comparison with aquatic environments, there are only two reports of cultivation-independent observation or enumeration of indigenous viruses within soils. TEM has been used to show that there is an abundance of virus particles within British agricultural soils ($1.5 \times 10^7\,g^{-1}$). Corresponding bacterial direct counts, estimated by EFM and acridine orange staining, are *c.* 200-fold higher. However, control experiments indicate a 40-fold loss of bacteriophage particles to binding with the soil matrix. Adjusting for extraction efficiency, actual viral abundance is estimated to be 10-fold higher, yielding a of 0.04. This low VBR estimate for soil samples is among the lowest reported by direct counting.

EFM using SYBR Gold staining has been a robust means of estimating VLP abundance in 0.2-μm filtered extracts of Delaware agricultural soils. In an investigation of suitable methods for extraction of autochthonous viruses from soils, a grand mean of 4.3×10^8 VLP g^{-1} dry soil has been observed using EFM. Brightly staining particles are almost entirely encapsulated double-stranded DNA (dsDNA) because treatment with heat followed by DNase digestion completely eliminates VLPs (Figure 1). Treatment with DNase alone removes a statistically insignificant number of particles. While the DNase experiment demonstrates that SYBR Gold-positive particles contain dsDNA, conclusive evidence that VLPs seen using EFM are actually viruses requires TEM examination of soil extracts. Initial attempts to examine directly 0.2-μm filtered soil extracts using TEM have failed owing to excessive interference from particulate matter. Subsequently, viral extracts of soil are purified by CsCl gradient centrifugation prior to TEM. With this combination of techniques (extraction, 0.2-μm filtration, and CsCl gradient centrifugation), a morphologically diverse range of virus particles have been detected in soil samples (Figure 2). TEM examinations of both Matapeake and Evesboro soil samples indicate a viral abundance of $1.5 \times 10^8\,g^{-1}$ dry soil (grand mean). Both EFM and TEM direct counts of VLPs in Delaware soil samples exceed those reported from British soils by 10 and 28 times; however, the significantly greater abundance of viruses in Delaware soils is probably due to methodological differences in extraction of virus particles.

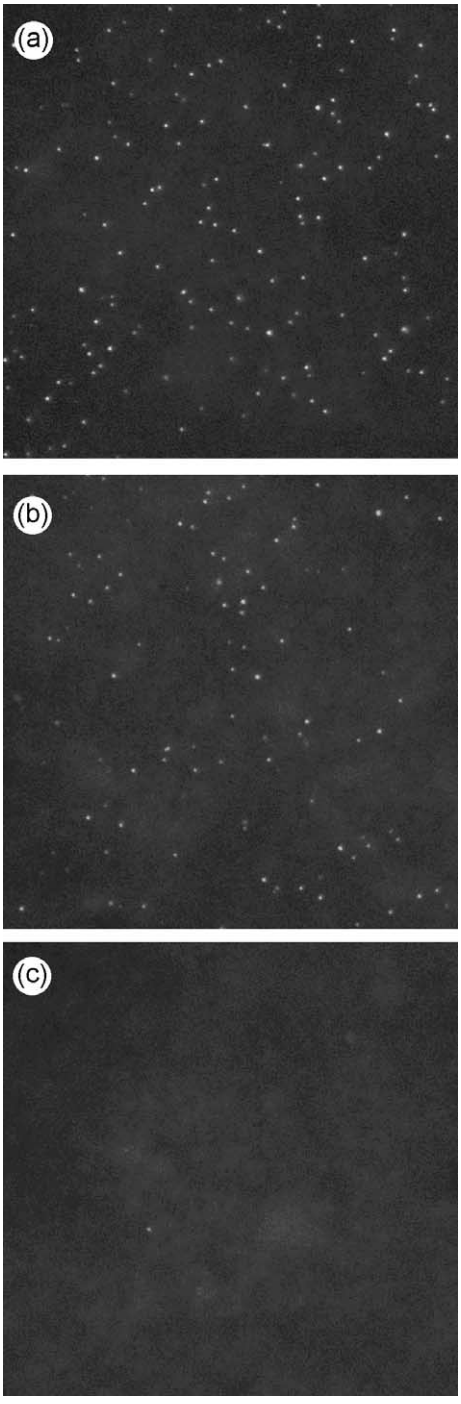


Figure 1 Epifluorescence micrographs of soil viruses stained with SYBR Gold: (a) untreated control; (b) pretreatment with DNase (1 h, 20°C); (c) pretreatment with heat (95°C, 20 min) then DNase.

TEM-based abundance is approximately five times lower than corresponding SYBR Gold epifluorescence counts for Delaware soils. The tendency of

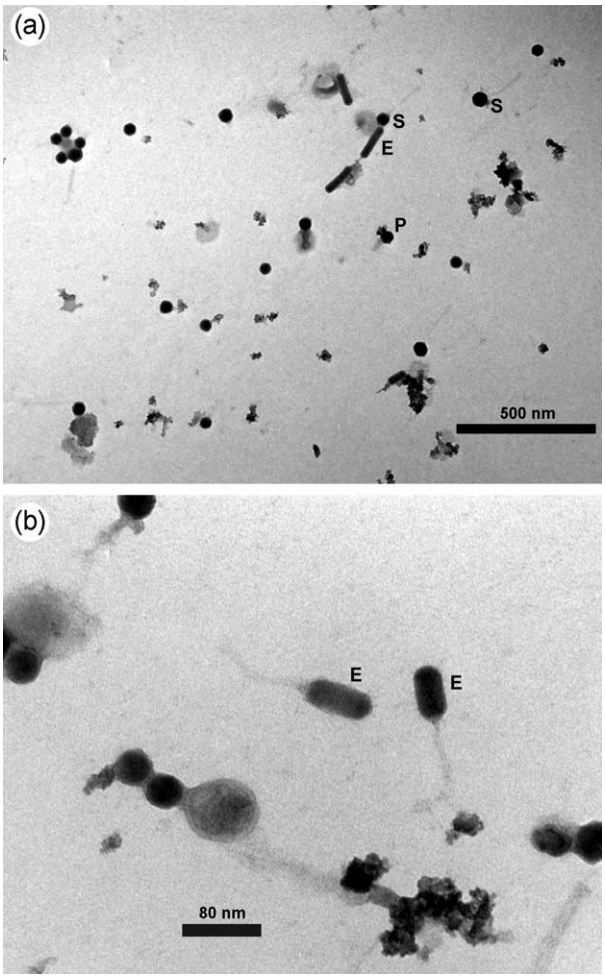


Figure 2 Transmission electron micrographs of virus-like particles from Matapeake soil: (a) ×50 000; (b) ×85 000. S, Syphoviridae; P, Podoviridae; E, elongate capsid phage.

EFM counts significantly to exceed TEM counts of virus particles is well-known for water and sediment samples and thus it is not surprising that this trend has been observed in Delaware agricultural soils. Approximately 65% of bacteriophage observed in Matapeake soil samples are tailed (49% short tails: Podoviridae; and 16% long tails: Myoviridae and Siphoviridae), and capsid diameters range from 27 to 114 nm, with a mean value of 49.6 nm ($N = 260$; from two sampling dates). Similar to aquatic viral communities in Chesapeake Bay, soil samples show a predominance of short-tailed bacteriophages as well as capsid diameters from 40 to 60 nm. Nearly 10% of soil viruses have elongated capsids (mostly morphotypes A3 (Myoviridae) and B3 (Siphoviridae)) (Figure 2). It is unlikely that the presence of elongate viruses has ever been documented for aquatic environments. Elongated morphotypes (A3, B3, and C3) comprise less than 2% of known phages, thus the observation of relatively abundant elongate bacteriophage is striking. Is it possible that these elongate viruses are specific to soils? And that such elongated capsids contain a relatively large (more than 200 kb) viral genome? This preliminary information is tantalizing evidence that soils contain morphologically diverse virus populations that differ dramatically from those occurring in better-studied, aquatic environments.

Genomic Approaches to Viral Diversity

While direct-counting approaches are an essential component to ecologic studies of microorganisms, true appreciation of extant diversity within microbial communities can only come through application of molecular genetic tools at the gene level. Approaches to studying prokaryotic diversity based on the sequence of the SSU (16S) rRNA have been critical to the formation of a universal taxonomy. In addition, techniques such as denaturing gradient gel electrophoresis (DGGE), terminal restriction fragment length polymorphism (TRFLP), and automated ribosomal intergenic spacer analysis (ARISA) applied to 16S rDNA polymerase chain reaction (PCR)-amplified gene products reveal a molecular fingerprint of prokaryotic diversity which can, in some cases, be linked to species composition. Application of molecular-fingerprinting tools to analysis of viral communities is somewhat more difficult in that universal genetic markers such as 16S rDNA do not exist for viruses. Nevertheless, certain subgroups of viruses, such as phycoviruses and cyanophages, have been examined by DGGE and TRFLP using genes (e.g., DNA polymerase, and a capsid gene) conserved among these viruses.

With the increasing availability of high-throughput DNA sequencing, it is now possible to explore, through shotgun sequencing, the genetic composition of whole microbial populations. Recently, such metagenomic approaches to the study of marine microorganisms have revealed entirely novel physiologies (i.e., phototrophy) within the pelagic ocean. Among all molecular genetic approaches used in microbial ecology, high-throughput shotgun sequencing of entire microbial communities holds the greatest promise for exploration of the vast genotypic diversity of prokaryotes. The conviction that metagenomics is the key to deeper understanding of the role of oceanic microbial communities in the global carbon cycle has led the US Department of Energy to fund a 3-year initiative to obtain the entire genetic sequence of microorganisms inhabiting the Sargasso Sea. Application of a metagenomics approach to characterization of marine viral communities demonstrates that the vast majority of viral genetic diversity is unknown. Early indications from an investigation of a Delaware agricultural soil indicate that soil

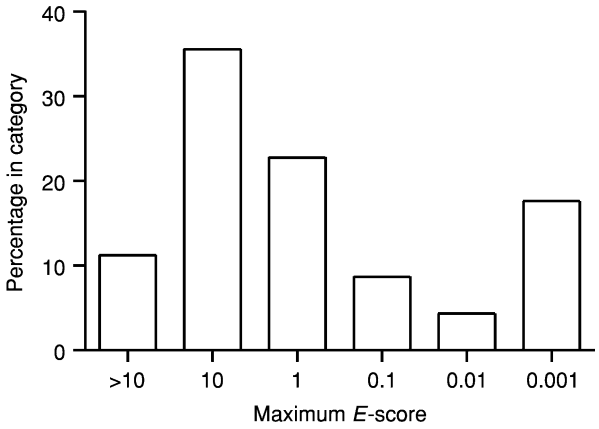


Figure 3 Distribution of BLASTX results for a soil virus metagenome library according to expectation value (*E*). Categories 10–0.1 are not significant and greater than 10 is considered 'no result.'

viruses may contain the greatest amount of unknown sequence (i.e., lacking significant homology to known sequence) of any microbial source investigated prior to 2003. Moreover, this soil-virus metagenomic library contains a larger number of sequences than any previously reported viral metagenome library. The library has been constructed from *c.* 10^9 soil VLPs using PCR-based approach for wholesale and unbiased amplification of viral DNA. After analysis of the metagenome library using PHRED/PHRAP, 9589 sequences were compared with the GenBank database using BLASTX and a minimum *E*-value of 10. Of these sequences 78% have no significant homology to known sequence at a permissive *E*-value of greater than 0.01 (Figure 3). Among sequences with significant hits ($E < 0.01$), only 33% align with known viral species, while most of those remaining align with bacterial sequences. Using a stricter criterion of $E < 0.001$, a viral metagenome library from California coastal seawater contains only 65% unknown sequence, with *c.* 34% of sequences aligning with viral sequences.

Given the prevalence of unknown sequence within metagenome libraries, it seems at first glance that this approach will offer little information concerning viral diversity. However, as more viral genomes and metagenomic sequence data become available, it is probable that significant homologies between unknown viral sequences will arise. In a way loosely similar to the discovery of new clades of bacteria through divergent SSU rRNA sequence, homologies between unknown viral sequences may lead to new genetic markers for exploration of viral evolution, taxonomy, and diversity.

Summary

In comparison with marine environments, the distribution, abundance, and ecology of soil viruses has not been thoroughly explored, yet the few initial studies available suggest that viruses are abundant members of soil ecosystems and probably play a significant role in modulating the structure, function, and productivity of soil communities. As seen in marine ecosystems, mortality of microorganisms and potentially higher organisms induced by viral infection has the potential significantly to affect the flow of energy and the biogeochemical cycling of nutrients through the soil food web. Viral transduction among indigenous host bacteria has not been definitively demonstrated in soil ecosystems but probably plays a significant role in the evolution of soil bacteria. Lack of progress in understanding these processes and the overall diversity of soil viruses has largely occurred owing to methodological barriers associated with the heterogeneous nature of the soil matrix and the general difficulties related to culture-independent surveys of biological diversity. As these barriers are broken, future research endeavors will undoubtedly discover the significance of viruses in the soil ecosystem.

List of Technical Nomenclature

EFM	Epifluoresence microscopy
HGT	Horizontal gene transfer. The transfer of genetic information from a donor cell to a recipient cell by transduction, transformation, or conjugation
TEM	Transmission electron microscopy
VLP	Virus-like particle

See also: **Biodiversity**; **Microbial Processes:** Environmental Factors; Community Analysis; Kinetics

Further Reading

Ackermann HW (1996) Frequency of morphological phage descriptions in 1995. *Archives of Virology* 141(2): 209–218.

Ashelford KE, Day MJ, and Fry JC (2003) Elevated abundance of bacteriophage infecting bacteria in soil. *Applied Environmental Microbiology* 69: 285–289.

Barksdale L and Arden SB (1974) Persisting bacteriophage infections, lysogeny, and phage conversions. *Annual Review of Microbiology* 28: 265–299.

Bergh O, Borsheim KY, Bratbak G, and Heldal M (1989) High abundance of viruses found in aquatic environments. *Nature* 340: 467–468.

Bratbak G, Heldal M, Naess A, and Roeggen T (1993) Viral impact on microbial communities. In: Guerrero R

and Pedros-Alio C (eds) *Trends in Microbial Ecology*, pp. 299–302. Barcelona, Spain: Spanish Society for Microbiology.

Breitbart M, Salamon P, Andresen B *et al.* (2002) Genomic analysis of uncultured marine viral communities. *Proceedings of the National Academy of Science of the USA* 99(22): 14250–14255.

Danovaro R, Dell'anno A, Trucco A, Serresi M, and Vanucci S (2001) Determination of virus abundance in marine sediments. *Applied Environmental Microbiology* 67(3): 1384–1387.

Jin Y and Flurry M (2002) Fate and transport of viruses in porous media. In: Sparks DL (ed.) *Advances in Agronomy*, vol. 77, pp. 39–102. Amsterdam, the Netherlands: Academic Press.

Kepner RL Jr and Pratt JR (1994) Use of fluorochromes for direct enumeration of total bacteria in environmental samples: past and present. *Microbiological Reviews* 58(4): 603–615.

Kokjohn TA (1989) Transduction: mechanism and potential for gene transfer in the environment. In: Levy RV (ed.) *Gene Transfer in the Environment*, pp. 73–93. New York: McGraw-Hill.

Noble RT (2001) Enumeration of viruses. *Methods in Microbiology* 30: 43–50.

Paul JH and Kellogg CA (2000) Ecology of bacteriophages in nature. In: Hurst CJ (ed.) *Viral Ecology*, pp. 211–246. San Diego, CA: Academic Press.

Suttle CA (2000) Ecological, evolutionary, and geochemical consequences of viral infection of cyanobacteria and eukaryotic algae. In: Hurst CJ (ed.) *Viral Ecology*, pp. 248–296. San Diego, CA: Academic Press.

Williams ST, Mortimer AM, and Manchester L (1987) Ecology of soil bacteriophages. In: Goyl SM, Gerba CP, and Bitton G (eds) *Phage Ecology*, pp. 157–179. New York: John Wiley.

Woese CR (1987) Bacterial evolution. *Microbiological Reviews* 51(2): 221–271.

Wommack KE and Colwell RR (2000) Virioplankton: viruses in aquatic ecosystems. *Microbiology and Molecular Biology Reviews* 64: 69–114.

Zhong Y, Chen F, Wilhelm SW, Poorvin L, and Hodson RE (2002) Phylogenetic diversity of marine cyanophage isolates and natural virus communities revealed by sequences of viral capsid assembly protein g20. *Applied Environmental Microbiology* 68(4): 1576–1584.

BIOCONTROL OF SOIL-BORNE PLANT DISEASES

C E Pankhurst, CSIRO Land and Water, Townsville, QLD, Australia
J M Lynch, University of Surrey, Guildford, UK

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Soil-borne root diseases are one of the more intractable problems associated with achieving the sustainability of agriculture. Their occurrence is generally a sign of a biological imbalance within the soil ecosystem, where the natural enemies, predators, competitors, or antagonists of root parasites or disease-causing organisms are low in number and/or activity. This biological imbalance or loss of natural suppressiveness of the soil toward disease-causing organisms can, in many instances be associated with conventional agricultural management practices such as intensive soil cultivation, overuse of fertilizers and other agrichemicals, and continuous cropping. All of these practices tend to have a negative impact on the physical, chemical, and biological attributes of soil quality/health. In some cropping systems such as high-value vegetable and fruit crops, soil fumigation (or solarization) may be used to sanitize the soil and still permit growers to plant the same crop in the same field time after time. However, such practices, which work against the soil biota and debilitate its natural suppressive qualities, are not sustainable in the longer term.

Biological management or control of soil-borne root diseases has been a fruitful area of research for the past 40 years. There have been two basic research approaches. The first approach involves enhancement of the natural suppressiveness or biocontrol capability of microbial communities in the soil toward root disease organisms, through the use of alternative agricultural practices. These practices include the greater use of crop rotations and the use of organic amendments to stimulate soil microbial activity and provide a source of plant nutrients. The second approach involves the deliberate use of specific antagonist organisms for prevention and management of specific root diseases. These microbial biocontrol agents are usually single strains of bacteria or fungi which have been isolated from a soil and shown to have antagonistic properties toward the targeted root disease organism. They have been used in a number of ways, including direct addition to soil to reduce inoculum levels of pathogens in the soil and applied to seeds to protect against seed and root infections. The potential use of microbial biocontrol agents in agriculture has

been hampered by numerous technical difficulties (inconsistent performance, formulation, and delivery) and complex and expensive regulatory protocols.

Enhancement of Biocontrol by Agronomic Practices

The value of enhancing the biocontrol of soil-borne plant diseases by agronomic practices has been the subject of major debate between advocates of so-called conventional agriculture and those who practice so-called alternative agriculture. Modern farming in developed and many developing countries has moved increasingly toward less dependency on naturally occurring biological controls and greater dependency on synthetic pesticides. In contrast, alternative agriculture gives emphasis to biological interactions and natural biological cycles, making them relevant rather than irrelevant to the farming system.

There are two agronomic practices frequently considered for pest control. These are (1) the greater use of crop rotations, and (2) the greater use of organic amendments and/or green-manure crops in the farming system.

Crop Rotations

The use of crop rotations in the farming system allows time for the soil microbiota to displace, weaken, or destroy the propagules of soil-borne pathogens of any one crop while another, usually unrelated crop is growing. In general, soil-borne plant pathogens multiply in the presence of their preferred host plant(s) and decline when the host plant is absent. Most of these pathogens could survive in the soil in the absence of the host plant if it were not for the combined action of competition, antibiosis, and predation/parasitism imposed by the associated soil microbiota. Whilst crop rotations do successfully reduce populations of some soil-borne pathogens, e.g., root disease of cereals caused by *Gaeumannomyces graminis* var. *tritici* (Take-all) and *Heterodera avenae* (cereal cyst nematode) is reduced by rotation with noncereal crops such as peas and lentils, there are instances, because of the broad host range of the pathogen, where this is less successful. For example, *Rhizoctonia solani*, the fungal root pathogen which causes 'bare patch' of cereals, has a wide host range and thus affects the growth of many plant species. It is also an efficient saprophyte, readily colonizing particulate soil organic matter. Thus, whilst there is evidence that inoculum levels of this pathogen decline following medic or pea crops, it is clearly more difficult to achieve effective control of this pathogen through crop rotations alone.

Yield decline of sugarcane in Queensland, Australia, is a good example where growth of a crop as a monoculture, coupled with intensive cultivation of the soil prior to crop establishment and mechanized harvesting, can lead to the buildup of root-pathogenic fungi (e.g., *Pachymetra chaunorhiza*) and nematodes (e.g., *Pratylenchus zeae*) which reduce sugarcane yields. The introduction of a rotation crop such as soybeans at the end of the cane-growing cycle is effective in reducing populations of these pathogens to low levels and increasing the yield of the following cane crop.

Despite the potential biological control benefits of crop rotations, there are often economic disincentives associated with this practice seen by farmers. These include the economics of growing more than one crop – the rotation crop may provide less economic return than the primary crop – and the time required to achieve effective disease reduction – it may be necessary to grow the rotation crop for one year or more. In many instances therefore, the benefits of crop rotation have given way to practices such as increased tillage, burning of crop residues, and soil fumigation, which are aimed at reducing pathogen levels in the soil. However, these practices mostly work at cross-purposes with the goal of making agriculture more sustainable.

Organic Amendments

Organic amendments, including animal manure, composts, mulches, green-manure crops, and municipal and industrial by-products, are being increasingly used in agricultural systems to recycle nutrients and energy as well as improve soil conditions for plant growth. Many organic amendments have also been shown to suppress soil-borne fungal pathogens and/or the disease they cause and several have been effectively used for control of plant-parasitic nematodes.

There are several mechanisms whereby organic amendments reduce populations of soil-borne plant pathogens. The most widely reported is the stimulation of the general microflora, including organisms that are suppressive toward the pathogen. Other mechanisms include the production of compounds toxic to pathogens (e.g., ammonia, nitrous acid, volatile fatty acids), after degradation of the amendments by soil microorganisms, and the induction of systemic resistance to pathogens. Production of compounds toxic to root pathogens is common when high nitrogen-containing organic amendments are added to soil. For example, the microsclerotia of *Verticillium dahliae*, which causes verticillium wilt of potato, are killed by ammonia and nitrous acid generated following the addition of bone meal, soymeal, poultry manure, or liquid swine manure to soil. Little is known about how organic amendments induce

systemic resistance to root disease, and only a limited number of organic composts have been shown to cause this. For example, it is sufficient to expose only part of the root system of cucumber plants to a compost-amended horticultural potting mix suppressive to *Pythium* root rot in order to induce protection to the disease in the entire root system.

Enhancement of microbial populations suppressive toward soil-borne root pathogens has been demonstrated with a wide range of organic amendments in both horticulture and agriculture. Many reports show that *Pythium* and *Phytophthora* root rots are readily controlled by natural composts, whether applied as mulches to the soil surface, incorporated as soil amendments, or added as a component of potting mixes. Many types of microorganism appear to contribute toward the suppression of *Pythium* and *Phytophthora* spp. in these amended soils and potting mixes. Twenty percent of all bacterial strains recovered on 0.1-strength tryptic soy agar (TSA) from the rhizosphere of cucumber sown in a composted pine bark-amended potting mix induced biological control of *Pythium* damping-off when applied as seed treatments. Fluorescent pseudomonads, *Pantoea* and *Bacillus* spp. were the most effective in biocontrol and also the most abundant bacterial species present in the compost-amended potting mix.

Control of pathogens such as *R. solani* with organic amendments appears to be more variable than control of other pathogens such as *Pythium* spp. The basis for this difference is that *R. solani* is controlled by a much narrower spectrum of biocontrol agents and this microflora does not consistently colonize composts. This raises the important issue of the relationship between the quality (C:N ratio) and the degree of decomposition of the organic amendment and its efficacy in augmenting disease suppression. Generally speaking, fresh compost material such as green manures with a low C:N ratio can serve as a food source for both potential biocontrol agents and plant pathogens with high saprophytic ability. During this early phase in the decomposition process, the organic amendment may often increase root disease. However, in the next phase in the decomposition process, when the organic matter is fully colonized by microorganisms, the pathogen cannot effectively compete for resources and disease is suppressed. Later in the decomposition process, when the organic matter starts to become humified, the availability of resources to the biocontrol agents also becomes limiting, and this represents the phase when suppression begins to decline. Thus composted organic amendments with a high C:N ratio are more likely to be effective in pathogen control than fresh organic amendments with a low C:N ratio.

In horticulture the use of organic amendments is typically associated with the incorporation of green-manure crops or other readily available organic matter sources such as urban organic wastes. A comparison of organic and conventional farms in the central valley of California has shown that the severity of corky root of tomatoes, caused by *Pyrenochaeta lycopersici*, is less in organic farms than in conventional farms. The organically managed soils, which are fertilized by various composts and manures, have significantly higher microbial activity, which is positively correlated with lower corky root severity. In another case, municipal solid waste (fresh or composted), added 24 months previously to an arid Mediterranean soil, enhanced soil microbial activity in the soil, leading to improved biocontrol activity against *Pythium ultimum*.

Biocontrol with Introduced Microbial Biocontrol Agents

The deliberate introduction of microbial biocontrol agents into the soil to maintain the population of a targeted soil-borne plant pathogen at or below some economic threshold can be regarded as the more traditional or classic approach to biocontrol. This has been a productive area of research over the last 20 years and there are currently about 30 commercial products available worldwide for the biocontrol of soil-borne plant diseases. Some examples of commercial products of bacterial and fungal biocontrol agents used against soil-borne plant diseases are given in Table 1. The most common bacterial biocontrol agents are members of the genera *Bacillus*, *Burkholderia*, *Enterobacter*, *Pseudomonas*, and *Streptomyces*, and the most common fungal biocontrol agents are members of the genera *Gliocladium* and *Trichoderma*. Embodied in the classic approach to biocontrol are well-developed procedures for the selection, evaluation, and delivery of biocontrol agents to the soil, seed, or rhizosphere.

Selection, Evaluation, and Delivery of Microbial Biocontrol Agents

Several approaches have been used for the selection of microbial biocontrol agents, with no system apparently more successful than another. One approach has been to isolate potential microbial antagonists from the intended environment of use such as soils, seeds, or roots. This is based on the premise that any antagonist will be ecologically adapted to this environment and be able to survive and express activity when reapplied as a biocontrol agent. Another approach has been to isolate antagonists from soils suppressive to a particular pathogen. This approach has been

Table 1 Commercial biocontrol products for use against soil-borne plant diseases

Agent/product	*Target pathogens*	*Target crops*	*Application method*	*Manufacturer/distributor*
Bacteria				
Agrobacterium radiobacter				
Nogall	*A. tumefaciens*	Trees	Root dip	Bio-Care Technology, Australia
Bacillus subtilis				
Kodiak (HB, AT)	*Pythium/Rhizoctonia/ Fusarium*	Cotton/legumes	Seed treatment	Gustafson Inc., USA
Burkholderia cepacia				
Intercept	*Pythium/Rhizoctonia/ Fusarium*	Field crops/vegetables	Seed treatment/drench	Soil Technologies, USA
Pseudomonas aureofaciens				
Spotless	Turf diseases	Turf	Irrigation	Eco Soil Systems, USA
Streptomyces griseoviridis				
Mycostop	Broad	Field/vegetables/ornamental	Drench/spray	Kemira Agro OY, Finland
Fungi				
Coniothyrium minitans				
Contans	*Sclerotinia*	Field crops/vegetables	Spray	Prophyta, Germany
Fusarium oxysporum				
Biofox C	Pathogenic *Fusarium*	Vegetables/ornamental	Seed treatment/soil incorporation	SIAPA, Italy
Gliocladium virens				
SoilGard	*Pythium/Rhizoctonia*	Greenhouse crops	Granule incorporation	Grace-Sierra Co., MD, USA
Tricohoderma harzianum				
Root Shield	*Pythium/Rhizoctonia/ Fusarium*	Various	Granule incorporation/ drench	Bioworks Inc., USA

Complied from data in Whipps JM (1997) Developments in the biological control of soil-borne plant pathogens. *Advances in Botanical Research* 26: 1–135 and Stewart A (2001) Commerical biocontrol – reality or fantasy? *Australasian Plant Pathology* 30: 127–131.

used for the isolation of *Streptomyces griseoviridis*, a biocontrol agent for the control of *Fusarium* and *Pythium* spp. infections of ornamentals and vegetables. It is also used for the isolation of nonpathogenic *Fusarium oxysporum*, an effective biocontrol agent against *Fusarium* wilt of sweet potato and tomatoes. Alternatively, propagules or mycelia of pathogens have been placed in soils as baits from which antagonists have been isolated. In this case, antagonists have the potential to attack the pathogen and be adapted to the environment where the pathogen is active. This procedure has been used to obtain antagonists from sclerotia of several pathogens, including species of *Sclerotinia* and *Sclerotium*.

Having obtained a collection of antagonists, they are then screened for reproducible biocontrol activity. This usually involves a bioassay, which attempts to reproduce to some extent the conditions where biocontrol is required to act. As this often involves the screening of many hundreds of isolates, it usually involves some form of seedling bioassay carried out under controlled conditions. Successful isolates from the primary screens are then subjected to more rigorous evaluation and subsequent large-scale glasshouse or field trials.

Once a biocontrol agent has shown reproducible activity in a series of screening trials, methods for inoculum production, formulation, and application need to be considered in relation to the crop, disease, and environment of use. The production of suitable quantities of viable and active cells, spores, or biomass is the first step in this procedure. The most commonly used method is liquid fermentation utilizing a range of different and often inexpensive substrates such as molasses, brewer's yeast, and corn steep liquor. Liquid fermentations have the advantage in allowing control of nutrients, pH, temperature, and other environmental parameters, which can help optimize biomass production, spore quantity, and quality, and reduce contamination. An alternative method is solid substrate fermentation, involving a range of agricultural waste products, including lucerne powder, sugarcane bagasse, sawdust, various grains, bran, and peat.

Various methods of delivery of biocontrol agents have been developed. The method of choice for a particular agent is very much dependent on the nature of the biocontrol agent, the target pathogen, the target crop, and its method of cultivation. Methods include liquid formulations of the biocontrol agent,

which are applied to the soil or growing medium as a drench, spray, or via irrigation. This is common with bacterial biocontrol agents. For example the biocontrol product Spotless, based on *Pseduomonas aureofaciens*, for control of turf diseases, is delivered via irrigation. Other methods include incorporation of the biocontrol agent into various kinds of granule formulations, which can be added to the soil at or before the time of planting, or applying the biocontrol agent to the seed before planting using specialized seed-coating or pelleting procedures.

Mode of Action of Biocontrol Agents

Several modes of action of microbial biocontrol agents have been identified, none of which are mutually exclusive. These can involve interactions between the antagonist and pathogen directly, either associated with roots or seeds, or free in the soil. Three direct modes of action are known: competition, where antagonist and pathogen compete for nutrients and/or space; antibiosis, where antagonists secrete metabolites harmful to pathogens; and parasitism, where the biocontrol agent infects the pathogen. In addition, indirect interactions are known where the plant itself responds to the presence of the antagonist, resulting in induced resistance or plant growth promotion. Often one antagonist may exhibit several modes of action simultaneously or sequentially. Also, in the case of natural biocontrol in some suppressive soils, several antagonists exhibiting a range of modes of action may act together to control disease.

Understanding the mode of action of biocontrol agents has been a fertile area of research in recent years. It holds the key for the improved selection and screening of new biocontrol agents as well as offering the possibility of improving biocontrol activity directly via genetic manipulation.

Competition Competition between antagonist and pathogens as a mechanism of biocontrol may occur at different levels. For example, competition for space or specific infection sites on roots or seeds has been proposed as a mechanism of biocontrol of pathogenic *F. oxysporum* by nonpathogenic strains of *F. oxysporum* and of pathogenic strains of *R. solani* by nonpathogenic *Rhizoctonia* spp. In both instances the pathogen is excluded by the more rapid and extensive colonization of the root surface by the biocontrol strain. Competition between microorganisms for carbon, nitrogen, and other nutrients in the rhizosphere is another well-researched mechanism of biocontrol. Competition for iron, mediated by production of iron-chelating siderophores, has been conclusively demonstrated as a mechanism of biocontrol by several species of bacteria in soils where iron is limiting. This is a widely recognized mechanism of biocontrol by fluorescent *Pseudomonas* spp., which produce a range of siderophores including pseudobactins and pyoverdines. The siderophores are thought to sequester the limited supply of iron that is available in the rhizosphere to a form that is unavailable to pathogenic fungi and other deleterious microorganisms, thereby restricting their growth.

Antibiosis Antibiosis is generally mediated by the production of low molecular weight antibiotics by the antagonist, which can inhibit the growth of the pathogen. Evidence for a role of antibiotics in biocontrol by both fungi and bacteria includes strong correlations between antibiotic production and biocontrol efficiency, demonstration that the purified antibiotic produced can mimic the effect of the biocontrol agent and demonstration that antibiotic-deficient mutants exert less biocontrol activity than wild-types. Antibiotic production by fungi exhibiting biocontrol activity has been most commonly reported for isolates of *Gliocladium* and *Trichoderma*. For example, production of gliotoxin by *G. virens* is implicated as the key factor in its biocontrol activity against *Pythium ultimum* and *R. solani*. There are large numbers of bacterial biocontrol agents that produce antibiotics and other secondary metabolites that appear to be important for the control of different fungal pathogens. One well-known example is the production of phenazine-carboxylic acid by *Pseudomonas fluorescens* strain 2-79, which has been demonstrated to be involved in the biocontrol activity of this strain against take-all of wheat caused by *Gaeumannomyces graminis* var. *tritici*.

An antibiotic produced by *P. fluorescens* (2,4-diacetylphloroglucinol) is especially active against the damping-off of sugar beet. In a recent study, its activity on *Pythium*-infected peas was comparable with that of *P. fluorescens* Q2-87, although the HCN-producer *P. fluorescens* CHAO and the competitive-excluder *P. fluorescens* SBW25 performed better.

Antibiosis also features in fungal biocontrol. For example, *Trichoderma harzianum* produces 6*n*-pentyl-2*H*-pyran-2-one and harzianopyridone, while *Gliocladium virens* (now *T. virens*) produces gliovirin. It should also be noted that *Trichoderma* spp. produce metabolites which stimulate plant growth directly.

Parasitism Parasitism of plant pathogens as a mechanism of biocontrol is usually associated with fungal biocontrol agents. Most evidence for this comes from field observations of infected fungal propagules such as spores or sclerotia. For example, oospores of *Phytophthora* and *Pythium* spp. are frequently found to

be infected by *Olpidiopsis gracilis*, whilst sclerotia of *R. solani* are infected by the obligate sclerotial mycoparasite *Verticillium biguttatum*. The interaction between the mycoparasite and its host involves a sequence of processes encompassing location, contact, recognition, localized lysis, penetration, intracellular growth, and exit. Various chemical interactions are implicated in these processes, including involvement of lectins during the initial contact, and recognition between mycoparasite and the host fungus and a suite of different cell wall-degrading enzymes (e.g., β-1,3-glucanases, chitinases, proteinases, and lipases) during the penetration process. Other mechanisms of parasitism are associated with fungi such as *Verticillium chlamydosporium* and *Paecilomyces lilacinus*, which can infect the egg masses and cysts of the cereal cyst and root knot nematodes.

Induced systemic resistance Induced systemic resistance (ISR), also referred to as systemic acquired resistance, refers to the situation where the plant acquires resistance to infection from a pathogen following some initial triggering inoculation with a microorganism or some abiotic agent. Importantly, development of ISR is generally accompanied by the expression of a set of genes within the plant, including those that encode for pathogenesis-related proteins such as chitinases and β-1,3-glucanases. Most of the research on ISR has concerned foliar pathogens, and only relatively recently has the potential of this mechanism been recognized for biocontrol of soil-borne plant pathogens. Using split-root systems, both bacteria and fungi have been shown to induce resistance in several plants when applied to roots. For example, *Pseudomonas putida*, *Serratia marcescens*, and nonpathogenic isolates of *F. oxysporum* have been shown to induce systemic resistance to *F. oxysporum* f. sp. *cucumerinum* in cucumber.

The phenomenon of treating seeds, roots, or cuttings with inducing bacteria or fungi and achieving ISR to subsequent stem or foliar infection by a range of viral, bacterial, and fungal pathogens is also known, suggesting an important role for ISR in biocontrol in general.

Plant growth-promoting rhizobacteria Rhizobacteria which exert a beneficial effect on plant growth have been termed plant growth-promoting rhizobacteria (PGPR). The beneficial effect of PGPR may result from the direct biocontrol of root pathogens by any of the mechanisms described above (e.g., antibiosis), to indirect mechanisms including the direct promotion of growth via production of plant growth hormones, increased nutrient availability, stimulation of *Rhizobium* nodulation, and ISR. Growth promotion by PGPR may enable the plant to tolerate or escape the disease-causing pathogens, e.g., if the PGPR cause an enhanced seedling emergence rate, the time the plant is susceptible to preemergence damping-off pathogens such as *Pythium* spp. may be reduced.

Future Prospects

Soil-borne plant diseases are a particular problem with intensive agriculture. To a large degree, the problems they present can be associated with the cultural practices used to grow our crops including excessive tillage of the soil, continuous cropping, and reliance on synthetic inorganic fertilizers and pesticides. As a result, the health of our soils has declined, and with it most of the soils' natural capacity to suppress indigenous and exotic soil-borne plant diseases. There would appear to be two parallel drivers of change to overcome this situation. The first is based on improving soil health, by utilizing soil-management procedures that maintain soil community relationships, optimize soil organic matter levels, and encourage soil structural development. Here the objective is to capture the natural benefits of the soil biota in disease control through better crop management practices, including the improved use of crop rotations and improving the organic matter status of the soil through the use of natural and composted soil amendments. The second driver is based on the public's concern for maintenance of a clean and safe environment and hence interest in replacing chemical pesticides for disease control with biocontrol agents. However, it should be noted that, as with pesticide chemicals, the objective behind the use of biocontrol agents is the control rather than the management of root disease problems. Given that many root disease problems are caused by a suite of root pathogens (e.g., yield decline of sugarcane), some will argue that it may simply not be possible to control such problems with introduced biocontrol agents and that greater emphasis should be placed on developing management solutions to root disease problems rather than developing short-term control measures. However, the reality is that both approaches are equally valid in the quest for a more sustainable agriculture and that there will be many instances where one or both approaches will be more applicable.

Problems inherent in the quest for effective biocontrol agents of soil-borne diseases have been reviewed recently. The problems include inconsistency in performance, lack of broad-spectrum disease control activity, high costs associated with the registration of biocontrol products, and public concern over the potential risks of introducing exotic biocontrol agents

into the environment. Inconsistency in the performance of biocontrol agents for soil-borne plant diseases is not surprising in view of the multitude of abiotic and biotic factors that can affect microbial survival plus the inherently large buffering capacity of the resident soil microbiota, which is more adapted to the soil conditions. A single introduced strain may flourish for a short period of time by virtue of its high initial inoculum level, and in some circumstances this longevity may be sufficient to give biocontrol (e.g., for damping-off diseases). However, the introduced strain will progressively become restricted to specific microenvironments to which it is more suited than the resident microbes and this will undermine its ability to elicit a biocontrol effect. Strategies to address this problem and also the lack of broad-spectrum control of most biocontrol agents include the use of mixtures of biocontrol agents. The use of combinations of *Pseudomonas* strains, *Phialophora graminicola*, and *Idriella bolleyi* to provide better control of cereal take-all has been proposed because of their ability to distribute themselves preferentially at different sites along the root. Multiple strains or mixtures of strains might also be expected to provide better breadth of activity by targeting multiple pathogens, and may also provide more stable and robust biocontrol across different sites and seasons. This may also prove more attractive to biotechnology companies who are reluctant to enter the biocontrol market because of the inconsistent disease control given by microbial products and the current limited size of the commercial market. The concept of combining induced resistance and treatment with a microbial biocontrol agent also warrants further investigation.

The registration of biocontrol products for soil-borne plant diseases has been hampered by them being classified as microbial pesticides. This generally means that the same time-consuming and extremely expensive efficacy–safety data package that is required for registration of a chemical pesticide is also required for registration of microbial biocontrol products. To change this situation in the foreseeable future, regulatory bodies need to be provided with more information in order to make valid judgments about the safety and environmental impact of microbial biocontrol agents. In addition distinctions may need to be made between those biocontrol agents whose mode of action is via competition, exclusion, or some other nonchemical-based mechanism and those where the primary mode of action is via the production of a toxic metabolite.

Concerns about the risks of introduced biocontrol agents to the environment have received considerable attention in recent years. Specific concerns range from the potential effect the introduced microorganism may have on nontargeted microorganisms, to the exchange and spread of genetic material between microorganisms, especially if the biocontrol agent has been genetically modified in any way. Conventional mutagenesis, protoplast fusion, and genetic modification have all been successfully used to enhance the biocontrol efficacy of biocontrol agents such as *Pseudomonas* spp., *Agrobacterium* spp., *Trichoderma* spp., *Gliocladium virens*, and saprophytic *Fusarium* spp. However, despite the fact that there has not been a single case of clear adverse effects to the environment due to the release of a genetically modified microorganism, there is a broad consensus that releases of 'novel' organisms must be preceded by careful examination of their threat to the environment. Regulations in many countries now require an analysis of environmental impact as part of an application for the registration and commercial development of not only genetically modified but also unmodified biocontrol agents.

Conclusions

Biocontrol of soil-borne plant diseases is continuing to develop strongly, utilizing approaches inherent in alternative agricultural practices where the emphasis lies in enhancing the natural biocontrol activities of microbial communities, and approaches where individual microbial biocontrol agents are introduced into the soil environment to control specific plant pathogens. Both approaches seek to harness the various forms of biocontrol that have been identified in the soil microbiota (competition, antibiosis, parasitism, induced plant resistance). With increased public awareness of the need to improve the health and safety of the environment, and to use more sustainable and environmentally friendly cultural practices in agriculture, biocontrol of soil-borne plant diseases is likely to become an increasingly important facet of food production.

Further Reading

Akhtar M and Malik A (2000) Roles of organic soil amendments and soil organisms in the biological control of plant-parasitic nematodes: a review. *Bioresource Technology* 74: 5–47.

Baker RR and Dunn PE (eds) (1990) *New Directions in Biological Control*. New York: Alan R. Liss.

Boland GJ and Kuykendall LD (eds) (1998) *Plant–Microbe Interactions and Biological Control*. New York: Marcel Dekker.

Cook RJ (1994) Introduction of soil organisms to control root diseases. In: Pankhurst CE, Doube BM, Gupta VVSR, and Grace PR (eds) *Soil Biota: Management in*

Sustainable Farming Systems, pp. 13–22. Melbourne: CSIRO Press.

Cook RJ and Baker KF (1983) *The Nature and Practice of Biological Control of Plant Pathogens*. St Paul, MN: American Phytopathological Society Press.

Dowling DN and O'Gara F (1994) Metabolites of *Pseudomonas* involved in the biocontrol of plant disease. *Trends in Biotechnology* 4: 239–249.

Hoitink HAJ and Boehm MJ (1999) Biocontrol within the context of soil microbial communities: a substrate-dependent phenomenon. *Annual Review of Phytopathology* 37: 427–446.

Hokkanen HMT and Lynch JM (eds) (1995) *Biological Control: Benefits and Risks*. Cambridge, UK: Cambridge University Press.

Hornby D (ed.) (1990) *Biological Control of Soil-borne Plant Pathogens*. Wallingford, UK: CAB International.

Kloepper JW (1993) Plant growth-promoting rhizobacteria as biological control agents. In: Metting FB Jr. (ed.) *Soil Microbial Ecology – Applications in Agricultural and Environmental Management*, pp. 255–274. New York: Marcel Dekker, Inc.

Lewis JA, Papavizas GC, and Lumsden RD (1991) A new formulation system for the application of biocontrol fungi to soil. *Biocontrol Science and Technology* 1: 59–69.

Lynch JM (1992) Environmental implications of the release of biocontrol agents. In: Tjamos EC, Papavizas GC, and Cook RJ (eds) *Biological Control of Plant Disease – Progress and Challenges for the Future*, pp. 389–397. New York: Plenum Press.

Rovira AD, Elliott LF, and Cook RJ (1990) The impact of cropping systems on rhizosphere organisms affecting plant health. In: Lynch JM (ed.) *The Rhizosphere*, pp. 389–436. Chichester, UK: Wiley-Interscience.

Stewart A (2001) Commerical biocontrol – reality or fantasy? *Australasian Plant Pathology* 30: 127–131.

Whipps JM (2001) Microbial interactions and biocontrol in the rhizosphere. *Journal of Experimental Botany* 52: 487–511.

BIODIVERSITY

D H Wall, Colorado State University, Fort Collins, CO, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Soils teem with life. Representatives of almost every phylum of organism known aboveground also occur in soil. In less than a handful of soil, there can be billions of types of microbes and hundreds of species of microscopic invertebrates. The identities and natural histories of these microscopic flora and fauna and many of the larger, visible soil fauna are the least-known biota in terrestrial ecosystems. The major contributions of soil organisms to the maintenance of life on Earth are now being recognized by scientists and the public, opening a new frontier for exploration and a rising concern about the increasing degradation of the soil habitat. Soil biota are the primary drivers of numerous ecosystem processes. These activities provide a wealth of essential ecosystem services for humans, such as decaying organic matter, filtering water, stabilizing soil, generating and renewing soil fertility, providing nutrients for plant growth, modifying the hydrologic cycle (including mitigating floods and controlling erosion), and controlling pest and pathogens of plants and animals. Soil biodiversity is intimately linked to all ecosystems, terrestrial and aquatic, and to the atmosphere. Understanding how soil biodiversity will change with the rapid degradation of soils globally is important knowledge for policymakers and members of the public who need to develop and implement strategies for the future.

Diversity and Abundance

Soil biodiversity includes a plethora of life that is hidden from our everyday view: viruses; bacteria; actinomycetes; fungi; algae; the protozoans (single-celled eukaryotes); microscopic invertebrates such as rotifers, tardigrades (water bears), soil planaria, (flatworms), and nematodes (roundworms); the microarthropods, Acari (mites) and Collembola (springtails); and larger invertebrates, easily seen by the naked eye such as terrestrial gastropods (snails and slugs), isopods (pill bugs, sow bugs), Oligochaeta (enchytraeids and earthworms), spiders, scorpions, beetles, centipedes, millipedes, crustaceans, ants, and termites. Because of this diversity, the invertebrates are frequently grouped by size (body width), microfauna, mesofauna, and macrofauna (Table 1). Vertebrates also depend on the soil as a habitat, for example, moles, prairie dogs, meercats, wombats, small rodents, and some species of lizards, snakes, frogs, and even birds. Many aboveground invertebrates may be temporary inhabitants (such as nematodes parasitic on insects; immature flies in the Diptera such as tipulid crane flies and tephritid fruit

flies; cicadas; and coleopteran scarab beetles) living in the soil for only one stage of their life cycle or occurring seasonally. Thus soil diversity is dynamic and tightly coupled to the aboveground biotic and climatic systems.

Table 1 Some invertebrates found in soil grouped by body width

Microfauna (<100 μm diameter)	*Mesofauna (100 μm to 2 mm diameter)*	*Macrofauna (>2 mm diameter)*
Nematoda (roundworms)	Acari (mites)	Opilones (daddy longlegs)
Protozoa	Collembola (springtails)	Isopoda (pill bugs, sow bugs)
Rotifera	Enchytraeidae (pot worms)	Chilopoda (centipedes)
Tardigrades	Isoptera (termites)	Diplopoda (millipedes)
	Formicoidea (ants)	Lumbricidae (earthworms)
		Coleptera (beetles)
		Arachnida (spiders and scorpions)
		Mollusca (snails and slugs)

Adapted from Swift MJ, Heal OW, and Anderson JM (1979) *Decomposition in Terrestrial Ecosystems*, Oxford, UK: Blackwell; Wall DH, Adams G, and Parsons AN (2001) Soil biodiversity. In: Chapin FS III, Sala OE, and Sannwald EH (eds) *Global Biodiversity in a Changing Environment*, pp. 48–49. New York: Springer-Verlag.

The diversity of species in soil appears so immense that knowledge of the numbers of species (species richness) in just two groups, the microarthropods and nematodes, could substantially increase global biodiversity estimates. Yet our knowledge of the numbers of soil species globally, even within any one group, is surprisingly poor. Data suggest 90–95% of the soil biota are unidentified. There is no one location where the total number of species or molecular types of microbes, invertebrates, or vertebrates has been assessed, although up to 1000 invertebrate species have been enumerated in a square meter of forest soil. Global estimates are based on many studies of varying numbers and sizes of soil samples from a variety of ecosystems and generally only to a shallow (0–20 cm) depth. This contributes to the underestimate of soil biodiversity given that mites and nematodes occur to 15-m depths in deserts and bacteria to greater depths. Global estimates of the total numbers of species identified so far from soils include 600 species of enchytraeids, 1500 species of soil protozoa, 5000 species of nematodes, and 20 000–30 000 species of mites (Figures 1 and 2). And, as in aboveground systems, many of the species identified in a soil sample are rare and have a limited distribution. Identification of species is complicated by the sheer numbers of organisms in the soil. Consider the billions of bacteria, meters of fungal hyphae, millions of protozoa, 10 million nematodes, 45 000 oligochaetes,

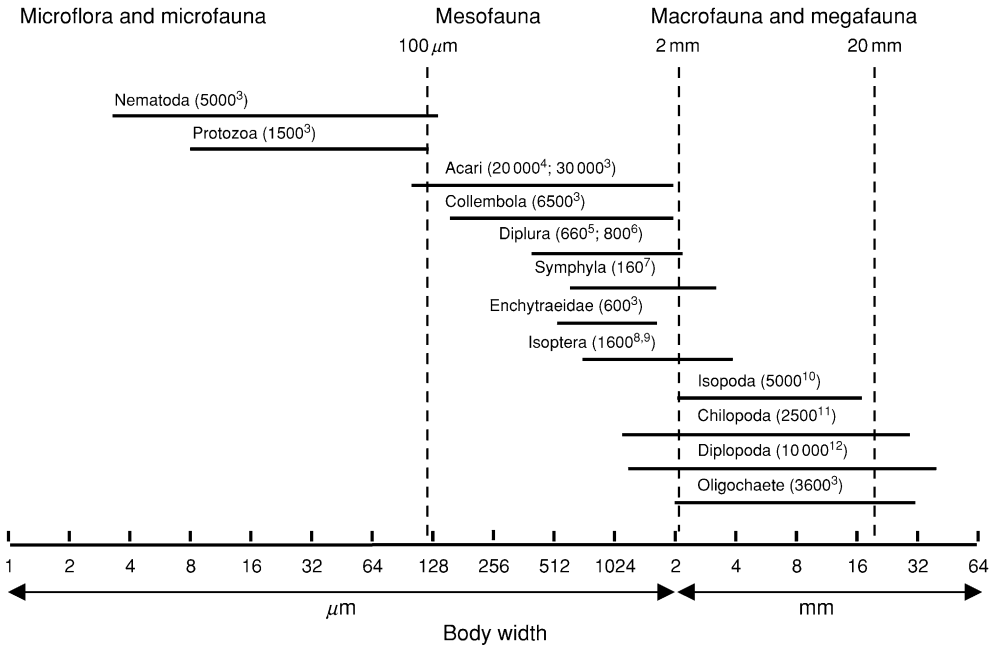


Figure 1 Invertebrate biodiversity represented by body width, with number of species described (as of 2003) for each group. (Source: Wall DH and Virginia RA (1997) The world beneath our feet: soil biodiversity and ecosystem functioning. In: Raven PH (ed.) *Nature and Human Society: The Quest for a Sustainable World*. Washington, DC: National Academy Press.)

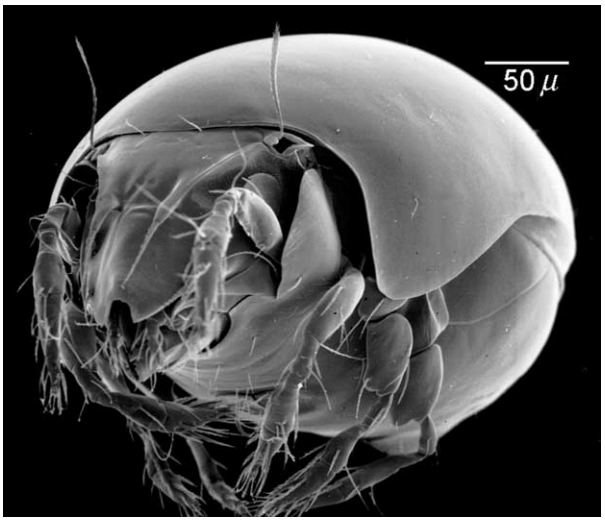


Figure 2 *Ceratozetes borealis*: a soil mite found in litter and soil in the Yukon and Alaska. (Courtesy of Valerie Behan-Pelletier.)

and up to a million arthropods that can inhabit a square meter of soil.

Even with the enormity of soil biodiversity, there is excitement about the progress and plans to overcome the difficulties of research on these and other species yet to be identified. Advancements in technology such as molecular technology, biochemical methods, and microscopic advancements (scanning electron microscope, image analysis) are enabling faster identification and/or characterization of organisms. Scientists, distributed globally with expertise in specific soil biotic groups, have accelerated our knowledge of the number of species, by sharing information, images and results via the internet, and creating web pages for the public. Today there are a very few taxonomists and systematists who specialize in the identification, natural history, and phylogenies of the soil biota. These specialists are also versed in the proper techniques for collecting, extracting, and identifying the various organisms, information that is needed for analysis of different ecosystems. As methods for the collection and extraction of soil fauna are primarily dependent on motility and size of the organism, there are no standard techniques for the collection or extraction of even a single group of fauna. This is comprehensible when one considers that collecting a handful of soil might suffice for extracting and identifying many species of microscopic nematodes, mites, protozoa, and microbes, but would not be adequate for earthworms. Technology, new educational tools, and training, and the excitement of conquering these challenges are attracting students to this new frontier of science.

Distribution

Soils, like air and water, are nonrenewable resources that have developed over geologic time. Both the type of soil and the evolutionary adaptation and dispersal mechanism of the organism to a particular location play a role in establishing the structure of a soil biotic community. Soil chemical and physical factors combine to make each soil a unique habitat. Chemical (e.g., pH, salinity, type and amount of organic matter and nutrients) and physical (e.g., texture (amount of sand, silt, and clay), pore size, moisture, temperature) variations, along with the composition of vegetation and geologic history all contribute to characterizing the diversity of species in soil. These components are all intertwined, and changes in any one factor can result in a new habitat for the biota. For example, an increase in organic matter generally has a ripple effect, increasing the soil water-holding capacity, which tends to create a darker soil and higher soil temperatures. Soil texture is also linked to microclimate. As clay content increases, soil fertility, soil moisture, and soil temperature can increase. Clay content is in turn largely dependent on the parent material (granite related to sandy soils versus limestone related to clayey soils) or landscape position (a floodplain is more fine-grained, a ridge top more coarse-grained). Thus, beneath our feet, there are thousands of soil habitats due to many combinations of factors such as pH, fertility, organic matter, soil moisture, climate, and geologic history.

Soil biota occupy remarkably different habitats within the soil: some species living only in decaying leaves and dead wood, or in the rich organic matter horizon, while others occur several centimeters to meters deeper in the mineral soil. For example, some species of earthworms occur in litter-organic horizons and others exist in deeper mineral soil. Species of organisms may be very different even at millimeter scales near a growing root than further away in bare ground. Increasing to larger scales, within a backyard or agricultural field, the diversity (species richness) might be similar. However, comparison of an agricultural field to a nearby natural area will generally show a very different community. At a global scale, primary factors affecting the biogeography of soil organisms are climate, the type (trees, grasses, forbs) and amount of vegetation (organic matter), and soil texture (composition of sand, silt, and clay). Earthworms are generally absent from dry ecosystems, illustrating an evolutionary adaptation to moist soils. Termite and ant diversity decreases with distance from the equator, but other invertebrates such as earthworms, enchytraeids, mites, and nematodes follow different patterns. As data are compiled

on species and their natural history and combined with geographic information systems (GISs) maps of soil and vegetation, more reliable estimates of the global biogeography for soil microflora and fauna will become feasible.

Dispersal mechanisms of biota are a key to their distribution at local, regional, and global scales. Soil organisms are transported with soil: by humans transporting plants; or by cars, trucks, or other machinery; in rainwater, floods, seawater; carried externally and internally by animals such as birds, ungulates, and insects; in plant debris; and by wind. Protozoa, rotifers, tardigrades, and nematodes are aquatic biota, living in water films around soil particles, but as soil dries they enter a dormant state, anhydrobiosis (meaning life without water), reduce their body surface area, water content, and metabolism, and can live for years until they revive in water. While in anhydrobiosis, they can be wind-transported over hundreds of miles. Recent satellite photos from the year 2000 show large dust storms from Africa moving across the Atlantic ocean and transporting soil bacteria and fungi, including plant and animal pathogens into the North American soils. Collembola live in air pores of soil and in organic matter and were long thought to be confined to small geographic areas. However, they are now known to survive and move to other continents in seawater, because of a physiological advantage, their hydrophobic (unwettable) cuticles. This finding on a new means of dispersal will accelerate hypotheses on the global distribution patterns of Collembola. Other dispersal mechanisms are based in geologic time. For example, earthworm distribution in North America has been patterned largely by glaciation.

Soil Biodiversity and Ecosystem Functioning

Soil organisms moderate numerous ecosystem processes (e.g., decomposition, C and N transformation, hydrologic cycles) that are a major component of global cycles. Land-use change, climate change, atmospheric change, and an increase in invasive species are all components of 'global change.' Whether this global change results in the alteration of species composition within the soil community, and how this may impact ecosystem function, is an important question in the research of soil biology today. Little knowledge exists about the functional attributes of the majority of individual species, but there is considerable information on the consequence of larger species on soil processes. Macrofauna (earthworms and termites) species have a significant effect on soil structure and hydrology, and transfer and reconfigure organic matter, affecting nutrient mobility. The influence of these animal species on surface-soil plant litter can be measured at local and larger regional scales; but for smaller species scientists have used other approaches to determine their influence on ecosystem processes. Scientists lump species with similar morphologies and feeding sources into functional groups and measure the transformations of carbon and nitrogen through each functional group within the soil food web. A couple of examples illustrate this.

Perhaps some of the greatest diversity is involved in the decomposition (decay) component of the soil food web. In natural terrestrial ecosystems, the bulk of the world's dead plant material (leaves, stems, dead roots, grasses, also known as detritus) falls to the ground to be decomposed. The rate of decomposition in an ecosystem is affected by climate, the quantity and quality (type or chemical composition, generally C, N, and lignin) of the substrate or detritus (leaf versus wood versus animal, for example), and a 'detritivore' functional group in the food web that includes many species of generalist feeders from numerous phyla. If the substrate is more recalcitrant (has more lignin, higher C:N ratio, e.g., wood of trees) the decomposition will be primarily by fungi and fungal consumers. If the organic matter is more labile (lower lignin, lower C:N, e.g., some grasses) bacteria and their consumers are the major decomposers.

The breakdown of detritus is a succession of the species of invertebrates and fungi, actinomycetes, and bacteria, all with specific roles in breaking down complex organic substrates into the inorganic nutrients that are then available for plant growth. For example, some faunal species (e.g., arthropods) shred leaves into smaller particles; others channel through the soil, increasing the distribution of microbes and the organic matter while affecting soil porosity in the soil profile; other fauna ingest the pieces of organic matter, further exposing it to enzymes that aid the decay, and then pass it from their bodies in a different chemical form. Smaller microscopic animals, nematodes, mites, and Collembola prey on the decomposers, bacteria and fungi, maintaining them in an accelerated growth phase, speeding the transfer of nutrients to soil. As with other biotic food webs, predators are at the top of the soil food web. Interestingly, predators in soil food webs can be microflora (e.g., fungi that prey on nematodes) and microfauna (predaceous nematodes, or mites or tardigrades feeding on nematodes), as well as macrofauna. Determining predator–prey relationships of species contributes to biocontrol management of soil pathogens of humans and plants.

Global Change and Impacts on Soil Biodiversity

Land-use change is the major driver affecting soil biodiversity and future soil sustainability. Land-use changes (tillage, erosion, dams, change in plant species) affecting soil physical and chemical properties, soil structure, and the base of the soil food web (chemical composition of plants, amount of organic matter, oil, pollution, manure) have direct impacts on species composition. The conversion of a natural grassland or forest to a managed system for agriculture, pasture, urban, or industrial use, changes the determinants of soil biodiversity, the vegetation, soil structure, and microclimate. The disruption to the natural vegetation and the soil habitat with land-use change decouples the nutrients provided by the decomposition food web from plant uptake. The result is a loss in soil fertility provided by the original soil and its inhabitants. Additional fertilizer is required, and pesticides may be necessary, especially with more intensive agriculture, because the new soil food web generally has fewer predators, resulting in a change in or loss of biocontrol of plant pathogens. While tillage methods are beneficial in conserving carbon in soils and creating a food web that is more detritus-based, some lower level of herbicides and pesticides is generally used. There are examples from tropical wet and dry forests, grasslands, deserts, and other ecosystems showing that these land-use changes affect the total soil biota (macrofauna, microfauna, and microflora), generally reducing species diversity. Desertification resulting from land-use change has a considerable impact on soil processes, including soil carbon, soil structure, soil biota, and soil fertility.

Invasive species of plants and animals also influence changes in soils and biotic communities. The movement of plants, people, and machinery has not only increased the numbers of species invasions into soils globally, but has significantly increased impacts on soil biodiversity and some characteristic of the ecosystem (e.g., nutrient cycling, species richness and or abundance, plant factors, soil physical or chemical factors) and increased costs of eradication. An earthworm species introduced to New York has changed forest-floor litter quality and composition, soil chemistry, and water infiltration rates. European nations are examining ways to eradicate the Australian planarian flatworm, *Artioposthia triangulata*, a predator of earthworms. This invasive species has impacts aboveground (removing the food source for birds) and belowground (removing an animal species that influences organic matter transformation, soil hydrology, and structure). Invasive plant species with differing rooting depths and plant chemical composition can have repercussions for soil communities. Woody plant invasions into grasslands of the Great Plains of the USA have greater rooting depths (affecting soil carbon storage) and, at these depths, a more depauperate nematode community.

Increasing levels of CO_2 and climate change (temperature and rainfall patterns, and extreme, infrequent events such as droughts and rainfall) indirectly affect soil biodiversity through impacts on plant community composition and/or chemistry. Although data vary with ecosystem, modifications in the soil community and changes in belowground herbivory and decomposition pathways for soil food webs have been documented. Even in the Antarctic Dry Valleys where there are no visible plants, climate change has decreased soil nematode populations. Direct effects of elevated CO_2 are considered to be less important, because soils have high ambient levels of CO_2 owing to root and microbial respiration.

Soil food webs can be modeled to simulate the loss of species due to global change and to examine the effect, if any, on nutrient cycling and plant production. For example, in the short grass steppe of Colorado, USA, a model has been assembled using field data on the microbes and micro- and mesofauna. Other laboratory studies and results glean from the literature-generated information on the life cycles, generation time, biomass, and energetics for the dominant invertebrate species. The model accounted for transfers of carbon and nitrogen through the soil food web in 15 functional groups including plant symbionts (mycorrhizal fungi), herbivores (plant parasitic nematodes), decomposers (bacteria, saprophytic fungi), bacterial feeders (protozoa, nematodes), fungal feeders (nematodes, mites), omnivores, and predators (mites, nematodes). A global change scenario (elevated CO_2 and resulting change of plant species composition), which might indirectly result in the loss of biodiversity, was simulated for each of the 15 functional groups in the soil food web. For example, loss of bacteria and fungi led to extinctions of other groups in the soil food web. Removal of six of 15 groups impacted the abundance of other groups, and three of these groups (bacteria, saprophytic fungi, and root-feeding nematodes) caused up to a 10% change in two ecosystem processes, nitrogen mineralization, and plant primary production. This suggests that ecosystem functioning may be affected by the loss of soil biotic functional diversity, a result that needs further testing. Some scientists postulate that because of the high species diversity in soils, if species are lost, other species will perform the same function in the food web, i.e., there is a lot of redundancy of species in soils. Additional experimental results show that the effects of the change on soil

community composition are idiosyncratic, or dependent on the ecosystem, vegetation, climate, and organisms examined. None of the studies so far have accounted for: (1) long-term loss of species; (2) infrequent climatic events; (3) highly variable changes in land use; (4) long-term effects of elevated CO_2 or increased nitrogen deposition, and other atmospheric changes; and (5) invasive species or combinations of these factors and soil disturbance.

Soil food web models may not reflect the complexity of individual species interactions and the underlying habitat but are critical for scientific analysis and forecasting for future scenarios of C and N cycling. The C and N cycling of the soil food web ties the soil to the aboveground ecosystem by relating the soil biota to soil carbon storage, nutrients available for plant growth and to atmospheric CO_2 and trace gas fluxes. Models, along with studies at species and functional group resolution, are currently being examined in experiments that manipulate biodiversity in field, laboratory (microcosms with generally less than 10 species), and in large soil experiments in plant growth facilities. Together these provide a theoretical and quantitative basis for integrating the soil biota into global ecosystem studies.

Sustaining Soil Biodiversity

Soil biodiversity is an important resource that provides ecosystem processes essential to the functioning of natural and global systems. Our understanding of the species, their interactions, and effect on processes occurring in the soil food web in natural systems are an important contribution to management of land, particularly agriculture. The link between aboveground and belowground diversity is strong, although occurring at different temporal scales for organisms, and changes affecting aboveground diversity and function are reflected in belowground ecosystems. An immediate effect is a decrease in the biological capacity of soils and a change in the regulation of interactions and processes. Knowledge on whether all or a few key taxa are important in this regulation of ecosystems processes is a high priority for planning for future sustainability.

Current priorities in the study of soil biodiversity include: (1) understanding the functions of rarely sampled fauna so that species estimates and global distributions are based on an accurate and current database; (2) determining which habitats are most vulnerable for soil biodiversity loss, e.g., where the 'hot spots' of biodiversity are, and which habitats and what time frames are most amenable to restoration; (3) synthesizing data and determining which invertebrate and microbial 'species' are key to ecosystem processes; (4) identifying gaps in knowledge of multispecies interactions and their influence on ecosystem functioning; (5) collaborating on long-term (more than 3 years) experiments to examine effects of global changes on aboveground–belowground biodiversity linkages and ecosystem functioning; and (6) determining, based on natural history information, which species are more likely to be invasive if introduced, and using this potentially to reduce spread and to identify threats to other species.

See also: **Archaea**; **Bacteria:** Plant Growth-Promoting; **Crusts:** Structural; **Fauna**; **Food–Web Interactions**; **Fungi**; **Microbial Processes:** Community Analysis; **Nematodes**; **Protozoa**

Further Reading

Coleman DC and Crossley DAJ (1996) *Fundamentals of Soil Ecology*. San Diego, CA: Academic Press.

Hendrix PF (ed.) (1995) *Earthworm Ecology and Biogeography in North America*. Boca Raton, FL: Lewis Publishers.

Hunt HW and Wall DH (2002) Modelling the effects of loss of soil biodiversity on ecosystem function. *Global Change Biology* 8: 33–50.

May RM (1988) How many species are there on earth? *Science* 241: 1441–1449.

Pace NR (2000) Microbial diversity and the biosphere. In: Raven PR and Williams T (eds) *Nature and Human Society: The Quest for a Sustainable World*, pp. 117–129. Washington, DC: National Academy of Sciences and National Research Council.

Schaefer M and Schauermann J (1990) The soil fauna of beech forests: comparison between a mull and a moder soil. *Pedobiologia* 34: 299–314.

Wall DH and Virginia RA (2000) The world beneath our feet: soil biodiversity and ecosystem functioning. In: Raven PR and Williams T (eds) *Nature and Human Society: The Quest for a Sustainable World*, pp. 225–241. Washington, DC: National Academy of Sciences and National Research Council.

Wall DH, Adams G, and Parsons AN (2001) Soil biodiversity. In: Chapin FS III, Sala OE, and Sannwald EH (eds) *Global Biodiversity in a Changing Environment: Scenarios for the 21st Century*, pp. 47–82. New York: Springer-Verlag.

Wall DH, Snelgrove PVR, and Covich AP (2001) Conservation priorities for soil and sediment invertebrates. In: Soule ME and Orians GH (eds) *Conservation Biology: Research Priorities for the Next Decade*, pp. 99–124. Washington, DC: Island Press.

BUFFERING CAPACITY

B R James, University of Maryland, College Park, MD, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Buffering is the resistance of a system to change in response to a perturbation, and it is a key attribute of soils from the molecular level to the landscape scale. Chemical, physical, and biological processes may raise or lower solute concentrations (or activities) as intensive variables in soil solution, thereby temporarily disturbing dynamic equilibria or steady-state conditions between soil water and solid or gas phases of the soil. In response to such a perturbation, one or more processes may release solute to soil solution or remove it to restore wholly or partially the original concentration. The reservoir of such solutes in solution, in solid phases, or in the soil atmosphere for restoration of soil solution chemical composition is known as the 'capacity factor.' The measurable change in the capacity factor per unit change in the intensive variable is called the 'buffer index,' 'buffer intensity,' or, more commonly, 'buffer capacity.' Myriad chemical reactions (buffering mechanisms) govern such release or uptake of ions and molecules between soil solution and solid or gas phases, including cation and anion exchange, oxidation–reduction reactions, dissolution–precipitation processes, and metal–organic ligand complexation. Understanding and quantifying buffer capacities and buffering mechanisms at the colloid and molecular level of soils can aid in predicting the sensitivity and resilience of soil-water systems to anthropogenic and natural perturbations of ecosystems.

New ecological theory related to succession dynamics of disturbed ecosystems identifies that natural systems, including soils, are constantly and naturally recovering from regular and irregular disturbances of various severities. The processes and mechanisms of recovery determine the biodiversity and stability of ecosystems, and buffering in soils and natural waters is a key control of nutrient availability and pollutant bioavailability. Buffering reactions influence water quality and community regrowth, migration, and recruitment in disturbed patches on the landscape. The success of human efforts to restore disturbed ecosystems and soils is also determined in part by the scale and nature of soil-buffering processes, as related to natural ones that govern element transformations.

The Samovar Analogy for Buffering in Soils

The concepts of the intensive variable, capacity factor, buffer capacity, and buffering mechanisms for soils can be modeled and visualized by analogy with old-fashioned, Russian samovars used to make large volumes of tea (Figure 1). Each samovar comprises a large copper reservoir to boil water and a narrow glass tube on the outside of the samovar's tank to indicate how full the tank is. As boiling water is drawn out of the tank through the spigot, the indicator tube empties quickly and temporarily, but is refilled to the new tank level when the spigot is shut off (as indicated by the arrows between the tank and indicator tube in Figure 1). To decrease the volume of a samovar tank by a given fraction of its total capacity (e.g., from level 1 to level 2 in the indicator tube), a greater volume of water must be withdrawn from the larger samovar than from the smaller one.

Since the level of water in the indicator tube is proportional to how full each tank is, it is a parameter that is independent of the volume of the samovar. The height of the water column is analogous to an intensive variable in soil solution such as pH, a measure of how acidic an aqueous sample is, but not an indicator of its volume. The volume of the samovar represents the capacity factor (e.g., total soil acidity), and the

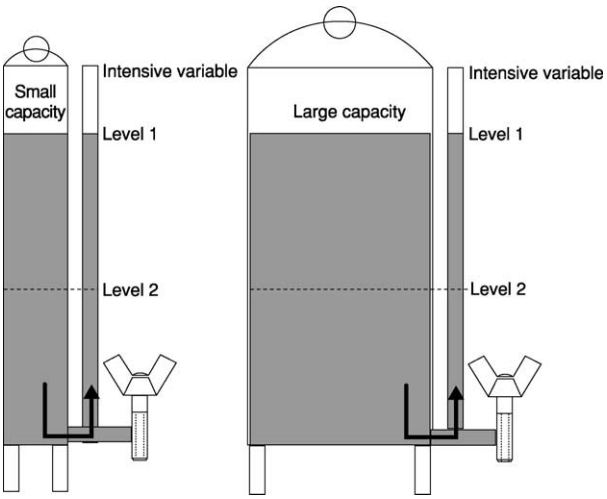


Figure 1 Samovar models representing small and large buffer capacities (tank sizes) in equilibrium with the same level of intensity variable (represented by the height of liquid in the narrow indicator tube outside the tank). (Adapted from Brady NC and Weil RR (2002) *The Nature and Properties of Soils*, 13th edn. Upper Saddle River, NJ: Prentice-Hall.)

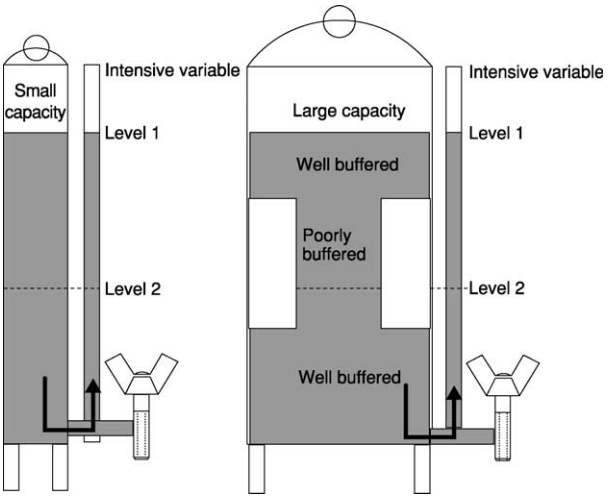


Figure 2 Samovar models representing small and large buffer capacities (tank sizes) in equilibrium with the same level of intensity variable (represented by the height of liquid in the narrow indicator tube outside the tank). The larger tank with the hourglass shape models a buffered system in which the buffer capacity (change in tank volume per unit change in the height of the liquid in the indicator tube) is not constant and varies during titration (drainage of the tank). (Adapted from Brady NC and Weil RR (2002) *The Nature and Properties of Soils*, 13th edn. Upper Saddle River, NJ: Prentice-Hall.)

change in the tank volume per unit change in the level of the indicator tube is the buffer capacity. The restoration of the water level in the indicator tube from the samovar after shutting off the spigot is due simply to the flow of water, i.e., the buffering mechanism in this model.

In this simple comparison of two samovars of different volumes, but with identical fractions of their capacities filled with water, as the boiling water is drained completely, the buffer capacity (change in tank volume per change in water level) remains constant. In complex soil colloid-water systems, however, buffer capacities do not usually remain constant as an intensive variable is changed, and the magnitude of the buffer capacity will change accordingly. This is modeled by a peculiar samovar with a narrow middle section and wide sections at the top and bottom (Figure 2). In this samovar model, the system would be well buffered at the beginning and end of draining the tank, but would be poorly buffered in the middle. In soils, measurements of intensive variables are often made routinely (e.g., pH, Ca concentration, partial pressure of O_2, and oxidation–reduction potential), and are used to indicate the energy state or 'ability of the system to do work' in affecting other coupled systems such as groundwater or plants rooted in the soil.

The size of the capacity factor and buffering mechanisms linking the intensive and capacity parameters are often not known with the same accuracy as are measurements of intensive variables in soils (i.e., the samovar tank is invisible). Therefore, investigations of the nature and scale of buffering processes in soils are needed to understand controls on the value of an intensive variable and what chemical and biological processes control its resistance to change.

The samovar analogy for soil buffering will be used to explain similarities and differences among several soil chemical processes responsible for buffering of soil acidity, oxidation–reduction status, and metal ion activities controlled by dissolution–precipitation and organic complexation by ligands. Soil acidity and oxidation–reduction status of soils are considered master variables that control the steady-state chemical composition of soils and the kinetics of approach to chemical equilibrium. As a result, predicting their resistance to change in response to disturbance of soil conditions is relevant to many environmental quality, plant nutrition, and human health impacts of soils. Dissolution and precipitation reactions are important in weathering, soil development, and leaching processes, as well as in controlling nutrient and pollutant solubility, speciation, bioavailability, and cellular absorption. Organic and inorganic ligand complexation, typically of cations, controls ion activity in equilibrium with soluble and insoluble forms of the element. Each of these processes is a coupling of an ion activity (intensive variable), a capacity factor (reserve of ion), and a buffering capacity and buffering mechanism that govern how the intensity and capacity factors interrelate in a soil system.

Buffering of Soil Acidity

The intensive variable for soil acidity is conceptually defined as the hydrogen ion activity of soil pore water or 'active acidity,' and is operationally defined by measuring pH in a soil-water or dilute salt suspension with a glass-reference electrode system, where pH is $-\log(H^+)$ and (H^+) is the hydrogen ion or proton activity in units of moles per liter. To neutralize the active acidity (e.g., raising pH from 5.0 to 7.0) in the soil solution of the top 15 cm of a hectare of loamy soil at field capacity moisture (−10 kPa water potential) containing 200 g montmorillonite and 30 g organic matter per kilogram of soil would require approximately 500 g of $CaCO_3$. This is just a handful of lime, compared with the 10 000 kg $CaCO_3\,ha^{-1}$ that might be needed to raise the whole soil pH to 7.0. The 20 000:1 ratio of lime needed to neutralize the capacity factor or 'reserve acidity' of the soil to that needed to neutralize the active acidity reflects the relative size of the metaphorical samovar tank and its indicator tube for soil acidity.

Soil properties that determine the buffer capacity for soil pH include initial pH, aluminosilicate clay content, and mineralogy; organic-matter content; and Al(III), Fe(III), and Mn(III,IV)(hydr)oxide contents. Soil pH buffer capacity is determined by adding increasing quantities of acid or base to a given mass of soil and measuring pH after allowing sufficient equilibration time for pH to stabilize. It is often designated:

$$\beta = \mathrm{dC/dpH} \qquad [1]$$

where β is buffer index or capacity in units of moles of H^+ or OH^- added per kilogram of soil (dC) per unit pH change, dpH. The buffer capacity is quantified as the reciprocal of the slope of a linear relationship between measured pH plotted on the ordinate versus moles of H^+ or OH^- added per kilogram of soil on the abscissa.

Other, quicker indicator tests for reserve or total acidity as the capacity factor have been developed to estimate 'lime requirement' for acidic agricultural soils used to grow crops that are sensitive to acid soil conditions. Examples of these ways to estimate reserve acidity include chemical extractions of exchangeable and reactive Al^{3+}, measures of cation exchange capacity (CEC) and the fraction of exchange sites occupied by 'basic' cations (predominantly Ca^{2+}, Mg^{2+}, K^+, and Na^+), measurements of decreases in the pH of well-buffered solutions added to the soil sample, and equilibrations of a soil sample with a $BaCl_2$ solution buffered at pH 8.2, followed by titration to pH 5 to determine the quantity of acid that reacted with the buffer. Measures of free $CaCO_3$ content can be used to estimate the total alkalinity of a soil and its buffer capacity against acidification.

Diverse chemical reactions occurring on colloid surfaces of clay, oxides, and organic matter are responsible for pH buffering. They involve cation exchange reactions, dissolution and precipitation of sparingly soluble compounds, and surface charge changes in response to pH. In most soils, pH buffer capacity and buffering mechanisms are due to combinations of these reactions, and titrations and quick tests for total acidity do not distinguish between them. None the less, understanding the relative importance of these processes and their chemical basis allows predictions of the buffering behavior of diverse soils, and it allows extrapolation to unstudied soil systems and environments, particularly as related to ecological processes, soil contamination, and soil remediation.

Dissolution and Precipitation of $CaCO_3$ and $Al(OH)_3$

Soils containing free $CaCO_3$ have a pH between 7.0 and 9.5 (typically 8.0–8.5) as governed by the dissolution and precipitation of $CaCO_3$ in water containing partial pressures of CO_2 as high as 10 times that of atmospheric levels. Acidity produced by the hydration of dissolved CO_2 and by other acid-generating reactions reacts with $CaCO_3$ in accordance with:

$$CaCO_3 + H_3O^+ = Ca^{2+} + HCO_3^- + H_2O \qquad [2]$$

In the presence of high levels of soluble or exchangeable Mg^{2+} or Na^+, soluble $MgCO_3$ or Na_2CO_3 will form and allow the soil pH to rise to values higher than that controlled by $CaCO_3$–CO_2 equilibria (sometimes to pH >10). Addition of base or acid will be buffered by the lime in the soil, thereby maintaining its pH at approximately 8.2.

Under strongly acid soil conditions, Al^{3+} in soil solution and on cation exchange sites hydrolyzes and generates acidity, while the dissolution of $Al(OH)_3$ neutralizes it, leading to large β values at pHs <4.5:

$$Al^{3+} + 3H_2O = Al(OH)_3 + 3H^+ \qquad [3]$$

Soluble and exchangeable Al^{3+} is released from aluminosilicate clays upon weathering of Si-dominated tetrahedral sheets (releasing soluble monosilicic acid, $Si(OH)_4$), thereby exposing Al-dominated octahedral sheets to H^+ attack. As a result, the buffer capacity of strongly acid soils (pH <5) is dominated by Al chemistry.

Cation Exchange Reactions and pH Buffering

In the intermediate pH range between 4.5 and 6.5, cation exchange reactions on permanent aluminosilicate cation exchange sites, and on pH-dependent sites of organic matter and (hydr)oxides govern uptake and release of H^+ and OH^- as controls on the buffering-capacity size and mechanism for pH control. Permanent, negatively charged cation exchange sites on aluminosilicate clay minerals are due to isomorphous substitution of lower charge cations for Si^{4+} (e.g., Al^{3+}) in tetrahedral sheets and for Al^{3+} (e.g., Mg^{2+} and Fe^{2+}) in octahedral positions. Cation exchange sites created in this way during clay mineral formation are dominated by Ca^{2+} at and above pH 7 and become increasingly satisfied by H^+ and Al^{3+} at lower pH values. In this way, the so-called base saturation (an intensive variable reflecting the fullness of the acidity samovar) decreases as exchange sites are occupied by Al^{3+} and H^+.

Organic matter functional groups (principally carboxylic acids, represented by R—COOH, and phenolic acid groups, such as —OH on aromatic rings) are weak acids with association constants (pK_as) in the range of 2–7 for carboxylic acids and 7–10

for phenolic acid groups. These groups deprotonate at pHs $>pK_a$, and protonate at pHs $<pK_a$. When $pH = pK_a$, maximal pH buffering is observed. The complexity of humic and fulvic acid molecular aggregates and functional groups leads to overlapping pK_as that are not independent and which change as adjacent groups are protonated or deprotonated. In addition, conformational changes in the folding patterns of humic acid polymers as the suite of cations changes during titration result in changeable pK_as for the cation exchange sites responsible for pH buffering. As a result, titration curves of soil organic matter are often linear and do not reflect sharp inflection points characteristic of mono- or polyprotic simple organic acid titration curves. In addition to exchange reactions on organic acids between H^+, Al^{3+}, and Ca^{2+}, some precipitates, such as calcium oxalate, may exist in organic matter-rich horizons, such as in forest soils. These Ca–organic C precipitates may contribute to pH buffering via dissolution and precipitation reactions similar to those of $CaCO_3$, but at soil pHs <4.

pH-Dependent Surface Charge and pH Buffering

Oxide and hydroxide coatings of Fe(III), Mn(III,IV), and Al on sand, silt, and clay minerals are dominated by oxygen- and hydroxide-rich planes exposed to soil solution, and changes in pH protonate and deprotonate these functional groups, thereby creating variable or pH-dependent charge colloid surfaces that contribute to buffer capacity:

$$\begin{aligned} Me^{III}{-}OH_2^{+1} + OH^- &= Me^{III}{-}OH^0 + OH^- \\ &= Me^{III}{-}O^{-1} \end{aligned} \quad [4]$$

where Me(III) represents a trivalent metal structurally associated with the mineral surface, and the charges on the surface become increasingly negative as pH is raised. Highly weathered soils that contain few aluminosilicate clay minerals or organic matter are often dominated by these variable-charge colloids, and they contribute to pH buffering through positive- and negative-charge creation.

The relative importance of $CaCO_3$ dissolution and precipitation, Al hydrolysis, cation exchange on permanently charged aluminosilicate minerals, organic-matter functional group protonation and deprotonation, and variation in negative and positive charges on variably charged oxide coatings depend on the relative abundance of these colloids in a given soil. In many soils, more than one of these soil chemical processes are responsible for pH buffering, and titrations of whole-soil materials reflect an integration of the several processes. Table 1 summarizes the relative magnitudes of the buffer capacities associated with these processes.

Table 1 Approximate buffer capacities of soil materials and horizons in the pH range 3.0–10.0

Soil constituent or horizon	*Buffering pH range*	*Buffering index (mmol kg^{-1} pH unit^{-1})*	*pH-buffering mechanism*
Aluminosilicate clays	3–10		
Smectite	3–10	178–333	Cation exchange; mineral dissolution/precipitation; Al hydrolysis
Vermiculite	3–10	333–444	Cation exchange; mineral dissolution/precipitation; Al hydrolysis
Illite	3–10	44–88	Cation exchange; mineral dissolution/precipitation; Al hydrolysis
Kaolinite	3–10	2–11	Cation exchange; mineral dissolution/precipitation; Al hydrolysis
Soil organic matter	5–8	360–444	Protonation–deprotonation of weak acid functional groups; conformational changes
Allophane and imogolite	3–10	44–111	Protonation–deprotonation of weak acid functional groups; conformational changes
Fe(III) and Al(III) (hydr)oxides	3–10	11–89	Protonation–deprotonation of weak acid functional groups; conformational changes
$CaCO_3$ and $MgCO_3$	>7	4444	Dissolution/precipitation
Forest floor organic horizons	3–7	180–360	Metal–ligand complexation/decomplexation; cation exchange; protonation/deprotonation; of weak acid functional groups
Mineral horizons of agricultural and forest soils	4–10	53	Cation exchange, mineral dissolution/precipitation, Al hydrolysis, protonation–deprotonation of weak acid functional groups, conformational changes, and metal–ligand complexation/decomplexation; principally associated with soil organic matter

Source: Bloom PR (2000) Soil pH and pH buffering. In: Sumner M (ed.) *Handbook of Soil Science*, pp. 333–352. Boca Raton, FL: CRC Press, with permission; McBride MB (1994) *Environmental Chemistry of Soils*. New York, Oxford University Press, with permission; James BR and Riha SJ (1986) pH buffering in forest soil organic horizons: relevance to acid precipitation. *Journal of Environmental Quality* 15: 229–234, with permission.

Buffering of Soil Oxidation–Reduction Status

The concept of 'electron activity' of soils refers to the thermodynamic tendency for electrons to be transferred from reductants (e.g., Fe(II)) to oxidants (e.g., O_2), and it is operationally defined by Pt electrode potentials versus a standard reference electrode. When the measured potential is corrected for the reference electrode potential relative to the hydrogen electrode ($E^0 = 0.0$ V), it is designated E_h. This voltage can be converted to pE, a parameter analogous to pH, by dividing E_h in volts by 0.059 (the Nernstian slope factor relating electrode voltage to electron activity). The E_h or pE is the intensity variable that is a measure of electron activity or 'electron pressure' in a soil system. Buffering of electron activity is called poise, and is analogous to proton buffering, except that the processes responsible for resistance to change in pE are due to electron transfer reactions, many of which are microbial and enzymatically catalyzed.

Heterotrophic microbial respiration in soils uses organic C as the electron source (reductant) and an array of electron acceptors as oxidants. Strict aerobes use dissolved O_2 as the required oxidant to derive energy from the oxidation of organic C, the most energy-efficient process for cellular respiration. In flooded soils in which all pores are filled with water, the diffusion rate for dissolved O_2 is 10 000 times slower than it is in air. As a result, microbial respiration may deplete available O_2 faster than it is replenished, thereby leading to the onset of anaerobic conditions, with a decrease in the measured pE. Under anaerobic conditions, facultative anaerobes and strict anaerobes use alternative electron acceptors, with decreasing metabolic efficiency in the order (depending on pH): NO_3^-, Mn(III,IV)(hydr)oxides, Fe(III)(hydr)oxides, SO_4^{2-}, CO_2, and H^+.

The quantities of each of these electron acceptors (the capacity factor for electron activity) in the soil will 'poise' the E_h at a given level as governed by the free energy of reaction associated with its reduction, as coupled to the oxidation of organic C. The expected pE and E_h values at which each of these couples would poise the soil are shown at pH 5 and 7 in Table 2. Since the molecular reaction mechanism for many reduction reactions involves the transfer of the electron with a proton to the oxidant (equivalent to the addition of a H atom), the overall reaction raises the soil pH in many cases. As a result, the higher the pH of a soil, the lower the E_h or pE at which a given reduction reaction is expected to occur and at which the soil system will be poised (i.e., at a lower (H^+), a greater (e^-), or 'electron pressure' is needed to effect the given reduction).

Table 2 Common reduction half-reactions that poise soils at designated pE values at pH 5 and 7; the log *K* value is the pE for the reaction at pH 0

Half-reaction (for 1 electron reduction)	*log* K	*Poising pE*	
		pH 5	*pH 7*
$1/4O_2 + e^- + H^+ = 1/2H_2O$	20.8	15.6	13.6
$1/5NO_3^- + e^- + 6/5H^+ = 1/10N_2 + 3/5H_2O$	21.1	14.3	11.9
$1/2Mn_3O_4 + e^- + 4H^+ = 3/2Mn^{2+} + 2H_2O$	30.7	16.7	8.7
$Fe(OH)_3 + e^- + 3H^+ = Fe^{2+} + 3H_2O$	15.8	4.8	−1.2
$1/8SO_4^{2-} + e^- + 5/4H^+ = 1/8H_2S + 1/2H_2O$	5.2	−1.0	−3.5
$1/8CO_2 + e^- + H^+ = 1/8CH_4 + 1/4H_2O$	2.9	−2.1	−4.1
$e^- + H^+ = 1/2H_2$	0	−5	−7
$1/4CO_2 + e^- + H^+ = 1/24C_6H_{12}O_6 + 1/4H_2O$	−0.21	−5.9	−7.9

Buffering of Ion Activities via Dissolution–Precipitation, Ion Exchange, and Ligand Complexation

The soil solution activities of ions other than H^+ and OH^- are intensity measurements that may be controlled by dissolution–precipitation, ion exchange, and ligand complexation reactions responsible for buffering in soils. Solubility, exchange, and complexation equilibria maintain ion activities in soil solutions as leaching, and plant and microbial uptake; and other biotic and abiotic processes occur and increase or decrease particular ion activities. The disturbance of these equilibria will induce restorative shifts in accordance with the LeChatelier principle, which states that the balance of products and reactants in a dynamic chemical equilibrium will shift in response to a perturbation in order to restore the original balance of reactants and products. For example, if sparingly soluble $PbCrO_4$ is present in a soil, Pb^{2+} and CrO_4^{2-} concentrations of approximately 5.3×10^{-7} mol L^{-1} will be maintained in soil solution as controlled by the solubility product, Ksp, of 2.8×10^{-13}. Any process that depletes the soluble Pb^{2+} or CrO_4^{2-} will induce the dissolution of a small amount of solid $PbCrO_4$ to restore the equilibrium activities of the Pb^{2+} and CrO_4^{2-}. Similarly, additions of these ions to the soil solution will induce precipitation of more $PbCrO_4$. In this way, the ion activities are buffered and maintained in soil solution, as shown in eqn [5]:

$$PbCrO_4 = Pb^{2+} + CrO_4^{2-} \qquad [5]$$

Similar reactions occur via cation and anion exchange on charged colloids to buffer ion activities,

as discussed above for soil acidity buffering. The suite of exchangeable cations in soils is dominated by Ca^{2+}, Mg^{2+}, K^+, Na^{2+}, H^+, and Al^{3+}, depending on the pH of the soil, the mineralogy of the soil parent material, the aluminosilicate clay mineralogy, plant uptake, microbial processes, and human-induced changes in the soil conditions. Multiple equilibria are established in any soil, but all comprise a capacity factor, represented by the exchangeable cations, and an intensity factor. The capacity factor is sometimes called the 'quantity' term, and the quantity-to-intensity ratio (Q/I) is a measure of the nature of the cation or anion exchange buffering reactions responsible for maintaining ion activities in soil solution. Eqn [6] represents a Ca–K cation exchange reaction for such a Q/I relationship:

$$\text{Ex-Ca} + 2K^+(aq) = \text{Ex-K}_2 + Ca^{2+}(aq) \qquad [6]$$

in which Ex-Ca represents solid-phase, exchangeable Ca, $K^+(aq)$ is soluble K^+, Ex-K_2 is exchangeable K^+, and $Ca^{2+}(aq)$ is soluble Ca^{2+}. Any environmental conditions that change the activity ratio of K^+ to Ca^{2+} in soil solution (e.g., changing water content, preferential plant uptake of one cation over the other, precipitation of Ca^{2+}, or addition of Ca- or K-containing materials to the soil) will result in K-for-Ca or Ca-for-K exchange reactions that restore and buffer the ion activities in soil solution. The balance between the ratios of exchangeable Ca-to-K and soluble Ca-to-K in solution can be described by an equilibrium constant, a measure of the buffer capacity between exchangeable and soluble forms of the cations.

The activities of many ions in soil solution are also in equilibrium with complexed or chelated forms of the metals that are also soluble, but in forms in which the positive charge of the cation has been neutralized by negatively charged ligands. Any processes that increase or decrease the activity of the hexaquo form of the 'free' form of the cation, e.g., $Fe(H_2O)_6^{3+}$, that is in equilibrium with complexed forms, e.g., complexed with carboxylic acid groups, hydroxyl ions, or other Lewis bases, will induce a shift in the equilibrium to restore the balance of the complexed and free form of the ion. In this way, the 'free' form of the cation (its activity or intensity) is maintained through the equilibrium with the capacity factor, the complexed form. Similar to solubility products and cation exchange equilibria, a stability constant quantifies the relative thermodynamic stability of the complexed and free forms of a metal in a quantity–intensity relationship.

See also: **Acid Rain and Soil Acidification**; **Biodiversity**; **Environmental Monitoring**; **Eutrophication**; **Forest Soils**; **Microbial Processes:** Environmental Factors; **Nuclear Waste Disposal**; **Organic Soils**; **Pollutants:** Persistent Organic (POPs); **Pollution:** Groundwater; Industrial; **Remediation of Polluted Soils**; **Waste Disposal on Land:** Liquid; Municipal; **Wetlands, Naturally Occurring**

Further Reading

Bloom PR (2000) Soil pH and pH buffering. In: Sumner M (ed.) *Handbook of Soil Science*, pp. 333–352. Boca Raton, FL: CRC Press.

Brady NC and Weil RR (2002) *The Nature and Properties of Soils*, 13th edn. Upper Saddle River, NJ: Prentice-Hall.

James BR and Bartlett RJ (2002) Redox phenomena. In: Sumner M (ed.) *Handbook of Soil Science*, pp. 169–194. Boca Raton, FL: CRC Press.

James BR and Riha SJ (1986) pH buffering in forest soil organic horizons: relevance to acid precipitation. *Journal of Environmental Quality* 15: 229–234.

Lindsay WL (1979) *Chemical Equilibria in Soils*. New York: Wiley-Interscience.

Marschner H (1995) *Mineral Nutrition of Plants*, 2nd edn. London, UK: Academic Press.

McBride MB (1994) *Environmental Chemistry of Soils*. New York: Oxford University Press.

Reice SR (1994) Nonequilibrium determinants of biological community structure. *American Scientist* 82: 424–435.

Stumm W and Morgan JJ (1996) *Aquatic Chemistry*, 3rd edn. New York: Wiley-Interscience.

Bulk Density *See* **Porosity and Pore-Size Distribution**

C

CALCIUM AND MAGNESIUM IN SOILS

N S Bolan and P Loganathan, Massey University, Palmerston North, New Zealand
S Saggar, Landcare Research, Palmerston North, New Zealand

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Calcium (Ca) and magnesium (Mg) are the fifth and the eighth most abundant elements in the Earth's crust, respectively. While Ca is an important component of plant cell wall, Mg is the central component of chlorophyll. Both elements, regarded as secondary plant nutrients, undergo almost similar reactions in most soils and are not always considered explicitly in fertilizer-recommendation programs. One of the reasons for their exclusion in fertilizer programs is that Ca and Mg are added to soil as accessory elements in many fertilizers and liming materials. Increasing use of Ca-free ammonium phosphate fertilizers and reduced use of liming materials have resulted in increasing incidences of Ca deficiency in soils. Similarly, increasing incidences of Mg deficiency are attributed to the decreased use and reduced concentration of Mg in other major fertilizers (e.g., kainite). Further, accelerated soil acidification caused by modern agricultural practices has exacerbated Ca and Mg deficiency in soils. This article describes the inputs of Ca and Mg in soils, their plant and animal requirements, and modeling of the reactions of their compounds in soils.

Input to Soils

Calcium in soils is found mainly in minerals such as feldspar, calcite, dolomite, apatite, and hornblende. Calcium sulfate (gypsum) and calcium carbonate (calcite), which occur in arid and calcareous soils, respectively, control Ca concentration in these soils. Soils developed from calcite are generally alkaline in reaction. A high pH and the presence of Ca favor the formation of Ca humate complexes, which account for the dark color of these soils. The Ca content of soils depends on the type of parent materials and the extent of weathering. Although most soils contain 1.0–50 g kg^{-1} Ca, some of the calcareous soils contain more than 200 g kg^{-1} Ca.

Magnesium is a normal component of both igneous and sedimentary rocks, and of the soils developed from such rocks. Soils developed from basic rocks (diabase, basalts, limestone, and serpentine) generally contain high levels of Mg (2.7–28.6 g kg^{-1}) and those developed on coastal sand and granite and sandstones contain low levels of Mg (0.1–3.4 g kg^{-1}). In most soils, Mg is present in primary minerals such as biotite, serpentine, olivine, augite, and hornblende, and in secondary silicate clay minerals, chlorite, vermiculite, illite, and montmorillonite (Table 1), in organic matter as exchangeable cation, and also in soil solution. However, the majority of soil Mg is present in forms that are not readily available to the plant.

Calcium deficiency in soils can be overcome by adding Ca-containing compounds. Traditionally, superphosphates (single superphosphate, SSP; triple superphosphate, TSP) have been used as the major source of phosphorus, but they also supply Ca. The Ca in superphosphates is present in readily soluble gypsum ($CaSO_4 \cdot 2H_2O$ in SSP) and monocalcium phosphate ($Ca(H_2PO_4)_2$ in SSP and TSP) forms. The other two most commonly used Ca compounds are lime and gypsum. Lime is added mainly to overcome the problems associated with soil acidification; gypsum is used both as a sulfur (S) source and as an amendment to improve the physical conditions of soils.

A range of liming materials, which vary in their ability to neutralize the acidity, can supply Ca and Mg to soils. These include calcite ($CaCO_3$), burnt lime (CaO), slaked lime ($Ca(OH)_2$), dolomite ($CaMg(CO_3)_2$), and slag ($CaSiO_3$). The acid-neutralizing value of liming materials is expressed in terms of calcium carbonate equivalent (CCE), defined as the acid-neutralizing capacity of a liming material expressed as a weight percentage of pure $CaCO_3$. A neutralizing value of greater than 100 indicates greater efficiency of the material relative

Table 1 Calcium and magnesium minerals in soils

Mineral	*Chemical formula*	*Calcium content ($g\,kg^{-1}$)*	*Magnesium content ($g\,kg^{-1}$)*
Actinolite	$Ca(Mg,Fe)_3Si_4O_{12}$	40–70	100–160
Augite	$CaMg(SiO_3)_2$	90–120	45–100
Diopside	$CaMg(SiO_3)_2$	75–185	20–140
Hornblende	CaMg metasilicate	50–80	10–90
Gypsum	$CaSO_4 \cdot 2H_2O$	200–250	–
Calcite	$CaCO_3$	300–500	–
Fosterite	Mg_2SiO_4	–	320–350
Pyrope	$3MgO \cdot Al_2O_3 \cdot 3SiO_2$	–	60–130
Iolite	$H_2(Mg,Fe)_4Al_8Si_{10}O_{37}$	–	50–80
Enstatite	$MgSiO_3$	–	180–220
Serpentine	$H_4Mg_3Si_2O_9$	–	19–26
Talc	$H_2Mg_3Si_4O_{12}$	–	160–200
Phlogopite	$H_3Mg_3Al(SiO_4)_3$	–	130–180
Biotite	$(H,K)_2(Mg,Fe)_2Al_2Si_3O_{12}$	–	10–160
Clinochlore	$H_8(Mg,Fe)_5Al_2Si_3O_{18}$	–	100–120

Table 2 Calcium and magnesium fertilizers

Fertilizer	*Chemical formula*	*Solubility ($g\,l^{-1}$)*	*Solubility product (pK_{sp})*	*Calcium content ($g\,kg^{-1}$)*	*Magnesium content ($g\,kg^{-1}$)*
Monocalcium phosphate	$Ca(H_2PO_4)_2$	18	1.14	171	–
Dicalcium phosphate	$CaHPO_4$	0.14	6.6	294	–
FGD gypsum	$CaSO_4 \cdot 2H_2O$	–	–	232	–
Mined gypsum	$CaSO_4 \cdot 2H_2O$	–	–	387	–
Tricalcium phosphate	$Ca_3(PO_4)_2$	0.02	24.0	398	–
Hydroxyapatite	$Ca_{10}(PO_4)_6(OH)_2$	Insoluble	55.9	374	–
Carbonate apatite	$Ca_{10}(PO_4)_{6-x}(CO_3)_xF_2$	Insoluble	108.3	396	–
Fluorapatite	$Ca_{10}(PO_4)_6F_2$	Insoluble	110.2	217	–
Dolomite	$MgCO_3 \cdot CaCO_3$	0.038	17.09	258	100
Calcined dolomite	$MgO \cdot CaCO_3$	–	–	350	160
Hydrated dolomite	$MgO \cdot Ca(OH)_2$	–	–	–	170
Magnesite	$MgCO_3$	0.076	8.24	–	260
Brucite	$Mg(OH)_2$	0.091	11.41	–	360
Magnesia	MgO	0.150	–	–	560
Kieserite	$MgSO_4 \cdot H_2O$	710	–	–	160
Epsom salt	$MgSO_4 \cdot 7H_2O$	1270	–	–	90
Kainite	$MgSO_4 \cdot KCl \cdot 3H_2O$	–	–	–	70
Langbeinite	$2MgSO_4 \cdot K_2SO_4$	–	–	–	110
Fosterite	Mg_2SiO_4	0.067×10^{-3}	28.11	–	320–350

FGD, flue gas desulfurization.

to pure $CaCO_3$. The amount of liming material required to rectify soil acidity depends on the neutralizing value of the liming material and pH-buffering capacity of the soil. Recently the potential value of other Ca-containing compounds in overcoming the problems associated with soil acidification has been evaluated. Some of these materials include phosphate rocks, flue gas desulfurization (FGD) gypsum, fluidized bed boiler ash, fly ash, and lime-stabilized organic composts.

Magnesium deficiency in soils can be overcome by adding Mg fertilizers such as serpentine superphosphate, epsom salt, kieserite, dolomite, and calcined magnesite (magnesia) (Table 2). Epsom salt and kieserite are fast-release Mg sources, used both for soil and foliar applications. The other water-insoluble fertilizers are used as slow-release sources. Dolomite, which contains both Ca and Mg, is more effective in acid soils, because the Mg is brought into solution by the acid soil. Dolomite is the most widely used source of Mg, both as an ingredient of mixed fertilizers and as a separate amendment for liming. There will rarely be any need for additional Ca and Mg for any crop where continuous acidification of soils (e.g., legume-based pastures) requires a regular liming program where dolomite is the major liming material.

Reactions in Soils

Calcium and Mg are present in three major forms in soils: in solution, on exchangeable sites, and in minerals. This arbitrary division accounts for differences in their bioavailability in soils, which follows the pathway: solution > exchangeable > mineral. Unlike potassium (K^+) and ammonium (NH_4^+) ions, Ca^{2+} and Mg^{2+} ions do not get fixed in the interlayers of 2:1 silicate clay minerals. Only a small fraction of the total Ca and Mg is present in soil solution and, depending on the soil type, the majority of these elements are present in other forms. Plants absorb these two cations from soil solution, which is buffered by the readily exchangeable forms that, in turn, are slowly replenished by soil reserves containing slowly exchangeable and structural forms. There is equilibrium between the various forms of Ca and Mg that allows the release of these two from less-available forms to more-available forms (i.e., labile form).

Calcium and Mg^{2+} ions released from fertilizers undergo exchange reactions similar to K^+ ions. The adsorption sites of the soil mineral colloids are not very selective for Ca^{2+} and Mg^{2+} ions, though Ca^{2+} is slightly more preferred than Mg^{2+} because of the lower diameter of hydrated Ca^{2+} ion compared with hydrated Mg^{2+} ion. The adsorption of Ca^{2+} to organic colloids such as humic acids is, however, very specific. Thus in soils containing large amounts of Ca (calcareous soils), the humic acids are mainly present in the form of Ca humate. Ca^{2+} and Mg^{2+} ions adsorbed on to both mineral and organic colloids tend to equilibrate with these ions in solution. Most mineral soils contain sufficient concentration of Ca^{2+} in solution, and their exchangeable sites are well saturated with Ca^{2+} to meet crop demand adequately. However, in acid peat soils, the native Ca and Mg contents can be so low that plants suffer from their deficiency, requiring Ca- and Mg-containing fertilizers. In acid soils, most of the Ca and Mg would exist in soluble ionic forms. Most acid soils contain adequate Ca for most plants and only highly leached, low-CEC acid soils (sands, oxisols) may show Ca-deficiency symptoms. Plant root growth in highly acidic soils can be affected because of a high Al^{3+}-to-Ca^{2+} ratio in soil solution.

As with other cations, the requirements for Ca and Mg application to plants are based on the exchangeable soil Ca and Mg tests. Field calibrations of Ca soil tests are not available. However, field calibrations for a Mg soil test are available that enable the soil-testing service to predict the Mg status of the soils and to make necessary Mg fertilizer recommendations. A number of soil-test methods are used to predict Mg availability to plants, which include 1N ammonium acetate exchangeable Mg and percentage CEC saturated with Mg. The ability of these indices to predict the availability of Mg to plants varies depending on the relative concentration of Mg^{2+}, Ca^{2+}, and K^+ ions in soil solution. A simple guide that may be useful to overcome Ca and Mg deficiency in soils is to maintain an adequate saturation of the exchange sites. Critical Ca and Mg saturation levels are considered to be 30–50% and 5–10% of CEC, respectively. However, for perennial plants (e.g., forestry), soil tests based on dilute acid extractions have been found to predict Mg requirements better than the exchangeable-Mg soil test.

Leaching of nitrate (NO_3^-) sulfate (SO_4^{2-}) and chloride (Cl^-) ions induces the leaching of basic cations such as Ca^{2+} and Mg^{2+}. When an excessive amount of K fertilizer is added to soil, it is possible that some of the Ca^{2+} and Mg^{2+} ions retained on to the cation exchange sites will be replaced by K^+ ions, inducing the leaching of Ca and Mg. In addition, induced deficiency may occur in some soils as a result of nutrient imbalances. Soils that are heavily fertilized, particularly with materials lacking in Mg^{2+} or high in Ca^{2+}, K^+, and NH_4^+ can also induce Mg deficiency. By growing crops that require high levels of Mg throughout the growing season (e.g., tobacco, citrus, potato, cotton, and soybeans), Mg deficiency in soils may be exacerbated or induced.

Modeling the Dissolution Reactions of Calcium and Magnesium Compounds in Soils

Most of the modeling work on Ca and Mg compounds in agricultural soils is concentrated on their use as liming materials rather than essential plant nutrients. Liming materials are used to overcome Ca and Mg deficiency, and the problems associated with soil acidification. They produce alkaline hydroxyl ions (OH^-) that neutralize the acid H^+ ions, thereby decreasing the activity of Al^{3+} and Mn^{2+}, and increasing concentration of basic cations such as Ca^{2+} and Mg^{2+}. The amount of liming material required to correct soil acidity depends on a number of factors, which include the neutralizing value of the liming material, the pH-buffering capacity of the soil, the form and amount of fertilizer used, and the production capacity (e.g., stocking rate in grazed pastures) of the farm.

Based on the concept of equal-diameter reduction, a mathematical model has been developed to predict the rate of lime particle dissolution. The equal-diameter reduction hypothesis assumes that (1) the initial mass of limestone is present as spheres of uniform size, density and composition, and (2) the rate of

loss of mass is directly proportional to the instantaneous surface area of the spheres. This suggests that as the surface area decreases with dissolution, the rate of dissolution per unit surface area remains constant, i.e., the rate of dissolution is proportional to the surface area, as represented in the following equation:

$$dm/dt = -kS_a \quad [1]$$

where m is the mass that has dissolved from the surface after time t, k is the dissolution rate constant, and S_a is the total area of the spheres.

The above equation can be solved for conditions required by model of lime particles of increasing complexity of shape and size. For a simple system of particles of spherical size with equal size, the equation can be simplified to:

$$M/M_o = (1 - kt/rD_o)^3 \quad [2]$$

where M_o is the initial mass, M is the mass at time t, r is the density, and D_o is the initial radius of the spherical particle.

Commonly known as the 'cube root equation,' this equation is widely used to predict the dissolution of other insoluble fertilizer materials such as elemental sulfur and Mg fertilizers. A conceptual model for the dissolution of lime in soils has been developed incorporating the following four chemical reaction systems in soils: (1) dissolution of lime in stagnant aqueous system; (2) cation exchange reactions with Ca^{2+} and exchangeable acidity; (3) leaching and accumulation of dissolved components; and (4) the carbonate equilibrium system.

The cube root equation (eqn 2) to predict the dissolution of a range of Mg fertilizers has indicated that, within each size class, the changing particle surface area controls the rate of dissolution of dolomite. The specific dissolution rates increase with decreases in particle size.

Plant Requirements and Deficiency Symptoms

Ca is an important constituent of the plant cell wall and is involved in maintaining the turgidity of plants. Ca is required for cell elongation and cell division; it activates enzymes, particularly those that are membrane-bound, and is important in membrane permeability and the maintenance of cell integrity. Therefore, Ca provides protection against drought, salinity, mechanical stress, and toxic elements. Low Ca levels in storage organs induce high membrane permeability and allow solute diffusion in these tissues, an important mechanism for accumulating large amount of sugars from phloem in fruits and other plant storage organs. Ca moves very slowly in plants and is not transported from older leaves to younger leaves. Therefore Ca-deficiency symptoms are often noticed first in the growing tips, which require a continuous supply of Ca.

To a large extent, the Ca level in plants is genetically controlled and is little affected by Ca fertilization in the root medium, provided Ca availability is adequate for normal plant growth. Calcium deficiency is characterized by a reduction in growth of meristematic tissues, where the affected tissue becomes soft due to the dissolution of the cell walls. Calcium deficiency is rarely seen in field crops but is often observed in fruit crops such as apples. Of all the mineral elements, Ca has the greatest impact on the postharvest quality of pipfruit. In apple, the disease is called 'bitter pit,' and in tomato it is known as 'blossom end rot.'

Mg is the central unit and the only metal ion of chlorophyll in plant leaves, and it cannot be substituted by other metal ions. An insufficient supply of Mg reduces chlorophyll formation, which is likely to affect the photosynthetic ability of the plants. Because Mg is a structural element of chlorophyll, it is assigned a dominant role in the life of the plant. But it is not only on this account that Mg is indispensable to plants; its capacity for forming complexes with water can be of even greater significance, since in this way Mg has a controlling action on the swelling of plasma. Magnesium is involved in the production of starch during photosynthesis and plays a major role in the functions of many enzymes in plants.

Mg-deficiency symptoms generally show up in the leaves, since the leaf is the primary plant organ in which assimilation of photosynthate occurs. As a rule, Mg deficiency leads to chlorosis of the leaves, the formation of pale yellow spots and streaks that result from local failure in the formation of green leaf pigment. The deficiency is initially characterized by an interveinal chlorosis, although, in acute stages, the leaf may be generally deficient in both green and yellow pigments (i.e., variegated), and there may be necrosis in the areas of the leaf first affected by the deficiency. A deficiency of Mg also induces the formation of anthocyanins in some plant species such as cotton. One of the most important Mg-deficiency diseases, commonly referred to as 'sand drown,' was identified in tobacco leaves with less than $2.5\,g\,kg^{-1}$ Mg. Similarly, pastures with less than $2.0\,g\,kg^{-1}$ Mg lead to 'grass tetany' in grazing animals.

Unlike Ca, Mg is readily mobile in plants; it moves from older to younger leaves under Mg deficiency, and the deficiency symptoms therefore show first on the older leaves of the plant. Plants differ markedly in

their response to a Mg deficiency in soil. In general, buckwheat is sensitive, corn intermediate, and small grains, grasses, and clover are only slightly responsive to Mg fertilization.

Plant Uptake

Plants take up Ca and Mg as their respective cations (Ca^{2+} and Mg^{2+}). For plants to utilize Ca and Mg, these ions need first to move to the root surface, followed by uptake by the roots. As the concentration of Ca^{2+} in soil solution is generally high (0.25–12.5 mmol l^{-1}), Ca^{2+} moves in soil solution predominantly by mass flow along with water, whereas Mg^{2+} ions move in soil solution by both mass flow and diffusion, depending on the concentration in soil solution (0.125–8.3 mmol l^{-1}) and transpiration rate. Magnesium uptake increases with increasing transpiration rate, especially when the concentration of Mg in soil solution is high, which is attributed to increased mass flow. However, the rate of ion movement in soil solution by both mass flow and diffusion depends largely on soil moisture content.

Once Ca^{2+} and Mg^{2+} ions reach the root surface, these ions are absorbed by both passive and active processes. Passive absorption occurs at high concentration along electrochemical potential gradients through the apoplasmic pathway, where the ions move freely through the free spaces in the cell walls and into the epidermis. Active absorption occurs against electrochemical potential gradients through the symplasmic pathway, and this process requires respiration energy and involves carrier molecules. Ca^{2+}, Mg^{2+}, and K^{+} ions compete with each other for absorption when passive absorption occurs at high concentration. However, when active absorption occurs at low concentration, competition between these ions is unlikely to occur, because different carrier molecules are involved in the absorption of these ions.

Interactions with Other Nutrients

Ca-deficiency symptoms are rarely seen in the field because of low levels of Ca required for plant growth functions. But most of the Ca in plants and soils acts as an excluder or detoxifier of other elements such as Al and heavy metals that might otherwise be toxic. Nutritional problems related to Ca and Mg are mainly caused by impaired translocation of, or antagonism in, the uptake of NH_4^+, K^+, and Al^{3+} ions rather than by a simple Ca- or Mg-deficiency. For example, excessive use of K through fertilizer and effluent application has been shown to result in Ca and Mg deficiency in pasture.

While the addition of Mg-free liming materials (e.g., calcite) increases the supply of Ca, it has the opposite effect on the availability of Mg. Lime-induced reductions in tissue Mg level and Mg uptake by plants result from an increase in Mg adsorption due to an increase in pH-dependent adsorption sites in soils containing variable-charge components. Liming not only creates new, amorphous Al polymers, but also changes the charge character of their surfaces. More negative-charge sites are formed on such variable surface-charge materials, and cation adsorption is favored. Mg is favored over other cations for continued adsorption by such materials because of the presence of a Mg sink. However, excessive addition of Ca-containing materials such as calcite and gypsum has been shown to increase the leaching potential of Mg, mainly because the Ca^{2+} added through lime exchanges with Mg^{2+} on the soil surfaces, leading to the leaching of Mg^{2+} in the soil solution.

Animal Requirements

Calcium, as one of the constituents of hydroxyapatite mineral, forms the matrix of bone and teeth and is involved in nerve function, contraction of muscles, and blood clotting in animals. Calcium-binding proteins such as calmodulin play a central role in cellular regulation. A decrease in blood Ca levels in recently calved dairy cattle causes a disorder known as 'milk fever' (i.e., hypocalcemia), which is characterized by restlessness, muscle tremors, and sometimes coma. Animals with milk fever are treated with calcium borogluconate, administrated subcutaneously or intravenously. The ratio of the different basic cations in herbage influences animals' relative absorption of these cations. For example, increasing intake of K has been identified as the major cause of the decrease in the absorption of Ca in grazing animals that results in milk fever.

The supply of Ca in animals is monitored from the dietary cation–anion difference (DCAD) values in the pasture:

$$\mathrm{DCAD} = \left(\left[\mathrm{Na}^+\right] + \left[\mathrm{K}^+\right]\right) - \left(\left[\mathrm{Cl}^-\right] + \left[\mathrm{SO}_4^{2-}\right]\right) \qquad [3]$$

where concentrations are in milliequivalents per kilogram of dry matter. Dietary cation–anion balance is important in animals to maintain systemic acid–base balance and osmotic pressure in order both to protect the integrity of cells and membranes and to optimize biochemical and physiological processes. When herbage with excess cations over anions (positive DCAD) is fed to animals, the concentration of alkali ions, such as bicarbonate, increases in body fluids, resulting in alkalosis. Conversely, when feed

with a surplus of anions (negative DCAD) is ingested, then the concentration of acidic hydrogen ions increases and metabolic acidosis occurs. There have been conflicting reports on the optimum levels of DCAD values required for dairy cattle. It has, however, been shown that diets with high DCAD values tend to increase the incidence of milk fever, and the supplementation of precalving rations with anionic salts (low DCAD values) reduces the incidence of milk fever.

Magnesium plays an important role in the enzymatic metabolism of carbohydrates, lipids, proteins, and nucleic acids. Generally, Mg is an activator for the numerous enzymes such as phosphatases and the enzyme-catalyzing reactions involving adenosine triphosphate (ATP), which split and transfer phosphate groups. It is also involved in nerve conduction and muscular contraction. Deficiency of Mg in blood plasma causes a disorder known as 'hypomagnesemia' (grass tetany or staggers), which usually occurs in dairy cows in the early part of lactation. Mg-deficiency in animals can be overcome by regular use of Mg salts as a drench or water-trough treatment, as a lick, or in pasture after foliar application. Increasing intake of K has been identified as the major cause for decreased absorption of Mg in animals that results in tetany and convulsion. Irrigation of pasture with dairy-shed effluents rich in K has been shown to increase the incidence of Mg-deficiency, which leads to grass staggers.

In pasture and fodder crops, the grass staggers index (GSI) (eqn [4]) is used to predict the chances for the occurrence of Mg deficiency:

$$\mathrm{GSI} = [\mathrm{K}^{+}]/([\mathrm{Ca}^{2+}] + [\mathrm{Mg}^{2+}]) \qquad [4]$$

where concentrations are in milliequivalents per kilogram of dry matter. A GSI value of greater than 2.2 has been suggested to enhance the risk of grass staggers, a condition generally linked with animal serum Mg levels less than 1.0–1.5 mg per 100 ml, compared with normal levels of 1.7–3.0 mg per 100 ml. GSI values increase with increasing levels of K addition. Application of Mg fertilizers such as epsom salt is likely to decrease GSI values, mainly due to an increase in the concentration of Mg.

See also: **pH**

Further Reading

Adams F (1984) *Soil Acidity and Liming*. Madison, WI: Soil Science Society of America Publishing.

Bangerth F (1979) Calcium-related physiological disorders of plants. *Annual Review of Phytopathology* 17: 97–122.

Barber SA (1984) *Soil Nutrient Bioavailability – A Mechanistic Approach*. New York: John Wiley.

Christiansen MN and Foy CD (1979) Fate and function of calcium in tissue. *Communications in Soil Science and Plant Analysis* 10: 427–442.

Foy CD (1992) Soil chemical factors limiting plant root growth. In: Hatfield JL and Stewart BA (eds) *Advances in Soil Science*. Berlin, Germany: Springer-Verlag.

Geogievskii VI, Annenkov BN, and Samokhin VT (1981) *Mineral Nutrition of Animals*. London, UK: Butterworth.

Huttl RF and Schaaf W (eds) (1997) *Magnesium Deficiency in Forest Ecosystem*. London, UK: Kluwer Academic.

Mayland HF and Wilkinson SR (1989) Soil factors affecting magnesium availability in plant–animal systems: a review. *Journal of Animal Science* 67: 3437–3444.

McLaughlin SB and Winner R (1999) Tansley review no. 104 – Calcium physiology and terrestrial ecosystem processes. *New Phytologist* 142: 373–417.

Mengel K and Kirkby EA (1982) *Principles of Plant Nutrition*. Worblaufen-Bern, Switzerland: International Potash Institute.

Peverill KI, Sparrow LA, and Reuter DI (eds) (1999) *Soil Analysis – An Interpretation Manual*. Collingwood, Australia: CSIRO Publishing.

Saggar S and Bolan NS (2003) Secondary nutrients. In: Benbi DK and Nieder R (eds) *Handbook of Processes and Modeling in the Soil–Plant System,* pp. 561–588. New York: The Howarth Press.

Swartzenruber D and Barber SA (1965) Dissolution of limestone particles in soil. *Soil Science* 100: 287–291.

Terblanche JH and Wooldridge LG (1979) The redistribution and immobilisation of calcium in apple trees with special reference to bitter pit. *Communications in Soil Science and Plant Analysis* 10: 195–215.

Tinker PB and Lauchil A (eds) (1984) *Advances in Plant Nutrition*. New York: Praeger Publishers.

Wilkinson SR, Welch RM, Mayland HF, and Grunes DI (1990) Magnesium in plants – uptake, distribution, function, and utilization by man and animals. *Metal Ions in Biological Systems* 26: 33–56.

CAPILLARITY

D Or, University of Connecticut, Storrs, CT, USA
M Tuller, University of Idaho, Moscow, ID, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

The coexistence of gaseous, liquid, and solid phases in soil pores gives rise to a variety of interfacial phenomena that, for example, lead to spreading of liquid droplets on solid surfaces, liquid rising in capillaries and soil pores, or the entrapment of liquid in crevices. These phenomena, partially attributed to capillarity, determine retention and movement of water and solutes through soils. Hence they are of great importance in a variety of environmental and agricultural problems.

Liquid Properties

The phenomenon of capillarity in porous media results from two opposing forces: liquid adhesion to solid surfaces, which tends to spread the liquid; and the cohesive surface tension force of liquids, which acts to reduce liquid–gas interfacial area. The resulting liquid–gas interface configuration under equilibrium reflects a balance between these forces. The phenomenon of capillarity is thus dependent on solid and liquid interfacial properties such as surface tension, contact angle, and solid surface roughness and geometry.

Surface Tension

At the interface between water and solids or other fluids (e.g., air), water molecules are exposed to different forces than are molecules within the bulk fluid. For example, water molecules in the bulk liquid are subjected to uniform cohesive forces whereby hydrogen bonds are formed with neighboring molecules on all sides. In contrast, molecules at the air–water interface experience net attraction into the liquid because of lower density of water molecules on the air side of the interface, with most hydrogen bonds formed at the liquid side. The result is a membrane-like water surface that has a tendency to contract and reduce the amount of its excess surface energy. The surface tension reflects the amount of interfacial energy per unit area, or the energy required to bring molecules from the bulk liquid to increase the surface (it is also useful to express surface tension as force per unit length of interface). Different liquids vary in their surface tension σ (Table 1).

Surface tension also depends on temperature, usually decreasing linearly as the temperature rises. Thermal expansion reduces the density of the liquid and therefore also reduces the cohesive forces at the surface as well as inside the liquid phase.

Soluble substances can increase or decrease surface tension. If the affinity of the solute molecules or ions to water molecules is greater than the affinity of the water molecules to one another, then the solute tends to be drawn into the solution and to cause an increase in the surface tension. This is the effect of electrolytic solutes. For example, a 1% NaCl concentration increases the surface tension of an aqueous solution by $0.17\,\mathrm{mN\,m^{-1}}$ at 20°C. If, on the other hand, the cohesive attraction between water molecules is greater than their attraction to the solute molecules, then the latter tend to be relegated toward the surface, reducing its tension. That is the effect of many organic solutes, particularly detergents.

Contact Angle

When a liquid drop is placed on a solid surface, the angle formed between the solid–liquid (SL) interface

Table 1 Liquid–vapor interfacial tensions for various liquids

Liquid	*Temperature (°C)*	*Surface tension ($mN\,m^{-1}$)*
Water	20	72.94
	25	72.13
Methylene iodide	20	67.00
Glycerin	24	62.6
Ethylene glycol	25	47.3
Dimethyl sulfoxide	20	43.54
Propylene carbonate	20	41.1
1-Methyl naphthalene	20	38.7
Dimethyl aniline	20	36.56
Benzene	20	28.88
Toluene	20	28.52
Chloroform	25	26.67
Propionic acid	20	26.69
Butyric acid	20	26.51
Carbon tetrachloride	25	26.43
Butyl acetate	20	25.09
Diethylene glycol	20	30.9
Nonane	20	22.85
Methanol	20	22.50
Ethanol	20	22.39
Octane	20	21.62
Heptane	20	20.14
Ether	25	20.14
Perfluoromethylcyclohexane	20	15.70
Perfluoroheptane	20	13.19
Hydrogen sulfide	20	12.3
Perfluoropentane	20	9.89

Reproduced from Adamson AW (1990) *Physical Chemistry of Surfaces*, 5th edn. New York: John Wiley.

and the liquid–gas (LG) interface (Figure 1) is referred to as the equilibrium (or static) contact angle (γ). Two equivalent approaches are commonly used to describe the equilibrium contact angle on smooth and chemically homogeneous planar surfaces: (1) a force balance approach, and (2) an interfacial, free-energy minimization. The force balance formulation considers interfacial tensions (σ_{ij}) as forces per unit length; hence the force balance at the contact line of a drop resting on a solid surface under equilibrium requires the vector sum of the forces acting to spread the drop (outward) to be equal to opposing cohesion and viscous forces. The free-energy minimization approach regards interfacial tension as energy per unit area, and calculates changes in surface free energy (ΔF) due to infinitesimal displacement (ΔA):

$$\Delta F = \Delta A(\sigma_{SL} - \sigma_{GS}) + \Delta A \cos\gamma \, \sigma_{LG} \qquad [1]$$

The result is identical whether considering the minimization of free energy, with $\Delta F/\Delta A = 0$, or taking a balance of forces tangential to the solid surface; both cases yield the Young equation:

$$\sigma_{LG} \cos\gamma + \sigma_{SL} - \sigma_{GS} = 0 \qquad [2]$$

with L, G, and S indicating liquid, gas, and solid, respectively, and σ_{ij} the respective interfacial surface tensions. The equilibrium contact angle is therefore:

$$\cos\gamma = \frac{\sigma_{GS} - \sigma_{SL}}{\sigma_{LG}} \qquad [3]$$

Liquids that are attracted to solid surfaces (adhesion) more strongly than to other liquid molecules (cohesion) exhibit a small contact angle, and the solid is said to be 'wettable' by the liquid (Figure 1a). Conversely, when the cohesive force of the liquid is larger than the adhesive force, the liquid 'repels' the solid and γ is large (Figure 1b).

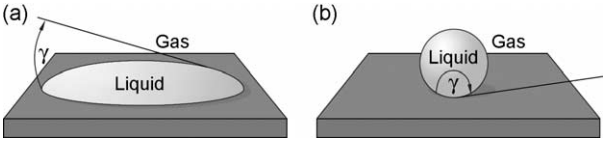


Figure 1 Liquid–solid–gas contact angles: (a) hydrophilic surface ($\gamma < 90°$) where liquid wets the surface; (b) hydrophobic surface ($\gamma > 90°$) where liquid 'repels' the surface.

Figure 2 illustrates differences in wettability of a silt soil. In Figure 2b a water droplet is resting on a soil surface that was treated to become water-repellent ($\gamma = 70°$). In contrast, Figure 2a depicts a wettable soil surface. In general, the contact angle of water on clean glass, and presumably on most soil minerals, is small, and for mathematical convenience is often taken as $\gamma = 0°$.

Curved Surfaces and Capillarity

When the forces that spread the liquid (adhesion and spreading on solids, or gas pressure within a bubble) are in balance with surface tension that tends to minimize interfacial area, the resulting liquid–gas interface is often curved. In porous media, the liquid–gas interface shape reflects the need to form a particular contact angle with solid(s) on the one hand, and the tendency to minimize interfacial area within the pore. A pressure difference forms across the curved interface, where the pressure at the concave side of the interface is greater by an amount that is dependent on the radius of curvature and the surface tension of the fluid. For a hemispherical liquid–gas interface having radius of curvature R, the pressure difference is given by the Young–Laplace equation:

$$\Delta P = \frac{2\sigma}{R} \qquad [4]$$

where $\Delta P = P_L - P_G$ when the interface curves into the gas (e.g., water droplet in air); or $\Delta P = P_G - P_L$ when the interface curves into the liquid (e.g., air bubble in water, water in a small glass tube). In many instances a bubble may not be spherical, or an element of liquid may be confined by irregular solid surfaces, resulting in two or more different radii of curvature such as water held in pendular rings between two spherical solid particles (Figure 3). The Young–Laplace equation for this case is given by:

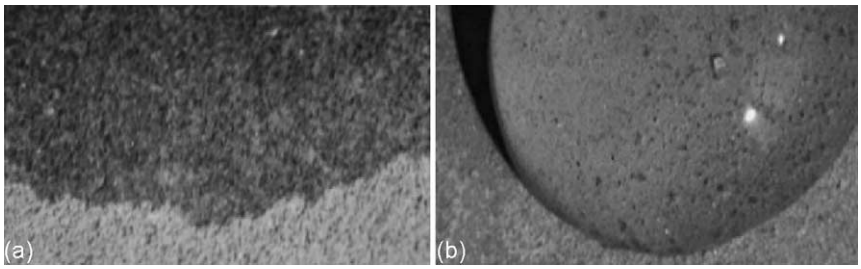


Figure 2 (a) Wettable silt surface ($\gamma \sim 0°$); (b) treated water-repellent silt soil surface ($\gamma = 70°$). (Reproduced from Bachmann J, Elliesb A, and Hartgea KH (2000) Development and application of a new sessile drop contact angle method to assess soil water repellency. *Journal of Hydrology* 231: 66–75.)

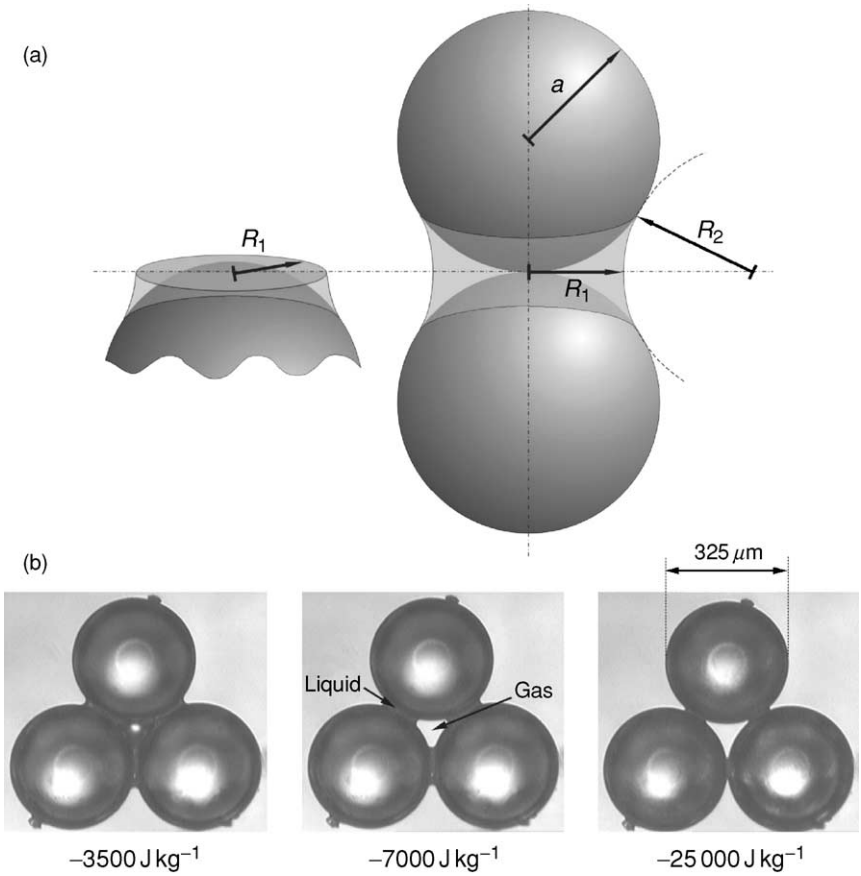


Figure 3 (a) Radii of curvature and shape of water held in pendular space between two spherical grains (note that for two equal spheres with radius a, the relationship between R_2 and R_1 is given as: $R_2 = R_1^2[2(a - R_1)]$. (b) Water menisci held between three spherical glass beads at different capillary pressures.

$$\Delta P = \sigma\left(\frac{1}{R_1} + \frac{1}{R_2}\right) \qquad [5]$$

Note that this equation reduces to eqn [4] for spherical geometry with $R_1 = R_2$, and the sign of R is negative for convex interfaces ($R_2 < 0$) and positive for concave interfaces ($R_1 > 0$). For an interface forming in a linear crevice or within a fracture, $R_2 \to \infty$, hence eqn [5] reduces to: $\Delta P = \sigma/R_1$, where R_1 equals half the fracture aperture.

The Capillary Rise Model

When a cylindrical glass tube of small diameter (capillary) is dipped into free water, a meniscus forms in the tube owing to the contact angle between water and the tube walls, and minimum surface energy requirements. The smaller the tube radius, the larger the degree of curvature and the pressure difference across the air–water interface (Figure 4). The pressure at the water side (P_W) is lower than atmospheric

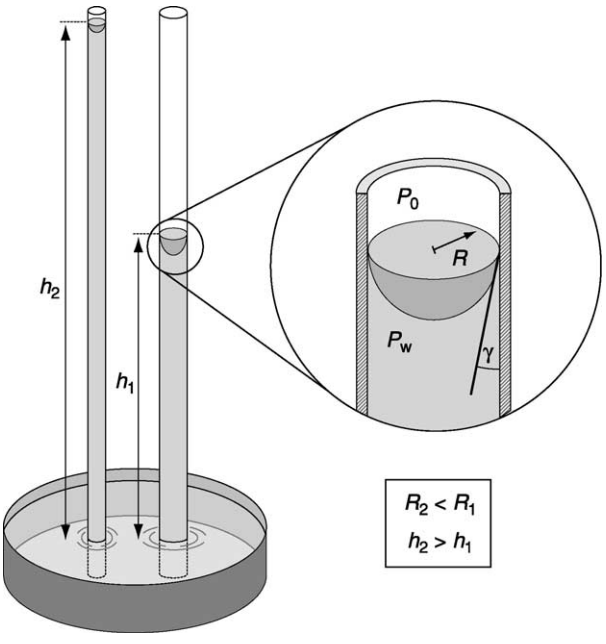


Figure 4 Capillary rise in cylindrical tubes with different radii.

pressure (P_0). This pressure difference causes water to rise into the capillary until the upward capillary force is balanced by the weight of the water column. In a cylindrical tube, the radius of meniscus curvature (R) is related to the tube radius r by $R = r/\cos\gamma$; consequently the equilibrium height of capillary rise in a cylindrical tube with contact angle γ is:

$$h = \frac{2\sigma\cos\gamma}{\rho_w g r} \qquad [6]$$

where g is the acceleration of gravity, and ρ_w is the liquid density. For water at 20°C in a glass capillary with $\gamma = 0°$, the capillary rise equation simplifies to: $h(\text{mm}) = 15/r(\text{mm})$.

Capillarity in Soils

The complex geometry of soil pore space creates numerous combinations of interfaces, capillaries, and wedges in which water is retained, and results in a variety of air–water and solid–water configurations. Water is drawn into and/or held by these interstices in proportion to the resulting capillary forces. In addition, water is adsorbed on to solid surfaces with considerable force at close distances. Due to practical limitations of present measurement methods, no distinction is made between the various mechanisms affecting water in porous matrices (i.e., capillarity and surface adsorption). Common conceptual models for water retention in porous media and matric potential rely on a simplified picture of soil pore space as a 'bundle of capillaries' (*See* **Water Retention and Characteristic Curve**). The primary conceptual steps made in such models are illustrated in Figure 5. The representation of soil pores as equivalent cylindrical capillaries greatly simplifies modeling and parameterization of soil pore space and relies heavily on the capillary rise equation (eqn[6]).

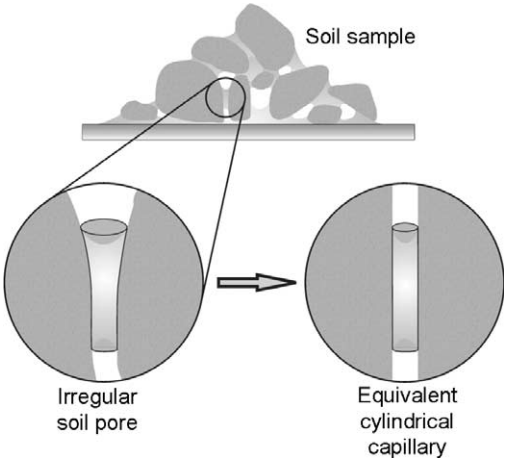


Figure 5 Idealization of the soil pore space as cylindrical capillaries.

Capillarity in Angular Pores

Cursory inspection of scanning electron micrographs of soils and other natural porous media (Figure 6) shows that pore spaces formed by aggregation of primary particles and mineral surfaces tend to be angular and slit-shaped, rarely resembling cylindrical tubes. Such observations and other shortcomings of the 'cylindrical capillary' model have led to development of new models for capillarity in angular and slit-shaped pores.

Capillarity in angular pores is quite different from the behavior in cylindrical pores with equivalent cross-sectional area. For example, when angular pores are drained, a fraction of the wetting phase (water) remains in the pore corners (Figure 7a). This aspect of 'dual occupancy' of wetting and nonwetting phases, not possible in cylindrical tubes, more realistically represents liquid configurations and the mechanisms for maintaining hydraulic continuity in porous media. Liquid-filled corners and crevices play an important role in displacement rates of oil and in other transport processes in partially saturated porous media. For all (regular and irregular) polygons with n corners, the total water filled area (A_{wt}) at a given matric potential is simply the sum of the

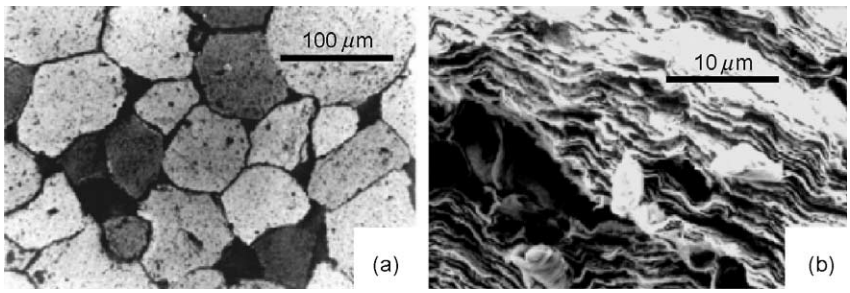


Figure 6 (a) Thin section of Devonian sandstone, revealing angular pore space. (Reproduced from Roberts JN and Schwartz LM (1985) Grain, consolidation and electrical conductivity in porous media. *Physical Review B* 31(9): 5990–5997.) (b) Scanning electron micrograph of calcium-saturated montmorillonite clay.

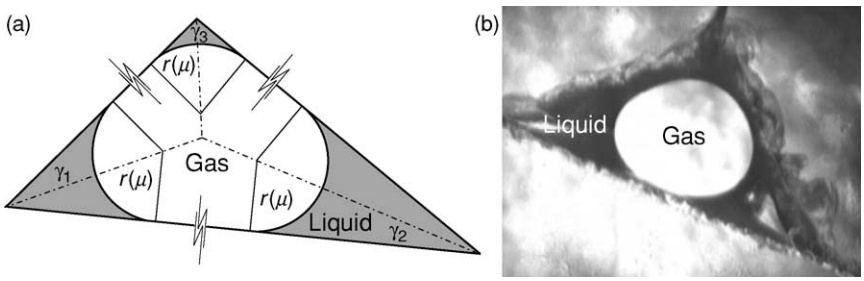


Figure 7 (a) Dual-occupancy of wetting and nonwetting phases in triangular pores; (b) liquid–vapor interfacial configuration in a triangular glass pore (~2 mm).

water-filled areas in each corner (Figure 7a). This sum is given by the simple equation:

$$A_{w_t} = r(\mu)^2 \cdot F(\gamma) \qquad [7]$$

with

$$F(\gamma) = \sum_{n=1}^{i} \left(\frac{1}{\tan\left(\frac{\gamma_i}{2}\right)} - \frac{\pi \cdot (180 - \gamma_i)}{360} \right) \qquad [8]$$

where μ is the matric potential and $F(\gamma)$ is a shape factor dependent on pore angularity (corner angles γ_i) only.

In contrast to a piston-like filling or emptying of circular capillaries, angular pores undergo different filling stages and spontaneous displacement in the transition from dry to wet or vice versa. Under relatively dry conditions (low chemical potentials) liquid accumulates in corners due to capillary forces. An increase in chemical potential leads to an increase in the capillary radius of interface curvature until the capillary corner menisci contact to form an inscribed circle. At this critical potential, liquid spontaneously fills up the central pore (pore snap-off). The radius of interface curvature at this critical point is equal to the radius of an inscribed circle in the pore cross-section. If an angular pore is drained, liquid is displaced from the central region first, leaving some liquid behind in corners. Subsequent decrease in chemical potential results in incrementally decreasing amounts of liquid in the corners. The critical potentials at spontaneous liquid displacement differ for imbibition and drainage. (*See* **Water Retention and Characteristic Curve.**)

For completeness, one must also consider the role of liquid films due to adsorption to solid surfaces. (*See* **Water Potential**; **Water Retention and Characteristic Curve.**)

Dynamic Aspects of Capillarity

Dynamics of Capillary Rise

The equilibrium height of fluid rise in a capillary (eqn [6]) does not contain any information regarding the rate of rise and the associated time scale, which is often of significant importance in many industrial and natural processes. A simple force balance can be employed between a driving capillary force F_σ:

$$F_\sigma = 2\pi \, R\sigma \cos\gamma \qquad [9]$$

and a retarding viscous force F_η (assuming Poiseuille flow):

$$F_\eta = 8\pi\eta x \frac{dx}{dt} \qquad [10]$$

to model the rate of capillary flow into a horizontal capillary. Inertial effects can be included, according to:

$$m\frac{d^2x}{dt^2} = F_\gamma - F_\mu \qquad [11]$$

where m is the mass of the liquid in the capillary, x is distance, and t is time. Substitution of the forces (eqns [9] and [10]) into eqn [11] and integration (neglecting higher-order terms) yields the so-called Lucas–Washburn–Rideal (LWR) equation:

$$x = \sqrt{\left(\frac{R\sigma}{\eta}\frac{\cos\gamma}{2}\right)t} \qquad [12]$$

which describes the rate of liquid penetration into a horizontal capillary with the dependency of x on $\sqrt{t}$. It is interesting to note that Washburn's neglect of inertial effects and Rideal's truncation of higher-order terms (r^{-n}, $n > 2$) in his series solution yield the same solution (eqn [12]). Exact solutions have been provided that fully account for inertial effects and expand the LWR expression to consider flows into horizontal grooves and other capillary shapes.

Analytical solutions for dynamic capillary rise with gravity present a mathematical challenge. Several simplified analytical solutions for the rate of capillary rise in vertical capillaries have been proposed, such as the following implicit solution:

$$\frac{\rho g R^2}{8\eta} t = z(t) - z_e \ln\left[1 - \frac{z(t)}{z_e}\right] \qquad [13]$$

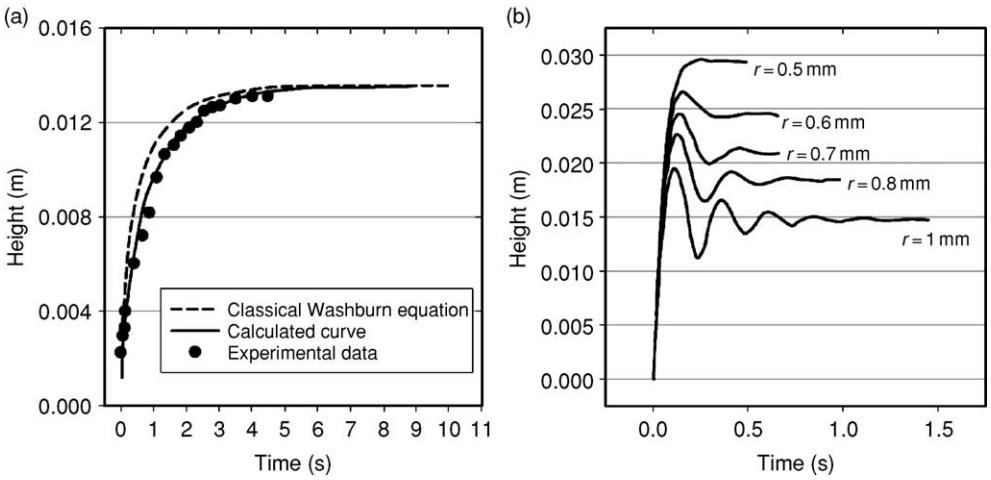


Figure 8 (a) Comparison of measurements and theoretical models for capillary rise dynamics of silicon oil (PDMS 10) in glass capillary with $r = 0.315$ mm (calculated curve from eqn [14]; classic Washburn equation from eqn [13]); (b) inertia-induced oscillations during capillary rise of water in different glass capillary sizes (numerical simulations). Note that inertial oscillations vanish for capillaries smaller than $r = 0.474$ mm according to eqn [15]. (Adapted from Hamraoui A and Nylander T (2002) Analytical approach for the Lucas–Washburn equation. *Journal of Colloid Interface Science* 250: 415–421.)

The solution diverges as $z(t)$ approaches the equilibrium capillary rise z_e (eqn [6]). Another approximate solution has been proposed, based on the introduction of a retardation coefficient (β):

$$z(t) = z_e\left(1 - \exp\left[-\frac{\sigma\cos\gamma}{\beta z_e}t\right]\right) \qquad [14]$$

The solution converges to the equilibrium capillary rise z_e (eqn [6]) for long periods of time. Figure 8a depicts comparison of eqns [13] and [14] with capillary-rise measurements of silicon oil (PDMS 10) in a glass capillary, with $r = 0.315$ mm ($\sigma = 20.1$ mN m^{-1}; $\eta = 10$ mPa; and $\rho = 0.935$ kg m^{-3}).

The nondimensional retardation coefficient for water in glass capillaries ranges from $\beta = 0.5$ for large radii ($r > r_c$), representing friction dissipation due to contact line motion and contact angle adjustment, to $\beta = 0.7$ for small radii ($r < r_c$) representing primarily viscous dissipation. The critical radius r_c is related to an interesting feature of capillary rise in the presence of gravity, namely inertia-induced oscillations in large capillaries, as depicted in Figure 8b. The inertial oscillations disappear in capillaries of radii smaller than r_c:

$$r_c = \frac{2\left(\sigma\cdot\cos(\gamma)\eta^2\rho^2 g^3\right)^{1/5}}{\rho g} \qquad [15]$$

Dynamic Contact Angle

The contact angle formed between a flowing liquid front (advancing or receding) and a solid surface is not constant but reflects the interplay between capillary and viscous forces. The relative importance of these forces is often expressed by the so-called capillary number, $\mathrm{Ca} = \eta v/\sigma$, with η the liquid dynamic viscosity, and v the contact line velocity. The dependency of the dynamic contact angle γ_D on the velocity of the contact line during complete wetting can be described by a nearly universal behavior according to the so-called Tanner law:

$$\gamma_D^3 = A\mathrm{Ca} \qquad [16]$$

where A is a constant ($\sim$94 for γ_D in radians). Eqn [16] fits the data of Hoffman for $\mathrm{Ca} < 0.1$ and $\gamma_D < 130°$ (Figure 9). The complete range of Hoffman's data fitted to the empirical expression:

$$\gamma_D(\mathrm{rad}) = \cos^{-1}\left\{1 - 2\tanh\left[5.16\left(\frac{\mathrm{Ca}}{1 + 1.31\mathrm{Ca}^{0.99}}\right)^{0.706}\right]\right\} \qquad [17]$$

is depicted by a continuous line in Figure 9.

For conditions of partial wetting ($\gamma_S > 0$), the relationships between contact angle and Ca are less universal. It has been postulated that at low Ca the apparent dynamic contact angle remains close to the static angle but rapidly deviates when Ca exceeds the value for γ_S (Figure 9). This postulate is formalized by the following expression:

$$\gamma_D^3 - \gamma_S^3 = A\mathrm{Ca} \qquad [18]$$

Additional examples of advancing and receding contact angle dependency on capillary number are shown in Figure 10. Note that for receding contact angle

there is a critical Ca above which the contact angle vanishes.

The theoretical basis for the postulate in eqn [18] was first derived by Voinov, using hydrodynamic approximations near the moving contact line, resulting in:

$$\gamma_D^3 - \gamma_S^3 = 9\text{Ca}\ \ln(Y/Y_m) \qquad [19]$$

where Y/Y_m is a ratio of macroscopic length over which the contact angle is defined (~mm) to molecular length where continuum theories fail (~nm). Application of eqn [19] with $Y/Y_m = 10^5$ to the data of Hoffman is depicted in Figure 9. A key shortcoming of such hydrodynamic models for a dynamic contact angle is the lack of consideration of the effects and interactions with solid surface properties.

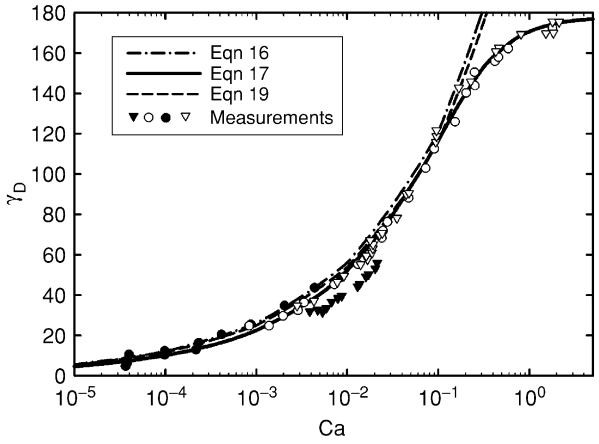


Figure 9 Experimental results of Hoffmann fitted with eqn [17] (Hoffmann RL (1975) A study of advancing interface. *Journal of Colloid Interface Science* 50: 228–241), and approximations given by eqns [16] and [19]. Note that, for water flow in soils, the capillary number Ca rarely exceeds the range of values between 10^{-6} and 10^{-4} (for $v = 1\,\text{mm s}^{-1}$, $\text{Ca} = 1 \times 10^{-5}$). (Adapted from Kistler SF (1993) Hydrodynamics of wetting. In: Berg JC (ed.) *Wettability,* pp. 311–429. New York: Marcel Dekker.)

Heterogeneous Surfaces and Microscale Hysteresis

Contact Angle on Chemically Heterogeneous and Rough Surfaces

Consider a chemically heterogeneous surface made up of patches of solids (or grains) with two different equilibrium contact angles γ_a and γ_b, and with the fraction of the area occupied by a solid given as f (Figure 11). The apparent equilibrium contact angle (γ_e) for the composite surface is given by the semi-empirical Cassie equation:

$$\cos\gamma_e = f\cos\gamma_a + (1-f)\cos\gamma_b \qquad [20]$$

An example of the Cassie law for contact angle of water on a sand surface with increasing amounts of hydrophobic grains is shown in Figure 12. The Cassie law (eqn [20]) is in remarkable agreement with experimental data for sand (Figure 12) and silt surfaces.

An interesting extension of the Cassie law for porous surfaces (soil, fabric, etc.) predicts that the apparent contact angle (γ_e) should be proportional to surface porosity (n):

$$\cos\gamma_e = (1-n)\cos\gamma_a - n \qquad [21]$$

The negative sign associated with porosity is due to the nonwetting properties of empty pores (i.e., air with $\cos\gamma_{air} = -1$).

These concepts of mixed wettability can be incorporated into the capillary rise model (eqn [6]) where capillary rise takes place in slits formed between two walls of different wettability. The same study applies

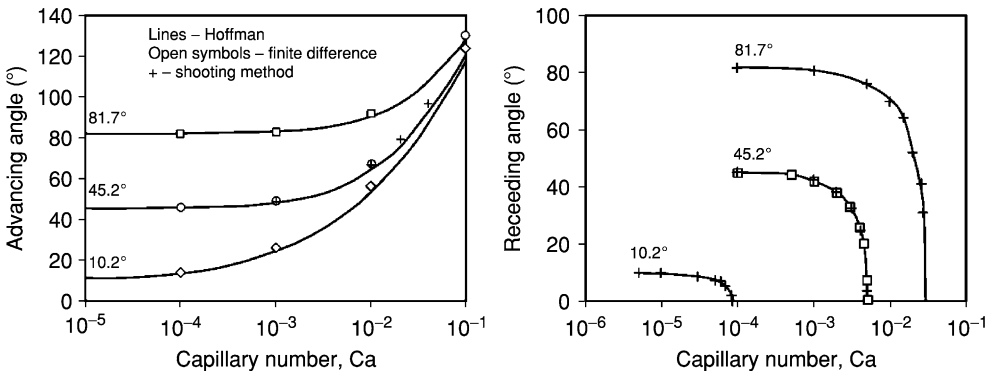


Figure 10 Finite difference computation versus eqn [18] and parameters from Kistler (Kistler SF (1993) Hydrodynamics of wetting. In: Berg JC (ed.) *Wettability,* pp. 311–429. New York: Marcel Dekker.) for advancing (left) and receding (right) contact angle as a function of Ca. (Adapted from Hirasaki GJ and Yang SY (2002) Dynamic contact line with disjoining pressure, large capillary numbers, large angles, and prewetted, precursor, or entrained films. In: Mittal KL (ed.) *Contact Angle, Wettability and Adhesion*, vol. 2, pp. 1–30.)

the Cassie law to liquid retention in porous media and demonstrates these effects on the hydraulic properties of unsaturated porous media with varying surface wettability.

In addition to surface chemical heterogeneity, the roughness of a surface is known to alter its wettability properties by increasing the wettable surface area per unit projected area, and by enabling a complex interplay between macroscopic contact angle and microscale geometry, leading to gas entrapment and a patchwork of microinterfaces underneath the wetting fluid. A spectacular demonstration of surface roughness-induced super hydrophobicity with a water drop resting on a fractal hydrophobic surface and forming a contact angle of about 170° is shown in Figure 13a. Such enhanced hydrophobicity is not only important for a variety of engineering and industrial treatments aimed at waterproofing of surfaces and fabrics, but it may also be important for explaining wettability properties of natural soil surfaces.

Assuming that surface roughness only affects the solid–liquid and solid–vapor interfacial areas, minimization of surface free energy results in the so-called Wenzel equation:

$$\cos\gamma_e = r\cos\gamma \qquad [22]$$

where γ is the static contact angle for a smooth surface of similar chemical composition (see scheme in Figure 11b).

The scope of surface influence is more complicated than predicted by simple expressions such as the Cassie and Wenzel equations. Other factors such as details of roughness geometry, interfacial pinning, and air trapping conspire to accentuate surface wetting properties as shown in Figure 13b. The scheme depicted in Figure 13b is based on experimental results showing the apparent contact angle on a rough surface plotted against the static contact angle on a smooth surface with similar chemical composition (to isolate the influence of surface roughness). Subsequent studies have shown a range of behaviors and asymmetry between the hydrophobic ($\cos\gamma < 0$) and hydrophilic ($\cos\gamma > 0$) sides of Figure 13b. It is interesting to note that certain roughness patterns

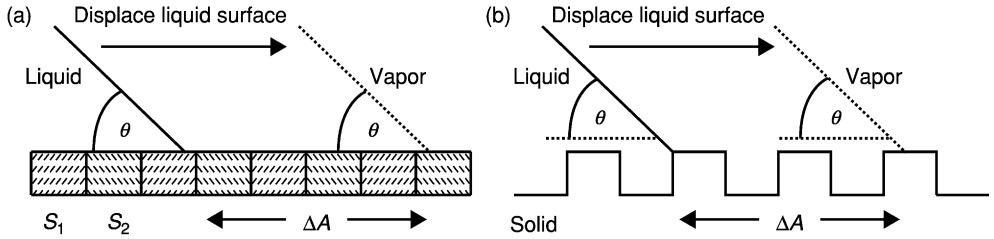


Figure 11 Definition sketch for contact angle formation on (a) chemically heterogeneous surface and (b) rough surface with $r = \Delta A/\Delta A_0$, where A_0 is the projected area over a smooth surface. (Reproduced from McHale G and Newton MI (2002) Frenkel's method and the dynamic wetting of heterogeneous planar surfaces. *Colloids and Surfaces A* 206: 193–201.)

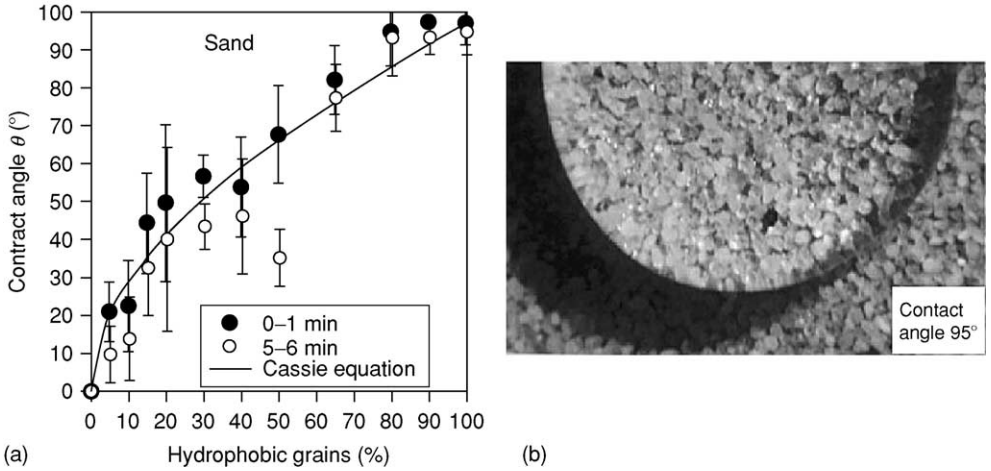


Figure 12 Application of the Cassie law to (a) experimental results of contact angle with sand surfaces containing different proportions of hydrophobic (treated) sand grains; and (b) an image of a water droplet on nonwetting sand forming a contact angle of 95°. (Adapted from Bachmann J, Elliesb A, and Hartgea KH (2000) Development and application of a new sessile drop contact angle method to assess soil water repellency. *Journal of Hydrology* 231: 66–75.)

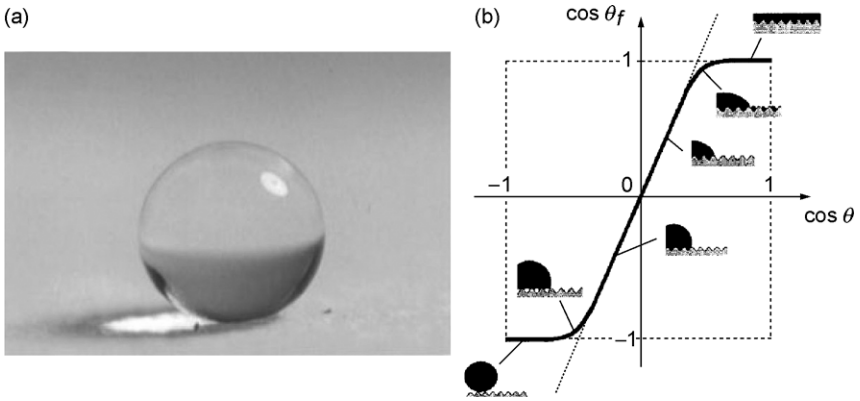


Figure 13 (a) Water drop ($r = 1$ mm) resting on fractal rough surface with $r = 4.4$ (eqn [22]); and (b) apparent contact angles as a function of surface microroughness for a range of surfaces with different wettability. (Reproduced with permission from Onda T, Shibuichi S, Satoh N, and Tsujii K (1996) Super-water-repellent fractal surfaces. *Langmuir* 12: 2125–2127.)

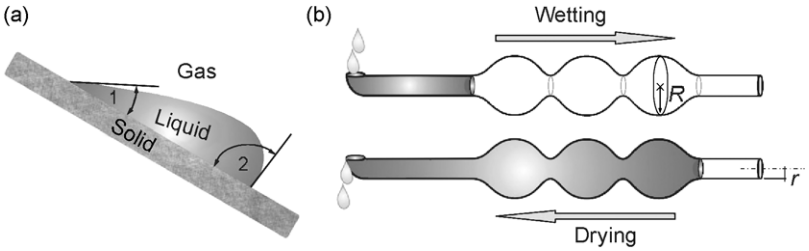


Figure 14 Two microscale mechanisms for hysteresis in capillary behavior: (a) differences between advancing and receding contact angle; and (b) the 'ink bottle' effect depicting two different amounts of liquid retained in identical pores under the same matric potential.

induce formation of air patches trapped underneath the liquid (similar to water drops resting on surfaces of some plant leaves).

Hysteresis

The amount of liquid retained in a porous medium is not uniquely defined by the value of the matric potential but is also dependent on the 'history' of wetting and drying. This phenomenon, known as hysteresis, is closely related to various aspects of pore geometry, capillarity, and surface wettability. The macroscopic manifestation of hysteresis in soil water retention (or soil water characteristic) is rooted in several microscale mechanisms, including: (1) differences in liquid–solid contact angles for advancing and receding water menisci (Figure 14a), which is accentuated during drainage and wetting at different rates; (2) the 'ink bottle' effect resulting from nonuniformity in shape and sizes of interconnected pores, as illustrated in Figure 14b, whereby drainage of the irregular pores is governed by the smaller pore radius r, and wetting is dependent on the larger radius R. Additional effects stem from pore angularity; (3) differences in air-entrapment mechanisms; and (4) swelling and shrinking of the soil under wetting and drying, respectively. From early observations to the present, the role of individual factors remains unclear, and hysteresis is a subject of ongoing research.

See also: **Water Potential**; **Water Retention and Characteristic Curve**

Further Reading

Adamson AW (1990) *Physical Chemistry of Surfaces*, 5th edn. New York: John Wiley.

Bachmann J, Elliesb A, and Hartgea KH (2000) Development and application of a new sessile drop contact angle method to assess soil water repellency. *Journal of Hydrology* 231: 66–75.

Bico J, Thiele U, and Quere D (2002) Wetting of textured surfaces. *Colloids and Surfaces A* 206: 41–46.

Blunt M and Scher H (1995) Pore-level modeling of wetting. *Physical Review E* 52(6): 6387–6403.

Dullien FAL, Lai FSY, and Macdonald IF (1986) Hydraulic continuity of residual wetting phase in porous media. *Journal of Colloid Interface Science* 109: 201–218.

Friedman SP (1999) Dynamic contact angle explanation of flow rate-dependent saturation-pressure relationships during transient liquid flow in unsaturated porous media. *Journal of Adhesion Science Technology* 13: 1495–1518.

Haines WB (1930) Studies in the physical properties of soil. V. The hysteresis effect in capillary properties, and the modes of moisture distribution associated therewith. *Journal of Agricultural Science* 20: 97–116.

Hamraoui A and Nylander T (2002) Analytical approach for the Lucas–Washburn equation. *Journal of Colloid and Interface Science* 250: 415–421.

Hirasaki GJ and Yang SY (2002) Dynamic contact line with disjoining pressure, large capillary numbers, large angles and pre-wetted, precursor, or entrained films. In: Mittal KL (ed.) *Contact Angle, Wettability and Adhesion*, vol. 2, pp. 1–30. Zeist, the Netherlands: VSP.

Hoffman RL (1975) A study of advancing interface. *Journal of Colloid Interface Science* 50: 228–241.

Kistler SF (1993) Hydrodynamics of wetting. In: Berg JC (ed.) *Wettability*, pp. 311–429. New York: Marcel Dekker.

Kool JB and Parker JC (1987) Development and evaluation of closed-form expressions for hysteretic soil hydraulic properties. *Water Resources Research* 23: 105–114.

Lucas R (1918) Ueber das Zeitgesetz des kapillaren Aufstiegs von Flussigkeiten. *Kolloid Zeitschrift* 23: 15–22.

Marmur A (1992) Wettability. In: Schrader ME and Loeb GI (eds) *Modern Approaches to Wettability: Theory and Applications*. New York: Plenum Press.

McHale G and Newton MI (2002) Frenkel's method and the dynamic wetting of heterogeneous planar surfaces. *Colloids and Surfaces A* 206: 193–201.

Morrow NR and Xie X (1998) Surface energy and imbibition into triangular pores. In: van Genuchten MT, Leij FJ, and Wu L (eds) *Proceedings of International Workshop on the Characterization and Measurement of the Hydraulic Properties of Unsaturated Porous Media*. Riverside, CA: University of California Press.

Nitao JJ and Bear J (1996) Potentials and their role in transport in porous media. *Water Resources Research* 32: 225–250.

Onda T, Shibuichi S, Satoh N, and Tsujii K (1996) Super-water-repellent fractal surfaces. *Langmuir* 12: 2125–2127.

Quere D, Raphael E, and Ollitrault J-Y (1999) Rebounds in a capillary tube. *Langmuir* 15: 3679–3682.

Rideal EK (1921) On the flow of liquids under capillary pressure. *Philosophical Magazine* 44: 1152–1159.

Rye RR, Mann JA Jr., and Yost FG (1996) The flow of liquids in surface grooves. *Langmuir* 12: 555–565.

Sciffer S (2000) A phenomenological model of dynamic contact angle. *Chemical Engineering Science* 55: 5933–5936.

Shibuichi S, Onda T, Satoh N, and Tsujii K (1996) Super water-repellent surfaces resulting from fractal structure. *Journal of Physical Chemistry* 100: 19512–19517.

Tuller M, Or D, and Dudley LM (1999) Adsorption and capillary condensation in porous media: liquid retention and interfacial configurations in angular pores. *Water Resources Research* 35(7): 1949–1964.

Ustohal P, Stauffer F, and Dracos T (1998) Measurement and modeling of hydraulic characteristics of unsaturated porous media with mixed wettability. *Journal of Contaminant Hydrology* 33: 5–37.

Voinov OV (1976) Hydrodynamics of wetting. *Fluid Dynamics* 11: 714.

Washburn EW (1921) The dynamics of capillary flow. *Physical Review* 17: 273–283.

CARBON CYCLE IN SOILS

Contents

Dynamics and Management
Formation and Decomposition

Dynamics and Management

C W Rice, Kansas State University, Manhattan, KS, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Carbon is the fundamental building block of all life on Earth. It is present in the atmosphere, plant and animal life, nonliving organic matter, fossil fuels, rocks, and is dissolved in oceans. Carbon is the sixth-most abundant element in the universe, after hydrogen, helium, oxygen, neon, and nitrogen. Concerns about increasing levels of carbon dioxide in the atmosphere have resulted in greater public and scientific interest in the global carbon cycle. Since the mid to late 1800s, fossil fuel use, expansion of cultivated agriculture, and forest clearing have led to an increase in atmospheric CO_2 from 260 ppm to current levels of approximately 370 ppm. Most of the recent increase in atmospheric CO_2 is attributed to burning of fossil fuels. Current levels of CO_2 are higher now than in the past 1000 years. The increase in CO_2 is of concern

because CO_2 is one of the three primary greenhouse gases, the other two being methane (CH_4) and nitrous oxide (N_2O). These three gases help retain heat that normally radiates from the Earth's surface. The concern is that elevated levels of CO_2 could increase the heat retained in the atmosphere thus leading to global warming. Observations have recorded a 0.6°C increase in the global temperature since 1900.

The basic carbon cycle of life is: (1) the conversion of atmospheric carbon dioxide to carbohydrates by photosynthesis in plants; (2) the consumption and oxidation of these carbohydrates by animals and microorganisms to produce carbon dioxide and other products; and (3) the return of carbon dioxide to the atmosphere (Figure 1). On a global level, the total carbon cycle is more complex and involves carbon stored in fossil fuels, soils, oceans, and rocks.

We can organize all the carbon on Earth into five main pools, listed in order of the size of the pool:

1. Lithosphere (Earth's crust): This consists of fossil fuels and sedimentary rock deposits such as limestone, dolomite, and chalk. This is much the largest carbon pool on Earth. The amount of carbon in the lithosphere is 66–100 million Pg (1 Pg = 10^{15} g). Of this amount, only 4000 Pg consists of fossil fuels;
2. Oceans: Ocean waters contain dissolved carbon dioxide, and calcium carbonate shells in marine organisms. The amount of carbon in oceans is 38 000–40 000 Pg;
3. Soil organic matter: The amount of carbon store in soil organic matter is 1500–1600 Pg;
4. Atmosphere: The atmosphere consists primarily of carbon dioxide, carbon monoxide, and methane. The amount of carbon in the atmosphere has increased from 578 Pg in 1700 to about 766 Pg in 1999, and continues to increase at the rate of about 6.1 Pg year^{-1};
5. Biosphere: The biosphere consists of all living and dead organisms not yet converted into soil organic matter. The amount of carbon that resides in the biosphere is 540–610 Pg.

In the terrestrial system, soils play a key role both in reservoirs and fluxes in the plant–soil–atmosphere continuum. Carbon in soils is present both in inorganic and organic forms. Carbon comprises up to 10% of the soil mass, with the exception of waterlogged soils such as Histosols that contain up to 30% carbon. The surface horizons of most soils contain less than 3% carbon. Most of the carbon that resides in soil is in the organic form, with the exception of arid soils and some soils formed from carbonate parent material. While carbon is a relatively minor component in terms of mass, it serves an important function in soil and the environment.

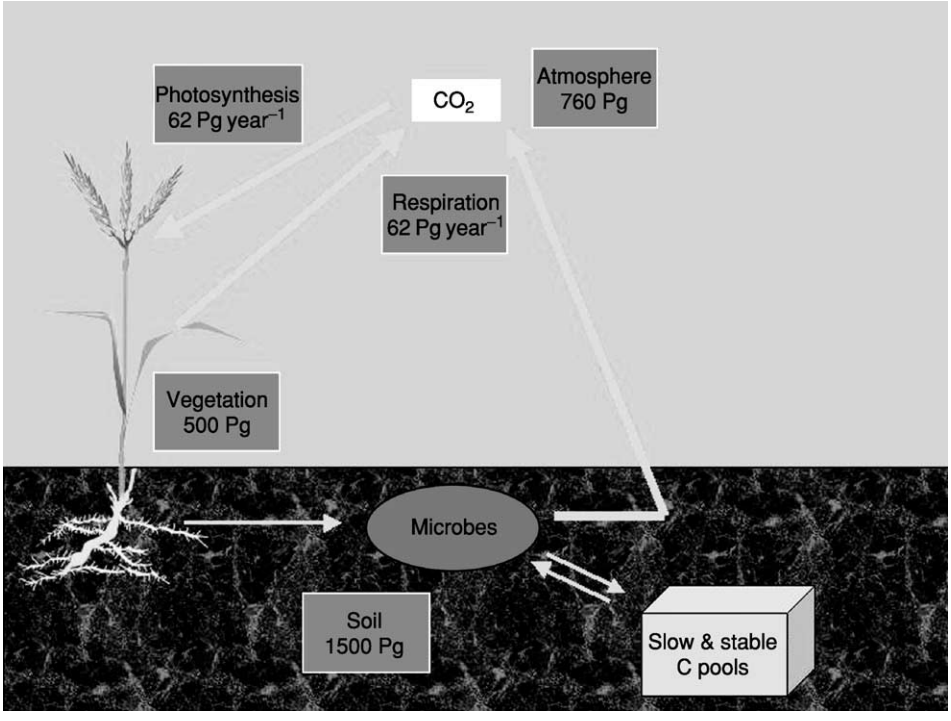


Figure 1 Simplified representation of the terrestrial carbon cycle between the atmosphere, plants, and soil.

Fluxes

The largest fluxes of the carbon cycle are those that occur between the atmosphere and the vegetation and oceans. Carbon enters the soil primarily through plants (Figure 1). During photosynthesis, plants convert CO_2 into organic carbon for growth. The estimated fixation by photosynthesis is approximately 62 Pg year^{-1}. This plant carbon enters the soil through root exudates, root turnover, and the addition of aboveground plant material. Upon their death, plant tissues decompose, primarily by soil microorganisms, and then much of the carbon in the plant material is eventually released back into the atmosphere as CO_2. Respiration is estimated to be 62 Pg year^{-1}, which balances photosynthesis. This compares with 5 Pg year^{-1} released by burning of fossil fuels; however the release of C by burning of fossil fuels is not offset by a sink, thus resulting in the dramatic increase in atmospheric CO_2. In some cases under anaerobic conditions, methane (CH_4) is produced by microorganisms. However, soil microorganisms also consume CH_4 under aerobic conditions. Approximately 7% of the atmospheric C is recycled between plant uptake and microbial respiration. Thus plant productivity and microbial respiration are the two key mediators between atmospheric CO_2 uptake and input.

Soil-Forming Factors

The level of organic carbon in soils is a function of climate, topography, biology (plant and microorganisms), parent material, and time. These are the same factors that affect soil development. The first few years after disturbance, soil carbon increases relatively dramatically but slows until equilibrium is attained under the conditions of the existing environment. This is easily seen after reclamation of mined lands or conversion of cultivated cropland to permanent grasslands. A change in the environmental conditions with climate or management can alter the equilibrium levels. The equilibrium is reached as plant inputs balance the outputs of microbial respiration. The soil organic C levels are a result of biological recalcitrance and physical and chemical stabilization of the carbon, to be discussed later.

Climate is probably the most important factor governing soil organic carbon levels. Climate through water and temperature affect plant productivity and microbial activity. Hot climates where water is restricted have the lowest levels of soil organic carbon, because plant production is limited. Deserts are characterized by this environment. In the USA, a west to east gradient from the Rocky Mountains east shows increasing levels of soil carbon due to greater precipitation and thus increasing plant productivity.

Vegetation type can also influence the soil organic carbon levels through compositions and production of the plant biomass. Plant materials vary in their decomposability due to chemical differences. Lignin and other polyphenolic substances decrease the decomposition of the plant material. Grasses tend to promote higher soil organic carbon. Grasses not only have high productivity but also allocate more photosynthate belowground. The high densities of grass roots tend to favor formation of soil organic carbon. While forests also sequester carbon, a greater proportion of that carbon is in woody biomass. There are differences among crop species in the ability to sustain soil organic carbon levels. In general high residue-producing annual crops such as corn, wheat, and grain sorghum replenish the supply of carbon to the soil. Low residue-producing crops such as cotton and soybeans tend to deplete soil organic carbon levels.

Parent material affects levels of soil carbon through the effect on texture, mineralogy, and pH. The amount of clay in the soils also influences the ability of the soil to form stable soil aggregates, which contribute to the protection and stabilization of soil organic carbon.

Topography affects soil organic carbon levels through effects on microclimate, runoff and erosion, and evaporation. Lower slope positions often have higher organic carbon levels because of deposition and greater plant production. Slope positions of excess water also restrict microbial decomposition, resulting in buildup of soil organic C.

Functions

Soil carbon is probably the most important component in soils as it affects the soil properties. Carbon as soil organic matter alters the physical, chemical, and biological properties of the soils (Figure 2a). Soil organic matter is a primary indicator of soil quality. Improvements in soil organic matter create a more favorable environment, leading to increases in plant growth.

Physically, soil organic matter improves aggregation of soil particles. Improved aggregation results in better soil structure, allowing for movement of air and water through the soil as well as better root growth. More stable soil structure results in less soil erosion, which retains nutrients on the land and protects water quality. Higher levels of soil organic carbon reduce bulk density, thus providing an improved rooting environment. In addition, soil organic matter holds soil water, which is an important

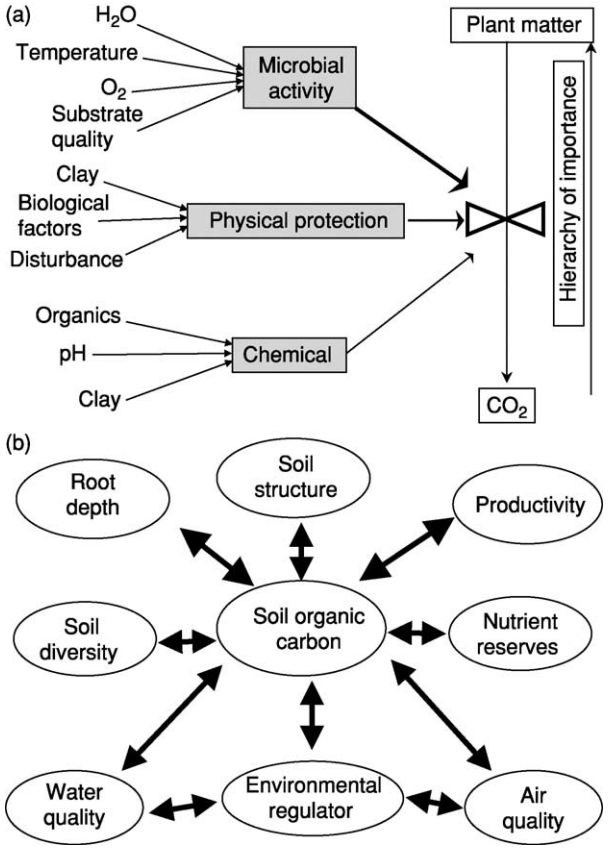


Figure 2 (a) Regulation of carbon stabilization in soil; (b) relationship between soil organic carbon and soil, water, and air quality. Reproduced with permission from Rice CW (2002) *Organic Matter and Nutrient Dynamics*. Encyclopedia of Soil Science. Dekker.

attribute for plant growth in arid and mesic environments. Higher soil organic matter decreases soil crusting and increases water infiltration rates which enhance plant productivity and thus the return of plant material to the soil.

Chemically, soil organic carbon increases the cation exchange capacity of the soil. Twenty to eighty percent of the cation exchange capacity of the soil is due to soil organic matter. These cation exchange sites are important for retention of nutrients. Associated with the organic carbon are organic-bound nitrogen, phosphorus, and sulfur which, upon decomposition, provide slow release of nutrients for plant production. In some cases, soil organic compounds enhance the chelation of metals, thus increasing the bioavailability of trace elements required for plant growth. Soil organic carbon often provides binding sites for many anthropogenic chemicals, thus minimizing leaching of hazardous chemicals through the soil profile or making them less available, which reduces toxicity.

Biologically, soil organic carbon is the source of carbon and energy for most soil microorganisms and fauna. Increased soil organic carbon enhances the biomass and diversity of the soil biota. Since the soil microbial community drives many of the microbial transformations in soil, plant nutrient availability is often enhanced with the increase in microbial biomass and activity of the soil. Some organic compounds in the soil also exhibit plant growth-promotion properties, further enhancing plant productivity.

Regulation of Soil Organic Carbon Dynamics

As much as 20% of plant carbon remains in the soil as organic matter, sometimes referred to as 'humus.' Some of this carbon can remain in soils for hundreds and even thousands of years. The quantity of organic carbon in soil is a function of the amount of plant material entering the soil, the decomposition rates of those residues, and the soil chemistry and mineralogy. Soil C may be stabilized due to biochemical recalcitrance, chemical stabilization, and physical protection (Figure 2b). Any environmental factor that affects microbial activity influences the decomposition of organic material. These factors include soil water, temperature, pH, and O_2.

Plant production and soil microbial activity are recognized as the biological processes governing soil carbon dynamics (Figure 2). The amount of plant material produced is a function of plant species, as well as climate (temperature and available water) and nutrients. The composition of the plant material entering the soil is also a key regulator of microbial decomposition. Typical plant components include soluble sugars, hemicellulose, cellulose, and lignin, which vary in proportions between plant species, within plant species, and within plant organs. Of these components, lignin is the most difficult to degrade. The quality of the carbon compounds (e.g., N content, C:N ratio, and lignin:N ratio) are important factors regulating decomposition processes (Figure 2). The C:N ratio often serves as a guide of decomposability; a ratio of more than 30 slows decomposition because of the lack of sufficient N for microbial growth; a ratio less than 20 provides sufficient N for microbial growth and allows microbial decomposition to proceed. Lignin content or lignin:N ratios are also used as a guide of organic matter degradability.

The organic carbon is transformed to greater biological recalcitrance as soil microorganisms process the C inputs from plant material. Soil carbon may be stabilized because of its biochemical recalcitrance, e.g., lignin derivatives or melanins produced by fungi and other soil organisms.

The rate of transformation is a function of temperature and soil water. As temperature increases microbial activity increases; generally, for every 10°C increase, microbial activity doubles ($Q_{10} = 2$). Soil-water content is also important; optimum microbial activity occurs at near 'field capacity,' which is equivalent to 60% water-filled spore space. As soils become waterlogged, decomposition slows and becomes less complete. Peat soils are a result of these waterlogged conditions. As soils dry below 60% water-filled pore space, decomposition is also slowed.

Soil structure plays a dominant role in controlling microbial access to substrates, and thus turnover (Figure 2). Relatively labile material may become physically protected from decomposition by incorporation into soil aggregates. Disruption of aggregates, either by natural forces (freeze–thaw, wet–dry) or human activity (tillage) stimulates decomposition of the organic C protected inside the aggregates. One of the primary causes of soil organic C loss is tillage. The role of soil structure protection of soil organic carbon from decomposition is demonstrated by increased C mineralization in disrupted aggregates relative to intact aggregates. Cultivation of soils strongly affects the structural stability and reduces the amount of soil organic C, and this loss reduces the proportion of soil macroaggregates.

Chemical stabilization of C refers to binding of organic C to clay surfaces or precipitation. Chemical stabilization by clay is regulated by clay type, pH, and other organics (Figure 3). Clay that have a high adsorption capacity such as 2:1 clays (e.g., montmorillonite), protect organic carbon to a greater extent than 1:1 clays (e.g., kaolinite), which have a lower adsorption capacity. Certain cations such as Ca, Al, and Fe are known to enhance stabilization of organic carbon through cation bridges with clay surfaces. Clay type is an inherent property of soil and therefore cannot be directly manipulated as easily as the biological and physical mechanisms of stabilization. Inorganic forms of carbon in soil are the result of equilibrium of CO_2 with water from carbonic acid (H_2CO_3), bicarbonate (HCO_3^{-1}), and carbonate (CO_2^{-3}) in the soil solution. Secondary minerals form in arid soils from the precipitation of Ca and Mg with carbonate. Dissolution of minerals also occurs depending on pH.

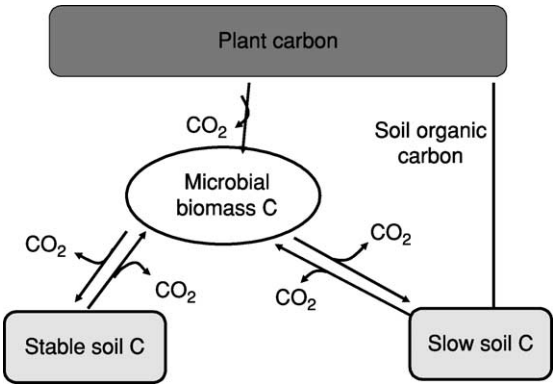


Figure 3 Conceptualization of soil carbon pools. (Adapted from Paul EA and Clark FE (1996) *Soil Biology and Biochemistry*. San Diego, CA: Academic Press, with permission.)

Composition of Soil Organic Carbon

The most widely accepted theory on the formation of soil humus carbon is that, as plant material is decomposed by microorganisms, the altered compounds and new compounds synthesized by soil microbes polymerize through chemical or enzymatic reactions. Soil organic carbon is undergoing constant transformation. Typically, most models separate soil organic carbon into pools of organic matter that differ in composition and decomposability. One model separates organic carbon into three pools (Figure 3). The 'active' pool, comprised of microbial biomass and labile organic compounds, makes up less than 5% of the soil organic C. The slow pool usually makes up 20–40% of the total organic C, and the recalcitrant pool makes up 60–70% of the soil C. The microbial biomass, while a smaller proportion of the organic C in the soil, is the processing agent. The recalcitrant pool is material that is difficult to degrade and contains what is referred in the literature as humic and fulvic acids – fractions obtained by chemical fractionation procedures. The active pool has turnover times on the order of months to years, the slow pool takes decades to turn over, while C in the recalcitrant pool takes from hundred to thousands of years to turn over completely. However, 2–5% of the recalcitrant pool is degraded annually. Since the recalcitrant pool is generally in equilibrium in natural systems, the rate for formation equals the rate of degradation.

Management

Much of the world's productive soils are now in cultivated agriculture. Historically, agriculture relied upon plowing the soil. In some cases, low crop yields and removal of crop residues reduced the amount of plant material returned back to the soil. This combination of agricultural practices resulted in reducing the replenishment of organic material (carbon) to the soil. As a result, soil C content has decreased by as much as 50% over a 50- to 100-year period in many agricultural soils. A typical soil carbon-loss curve is shown in Figure 4. Losses are rapid initially and then decline with time after the conversion to agriculture

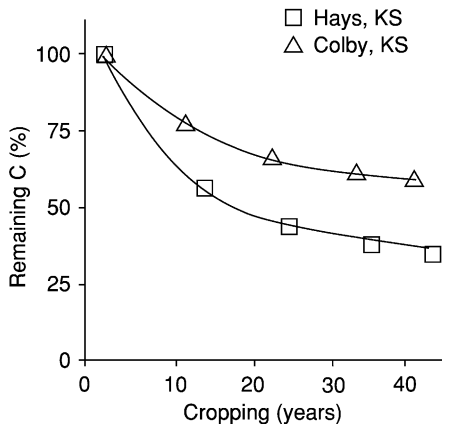


Figure 4 Loss of soil carbon at the onset of cultivation for two locations in Kansas, Hays and Colby.

Table 1 Management strategies for soil carbon sequestration

Land use	*Soil management*	*Crop management*
Cropland	Tillage management	Crop varieties
Rangeland	Residue management	Crop rotations
Forestry	Fertility management	Cover crops
	Water management	

Adapted from Lal R, Kimble JR, Follett RF, and Cole CV (1998) *The Potential of US Cropland to Sequester Carbon and Mitigate the Greenhouse Effect*. Ann Arbor, MI: Ann Arbor Press.

from native ecosystems. In recent decades, higher yields, return of crop residues, and development of conservation tillage practices have increased soil carbon. These and other advancements in crop and soil management practices have the potential to increase soil C. Table 1 lists several practices affecting the soil's ability to sequester C. These practices can be grouped into three guiding principles: (1) minimization of soil disturbance; (2) maximizing plant production through nutrients and water availability; and (3) maximizing return of plant biomass.

Cultivation or high tillage intensity is one of the primary reasons for lower organic carbon levels in soils. Tillage promotes disruption of soil aggregates, which exposes 'protected' carbon inside aggregates to microbial attack. In addition, tillage increases aeration, stimulating microbial activity, which results in conversion of organic carbon to CO_2. A reduction in tillage intensity, with no-tillage being the least disruptive in cropped soils, allows macroaggregate formation and reduces degradation of soil organic C. Minimum or conservation tillage also conserves carbon in soil. Losses of soil carbon by erosion are also reduced with conservation tillage.

Cropping intensity affects the equilibrium level of soil organic carbon. Crop rotations can impact soil carbon levels by varying the amount and composition

Table 2 Estimates of C sequestration potential of agricultural practices of US cropland

Agricultural practice	*(MTC ha^{-1} $year^{-1}$)*
Conservation Reserve Program	0.3–0.7
Conservation tillage	0.24–0.40
Fertilizer management	0.05–0.15
Rotation with winter cover crops	0.1–0.3
Summer fallow elimination	0.1–0.3

Adapted from Lal R, Kimble JR, Follett RF, and Cole CV (1998) *The Potential of US Cropland to Sequester Carbon and Mitigate the Greenhouse Effect*. Ann Arbor, MI: Ann Arbor Press.

of the plant material added to the soil. In agriculture, high residue-producing crops such as corn, wheat, and sorghum tend to sustain or increase soil C. Elimination of summer fallow, a practice used in the Great Plains, can increase soil C by providing more plant material on an annual basis. In those areas where sufficient soil water is available, double-cropping is a viable practice, where two crops are produced per year. In more humid regions, winter cover crops are also a viable option for returning plant material to the soil. In grassland systems, perennial crops also tend to increase soil C because of the lack of tillage and the addition of organic C through root turnover.

Those practices that directly affect the plant production are also important in increasing soil carbon. Improved water management alleviates short-term water stress during the growing season, which improves production of plant biomass and the return of plant material to the soil. In arid regions, irrigation can increase plant production and microbial activity. Water management also affects microbial activity, as discussed earlier. Fertility management eliminates nutrient limitations of plant growth. Proper management of nutrients provides sufficient nutrients for the plant while minimizing the excess that could lead to degradation of the environment.

For grazing lands, improved species, addition of legumes, and fertilizer management can increase soil C in tame pastures. For native lands, proper grazing rotation and management of burning are strategies to improve soil carbon, especially for previously poorly managed grazing lands.

Restoration of degraded soils and ecosystems offers high potential for storing carbon in soils. Marginal croplands could be reverted back to forests and grasslands. The Conservation Reserve Program of the US Department of Agriculture is designed to take marginal lands out of production (Table 2). Restoration of soils of mined land offers many opportunities for storing carbon in soil.

See also: **Carbon Emissions and Sequestration**; **Climate Models, Role of Soil**

Further Reading

Council for Agricultural Science and Technology (2004) *Climate Change and Greenhouse Mitigation: Challenges and Opportunities for Agriculture*. Task Force Report No. 141. Ames, IA: CAST.

Jastrow JD and Miller RM (1996) Soil aggregate stabilization and carbon sequestration: feed backs through organo-mineral associations. In: Lal R, Kimble JM, Follett RF, and Stewart BA (eds) *Soil Process and the Carbon Cycle*, pp. 207–223. New York: CRC Press.

Lal R, Kimble JR, Follett RF, and Cole CV (1998) *The Potential of U.S. Cropland to Sequester Carbon and Mitigate the Greenhouse Effect*. Ann Arbor, MI: Ann Arbor Press.

Paul EA and Clark FE (1996) *Soil Biology and Biochemistry*. San Diego, CA: Academic Press.

Paul EA, Paustian K, Elliot ET, and Cole CV (1997) *Soil Organic Matter in Temperate Agroecosystems: Long-term Experiments in North America*. Boca Raton, FL: CRC Press.

Rice CW and Owensby CE (2000) Effects of fire and grazing on soil carbon in rangelands. In: Follet R *et al.* (eds) *The Potential of U.S. Grazing Lands to Sequester Carbon and Mitigate the Greenhouse Effect*, pp. 323–342. Boca Raton, FL: Lewis Publishers.

Schlesinger WH (1997) *Biogeochemistry: An Analysis of Global Change*. San Diego, CA: Academic Press.

Smith JL, Lynch JM, Bezdicel DF, and Papendick RI (1992) Soil organic matter dynamics and crop residue management. In: Metting B (ed.) *Soil Microbial Ecology*, pp. 65–94. New York: Marcel Dekker.

Sundquist ET (1993) The global carbon dioxide budget. *Science* 259: 934–941.

Sylvia DM, Fuhrman JF, Hartel PG, and Zuberer DA (1998) *Principles and Application of Soil Microbiology*. Upper Saddle River, NJ: Prentice-Hall.

Formation and Decomposition

C A Cambardella, USDA Agricultural Research Service, Ames, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Earth's ecosystems are immense thermodynamic machines that run on energy derived from the sun. Energy enters the biotic portion of Earth's ecosystems through the process of photosynthesis, where plants use the energy of sunlight to transform CO_2, the predominant form of oxidized carbon in the atmosphere, to reduced forms of organic carbon. During photosynthesis, sunlight energy is transformed to chemical energy, and stored in the chemical bonds of organic carbon compounds produced in the plant. Photosynthetically fixed carbon, either directly or indirectly, provides the energy for all other forms of life on Earth. Living organisms use the energy tied up in the carbon bonds for growth, reproduction, and biochemical maintenance. The organic carbon and its associated energy continue to be transferred among the living components of the ecosystem and are eventually released to the atmosphere by respiration or combustion (**Figure 1**). Carbon and energy are linked as they move through ecosystems because the same processes govern their entry into, transfer through, and loss from ecosystems. Although aboveground plant and animal communities are the most obvious aspects of terrestrial ecosystems, these surface communities rely on the existence of a dynamic, efficiently functioning belowground ecosystem. This article will examine the interrelationship between plants and soil in a discussion of carbon cycling in the soil/plant system.

Sources of Carbon

Nearly all carbon entering the soil ultimately comes from plants, with small amounts coming from photosynthetic soil bacteria. Green plant tissues are, on average, 75% water and 25% dry matter. The herbaceous parts of plants are about 40–45% carbon on a dry-weight basis, regardless of age or type. Woody tissue is slightly higher, at 50% carbon. Plant dry matter also contains, on average, 42% oxygen, 8% hydrogen, and 8% ash. About half of the carbon that is photosynthetically fixed by plants is lost as CO_2 through plant respiration. Most of the remaining carbon is transferred to the soil as dead above- and belowground plant litter, and as root exudates, and is ultimately converted to soil organic carbon through the process of decomposition. Leaching can also move soluble carbon compounds from plant leaves on the soil surface into the soil, especially in environments with high rainfall. The types and average amounts of compounds associated with plant litter in decreasing order of decomposability are sugars, starches, and simple proteins (5%), protein (8%), hemicellulose (18%), cellulose (45%), fats and waxes (2%), polyphenols (2%), and lignin (20%). Compound type and amount differ among various plant parts (leaves, stems, roots, etc.) and from one plant species to another.

Anthropogenically produced organic substances represent a relatively small proportion of total carbon inputs to soil, but localized impacts of these substances on ecosystem function can be quite dramatic. This group of substances includes synthetically produced biochemicals, such as pesticides, polycyclic aromatic hydrocarbons from the burning of fossil fuels, and a variety of xenobiotics, such as plastics.

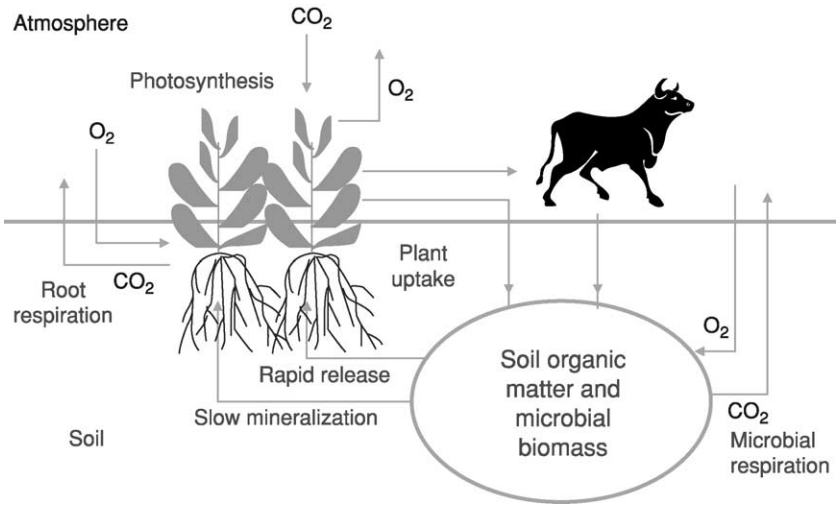


Figure 1 Soil carbon cycle. Source: USDA-NRCS, Soil Quality Institute, 2003.

Small amounts of atmospheric CO_2 can chemically react in the soil to produce carbonic acid and the carbonates and bicarbonates of calcium, potassium, and magnesium. The bicarbonates are readily soluble and can easily be lost through leaching. Carbonates produced in this manner represent a very small fraction of total carbon inputs to soil.

Aerobic Decomposition

Decomposition is the physical and chemical breakdown of organic material. In aerobic environments, decomposition is mediated by heterotrophic organisms, which derive their energy and carbon from organic matter produced by plants. Aerobic decomposition consumes oxygen, the preferred terminal electron acceptor, and returns carbon to the atmosphere as CO_2. Decomposition occurs both inside and outside living organisms and is the consequence of many interacting physical and chemical processes. The term mineralization refers specifically to the process that produces inorganic nutrient components from organic compounds. Most of the inorganic ions released by mineralization are readily available to higher plants and microorganisms.

The initial step in the decomposition process is physical fragmentation of the larger pieces of organic material into smaller ones by soil animals. There are three general categories of soil animals, classed by size. The largest are the soil macrofauna (>2 mm), such as insects, earthworms, and termites. The macrofauna impact decomposition directly through litter fragmentation and redistribution, and indirectly through burrowing. The smallest are the soil microfauna (<0.1 mm), such as flagellates, amebas, ciliates, and small nematodes. The microfauna are mobile organisms that inhabit the waterfilms on soil surfaces. They are very effective soil predators, feeding on bacteria, fungi, and other soil microfauna. Microfaunal activity impacts the temporal and spatial dynamics of nutrient mineralization and immobilization, which is an indirect, but important, control on decomposition. The soil animals that have the greatest direct effect on decomposition are the soil mesofauna (0.1–2 mm), such as collembola, mites, and other microarthropods. The soil mesofauna inhabit the air-filled pore spaces in soil, but, unlike the macrofauna, are unable to create new pore space. These organisms fragment and ingest plant litter, selectively feeding on material that has already been altered by microbial activity. They deposit large amounts of fecal material in the soil, which creates a more favorable environment for decomposition through enhanced soil water-holding capacity and surface area.

The next stage of decomposition is the biochemical alteration of the fragmented plant litter through the activity of soil bacteria and fungi. Since the compounds found in dead organic material are too large to pass through microbial membranes, the microbes secrete extracellular enzymes that convert the macromolecules into soluble products which can then be absorbed and utilized by the microbes. Numerous species of microorganisms are involved in the decomposition process because of the complex nature of the organic material.

Fungi are filamentous, aerobic soil organisms that constitute the major portion of the soil biomass. Soil fungi produce networks of filaments (hyphae) that enable them to grow into new substrates and transport materials through the soil. The hyphal network gives fungi a competitive advantage over bacteria in nutrient-poor environments, because they can

transport nutrients over distances in the soil. Fungi have enzyme systems that are capable of breaking down virtually all classes of plant compounds. They are the principal decomposers of fresh plant litter, because they secrete enzymes that enable them to penetrate the cuticle of dead leaves or the suberized exterior of roots. Only a few specialized soil fungi are capable of decomposing lignin because the chemical linkages among the phenolic rings in lignin molecules are so strong and varied. Fungi account for 60–90% of the microbial biomass in forest soils, where pH and nutrient concentrations are low and plant litter has a high lignin content. In grassland soils, there are equal amounts of bacteria and fungi, and bacteria tend to dominate in frequently disturbed agricultural soils.

The bacteria are the smallest and most numerous of the organisms in soil. There is a wide range of bacterial types in soil. Rapidly growing species specialize on labile substrates secreted by roots in the rhizosphere. Their small size and large surface-to-volume ratio enable them to absorb soluble substrates rapidly. Bacteria are also important in lysing and breaking down live and dead bacterial and fungal cells. Slower-growing actinomycete species have a filamentous structure similar to fungal hyphae and can break down relatively recalcitrant substrates. Most soil bacteria are immobile and exist in complex communities on the surfaces of soil aggregates. The communities often secrete a carbon-rich extracellular mucilage that functions as a protective coat and buffers the populations against severe environmental change. The mucilage is carbohydrate in nature and is composed primarily of low-molecular-weight polysaccharides. An important consequence of immobility is that the community eventually uses up the available food supply. When competition for food becomes severe, microbial activity declines, respiration rates become negligible, and the colony becomes inactive. At any given point in time, 50–80% of the bacteria in soil are inactive. However, these bacteria can reactivate in the presence of newly introduced labile substrates, such as when a farmer plows in fresh crop residues or when deciduous trees drop their leaves in the autumn.

Fresh inputs of above- and belowground plant litter are the primary stimulus for initiation of biochemical decomposition. Soil microorganisms first begin to oxidize the most readily decomposable compounds, incorporating a portion of the organic carbon into new microbial biomass, releasing a portion as CO_2 during respiration, and using a portion for synthesis of new organic compounds. The microbial biomass carbon can account for as much as one-sixth of the total soil carbon at the peak of microbial activity. Normally, the microbial biomass is on average about 4–5% of total soil carbon. When the easily decomposable compounds are depleted, microbial activity declines. As the opportunistic microbial populations begin to die of starvation, their bodies become part of the organic substrate available for decomposition. One year after the plant litter input, 20–30% of the plant litter carbon will remain in soil as soil organic matter. Some of this carbon will be protected from further decomposition by occlusion inside soil aggregates, by adsorption to clay surfaces, and by transformation into soil humus.

Anaerobic Decomposition

The rate of decomposition under anaerobic conditions is much slower than when oxygen is plentiful. Anaerobic soils and sediments occur in water-saturated environments, such as wetlands, swamps, estuaries, and rice paddies. Anaerobic microsites can also occur under generally aerobic soil conditions as rapid decomposition depletes the oxygen supply in soil microhabitats, such as a colony of bacteria growing around a small piece of straw or root fragment. In anaerobic environments, oxygen supply frequently limits decomposition rate, so organisms must use other electron acceptors, such as NO_3^-, SO_4^{2-}, and CO_2, to derive energy from organic matter. These alternative electron acceptors provide less energy return per unit of organic matter oxidized than oxygen, so energy production is less efficient in anaerobic systems. The dominant microorganisms in anaerobic environments are soil bacteria. The most common and widespread of these bacteria are the denitrifiers. Denitrifying bacteria use NO_3^- as a terminal electron acceptor in the decomposition of organic carbon substrates and produce nitrous oxide and di-nitrogen gas as waste products. As the nitrate supply is depleted, decomposition shifts to fermentation, where organisms break down organic material to simple organic compounds and hydrogen. These fermentation products are then used by methanogens, to transfer electrons to CO_2 to produce methane. Since the supply of SO_4^{2-} is limited in terrestrial environments, sulfate reduction is a relatively less important process than methane production. Methane and nitrous oxide gas are extremely effective at absorbing infrared radiation, and are much more potent greenhouse gases than CO_2. However, methane is an effective energy source for organisms that have access to oxygen because of its highly reduced oxidative state. Methanotrophic bacteria occur in the surface layers of anaerobic soils and use methane as an energy source through the same enzyme system that converts ammonium to nitrate. Most of the methane is consumed before it escapes to the atmosphere. Even in wetland

areas, methane accounts for a very small percentage of the carbon released to the atmosphere by decomposers.

Factors Controlling Decomposition

Decomposition is controlled by soil microclimate variables, such as moisture, temperature, and oxygen content; by intrinsic soil properties, such as soil texture and pH; and by the quality of the organic substrate available to the decomposers. In general, decomposition rates are highest when the temperature is between 25°C and 35°C and about 60% of the total soil pore volume is filled with water. Soil decomposers function best in soils with a pH that is close to neutral. Acidic soils generally have lower rates of decomposition compared with neutral or alkaline soils. Temperature directly affects decomposition through its effects on microbial activity. In general, microbial activity increases about 10-fold for every degree increase in temperature between about 5°C and 35°C. Typical soil temperature extremes seldom kill bacteria, and usually cause only temporary suppression of activity. Decomposition rates tend to decline if soil moisture levels are too low or too high. The optimum is approximately equal to the moisture content at field capacity, which is the point at which all of the larger soil pores have drained due to the force of gravity, and waters is held in the smaller capillary pores due to the matric forces in the soil. Soil microbes can function at soil moisture contents that are too low for plant growth but decomposition rates are reduced dramatically in wet, anaerobic soils. Considering the combined effects of moisture and temperature, soil carbon accumulation will be greatest in cool, wet soils, like those found in peat bogs and fens.

All else being equal, soils high in clay and silt are generally higher in organic matter than sandy soils. Finer-textured soils accumulate organic matter primarily because they lose less organic matter through decomposition than sandy soils. Clay particles bind with organic matter to form organomineral complexes that are very resistant to decomposition. Soils high in silt and clay tend to be structurally stable and can protect organic matter from decomposition through occlusion inside soil aggregates.

Substrate quality is defined as the susceptibility of the organic material to decomposition measured under standardized conditions. The quality of the organic substrate undergoing decomposition determines to a great extent the rate of biochemical oxidation. Differences in decomposition rate are a logical consequence of the types of carbon compound present in the organic material. Rapidly decomposing materials generally have higher concentrations of labile substrates and lower concentrations of recalcitrant compounds than do slowly decomposing materials. The ratio of carbon to nitrogen concentration, or C:N ratio, has frequently been used as an index of plant residue quality because residues with low C:N ratios generally decompose quickly. The C:N ratio in plant residue can vary widely with vegetation type. The range of C:N ratios of legumes and young green leaves is 10:1 to 30:1, whereas corn stalks, wheat straw, and other grass litter have a C:N ratio between 60:1 and 80:1. Wood can have C:N ratios as high as 600:1. In contrast, the C:N ratio of soil microorganisms varies between 5:1 and 10:1, with an average of 8:1. Bacteria generally contain more protein than soil fungi, and consequently have a lower C:N ratio.

Soil microorganisms require a balance of nutrients from which to build cells and extract energy. On average, soil microbes must incorporate into their cells about eight molecules of carbon for every molecule of nitrogen. Since soil microbes are not very efficient at converting substrate carbon to biomass C, only one-third of the carbon metabolized is incorporated into their cells. The remainder is lost as respired CO_2. This means that soil microbes need about 1 g of nitrogen for every 24 g of carbon that they metabolize. Most fresh plant residues do not have enough nitrogen to support the synthesis of new microbial tissue. If the C:N ratio of the plant residue is greater than 25:1, the soil microbes must scavenge the soil solution to obtain enough nitrogen. Assuming soil concentrations of inorganic nitrogen are sufficiently high, the incorporation of high C:N ratio residues will deplete the soil of soluble nitrogen. If soil concentrations of inorganic nitrogen are insufficient, then decomposition will stop, or at least be delayed.

Plant residues that are high in lignin and/or polyphenols are considered to be poor-quality substrates for the soil organisms that cycle carbon and nutrients. Lignin is degraded slowly because only some soil fungi produce the enzymes necessary to degrade lignin, and these enzymes are only produced when other more labile substrates are not available. The fungi invest more energy in the production of the enzymes to degrade lignin than they gain from the oxidation of the lignin molecule. Lignin is apparently degraded to provide access to labile compounds, such as cellulose and hemicellulose, protected within the interior of the lignin molecular structure. Phenolic compounds form highly resistant complexes with proteins in soil solution, which can slow the rates of both nitrogen mineralization and carbon oxidation.

The overall quality of plant residues can perhaps best be characterized by a combination of the indicators described in the previous paragraphs.

Poor-quality plant residues in general have a C:N ratio greater than 30:1, lignin contents greater than 20%, and polyphenol contents greater than 3%. This means that this material has a limited potential for microbial decomposition and subsequent mineralization of inorganic nutrients.

Soil Organic Matter Formation and Turnover

Soil organic matter is composed of living plant, animal, and microbial biomass, dead roots and other plant residues in various stages of decay, and soil humus. It is assumed that soil humus forms as a result of microbial activity but little is known about the exact mechanisms of humus formation. Humus contains numerous aromatic rings with attached phenolic and organic acid groups, but the remainder of the chemical structure is poorly understood. Some tentative models have been proposed for the complete molecular structure of humus but many scientists believe that a large portion of soil humus is amorphous, with no consistent molecular weight or repeating units of structure. Since humus comprises 60–80% of the total soil organic matter and has a relatively high nitrogen content, it constitutes a large reservoir of nitrogen in most ecosystems. Humus is tightly bound to the inorganic soil matrix and, as a result, decomposes extremely slowly because most of the structure is inaccessible to soil microbes and their enzymes.

The cycling of carbon in soils is the summation of the processes leading to rapid turnover of the majority of the plant litter in the surface soils, and processes leading to the slower production, accumulation, and turnover of humus deeper in the soil profile. Most decomposition occurs near the soil surface, where plant litter inputs are concentrated. Gravity carries most aboveground litter to the ground surface, where the initial decomposition and nutrient release occur. Since roots tend to grow into the surface soil to access these nutrients, most root litter is also produced primarily in the surface soil. There is a proportionally larger amount of aboveground plant litter carbon than root carbon that is lost as CO_2 during the transformation to stabilized soil organic matter. Soil mixing by animals and leaching of dissolved organic matter transfer surface carbon deeper into the soil profile. About half of the soil organic carbon is typically below 20 cm but deep-soil carbon is often older and more recalcitrant than surface soil carbon.

Soil disturbance increases soil organic carbon decomposition by exposing new soil surfaces to microbial attack and by increasing the oxygen content of the soil. The mechanism by which disturbance stimulates decomposition is basically the same at all scales. Disturbance disrupts soil aggregates so that the organic matter contained within them becomes more exposed to oxygen and microbial colonization. This disturbance effect explains why the introduction of European earthworms to the northeastern USA increased the decomposition rate of forest soils. It also explains why plowing causes rapid loss of organic matter from grassland and forest soils after conversion to agriculture, depleting the soil nutrient reserves and destroying soil structural integrity. However, it does not explain why disturbance results in a very rapid decline in soil organic matter in the first few years, followed by a much slower decline in subsequent years. In order to explain this observation, we must recognize that soil organic matter is a diverse mix of plant, animal, and microbial residues that vary in their susceptibility to microbial decomposition. This concept can be conceptualized by defining soil organic matter as a series of fractions that comprise a continuum based on decomposition rate. This can be extended further through the development of simulation models that mathematically describe the processes of organic matter formation and turnover. Models that best represent these processes include two to three organic matter pools that are kinetically defined with different turnover rates. Generally, these pools are conceptualized as one small pool with a rapid turnover rate (labile pool) and one to several pools of greater size and slower turnover rate (recalcitrant or stabilized pools).

Turnover rates of the labile pool of soil organic matter are measured in months to years. The size of the labile carbon pool can be increased very quickly through additions of fresh plant residues but carbon is also quickly lost from the labile pool when the soil is disturbed. The labile pool of organic matter is the first to be affected by soil disturbance and other changes in land management. Soil organic carbon is lost from this pool within the first few years after tillage of native grassland and forest soils. Intermediately labile pools of soil organic matter have turnover times typically measured in decades. These pools are an important source of mineralizable nitrogen and other plant nutrients, and provide a steady food source for soil microbial populations. Native grassland soils are rich in intermediately labile carbon, which accounts for the long-term fertility of these soils when they are brought under cultivation. Stabilized soil organic matter turns over every 100 to 1000 years and carbon losses from this pool are very gradual. This pool includes most of the soil humus and constitutes 60–90% of the soil organic matter in most soils. This pool is associated with the colloidal properties of humus and is involved in stabilization of soil structure.

Over the past few decades, considerable progress has been made towards understanding soil organic matter dynamics in temperate ecosystems. An important advance has been to integrate the structural and functional properties of soil organic matter and to relate them both to soil biological processes. Evidence suggests that soil organic matter turnover and consequently nutrient cycling and carbon storage potential are coupled to the formation and stabilization of soil aggregates. The most labile form of biologically active organic carbon is rapidly cycled in soil and represents a relatively small proportion of the total soil organic carbon (3–5%). The importance of the labile pool in maintaining soil fertility and controlling the supply of nutrients to plants has been extensively documented. However, the turnover time of the labile pool is so rapid that its direct effect on soil carbon storage potential may be limited. A significant proportion of biologically active organic carbon in soil is physically protected from decomposition through occlusion inside soil aggregates. Conceptually, this protected material comprises a pool of carbon that has an intermediate turnover time in soil of the order of 10–50 years and can represent up to 40% of the total soil carbon in native grassland systems. This intermediately labile pool of organic carbon is turning over at a timescale that is appropriate for assessing the impact of climate and land use change on ecosystem function, and is therefore the critical organic matter pool to monitor as a short-term relative indicator of longer term changes in soil carbon storage potential.

Further Reading

Brady NC and Weil RR (2002) *The Nature and Properties of Soils*, 13th edn. Upper Saddler River: Prentice Hall.

Chapin III FS, Matson PA, and Mooney HA (2002) *Principles of Terrestrial Ecosystem Ecology*. New York: Springer-Verlag.

Schlesinger WH (1997) *Biogeochemistry: An Analysis of Global Change*, 2nd edn, p. 588. San Diego: Academic Press.

Stevenson FJ (1986) *Cycles of Soil: Carbon, Nitrogen, Phosphorus, Sulfur, Micronutrients*. New York: John Wiley.

Tate III RL (1987) *Soil Organic Matter: Biological and Ecological Effects*. New York: John Wiley.

CARBON EMISSIONS AND SEQUESTRATION

K Paustian, Colorado State University, Fort Collins, CO, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Soils hold more than half of the carbon contained in all the earth's terrestrial ecosystems, approximately 1500 billion tonnes in organic forms and 900 billion tonnes as inorganic carbonates (measured to a depth of 1 m). Thus soils are a large repository of the carbon (C) that cycles through the atmosphere, vegetation, and soil and interacts with groundwater and aquatic ecosystems. The annual turnover of soil C is approximately 50 Pg year^{-1}, where this amount of C on average enters the soil through primary production, balanced by an approximately equal amount of C emitted from soils through decomposition. On an annual basis, this represents one of the larger fluxes in the terrestrial C cycle. Many factors, including climate and human land use and management can alter the rates and relative balance of inputs and emissions from soils, which can affect whether soils are a net source or a net sink for carbon. Thus, there is both concern that the effects of climate and land-use change on soils will exacerbate the problem of increasing CO_2 in the atmosphere and a potential that, through better management, soils can play a part in mitigating increasing CO_2 levels.

Soils and Carbon-Cycling

Carbon is added to soils mainly through the death and subsequent deposition of biomass – leaves, stems, roots – from higher plants, which assimilate CO_2 from the atmosphere through photosynthesis. The organic compounds making up plant tissues, i.e., cellulose, hemicellulose, sugars, proteins, nucleic acids, lipids, waxes, and other compounds, are the food source for a vast array of heterotrophic organisms residing in the soil. These organisms include bacteria, fungi, and other microorganisms as primary decomposers of plant materials and a diverse collection of micro-, meso-, and macrofauna that consume the microorganisms and each other. Through the metabolism of the ingested organic materials, by everything

from bacteria to earthworms, CO_2 is released in the process of respiration. (In oxygen-limited environments, carbon is also released in the form of methane (CH_4), through fermentation and anaerobic respiration.) Nearly all of the CO_2 released by respiration diffuses through the soil back to the atmosphere, although under certain conditions small amounts may enter groundwater as dissolved CO_2 or precipitate at lower soil depths to form secondary carbonates. Organic materials that are not fully metabolized (due to chemical or physical recalcitrance or inaccessibility to soils organisms), together with waste products and metabolites, can remain in soil and recombine into complex secondary compounds, generically referred to as humus. While much of the carbon entering the soil as plant (or other) biomass (collectively referred to as 'litter') is released back to the atmosphere as CO_2 within a year, some undecomposed material and especially secondary humic materials can reside in soils for hundreds to thousands of years. Sequestration refers to the buildup and storage of these C stocks in soils.

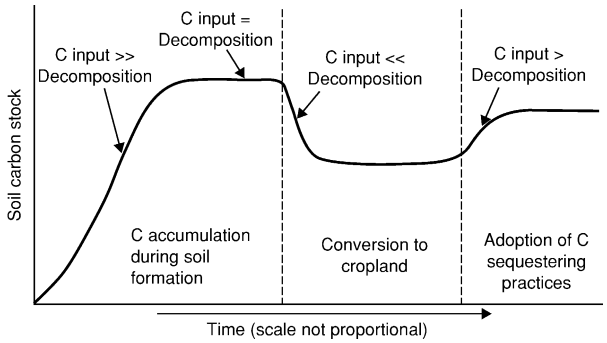


Figure 1 The buildup of soil C stocks initially during the course of soil formation and native vegetation succession, reaching an approximate equilibrium condition after several hundred to thousands of years. During this phase C inputs exceed C emissions from decomposition so that soil C stocks accumulate over time. Following human-induced disturbances such as deforestation and conversion to agriculture, soil C stocks can decline rapidly, as decomposition exceeds inputs, before leveling out toward a new, lower equilibrium state. Subsequent changes in land use (e.g., conversion back to perennial vegetation) or improved agricultural management practices can again change the balance in favor of C inputs exceeding emission, to increase soil C stocks.

Controls on Emissions and Sequestration

During the process of soil formation (pedogenesis), soils begin accumulating organic matter as vegetation becomes established and productivity and litter input increases. As ecosystems mature, soil organic C stocks tend to reach an approximate equilibrium or 'steady state,' when C inputs are balanced by C emissions. In simplified terms the rate of change in soil C stocks (C_s) can be expressed as:

$$dC_s/dt = I_c - kC_s, \quad [1]$$

in which I_c is the rate of C addition from litter fall and k is a specific rate of organic matter mineralization, implying that losses (kC_s) are directly proportional to the amount of organic C present. It is also apparent from eqn [1] that, at a fixed level of C input (which is determined by the productivity and type of vegetation), soil C stocks will tend toward an equilibrium point where $C_s = I_c/k$, where there is no change in C stocks (i.e., $dC_s/dt = 0$). At equilibrium, soils are neither a source nor a sink of C to or from the atmosphere. However, it is easy to see that if conditions are altered and C input rates and/or the specific organic matter decay rate is altered, the soil can be moved toward a new equilibrium state, along a trajectory of either soil C losses or gains (Figure 1). Shifts in environmental conditions, such as changing climate, as well as alterations in land use and management can induce changes in soil carbon and hence rates of emission or sequestration of carbon.

The environmental factors that influence soil C stocks, emissions, and sequestration are relatively well understood, at least in terms of explaining broad geographic and climatic trends. In native ecosystems, C inputs to soils are largely determined by the amount and distribution of primary productivity and the type and life cycle of the vegetation. Thus, C inputs tend to increase from dry to wet regions and from cold to warm regions. Within a given climate zone, soil fertility can vary greatly, creating smaller-scale variability in C additions. Vegetation type affects the amount and timing of inputs. For example, perennial grasses tend to allocate a high proportion of total net productivity belowground to roots, which turn over and regenerate on relatively short time scales (months to years). The long-lived, woody biomass of forests means that relatively less of the annual productivity is returned to soils in the short term. Depending on forest disturbance regimes, large pulses of carbon can be added over longer time scales with tree fall and stand replacement.

The activity of decomposer organisms also reacts to a wide variety of factors including climate as well as soil physical and chemical properties and disturbance regimes. Decomposition rates change along the same climate gradients as primary productivity, increasing from dry to wet and cold to warm. However, the effects of climate on decomposition tend to 'lag' behind productivity. Soil C stocks initially increase as conditions become warmer and wetter, but further along the climate gradient decomposition responds relatively more to increasing temperature and

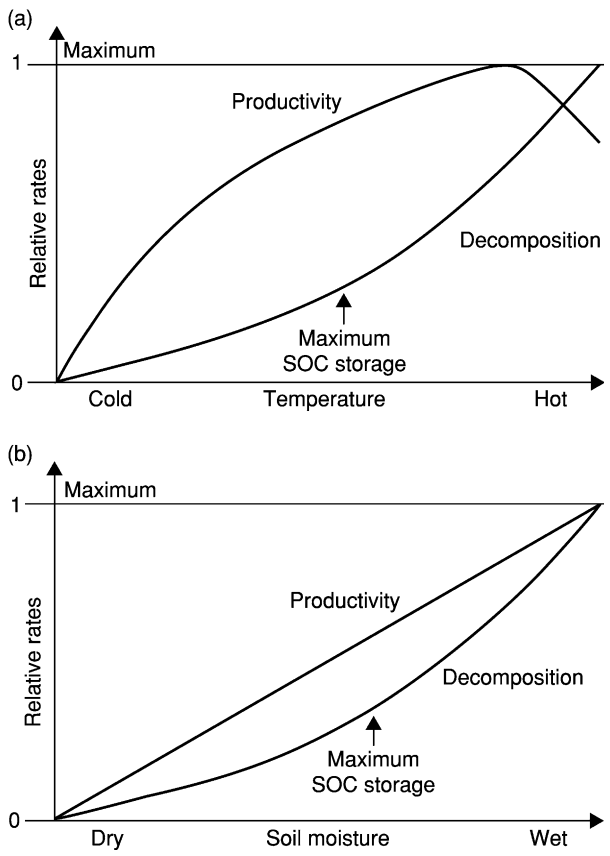


Figure 2 Climate controls on C additions and emissions through decomposition as determinants of soil C stocks. The relative responses of productivity (and C inputs) versus decomposition, across broad climate gradients relating to temperature (a) and moisture (b). Soil organic matter stocks tend to be greatest at points along the gradients where the greatest differences in the relative responses of productivity versus decomposition occur.

moisture, such that soil C stocks tend to level out or decline further along the gradients (Figure 2). Within this broad, climatically driven pattern of soil C stocks, productivity and decomposition rates (hence C storage and emissions) are influenced by soil chemistry, including nutrient availability and pH. Soil texture and mineralogy influence forces that bind organic matter to clay minerals, binding that tends to help stabilize organic matter in soils. Thus sandy soils tend to accumulate less organic matter than finer-textured soils containing more clay. Aeration provides one of the strongest controls on decomposition – in poorly drained (e.g., flooded) soils, decomposition rates are greatly reduced, which can lead to formation of so-called organic soils or peat that can extend to several meters in depth. These are some examples of the many environmental factors affecting soil C emission rates. The interaction of these factors, along with those controlling productivity and C inputs, can lead to a complex variability in soil C storage and C emissions at field to landscape scales.

Land Use and Land-Use Change

In managed ecosystems, human activities that manipulate vegetation- and soil-related processes such as land-use change and agricultural management can substantially alter C-cycling patterns of the ecosystem. Cropland represents a high degree of disturbance relative to native ecosystems, with wholesale vegetation removal (e.g., forest clearing) and replacement with annual crops and frequently intensive soil disturbance from tillage. The clearing of land for agriculture usually results in the rapid loss of organic matter from soils, often amounting to 20–40% of the total C in the top 30 cm of soil. Much of this loss occurs in the first years to decades following land conversion and plow out. On newly converted lands, management practices are often designed to exploit the nutrient reserves of soil, resulting in a 'mining out' of the organic matter. Historically, low crop productivity, harvest export, residue removal (including burning), intensive tillage, and bare fallowing have been the proximal causes of decreases in cropland soil C stocks. Wetland drainage and conversion to cropland, pasture, and forests are a special case in which decomposition and C emissions can be very high. Cultivated peat soils can lose as much as 20 Mg C ha^{-1} $year^{-1}$ in tropical and subtropical areas and 5–15 Mg C ha^{-1} $year^{-1}$ in temperate regions.

Over the past 200 years, agricultural areas have expanded greatly throughout North America and Europe, and especially in the past 100 years in tropical regions of Africa, Asia, and South America. Losses of C from these soils have amounted to as much as 50–100 Pg C since the introduction of agriculture. In many developing countries in the tropics, deforestation and land conversion to agriculture continues at a high rate, resulting in large net emissions of C. Global emissions from land-use conversion are estimated at 1–2 Pg C per year, mainly from biomass burning, but with soil C losses contributing also. In developed countries, in contrast, little new land is being converted to agriculture, and in parts of North America and Europe some cropland is being temporarily or permanently retired (i.e., government set-aside scheme), and urban and suburban development is the most significant land-use change involving agriculture. Thus, while the principles for managing land to reduce carbon emissions and sequester C in soils are similar across the globe, the most effective means and priority areas vary between developing and developed regions. In developing areas (predominantly in the tropics), this means reducing emissions

by avoiding deforestation through systems providing alternatives to slash-and-burn and more productive cropping systems. In developed countries, emphasis is on alternative production systems that rebuild organic matter stocks that were lost in previous decades and in converting marginal croplands to conservation land uses based on perennial grassland or forest vegetation.

Management Practices for Soil C Sequestration

Conditions for C sequestration in soils are most favorable where C stocks have been previously depleted but where the productivity capacity remains high. Such a condition is typical for many cropland soils. In general, C sequestration will be favored under management systems that (1) minimize soil disturbance and erosion, (2) maximize the amount of plant residue return, and (3) maximize water- and nutrient-use efficiency of plant production. Although it may be impossible to optimize all these system attributes simultaneously, practices that effectively sequester C share one or more of these traits.

Decreasing tillage intensity, especially by using no-tillage practices, has been found to promote C sequestration in many soils. Intensive tillage tends to break down soil structure, disrupting aggregates (or 'soil crumbs') that help to protect organic matter from rapid decomposition. If soil structure has been substantially degraded by intensive tillage, the resulting reductions in productivity and C inputs can also contribute to lower soil C storage. In many long-term experiments, conversion to no-till practices has been shown to increase soil C storage at rates of 0.1–0.7 Mg C ha^{-1} $year^{-1}$ or more, over periods of 20–30 years or more (Table 1). In semiarid regions, adoption of no-till can also enable higher crop C inputs by increasing water-use efficiency and allowing use of more intensified crop rotations, particularly elimination of bare (summer) fallows. Higher C inputs and more efficient water use helps to further increase soil C stocks under semiarid conditions.

Increasing the amount of residue returned to soil can be promoted through a variety of practices, including growing high residue-yielding crops, using hay crops in crop rotations, application of manure and biosolids, and by improved management of fertilizer, water, and pests. Most cropland soils show a clear response to increasing amounts of C return over time. The increase in soil C content is often directly proportional to the amount of C added to soil as crop residues, at least for a number of years or decades until a new equilibrium level is approached. Where production is water- or nutrient-limited, provision of these water and nutrient inputs can contribute to C sequestration. However, energy costs associated with manufacture and distribution of fertilizer, energy for irrigation pumping, as well as potential increased emissions of other greenhouse gases (N_2O and CH_4) from soil must be considered, for these costs may offset part or all the gains in C storage. The production-enhancing inputs (e.g., fertilizer, irrigation) are primarily used to meet objectives of food production and not as a means of mitigating greenhouse gas emissions. Practices promoting optimally efficient water and nutrient use, however, will probably have the greatest benefits in terms of decreased greenhouse gases.

Various management practices on grazing lands (pasture and rangeland) can increase soil C. On poorly managed grazing lands depleted of soil carbon, practices that increase production and C inputs can build up soil C. Such practices include improving grazing management, using improved species, sowing legumes, fertilizing, and irrigating (Table 1). As for annual crop systems, management of grazing lands for greenhouse gas mitigation needs to consider the net effects of practices on all greenhouse gases. For example, high nitrogen (N) fertilization rates in intensively managed pastures may cause large N_2O emissions that wipe out benefits from carbon sequestration, whereas phosphorus (P) fertilization and/or moderate N in highly P- or N-limited systems can yield large gains in productivity and C sequestration with little increase in N_2O emissions. Improvements in pasture productivity and forage quality through improved management can sequester C and also reduce methane emissions from grazing livestock by increasing forage quality.

Table 1 Representative soil C sequestration rates for some management practices in cropland, grazing lands, set-aside, and afforested cropland[a]

Management activities	*Sequestration (Mg C ha^{-1} $year^{-1}$)*
Adoption of no-tillage practices on cropland	0.1–0.7
Conversion of cropland to perennial grassland (e.g., set-aside)	0.5–1.5
Afforestation of cropland	0.1–0.5
Grazing-land improvement	0.1–1.0

[a]Estimates based on data syntheses reported by: Conant RT, Paustian K, and Elliott ET (2001) Grassland management and conversion into grassland: effects on soil carbon. *Ecological Application* 11: 343–355; Paustian K, Andren O, Janzen H *et al.* (1997) Agricultural soil as a C sink to offset CO_2 emissions. *Soil Use and Management* 13: 230–244; Post WM and Kwon KC (2000) Soil carbon sequestration and land-use change: processes and potential. *Global Change Biology* 4: 67–79; West TO and Marland G (2002) A synthesis of carbon sequestration, carbon emissions, and net carbon flux in agriculture: comparing tillage practices in the United States. *Agriculture, Ecosystems and Environment* 91: 217–232.

Table 2 Global, regional and country estimates of soil C sequestration potential

Area	*Land-use sector*	*Sequestration (Tg C year^{-1})*	*Source*
Global	Improved management of cropland	260	IPCC (2000)[a]
	Improved management of grazing land	470	
	Improved management of forest land	700	
	Agroforestry	630	
USA	Cropland	83	Sperow *et al.* (2003)[b]
USA	Cropland	75–174	Lal *et al.* (1999)[c]
USA	Grazing land (excluding cropland set-aside to grassland)	7–75	Follett *et al.* (2001)[d]
UK	Cropland	6	Smith *et al.* (2000)[e]
EU	Increased manure application to arable land	13	Smith *et al.* (1997)[f]
	Sewage sludge application to arable land	8	
	Increased straw application to cereal land	14	
	Afforestation of 30% surplus arable land	50	
	Conversion of arable land to ley-arable farming	40	

[a] International Panel on Climate Change (2000) *Land Use, Land Use Change, and Forestry. Intergovernmental Panel on Climate Change Special Report.* Cambridge, UK: Cambridge University Press.
[b] Sperow M, Eve M, and Paustian K (2003) Potential soil C sequestration on US agricultural soils. *Climatic Change* 57: 319–339.
[c] Lal R, Follett RF, Kimble J, and Cole CV (1999) Managing US cropland to sequester carbon in soil. *Journal of Soil and Water Conservation* 54: 374–381.
[d] Follett RF, Kimble JM, and Lal R (eds) (2001) *The Potential of US Grazing Lands to Sequester Carbon and Mitigate the Greenhouse Effect.* Boca Raton, FL: Lewis Publishers.
[e] Smith P, Milne R, Powlson DS *et al.* (2000) Revised estimates of the carbon mitigation potential of UK agricultural land. *Soil Use and Management* 16: 293–295.
[f] Smith P, Powlson DS, Glendining MJ, and Smith JU (1997) Potential for carbon sequestration in European soils: preliminary estimates for five scenarios using results from long-term experiments. *Global Change Biology* 4: 67–79.

Restoring degraded soils and converting marginal cropland to perennial vegetation such as grassland and forest can yield relatively high rates of soil C accrual. Elimination of soil disturbance, reduced erosion and greater proportion of primary productivity going back to the soil contribute to soil C increases. Average rates of soil C gain under afforestation (0.1–0.5 Mg C ha^{-1} year^{-1}) have been reported and somewhat higher rates may occur with conversion to grassland vegetation (Table 1).

Regional and Country Estimates of C-Sequestration Potential

Estimates of the potential for soil C sequestration on cropland and grazing lands have been made, globally and for some countries and/or regions, particularly in Europe and North America (Table 2). A global survey by the Intergovernmental Panel on Climate Change has estimated that as much as 2.3 Pg year^{-1} could be sequestered through widespread adoption of improved management and land-use practices on cropland, grazing lands, and forests. Soil C sequestration has in principle been included as an allowable option within the framework of the Kyoto Protocol for countries to meet part of their emission-reduction targets. However, there are many outstanding questions regarding quantification and accounting procedures, C credit trading, and issues such as permanence and leakage that need to be addressed before soil C sequestration is included as a policy option.

To date, most estimates of potentials have been based on highly aggregated data and thus have considerable uncertainty, which has not been formally assessed. Moreover, these biophysical potentials do not consider the economic factors that limit the adoption of C sequestering practices. On the other hand, the development of new technology to specifically enhance C sequestration rates and thus increase biophysical potentials is in its infancy. Thus, the amounts of C sequestration that can be attained are still poorly known and will depend to a large degree on economic and policy incentives.

In addition to C sequestration, increasing soil organic matter levels generally enhances several biological, chemical, and physical attributes of soils. These improvements include enhanced water storage capacity, increased water infiltration, reduced runoff (and erosion), increased soil buffering capacity, and increased storage of essential plant nutrients. Thus, promoting practices for increasing soil organic matter as part of strategies for mitigating increasing greenhouse gas concentrations can be highly beneficial for improving soil quality and the sustainability of managed lands. If these additional benefits can be 'bundled' with C sequestration, incentives for implementing C sequestration policies could probably be increased.

List of Technical Nomenclature

1 billion tonnes = 1 petagram (Pg) = 10^{15} grams (g)

1 million tonnes = 1 teragram (Tg) = 10^{12} grams (g)

See also: **Carbon Cycle in Soils:** Dynamics and Management; **Climate Change Impacts**; **Greenhouse Gas Emissions**; **Sustainable Soil and Land Management**

Further Reading

Antle JM, Capalbo SM, Mooney S, Elliott ET, and Paustian K (2002) Economic analysis of agricultural soil carbon sequestration: an integrated assessment approach. *Journal of Agricultural and Resource Economics* 26: 344–367.

Conant RT, Paustian K, and Elliott ET (2001) Grassland management and conversion into grassland: effects on soil carbon. *Ecological Application* 11: 343–355.

Follett RF, Kimble JM, and Lal R (eds) (2001) *The Potential of U.S. Grazing Lands to Sequester Carbon and Mitigate the Greenhouse Effect*. Boca Raton, FL: Lewis Publishers.

Intergovernmental Panel on Climate Change (2000) *Land Use, Land Use Change, and Forestry. Intergovernmental Panel on Climate Change Special Report*. Cambridge, UK: Cambridge University Press.

Kimble JM, Lal R, and Follett RF (eds) (2002) *Agricultural Practices and Policies for Carbon Sequestration in Soil*. Boca Raton, FL: Lewis Publishers.

Lal R, Follett RF, Kimble J, and Cole CV (1999) Managing U.S. cropland to sequester carbon in soil. *Journal of Soil and Water Conservation* 54: 374–381.

Paustian K, Collins HP, and Paul EA (1997) Management controls on soil carbon. In: Paul EA, Paustian K, Elliott ET, and Cole CV (eds) *Soil Organic Matter in Temperate Agroecosystems: Long-term Experiments in North America*, pp. 15–49. Boca Raton, FL: CRC Press.

Paustian K, Andren O, Janzen H *et al.* (1997) Agricultural soil as a C sink to offset CO_2 emissions. *Soil Use and Management* 13: 230–244.

Paustian K, Cole CV, Sauerbeck D, and Sampson N (1998) CO_2 mitigation by agriculture: an overview. *Climatic Change* 40: 135–162.

Paustian K, Elliott ET, Six J, and Hunt HW (2000) Management options for reducing CO_2 emissions from agricultural soils. *Biogeochemistry* 48: 147–163.

Post WM and Kwon KC (2000) Soil carbon sequestration and land-use change: processes and potential. *Global Change Biology* 6: 317–327.

Smith P, Powlson DS, Glendining MJ, and Smith JU (1997) Potential for carbon sequestration in European soils: preliminary estimates for five scenarios using results from long-term experiments. *Global Change Biology* 4: 67–79.

Smith P, Milne R, Powlson DS *et al.* (2000) Revised estimates of the carbon mitigation potential of UK agricultural land. *Soil Use and Management* 16: 293–295.

Sperow M, Eve M, and Paustian K (2003) Potential soil C sequestration on U.S. agricultural soils. *Climatic Change* 57: 319–339.

West TO and Marland G (2002) A synthesis of carbon sequestration, carbon emissions, and net carbon flux in agriculture: comparing tillage practices in the United States. *Agriculture, Ecosystems and Environment* 91: 217–232.

CATION EXCHANGE

L M McDonald, West Virginia University, Morgantown, WV, USA
V P Evangelou†, Iowa State University, Ames, IA, USA
M A Chappell, Iowa State University, Ames, IA, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Soil chemistry is the study of how the elements and their compounds are distributed between and within the three principal phases that comprise the soil, the solid, liquid, and gaseous phases. By studying cation exchange reactions, we seek to understand and predict how positively charged ions are distributed between the solid and liquid phases. Because this distribution plays an important role in the flocculation and dispersion of soils and suspended sediments, the availability and transport of nutrient and contaminant cations, and the regulation of soil acidity, cation exchange is an essential and unifying concept in soil science.

J.T. Way is credited with the first systematic studies of cation exchange reactions in soils. Building on the observation by H.S. Thompson that $CaSO_4$ was leached out when $(NH_4)_2SO_4$ was applied to soil columns, Way established that equivalent amounts of Ca^{2+} were removed from soils when leached with NH_4^+, K^+, and Na^+. Since that time, a significant amount of work has been done to apply the concept of cation exchange to model the availability of nutrient ions in soils, particularly K^+, NH_4^+, and Ca^{2+} exchange. The relative concentration of sodium on soil surfaces directly affects the degree of colloid dispersion and therefore the formation of soil crusts and soil hydraulic conductivity. Therefore, principles of Na^{2+}–Ca^{2+} exchange have been used to reclaim and manage saline-sodic soils. Considerable progress has been made into the effects of sodium and solution composition, pH, ionic strength, and mineralogy on

†Deceased.

the dispersive properties of soil. Aluminum–calcium exchange reactions have been used to study the effects of acid rain and other anthropogenic inputs on soil acidification. Cation exchange reactions have been, and continue to be, an active area of soil chemistry research, as evidenced by the numerous research articles that have been published on the subject. Several excellent reviews are available, including those with details on the history, experimental methods, and kinetic aspects of cation exchange.

Cation exchange occurs in soils because of two general phenomena that are simply described and easily understood. First, except for the very acid and/or highly weathered, most soils have a net negative charge. Second, all natural macroscopic systems are electrically neutral. As a result, the net negative charge of a soil particle must be balanced by a charge-equivalent number of cations at or near the particle surface (interfacial region). When salts are added to the soil by natural mineral weathering or organic matter decomposition processes, in irrigation water, as fertilizer, acid rain, or other anthropogenic input, some fraction of the added ions will accumulate in the interfacial region and displace a charge-equivalent number of ions from the interfacial region into the soil solution. The simplicity of these concepts belies the complexity of the cation exchange process when applied to a system as varied and heterogeneous as soil.

Cation Exchange Capacity

The negative charge in soils comes from permanently charged clays or variably charged minerals and soil organic matter. Permanent charge is a consequence of isomorphic substitution (e.g., Al^{3+} for Si^{4+} or Mg^{2+} for Al^{3+}) within the clay lattice and results in negative charge that is independent of soil-solution properties. Variable negative charge results from the deprotonation of surface functional groups on soil organic matter and the surfaces of oxides and hydroxides. The magnitude of the variable charge depends on soil-solution pH, ionic strength, temperature, and the presence of other specifically adsorbing (potential determining) ions. Cation exchange capacity (CEC) can be defined as either the total quantity of this negative charge per unit mass of soil or as the sum total of the exchangeable cations neutralizing this charge per unit mass of soil, depending on whether anion exclusion is to be considered part of the definition. It is expressed in units of centimoles of cation charge per kilogram ($cmol_c\,kg^{-1}$) or milliequivalents of charge per 100 g soil ($mEq\,100\,g^{-1}$). Because of the effects of ionic strength and pH on the surface charge of variable-charge minerals, and complications from soluble and sparingly soluble salts and carbonates, CEC is explicitly defined by the procedure used in the determination. Therefore, a meaningful determination of CEC should consider the uses for which CEC is being determined and the unique properties of the soils being measured, including potential complications from salts, carbonates, pH, and ionic strength.

The basic procedure for determining CEC is to saturate the soil with an 'index' cation (e.g., NH_4^+, Ba^{2+}) and then either (1) determine the concentration of Na^+, K^+, Mg^{2+}, Ca^{2+}, and Al^{3+} in the supernatant, or (2) determine the concentration of the index cation in the supernatant after it has been displaced by some other cation (e.g., K^+, Mg^{2+}). Supernatant concentrations are converted to units of charge per unit mass of soil by considering the valence of the index cation and the solid-to-solution ratio used in the procedure.

This procedure can be performed with unbuffered salts (e.g., $0.2\,mol\,l^{-1}$ ammonium chloride), pH-adjusted salts (e.g., calcium chloride at pH 8.2 or ammonium acetate at pH 7), or using the compulsive exchange technique. Unbuffered salts are preferred when the CEC at the native soil pH is desired (effective cation exchange capacity, ECEC). Buffered or otherwise pH-adjusted salt solutions are used for calcareous and gypsiferous soils, or soils otherwise containing soluble or sparingly soluble salts and carbonates. It is also used when a single procedure is needed to characterize a wide variety of soils so that lime and fertilizer practices will not bias the results for soils with variable-charge mineralogy. The compulsive exchange technique is the most time-consuming of the three but has the advantage that it accounts for ionic strength effects, an important consideration for highly weathered soils. Detailed procedures for these methods can be found in any reference work on methods of soil analysis.

Because CEC is in an important property of soils for many disciplines, it has been determined for a wide variety of soils and soil minerals. The CEC of clay minerals is proportional to the amount of accessible layer charge: low CEC on talc but high CEC on vermiculite and montmorillonite (Table 1). The muscovites have comparatively low CEC, because the high layer charge is satisfied by nonexchangeable K (Table 1). Oxides and hydroxides have generally low CEC, especially at the pH of most agricultural soils. Because of the importance of clays as exchange sites, CEC increases as the fine soil fraction increases (Table 1). Histisols have the highest CEC of any soil order because of the large contribution from soil organic matter (Table 1). For the other soil orders, CEC decreases as the amount of weathering increases (Table 1) due to replacement of 2:1 permanent charge

Table 1 Cation exchange capacity of common soil minerals, textural classes, and soil orders

Clay mineral[a]		*Textural class*[b,c]		*Soil order*[b,c]	
Mineral	*CEC ($cmol_c\,kg^{-1}$)*	*Texture*	*CEC ($cmol_c\,kg^{-1}$)*	*Order*	*CEC ($cmol_c\,kg^{-1}$)*
Talc	<1	Loamy sand	3.4	Ultisol	3.5
Kaolinite	2–15	Sandy loam	5.6	Alfisol	9.0
Biotite	10–40	Fine sandy loam	7.6	Spodosol	9.3
Chlorite	10–40	Silt loam	13.4	Entisol	11.6
Halloysite	10–40	Loam	14.2	Inceptisol	14.6
Muscovite	10–40	Clay loam	20.3	Aridisol	15.2
Dioctahedral vermiculite	10–50	Silty clay loam	27.5	Mollisol	18.7
Montmorillonite	80–150	Silty clay	32.5	Vertisol	35.6
Trioctahedral vermiculite	100–200	Clay	36.6	Histisol	128.0
Allophane	5–350				

[a]Reproduced from: Sparks DL (1995) *Environmental Soil Chemistry*, p. 46. San Diego, CA: Academic Press.
[b]Holmgren GGS, Meyer MW, Chaney RL, and Daniels RB (1993) Cadmium, lead, zinc, copper and nickel in agricultural soils of the United States of America. *Journal of Environmental Quality* 22: 335–348.
[c]Geometric mean for US soils.
CEC, cation exchange capacity.

minerals with variably charged 1:1 clays, iron, and aluminum oxides.

Qualitative Description of Cation Exchange

The first issue to be decided when describing cation exchange reactions is what is to be considered an exchange reaction. Historically, reactions that result in the formation of new solid phases or alterations of surface properties are not considered cation exchange reactions, and 'exchange' is taken to mean reversible adsorption as an outer-sphere or diffuse layer complex; for example:

$$\mathrm{ExMg} + \mathrm{Ca^{2+}(aq)} \leftrightarrow \mathrm{ExCa} + \mathrm{Mg^{2+}(aq)} \qquad [1]$$

where Ex indicates a solid-phase exchanging surface with a charge −2. However, because all ions have some potential to form inner-sphere complexes with surfaces, this distinction is as much conceptual as physical.

Cation exchange reactions are classified by (1) the number of ions considered, e.g., binary, ternary, or quaternary, and (2) the valence of the ions, either homovalent (Na^+–K^+, Mg^{2+}–Ca^{2+}, etc.) or heterovalent (Na^+–Ca^{2+}, K^+–Al^{3+}, Ca^{2+}–Al^{3+}, etc.). For a generalized, binary, homovalent exchange reaction such as eqn 1, an equilibrium expression can be written:

$$K_{\mathrm{Ca-Mg}} = \frac{(\mathrm{ExCa})}{(\mathrm{ExMg})} \cdot \frac{(\mathrm{Mg^{2+}})}{(\mathrm{Ca^{2+}})} \qquad [2a]$$

$$K_{\mathrm{Ca-Mg}} = \frac{\gamma_{\mathrm{ExCa}} \cdot [\mathrm{ExCa}]}{\gamma_{\mathrm{ExMg}} \cdot [\mathrm{ExMg}]} \cdot \frac{\gamma_{\mathrm{Mg}}[\mathrm{Mg^{2+}}]}{\gamma_{\mathrm{Ca}} \cdot [\mathrm{Ca^{2+}}]} \qquad [2b]$$

where the subscript on K indicates that Ca^{2+} in solution is replacing Mg^{2+} on a surface, the parentheses indicate activities, the brackets indicate concentration (solution, moles per liter; exchanger, moles per kilogram or centimoles of cation charge per kilogram) and γ represents activity coefficients. A generalized, binary heterovalent reaction can be written:

$$(1/2)\mathrm{Ex_2Ca} + \mathrm{Na^+} \leftrightarrow \mathrm{ExNa} + (1/2)\mathrm{Ca^{2+}} \qquad [3a]$$

$$K_{\mathrm{Na-Ca}} = \frac{\gamma_{\mathrm{ExNa}} \cdot [\mathrm{ExNa}]}{(\gamma_{\mathrm{Ex_2Ca}} \cdot [\mathrm{Ex_2Ca}])^{0.5}} \cdot \frac{(\gamma_{\mathrm{Ca}} \cdot [\mathrm{Ca^{2+}}])^{0.5}}{\gamma_{\mathrm{Na}} \cdot [\mathrm{Na^+}]} \qquad [3b]$$

where Ex is now defined as an exchanger with charge −1. Because of the difficulty of defining surface activity coefficients (γ_{ExCa}, γ_{ExMg} in eqn [2b] and γ_{ExNa}, $\gamma_{\mathrm{Ex_2Ca}}$ in eqn [3b]), the left-hand side of eqn [2b] and [3b] is called a selectivity coefficient rather than an equilibrium constant.

Selectivity is the tendency for a charged surface preferentially to adsorb one ion over another. If the exchanging surface in eqn 1 'preferred' Ca^{2+}, then the equilibrium would lie to the right and if the surface preferred Mg^{2+} then the equilibrium would lie to the left. Soil minerals (and soil organic matter) are not indifferent to the suite of ions in the soil solution and exhibit a preference for certain of these ions. This preference is a consequence of the properties of those ions, the bulk and surface properties of the soil particles, and the extent to which both of these are modified by the soil solution.

Properties of Ions

From Coulomb's Law, the interaction energy between a charged surface and an oppositely charged solute is proportional to the solute's charge and inversely

Table 2 Ionic and hydrated radii of some Group I and II cations, and ammonium

Cation	*Ionic radii (nm)*	*Hydrated radii (nm)*
Li^+	0.068	1.003
Na^+	0.098	0.790
K^+	0.133	0.532
NH_4^+	0.143	0.537
Rb^+	0.149	0.509
Cs^+	0.165	0.505
Mg^{2+}	0.089	1.080
Ca^{2+}	0.117	0.96
Sr^{2+}	0.134	0.96
Ba^{2+}	0.149	0.88

proportional to the solvated radius. Therefore, in the absence of any surface-specific effects, cation (group I and II) preference follows the lyotropic series:

$$Ba^{2+} > Sr^{2+} > Ca^{2+} > Mg^{2+} > Cs^+ > Rb^+ > K^+ > Na^+ > Li^+ \quad [4]$$

Because of periodic trends in polarizability and hydrated radii (Table 2), negatively charged surfaces prefer divalent cations over monovalent cations, and (within a periodic group) the more weakly hydrated Ba^{2+} over Mg^{2+}, and Cs^+ over Li^+. The lyotropic series assumes that exchangeable cations retain their hydration spheres (i.e., adsorbed as outer-sphere complexes) and that free metal ion activities are being compared. That is, preference for Ca^{2+} over Na^+ (on a concentration basis) will decrease when sulfate is the dominant anion (compared with the chloride system) because Ca has more potential to form ion pairs with sulfate than does Na. Typically this difference will disappear when ion activities and ion pairing are accounted for. If it does not, it may suggest either competition from charged ion pairs (e.g., $CaCl^+$) or surface modification due to inner-sphere complexation. Competition effects from charged ion pairs can be significant, especially when one of the exchanging cations is Pb^{2+} or Cd^{2+} in chloride-containing solutions.

Solution-phase single-ion activity coefficients can be calculated using either the Debye–Huckel equation or the Davies equation. The advantage of the Davies equation is that it has no ion-specific parameters and so can be readily incorporated into equilibrium-speciation computer models. This is, however, also a disadvantage of the Davies equation in that it would calculate the same activity coefficient for two different cations of the same valence (e.g., Ca^{2+} and Cu^{2+}). Both the Debye–Huckel and Davies equations are widely used and accepted by the soil chemistry research community.

Properties of Surfaces

Surface-specific effects are always present in soils and so the lyotropic series is not a universal relationship. Surfaces that act as strong bases will prefer the stronger Lewis acid when cations have the same valence. Surfaces that exhibit weak acid behavior (higher pK_α) show stronger preference for heavy metals than hard (Group I and II) metals compared with surfaces with stronger acid behavior (low pK_α). For example, illite or kaolinite shows stronger preference for Cu^{2+} or Cd^{2+} than montmorillonite. The anion with the highest potential to form surface complexes controls the sorption potential of heavy metals. For example, Ni^{2+} in the presence of $Ca(NO_3)_2$ exhibits greater adsorption potential than Ni^{2+} in the presence of $CaSO_4$, because sulfate will produce surface sites with high specificity for Ca^{2+}. A surface that has the potential to form inner-sphere complexes with certain monovalent cations shows stronger preference for these cations than any other cation. For example, vermiculite exhibits more preference for K^+ or NH_4^+ than for Na^+ or Ca^{2+}. For surfaces that do not have the potential to form inner-sphere complexes, preference depends on the magnitude of the surface electrical potential. For example, a surface with high electrical potential shows higher preference for divalent cations in the presence of a monovalent cation than a surface with low electrical potential. Surfaces that have the potential to undergo conformational changes, e.g., humic acids, prefer higher-valence cations (e.g., Ca^{2+}) over lower-valence cations (e.g., K^+).

Properties of Solvents

Little work has been done on cation exchange reactions in solutions other than water. Whether in nonaqueous solvents (benzene, trichloroethylene (TCE)) or in mixtures of water and miscible solvents (low-weight alcohols, acetonitrile, etc.), low-dielectric solvents have the potential to affect the properties of ions and exchanging surfaces. Low-dielectric solutions are likely to promote ion-pairing, especially for polyvalent ions, due to coulombic attractions. The effects of nonaqueous solvents on the surface properties of minerals have been variable, with reports of decreased (rutile) and increased (goethite) surface-charge density. The most pronounced effect of low-dielectric solvents is on the interlayer spacing of 2:1 minerals. At some critical solvent concentration, dielectric saturation occurs and clay interlayers collapse. The solvent concentration at which this collapse occurs is proportional to the expected complexing ability of the solvent with the interlayer cation.

Quantitative Description of Cation Exchange

To determine cation exchange selectivity coefficients, the terms on the right-hand side of eqns [2b] and [3b] are determined over a range of equilibrium solution-phase cation activities. The solution-phase and solid-phase exchangeable cation concentrations are readily determined experimentally, but the activity coefficients must be calculated and so require one or more assumptions. Solution-phase activity coefficients can be calculated using the Davies or Debye–Huckel equation, but there are no analogous universal equations to calculate solid-phase exchangeable cation activity coefficients, and so some convention must be accepted. The simplest of these is the Kerr convention (Table 3), which uses exchanger concentration units of moles per kilogram and assumes that the two solid-phase exchangeable cation activity coefficients cancel from eqns [2b] (γ_{ExCa}, γ_{ExMg}) and [3b] (γ_{ExNa}, γ_{Ex_2Ca}). While this assumption is approximately true for the homovalent Ca–Mg system Kerr studied (eqn [2]), it would not be true generally, especially for heterovalent exchange (eqn [3]). In the Vanselow convention (Table 3), exchanger activities are assumed equal to their mole fractions (the ideal solid solution theory), so that for eqn 3:

$$(ExNa) = X_{Na} = \frac{[ExNa]}{[ExNa] + [Ex_2Ca]} \qquad [5a]$$

$$(Ex_2Ca) = X_{Ca} = \frac{[Ex_2Ca]}{[ExNa] + [Ex_2Ca]} \qquad [5b]$$

Note that because the solid-phase exchangeable cation concentrations are expressed in moles per kilogram the denominator in eqn [5] is not a constant even though the CEC is constant (sum of Na and Ca on a charge-equivalent basis). For the Gapon convention (Table 3), it is assumed that solid-phase exchangeable cation activities are equal to concentrations when concentrations are expressed on a charge-equivalent basis. To do this, eqn [3a] needs to be rewritten as:

$$ExCa_{0.5} + Na^+ \leftrightarrow ExNa + (1/2)Ca^{2+} \qquad [6a]$$

where Ex still indicates an exchanger with an average charge of −1, and

$$K_{Na-Ca} = \frac{[ExNa]}{[ExCa_{0.5}]} \cdot \frac{[Ca^{2+}]^{0.5} \cdot \gamma_{Ca}^{0.5}}{[Na^+] \cdot \gamma_{Na}} \qquad [6b]$$

where only the first term on the right-hand side of eqn [3b] has been changed. Note that the distinction between the conventions is relevant only for heterovalent exchange. For homovalent exchange, all the conventions in Table 3 are equivalent to the Kerr convention. Although the Kerr, Vanselow, and Gapon conventions are the most widely used, there are others, including Gaines–Thomas, Krishnamoorthy–Overstreet, and Rothmund–Kornfeld. Details and references for these models can be found in most soil chemistry texts.

Homovalent Binary Exchange

The homovalent cation exchange isotherm is obtained by solving eqn [2b] for exchangeable Ca. Given that

$$((1/2)CEC) = ExMg + ExCa \qquad [7a]$$

when CEC has the units centimoles of cation charge per kilogram and ExMg and ExCa have units centimoles per kilogram, we can write:

$$CEC = ExMg_{0.5} + ExCa_{0.5} \qquad [7b]$$

to indicate that all terms have units centimoles of cation charge per kilogram. Note that the exchanger is still defined as having a charge −2. It is assumed that other cations, e.g., exchangeable K^+ and H^+, are present in negligible quantities and do not interfere

Table 3 Selectivity symbols and expressions for the three most commonly used cation exchange conventions

Selectivity coefficient	*Symbol*	*Homovalent exchange*[a]	*Heterovalent exchange*[a]
Kerr[b]	K_K	$\frac{[ExCa][Mg^{2+}]}{[ExMg][Ca^{2+}]}$	$\frac{[ExNa][Ca^{2+}]^{0.5}}{[Ex_2Ca]^{0.5}[Na^+]}$
Vanselow[b,c]	K_V	$\frac{(ExCa)(Mg^{2+})}{(ExMg)(Ca^{2+})}$	$\frac{(ExNa)(Ca^{2+})^{0.5}}{(Ex_2Ca)^{0.5}(Na^+)}$
Gapon[d]	K_G	$\frac{[ExCa][Mg^{2+}]}{[ExMg][Ca^{2+}]}$	$\frac{[ExNa][Ca^{2+}]^{0.5}}{[ExCa_{0.5}][Na^+]}$

[a]Square brackets indicate concentration, parentheses indicate activities for both the solution phase and the exchanger.
[b]For the exchange reaction as written in eqns [1] or [3a] solution-phase units, moles per liter; exchanger units, moles per kilogram.
[c]Solution-phase activity coefficients defined using either the Debye–Huckel or Davies equation. Exchanger activities defined using eqns [5a] and [5b].
[d]For the exchange reaction as written in eqn [6a]. Solution-phase units, moles per liter; exchanger units, centimoles of cation charge per kilogram.

with the exchange reaction being studied, or H^+ is tightly bound to the charged surface, giving rise only to pH-dependent charge. Using eqn [7], and the definition for the calcium activity ratio (AR_{Ca}) or the calcium concentration ratio (CR_{Ca}, calcium adsorption ratio):

$$AR_{Ca} = \frac{(Ca^{2+})}{(Mg^{2+})} \quad [8a]$$

$$CR_{Ca} = \frac{[Ca^{2+}]}{[Mg^{2+}]} \quad [8b]$$

where the square brackets indicate solution-phase units moles per liter or millimoles per liter, eqn [2b] becomes

$$ExCa_{0.5} = \frac{K_{Ca-Mg} \cdot CEC \cdot AR_{Ca}}{1 + AR_{Ca} \cdot K_{Ca-Mg}} \quad [9]$$

Equation [7b] essentially invokes the Gapon convention so that $\gamma_{ExCa} = \gamma_{ExMg} = 1$ in eqn [2b]. A plot of $ExCa_{0.5}$ versus AR_{Ca} will produce a curvilinear line, asymptotically approaching CEC. The path of the line from $ExCa_{0.5} = 0$ to $ExCa_{0.5} = CEC$ depends on K_{Ca-Mg} and CEC (Figure 1).

An important component of homovalent exchange is the magnitude of the exchange selectivity coefficient. With some exceptions, homovalent cation exchange reactions in soils or soil minerals exhibit a selectivity coefficient of approximately 1. This indicates that the surface does not show any particular adsorption preference for either of the two homovalent cations.

Cation preference can be demonstrated using fractional isotherms. Fractional isotherms are plots of equivalent fraction on the exchange phase versus equivalent fraction in the solution phase. To determine the nonpreference isotherm, it follows from eqn [2b] that if there is no preference for Ca over Mg, then $K_{Ca-Mg} = 1$, so that eqn [9] can be rewritten:

$$(ExCa_{0.5}/CEC) = \frac{CR_{Ca}}{1 + CR_{Ca}} \quad [10a]$$

This is the diagonal line in Figure 2 when CR_{Ca} is converted to a mole or equivalent fraction. An experimentally determined isotherm above the diagonal line reveals that the surface prefers Ca^{2+} ($K_{Ca-Mg} > 1$), while any line below the nonpreference line reveals that the surface prefers Mg^{2+} ($K_{Ca-Mg} < 1$). Stated another way, for a surface that prefers Ca^{2+}, a relatively small increase in the concentration of Ca^{2+} in solution results in a proportionally large concentration of Ca^{2+} on the surface. Introducing solution single-ion activities, γ_i, into eqn [10a]:

$$ExCa_{0.5}/CEC = \frac{([Ca^{2+}]/[Mg^{2+}]) \cdot \left(\gamma_{Ca}/\gamma_{Mg}\right)}{1 + ([Ca^{2+}]/[Mg^{2+}]) \cdot \left(\gamma_{Ca}/\gamma_{Mg}\right)} \quad [10b]$$

it can be shown that the homovalent nonpreference isotherm is independent of ionic strength because γ_{Ca} is nearly equal to γ_{Mg}.

Heterovalent Binary Exchange

The equation that is most commonly used to describe heterovalent cation exchange is the Gapon equation. For Na^+–Ca^{2+} exchange, eqn [8b] can be rearranged to solve for ExNa:

$$ExNa = \frac{K_G \cdot CEC \cdot SAR}{1 + K_G \cdot SAR} \quad [11a]$$

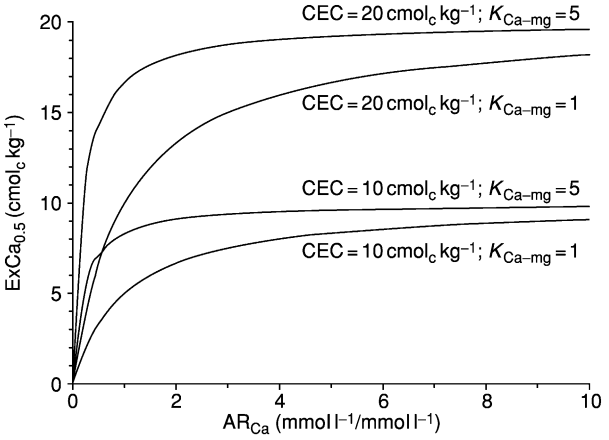


Figure 1 Effect of selectivity coefficient (K_{Ca-Mg}) and cation exchange capacity (CEC) on homovalent cation exchange isotherm shape. Plotted using eqn [9].

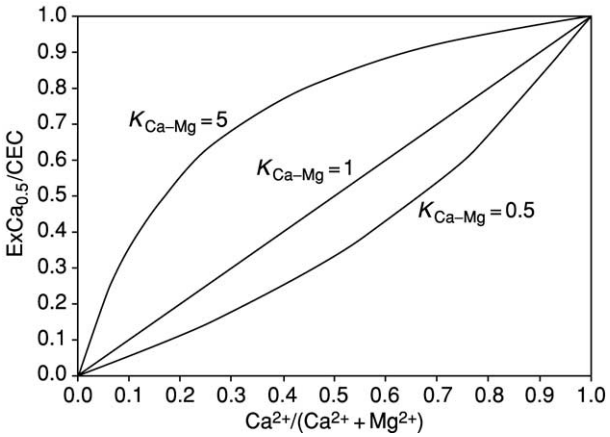


Figure 2 Calcium–magnesium fractional isotherms for three values of the homovalent cation exchange selectivity coefficient (K_{Ca-Mg}). Plotted using eqn [10a].

where K_G is the Gapon selectivity coefficient, SAR is the sodium adsorption ratio:

$$\text{SAR} = \frac{[\text{Na}^+]}{[\text{Ca}^{2+}]^{0.5}} \qquad [11b]$$

The square brackets denote concentration in moles per liter or millimoles per liter in the solution phase, and

$$\text{CEC} = \text{ExCa}_{0.5} + \text{ExNa} \qquad [11c]$$

with units centimoles of cation charge per kilogram. A plot of ExNa versus SAR will produce a curvilinear line, asymptotically approaching CEC. Because of the exponent in the denominator of eqn [11b], the K_G of a heterovalent exchange depends on solution concentration units, whereas the K_G of a homovalent exchange is independent of solution units. A heterovalent K_G obtained with solution units in millimoles per liter when multiplied by $(1000)^{0.5}$ gives the K_G in units of $(\text{moles per liter})^{-0.5}$.

Rearranging eqn [6b] gives:

$$(\text{ExNa}/\text{ExCa}_{0.5}) = K_G \cdot \left([\text{Na}^+]/[\text{Ca}^{2+}]^{0.5}\right) \qquad [11d]$$

Theoretically, a plot of $\text{ExNa}/\text{ExCa}_{0.5}$ or exchangeable sodium ratio (ESR) versus SAR will produce a straight line with a slope equal to K_G. The mean magnitude of K_G for arid US soils is approximately $0.015\,(\text{mmol}\,\text{l}^{-1})^{-0.5}$. However, experimentally determined K_G are found to depend on pH, salt concentration, clay mineralogy, and sodium load. As sodium load increases, K_G increases, and, as pH increases in variable-charge soils, K_G decreases. The K_G of Na^+–Ca^{2+} exchange may vary from 0.50 to $2(\text{mol}\,\text{l}^{-1})^{-0.5}$, whereas the K_G of K^+–Ca^{2+} exchange may vary from 2 to $30\,(\text{mol}\,\text{l}^{-1})^{-0.5}$, depending on mineralogy and pH. Illite has significantly greater affinity for K^+ than does smectite or organic matter.

Cation preference in heterovalent exchange is also demonstrated through fractional isotherms. Heterovalent fractional cation preference isotherms differ from homovalent cation preference isotherms in that the nonpreference line in heterovalent exchange is not the diagonal line and depends on ionic strength ('square-root effect'). The nonpreference line for a heterovalent exchange reaction is shown as one of the solid lines in **Figure 3**. Heterovalent exchange data occurring above the nonpreference line reveal that the surface prefers the monovalent cation, while exchange data occurring below the nonpreference line reveal that the surface prefers the divalent cation (**Figure 3**).

It is suggested that the Vanselow equation is consistent with thermodynamics of chemical reactions

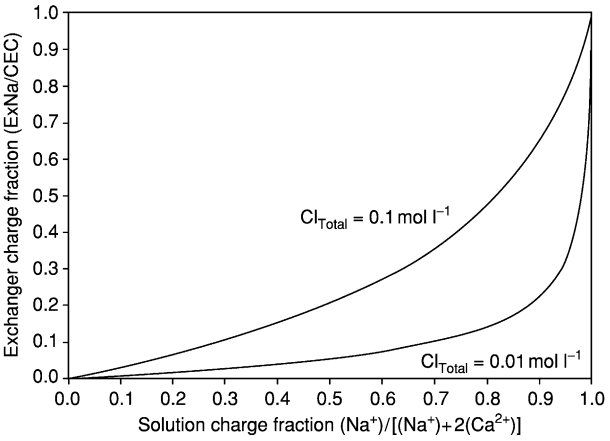


Figure 3 Nonpreference isotherms for heterovalent sodium–calcium exchange at two total chloride concentrations (Cl_{Total}). Plotted using eqn [14].

because it uses units of moles for solution phase and the exchanger. Solving the Vanselow equation for ExNa gives:

$$\text{ExNa} = \frac{\text{CEC}\cdot K_V \cdot \text{AR}_{\text{Na}}}{\left[4 + (K_V \text{AR}_{\text{Na}})^2\right]^{0.5}} \qquad [12a]$$

where $(\text{Na}^+)/(\text{Ca}^{2+})^{0.5}$ is the sodium activity ratio, AR_{Na}. Considering that for the nonpreference isotherm, as $I \to 0$, $K_v = 1$ and $\Delta G^0 = 0$, then:

$$(\text{ExNa}/\text{CEC}) = \frac{\text{AR}_{\text{Na}}}{\left[4 + (\text{AR}_{\text{Na}}{}^2)\right]^{0.5}} \qquad [12b]$$

Using a number of AR_{Na} values, as $I \to 0$, to cover the entire exchange isotherm, equivalent fractional loads for Na^+ can be estimated using eqn [12b]. A plot of ExNa/CEC versus $(\text{Na}^+)/[(\text{Na}^+) + (2\text{Ca}^{2+})]$ will produce a curvilinear line representing the nonpreference isotherm. Upon introducing activity coefficients in the $\text{Na}^+/[\text{Na}^+ + 2\text{Ca}^{2+}]$ term such that:

$$(\text{ExNa}/\text{CEC}) = \frac{(\text{Na}^+)/\gamma_{\text{Na}}}{(\text{Na}^+)/\gamma_{\text{Na}} + 2\cdot(\text{Ca}^{2+})/\gamma_{\text{Ca}}} \qquad [13]$$

(again assuming $K_V = 1$), it can be shown that as I increases, γ_{Na} and γ_{Ca} decrease disproportionately, with γ_{Ca} decreasing significantly more than γ_{Na}. Therefore, as I increases, eqn [13] decreases, indicating that the nonpreference isotherm is ionic-strength dependent. Alternatively, an equation for the heterovalent nonpreference isotherm has been derived:

$$\text{ExNa}/\text{CEC} = \left[1 + \left(\frac{2\gamma_{\text{Ca}}}{(\text{TN})\gamma_{\text{Na}}^2}\left(\frac{1}{(E'_{\text{Na}})^2} - \frac{1}{E'_{\text{Na}}}\right)\right)\right]^{-0.5} \qquad [14]$$

where E'_{Na} is the charge fraction of Na^+ in solution and TN is the total cation normality. Eqn [14] is plotted in **Figure 3** for two total cation normalities ($Cl_{Total} = Na^+ + 2Ca^{2+}$). Increasing the salt concentration in solution decreases the preference the surface shows for the divalent cation, Ca^{2+}. At sufficiently high salt concentrations, the surface may actually show slight preference for the monovalent ion. The influence of γ_i, induced by increasing I, on apparent ion preference, however, is very small in comparison with the square-root effect on apparent ion preference also induced by increasing I.

In general, the selectivity coefficient (K_v) for a binary exchange reaction depends primarily on the ionic strength of the solution, on the proportion of cations in the soil-absorbing complex, and the proportion of the cations in the soil solution phase. Exchange reactions of Na^+ with trace metal cations (Cd^{2+}, Co^{2+}, Cu^{2+}, Ni^{2+}, and Zn^{2+}) on Camp Berteau montmorillonite show that K_v is constant and independent of exchanger composition, suggesting ideal solid-solution behavior, up to an equivalent fraction of trace metal cations of 0.70. Sodium–calcium exchange studies on similar clay minerals show that there is a more pronounced selectivity of clay for Ca^{2+} at the calcium-rich end of the isotherm. That experimentally determined selectivity coefficients are not constant across the entire exchange isotherm indicates that the exchanger activity coefficients (γ_{ExNa}, γ_{Ex_2Ca} in eqn [3b]) change as a function of surface composition. Mass balance constraints mean that a change in any one exchanger activity coefficient must be compensated by an equal change in the other exchanger activity coefficient (the Gibbs–Duhem equation). These ideas lead to the derivation of the thermodynamic exchange equilibrium constant, K_{eq}, from the Vanselow convention as:

$$\ln K_{eq} = \int_0^1 \ln K_V \, dE_{Na} \qquad [15]$$

where E_{Na} is the equivalent fraction of Na on the exchanger (ExNa/CEC). Homovalent and heterovalent thermodynamic exchange constants (K_{eq}) for a number of exchangers and exchange reactions are given in **Table 4**.

The Future of Cation Exchange

The preceding discussion, and nearly all of the research into cation exchange reactions, has been conducted on binary exchange reactions. However, soils are multi-ion systems. In order for binary exchange reactions accurately to describe multi-ion exchange, the binary selectivity coefficients have to be independent of the exchanger composition, at least over the range of relevant solution and exchanger concentrations. In some studies, binary reactions sufficiently describe higher-order cation exchange; in others, they do not. Considerably more work needs to be

Table 4 Thermodynamic equilibrium exchange constants (K_{eq}) for binary homovalent and heterovalent exchange (all experiments conducted at 298 K except where noted)

	Homovalent		*Heterovalent*	
Exchanger	*Process*	K_{eq}	*Process*	K_{eq}
Calcareous soil	Ca–Mg	0.89–0.75	Ca–Na	0.38–0.09
Camp Berteau montmorillonite	Ca–Mg	0.95	Ca–Na	0.72
			Ca–NH_4	0.035
Wyoming bentonite	Ca–Cu	0.96		
	Na–Li	1.08		
	K–Na	1.67		
World vermiculite	Na–Li	11.42	Ca–Na	0.98
			Mg–Na	1.73
Chambers montmorillonite	Na–Li	1.15	Ca–K	0.045
	K–Na	3.41		
Kaolinitic soil clay (303 K)	Mg–Ca	0.65	K–Ca	16.16
Soil	K–Na	4.48–6.24	Ca–Na	0.42–0.043
	Mg–Ca	0.61	Mg–Na	0.75–0.053
			K–Ca	5.89–323.3
			K–Ca	12.09
			K–Ca	19.92–323.3
			K–Ca	0.46
			K–Ca	0.64–6.65
			K–Ca	6.42–6.76
			K–Mg	5.14

Adapted from Sparks DL (1995) *Environmental Soil Chemistry*. San Diego, CA: Academic Press.

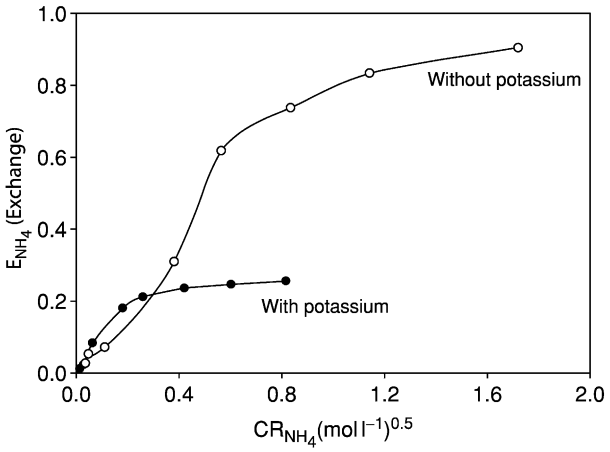


Figure 4 Effect of potassium on $NH_4^+ - Ca^{2+}$ exchange isotherm for vermiculite. Reproduced with permission from Evangelou VP and Lumbanraja J (2002) Ammonium–potassium–calcium exchange in the absence and presence of potassium on vermiculite and hydroxy-Al vermiculite. *Soil Science Society of America Journal* 66: 445–455.

done to identify the conditions where ternary exchange can be predicted from binary exchange data. One situation where binary data are not likely to be accurate is when surfaces undergo conformational changes. In circumneutral agricultural soils, Ca^{2+} and K^+ are present in large quantities, and NH_4^+ is added as fertilizer. Figure 4 shows an NH_4–Ca exchange isotherm for vermiculite, with and without added K^+. At low NH_4 concentrations (the agriculturally important region), adding K^+ increases the selectivity coefficient toward preference for NH_4^+. That is, in addition to the well-known effect of K^+ on ammonium fixation (nonexchangeable NH_4^+, K^+ appears to influence the availability of exchangeable NH_4^+.

As discussed in the Introduction, cation exchange is assumed to be a reversible, adsorption reaction as outer-sphere or diffuse-layer complexes. However, all cations possess some ability to form inner-sphere surface complexes and so a complete understanding of cation distribution in soils will require consideration of these reactions. Cation exchange equilibria can be incorporated into existing computer-based surface complexation models. The advantage to this approach is that ternary and quaternary exchange reactions can be included. Although the data and calibration requirements are substantial, this approach holds considerable promise for providing a comprehensive description of cation distribution in soils.

List of Technical Nomenclature

γ	Activity coefficient
ϵ	Dielectric constant
I	Ionic strength
K (Kelvin)	Temperature
z	Ion valence

See also: **Calcium and Magnesium in Soils**; **Chemical Equilibria**; **Crusts:** Structural; **Flocculation and Dispersion**; **Salt-Affected Soils, Reclamation**; **Sodic Soils**

Further Reading

Evangelou VP and Lumbanraja J (2002) Ammonium–potassium–calcium exchange on vermiculite and hydroxy-Al vermiculite. *Soil Science Society of America Journal* 66: 445–455.

Evangelou VP, Wang J, and Phillips RE (1994) New developments and perspectives in characterization of soil potassium by quantity–intensity (Q/I) relationships. In: Sparks DL (ed.) *Advances in Agronomy*. vol. 52, pp. 173–227. Orlando, FL: Academic Press.

Holmgren GGS, Meyer MW, Chaney RL, and Daniels RB (1993) Cadmium, lead, zinc, copper and nickel in agricultural soils of the United States of America. *Journal of Environmental Quality* 22: 335–348.

Rhee H and Dzombak (1998) Surface complexation/Gouy–Chapman modeling of binary and ternary cation exchange. *Langmuir* 14: 935–943.

Sparks DL (1989) *Kinetics of Soil Chemical Processes*. San Diego, CA: Academic Press.

Sparks DL (1995) *Environmental Soil Chemistry*. San Diego, CA: Academic Press.

Sposito G (1981a) Cation exchange in soils: a historical and theoretical perspective. In: Dowdy RH (ed.) *Chemistry in the Soil Environment*, pp. 13–30. Madison, WI: Soil Science Society of America.

Sposito G (1981b) *The Thermodynamics of Soil Solutions*. New York: Oxford University Press.

Suarez DL (1999) Thermodynamics of the soil solution. In: Sparks DL (ed.) *Soil Physical Chemistry*, pp. 97–134. Boca Raton, FL: CRC Press.

Sumner ME and Miller WP (1996) Cation exchange capacity and exchange coefficients. In: Sparks DL, Page AL, Helmke PA *et al.* (eds) *Methods of Soil Analysis, part 2. Chemical Methods*, pp. 1201–1229. Soil Science Society of America Book Series No. 5. Madison, WI: Soil Science Society of America and American Society of Agronomy.

Thomas GW (1977) Historical developments in soil chemistry: ion exchange. *Soil Science Society of America Journal* 41: 230–238.

CHEMICAL EQUILIBRIA

A P Schwab, Purdue University, West Lafayette, IN, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Although soils are very complex living systems, they are composed of identifiable minerals, amorphous compounds, and discrete ionic species that can be quantified. Solubilities, solid-phase transformations, dissolution of gases, and changes in oxidation states of these chemical constituents can be predicted by chemical thermodynamics. However, implicit in the application of these principles to soils is that the soil system (or, at the very least, the constituents in question) are in chemical equilibrium. If the assumption of equilibrium is valid, then we have very powerful tools at our disposal to describe the chemical behavior of the solid, liquid, and gas phases in soil.

Chemical Equilibrium

Before discussing chemical equilibrium in soils, we must clearly understand the concept. Perhaps the most famous discussions of equilibrium were presented in the 1880s by Henri LeChatlier, for whom LeChatlier's Principle was named. Two basic definitions of chemical equilibrium will be discussed: kinetic and free energy. The two definitions simply express the same phenomenon in different ways.

Kinetic Definition

Consider the generalized chemical reaction with reactants A and B, and forming products C and D with stoichiometric coefficients a, b, c, and d:

$$a\mathrm{A} + b\mathrm{B} \Leftrightarrow c\mathrm{C} + d\mathrm{D}$$

The rate of the forward reaction is given by:

$$v_\mathrm{f} = k_\mathrm{f}[\mathrm{A}]^a[B]^b$$

in which $[x]$ represents the solution concentration of component x. The rate expression for the reverse reaction is given by:

$$v_\mathrm{r} = k_\mathrm{r}[\mathrm{C}]^c[D]^d$$

When only the reactants are present, the forward reaction proceeds quickly. As products begin to accumulate, the reverse reaction begins to occur simultaneously. Eventually, the point will be reached in which the forward and reverse reactions are proceeding at the same rate and the concentrations of the reactants and products will not change with time. This is the point of a dynamic, chemical equilibrium. When the forward and reverse rates are equal:

$$k_\mathrm{f}[\mathrm{A}]^a[\mathrm{B}]^b = k_\mathrm{r}[\mathrm{C}]^c[D]^d$$

Rearranging

$$K_\mathrm{eq} = \frac{k_\mathrm{f}}{k_\mathrm{r}} = \frac{[\mathrm{C}]^c[D]^d}{[\mathrm{A}]^a[B]^b}$$

in which K_eq is the equilibrium constant.

Free-Energy Definition

The chemical potential of any species in solution (μ_i) is given as:

$$\mu_\mathrm{i} = \mu_\mathrm{i}^\mathrm{o} + RT\ln(i)$$

in which $\mu_\mathrm{i}^\mathrm{o}$ is the standard chemical potential of species i, R is the ideal gas constant, T is the temperature in Kelvin, and (i) is the activity of species i. The expression for the free-energy change of any chemical reaction is:

$$\Delta G = \sum_\mathrm{i} v_\mathrm{i}\mu_\mathrm{i}$$

in which ΔG is the change in Gibbs free energy during the course of the reaction, and v_i is the stoichiometric coefficient for the components of the equilibrium expression (v_i is positive for products, negative for reactants). Substituting:

$$\Delta G = \sum_\mathrm{i} v_\mathrm{i}\mu_\mathrm{i}^\mathrm{o} + RT\sum_\mathrm{i} v_\mathrm{i}\ln(i)$$

or:

$$\Delta G = \Delta G^\mathrm{o} + RT\ln\prod_\mathrm{i}(i)^{\mathrm{vi}}$$

where:

$$\Delta G^\mathrm{o} = \sum_\mathrm{i} v_\mathrm{i}\mu_\mathrm{i}^\mathrm{o}$$

and ΔG^o is the standard Gibbs free energy change of the reaction. The reaction quotient Q is represented as:

$$Q = \prod_\mathrm{i}(i)^{\mathrm{vi}}$$

such that:

$$\Delta G = \Delta G^\mathrm{o} + RT\ln Q$$

By definition, the free-energy change of reaction equals zero ($\Delta G = 0$) at equilibrium, and the value of Q at equilibrium becomes the equilibrium constant K_{eq}:

$$K_{eq}\ Q = \prod_{i} (i)_{eq}^{vi}$$

and:

$$\Delta G^{o} = -RT\ln K_{eq}$$

Thus, we have the free-energy definition of equilibrium ($\Delta G = 0$) and the thermodynamic derivation of the equilibrium constant.

Applicability of Chemical Equilibrium Principles to Soils

Invoking the assumption of chemical equilibrium in soils, particularly for soils in natural conditions, is difficult because soils are highly complex and forever changing. Soils continually have inputs of external energy, water, gases, and dissolved constituents. Losses of these same components are equally possible through radiation of heat, leaching, and biological activities. These changes perturb the soil system and upset any equilibrium that may be present. With this in mind, a reasonable question is, "Can the principles of chemical equilibrium be applied to soils?"

Despite the dynamic nature of soils and continuous inputs and removal from the system, chemical equilibrium can be applied to soils if care is taken, the limitations are clearly understood, and expectations are reasonable.

Limitations to Equilibrium in Soils

Most soils are open systems with frequent inputs and removal of mass and energy. Even the simple act of rain falling on a soil results in dilution of all dissolved soil constituents and disruption of all previous equilibria. Changes in temperature, plant assimilation of nutrients, evaporation of water, and diffusion of gases into or out of soil have a similar disruptive effect. These perturbations in equilibrium may be temporary or may have long-term implications, depending upon the kinetics of these reactions. For example, the half-life of complexation/dissociation reactions vary from 10^{-9} to 10^{3} s (Table 1). Thus, outside disturbances to equilibria will have only a small impact on complexation reactions. However, oxidation–reduction and solid-phase dissolution–precipitation reactions can require a very long time to achieve equilibrium. A classic example of this is the dissolution of gibbsite (γ-$Al(OH)_3$). Monitoring the activity of Al^{3+} and pH in a suspension of soil generally reveals that the concentration of Al^{3+} in the soil solution would be more than 10-fold less than equilibrium after 200 h, based on a comparison of the known equilibrium constant with measured parameters. Solubilities would be within a factor of 2 of equilibrium after 400 h and equilibrium should be attained after approximately 1000 h. Slight shifts in pH, temperature, or moisture content would result in nonequilibrium conditions for γ-$Al(OH)_3$(gibbsite) that would again require hundreds of hours to correct.

Table 1 Half-lives of selected complexation reactions in soil solutions

Reaction	*Half-life (s)*
$MnSO_4^o \Leftrightarrow Mn^{2+} + SO_4^{2-}$	10^{-9}
$Fe^{3+} + H_2O \Leftrightarrow Fe(OH)^{2+} + H^+$	10^{-7}
$Mn^{2+} + SO_4^{2-} \Leftrightarrow MnSO_4^o$	10^{-5}
$NiC_2O_4^o \Leftrightarrow Ni^{2+} + C_2O_4^{2-}$	10^{-1}
$CO_2 + OH^- \Leftrightarrow HCO_3^-$	10
$Al^{3+} + F^- \Leftrightarrow AlF^{2+}$	10^{3}

Data from Sposito G (1994) *Chemical Equilibria and Kinetics in Soils*. New York: Oxford University Press.

Some reactions are thermodynamically favorable but are associated with a large energy barrier (e.g., energy of activation). Unless that energy barrier can be overcome, equilibrium will not be realized. The conversion of N_2(g) to nitrate should be a spontaneous reaction, but the presence of approximately 78% N_2(g) in the atmosphere is clear evidence that this reaction does not proceed as predicted. Redox couples with oxygen-containing anions (e.g., $AsO_4^{3-}/AsO_3^{3-}, S^{2-}/SO_4^{2-}$) attain equilibrium slowly in the absence of a catalyst.

Measurement of critical system variables can be subject to serious limitations that restrict rigorous application of thermodynamics to natural soils. Detection limits of soluble constituents may preclude quantification, redox-sensitive electrodes have shortcomings, and distinguishing oxidation states of soluble elements can be troublesome. Redox is a crucial system parameter, but chemical limitations of redox electrodes are well documented, and not all oxidation–reduction couples establish reversible equilibria at the surface of the sensing electrode.

The final limitation is ensuring that all products can form under soil conditions. Thermodynamic data and/or equilibrium constants are available for hundreds of minerals found in soils, but many of these minerals can only be formed under conditions of high temperatures and pressures. This is particularly true of the primary minerals in soils. Thermodynamics might predict that these minerals form in soils, but this prediction is meaningless if the solids simply cannot form.

Soil Systems in Which Equilibrium Concepts May Apply

Chemical equilibrium principles can be applied with reasonable confidence to most (but not all) reactions occurring exclusively in the solution phase. As illustrated in Table 1, most reactions attain equilibrium in a matter of a few seconds or less; therefore, typical perturbations to equilibrium relax quickly. Notable exceptions are oxyanion redox couples and cleaving the $N_2(g)$ molecule to form nitrate.

Attempts to apply equilibrium to the solid phase in the soil require more consideration. The kinetics of dissolution of typical soil minerals are slow. Calcium phosphate and sulfate minerals can achieve equilibrium in hours; aluminum phosphates, carbonates, and most oxides in a matter of days or weeks; and complex aluminosilicates in years. If the soil solution has been in contact with soil minerals for the time necessary to establish equilibrium while experiencing minimal disturbances, application of equilibrium principles is possible. Complete equilibrium among all species in the soil solution and all soil minerals is not essential; however, these principles cannot be applied if even one component ion of the mineral being studied is out of equilibrium.

Consider the example of variscite, $AlPO_4{\cdot}2H_2O$. Phosphate ions may reach constant concentrations in a matter of hours in the soil solution, but Al^{3+} may require days or weeks due to the slow reaction kinetics of aluminum oxides. Therefore, the system will be considered to be in equilibrium only when both Al^{3+} and PO_4^{3-} have reached time-independent concentrations.

Reasonable Expectations from Application of Equilibrium

Universal adherence to chemical equilibrium in soils is unrealistic. However, employing thermodynamics and associated solubility and complexation constants in soil systems can be fruitful if one has reasonable expectations.

Predict endpoints of reactions Thermodynamics can be useful in determining possible endpoints versus impossible outcomes of chemical reactions in soils. For example, we can predict that metallic iron will readily convert to ferric oxides under surface soil conditions:

$$2Fe(s) + 1.5O_2(g) + 3H_2O \Leftrightarrow 2Fe(OH)_3(amor)$$

Determine ionic activities in soil solution The activities of individual species in the soil solution are often the driving forces behind many chemical and biological reactions in soils and can be calculated using chemical equilibrium constants.

Calculating final conditions for systems that attain equilibrium rapidly When changes to the soil system occur infrequently relative to the timescale required for equilibrium, then equilibrium concepts can be applied.

Experimental Approach to Equilibrium in Soils

Thermodynamics may be used to describe chemical reactions in soils regardless of whether the application is interpretation of experimental results or the prediction of changes in soil chemistry induced by environmental perturbations. In either case, the user must be acutely aware of the limitations of chemical equilibrium in soils and must have knowledge of the systems to which these principles may be applied with confidence. The steps taken are generally the same: critical system variables must be identified; key reactions of all the contributing and interacting components are established; reliable equilibrium constants are located and assigned to each reaction; and, finally, the system of equilibrium equations is set up as an mathematical array and solved. All steps in this process require careful consideration.

System Variables

The most important variables in a soil system are similar to those in any other geochemical environment. The list of parameters to be considered is driven in part by the interest and focus of the user and is strongly influenced by other variables that are intimately linked to the systems of interest. For example, if one is considering only the solubility of ferric oxides in soils with low pH, then the obvious system parameters are pH and soluble Fe. However, Fe^{3+} can form important solution complexes with Cl^-, SO_4^{2-}, and F^-, therefore, one must account for these ancillary anions before the Fe system can be fully defined.

Reactions and Associated Equilibrium Constants

For any given system, all relevant reactions and the corresponding constants must be identified. This is not an easy task and requires databases of reliable information. Much of this work was done in the development of geochemical models, as will be discussed below. For a very simple system such as amorphous Fe oxide in equilibrium in soil, at least the following solution species need to be considered: Fe^{3+}, FeF^{2+}, FeF_2^+, $FeCl^{2+}$, $FeOH^{2+}$, $Fe(OH)_2^+$, $Fe(OH)_3^o$, $Fe(OH)_4^-$, $FeHPO_4^+$, and $FeSO_4^o$. A typical equation and constant would be:

$$Fe^{3+} + H_2O \Leftrightarrow FeOH^{2+} + H^+$$

$$\log K_{eq} = -2.19$$

Equations and constants are assembled for all species, including the dissolution reaction for the controlling solid phase:

$$FeOH_3(amor) + 3H^+ \Leftrightarrow Fe^{3+} + 3H_2O$$

$$\log K_{eq} = 3.54$$

After all species have been associated with an independent equation or an actual measurement of a system parameter (e.g., total soluble Fe or pH), they are assembled in a mathematical array of simultaneous equations; the array is solved, and the system can be defined exactly. For small systems, this can be done by hand, but larger and more comprehensive systems require a full geochemical model to provide a solution.

Equilibrium Modeling of Soil Systems

Solving the array of simultaneous equations that results from defining a soil system in equilibrium provides a serious challenge. Hand solutions are highly time-consuming, and writing small computer programs for each new system is not productive. Fortunately, this problem was recognized in the mid 1970s, and several research groups set about the task of writing generalized computer models for geochemical systems.

Geochemical models In 1974, the WATEQ was published by the US Geological Survey. WATEQ was a fairly comprehensive model with a carefully reviewed database of thermodynamic constants. The program was large and somewhat slow to compile and execute. The MINEQL model, published in 1974, has an elegant mathematical structure and allows for very fast execution times and flexibility. However, the database had not been rigorously reviewed. Since then, geochemical modeling has evolved in many directions with many new models: PHREEQE; GEOCHEM; SOILCHEM; MINTEQ; and many others. Many review articles have been written comparing various models. The calculation of the equilibrium between the gas, solid, and aqueous phases differs only by speed of execution, the database of thermodynamic constants, and associated functions.

Model inputs Each model requires that the user define the systems to be modeled. The computer models are fully capable of accepting experimental data for the purpose of calculating ionic activities and comparing them with mineral solubilities. Similarly, the models may be used completely in a predictive mode in which assumptions are made concerning equilibrium with solid phases, pH, temperature, gas partial pressures, etc. The models then will report resulting activities, total elemental solubilities, residual masses of solid phases, and the distribution of mass of a given element across all phases.

Continuing with the amorphous Fe oxide example, let us assume that the system in question is in equilibrium with amorphous $Fe(OH)_3$(amor) at pH 5.5, with total soluble chloride, fluoride, and sulfate of $0.001\,mol\,l^{-1}$. Using this information as input into MINTEQA2 resulted in the data given in Table 2.

Table 2 Modeling results using MINTEQA2 and inputs of pH 5.5, equilibrium with $Fe(OH)_3$(amorphous), and total solubilities of $0.001\,mol\,l^{-1}$ for chloride, fluoride, and sulfate

Species	*log (Activity)*	*Species*	*Fe mass (%)*
Cl^-	−3.05	$Fe(OH)_2^+$	96.3
F^-	−3.05	$FeOH^{2+}$	1.3
SO_4^{2-}	−3.18	$Fe(OH)_3^o$	1.1
Fe^{3+}	−12.96		
Total soluble Fe	−7.59		

Application of Chemical Equilibrium to Soils: Examples

This section will explore the application of equilibrium to soils in which soil solution measurements were used as inputs into a geochemical model. The output of the model was then used to determine ion activity products and compare them with corresponding equilibrium constants. The first examples are successes, followed by apparent nonequilibrium examples and controversies.

Iron Oxides in Reduced Systems

An ideal system to study in soils is the Fe(II/III) oxide group because of well-characterized thermodynamics and rapid equilibrium. In a recent study, amorphous Fe(III)oxide was precipitated under reducing conditions and exposed to either small or large increments of O_2(g) to encourage oxidation. When O_2(g) was added slowly, equilibrium was established with Fe_3O_4(magnetite). When O_2(g) was added rapidly, equilibrium Fe^{3+} activities were consistently elevated relative to Fe_3O_4(magnetite), suggesting the formation of a different solid phase, Fe_3O_4(amorphous) (Figure 1). The proximity of the measured data to the predicted solubility products is striking. For Fe_3O_4(amorphous), the close correspondence is observed over a range from pE + pH 5 to pE + pH 12. Equilibrium with magnetite is suggested in the low-redox environments, and simultaneous equilibrium with $Fe(OH)_3$(amorphous) and magnetite is suggested at higher redox potentials. The formation of an amorphous mixed Fe oxide has been observed previously, but usually with higher solubility.

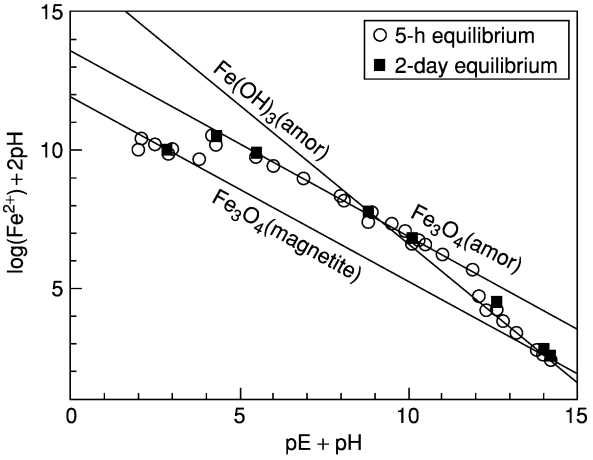


Figure 1 Measured solution activities of Fe (data points) as compared to theoretical solubilities (lines) of selected Fe solid phases. Reproduced from Brennan EW and Lindsay WL (1998) Reduction and oxidation effect on the solubility and transformation of iron oxides. *Soil Science Society of America Journal* 62: 930–937.

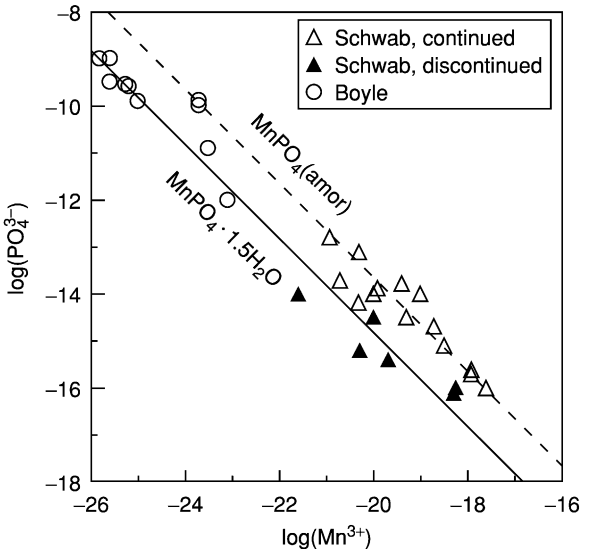


Figure 2 Measured solution activities of Mn^{3+} and PO_4^{3-} (data points) as compared to theoretical solubilities (lines) of $MnPO_4 \cdot 5H_2O$.

Manganese(III) Phosphate

Manganese is a critical element in soils in terms of plant and animal nutrition and can be a pollutant if Mn concentrations become elevated. Manganese phosphates are known to form in natural environments but are too soluble to persist, with the possible exception of $Mn(III)PO_4 \cdot 1.5H_2O$. Although the equilibrium constant for this mineral ($\log K_{eq} = -34.4$) suggests very low solubilities, Mn(III) is present in vanishingly small concentrations in soil (10^{-15} mol l^{-1} as Mn^{3+}). Nevertheless, two studies seem to suggest that the mineral can form in fertilized soils.

Activities of PO_4^{3-} and Mn^{3+} were determined in calcareous soils, and the correspondence between these activities and the theoretical solubility of $Mn(III)PO_4 \cdot 1.5H_2O$ was readily apparent (Figure 2). A similar study was conducted in nonalkaline soils, and two trends were observed in solubility, depending upon whether or not the soil had been recently fertilized. In soils fertilized annually, activities were slightly elevated relative to $Mn(III)PO_4 \cdot 1.5H_2O$, suggesting the presence of a less-crystalline, more-soluble form of the mineral. However, in those soils in which P applications had ceased several years prior to soil sampling, the correspondence between measured and theoretical activities for $Mn(III)PO_4 \cdot 1.5H_2O$ were closer.

Calcite in Soils: Equilibrium versus Nonequilibrium

Precipitation in soils of calcium carbonate, usually considered to be calcite, is easily reversible and quickly establishes equilibrium. Although many studies examining this mineral have been conducted and support the notion of calcite equilibrium, others have found varying degrees in nonequilibrium.

When at equilibrium, the solubility of calcite may be expressed by the following equation:

$$CaCO_3(\text{calcite}) \Leftrightarrow Ca^{2+} + CO_3^{2-}$$
$$\log K_{eq} = -8.48$$

Determining the single ion activities of Ca^{2+} and CO_3^{2-} in soil solution makes it possible to compare measured activity products to the thermodynamic constant. If care is taken in determining the ionic activities, then correspondence between measured and predicted activities can be observed. In one such study, the measured ion activity products ranged from $10^{-8.22}$ to $10^{-8.58}$ with a mean of $10^{-8.41}$.

Although some calcite-containing soils maintain apparent equilibrium with calcite, others show supersaturation. Despite following the diligent protocols to ensure accuracy of measurement, the measured ion activity products are as high as $10^{-7.0}$, indicating supersaturation by a factor of about 30. Nonequilibrium appears to have been induced by high levels of soluble organic matter that stimulates microbial activity (creating a dynamic system) and acts as a long-term source for soluble calcium.

The Outlook for Studying Chemical Equilibrium in Soils

Advances in geochemical modeling and computing capacity have established new ways of viewing and studying soils. Using numerical methods to predict the fate and transport of chemicals in soils is a

compelling notion, and models are being used worldwide. The applications are nearly limitless and include plant nutrition, fate of fertilizers, leaching of soluble constituents, examining the mobility of contaminant metals, and volatilization of organics from soils. Armed with the powerful geochemical models and the knowledge of the chemical and physical limitations of applying chemical equilibria in soils, today's soil scientists can be fully prepared to tackle the most challenging and interesting problems of chemical behavior in soil systems.

List of Technical Nomenclature

μ_i	Chemical potential of species *i*
μ_i^o	Standard chemical potential of species *i*
Equilibrium	A state of balance in a chemical reaction in which the concentrations of the reactants and products are no longer changing and in which minimum free energy has been attained
k	Rate constant
Redox	Oxidation/reduction potential
Solubility constant	Thermodynamic equilibrium constant for a reaction describing the dissolution of a solid phase
Steady state	A balance in an experimentally observed chemical reaction in which the concentrations of the reactants and products appear to be unchanging with time. Steady state differs from equilibrium in that the minimum free energy for a given reaction has not been achieved
Stoichiometry	A description of the relative quantities (in moles) of products and reactants in a chemical reaction as indicated by the coefficients in the balanced chemical equation

Further Reading

Boyle FW and Lindsay WL (1985) Preparation, X-ray-diffraction pattern, and solubility product of manganese (III) phosphate hydrate. *Soil Science Society of America Journal* 49: 758–760.

Brennan EW and Lindsay WL (1998) Reduction and oxidation effect on the solubility and transformation of iron oxides. *Soil Science Society of America Journal* 62: 930–937.

Crawford MB (1996) PHREEQEV: The incorporation of a version of model V for organic complexation in aqueous solutions into the speciation code PHREEQE. *Computers and Geosciences* 22: 109–116.

Gaskova OL, Azaroual M, Bodenan F, and Gaucher E (1999) Evidenced redox disequilibrium from iron and arsenic behavior in tailing leachate (Cheni Site, France). In: *Proceedings of the Ninth Annual V.M. Goldschmidt Conference*. Cambridge, MA.

Kittrick JA (1977) Mineral equilibria and the soil system. In: Dixon JB and Weed SB (eds) *Minerals in Soil Environments*, pp. 1–25. Madison, WI: Soil Science Society of America.

Lindsay WL (1979) *Chemical Equilibrium in Soils*. New York: Wiley Interscience.

Schwab AP (1989) Manganese–phosphate solubility relationships in an acid soil. *Soil Science Society of America Journal* 53: 1654–1660.

Sposito G (1994) *Chemical Equilibria and Kinetics in Soils*. New York: Oxford University Press.

Sposito G and Coves J (1988) *Soilchem, a Computer Program for the Calculation of Chemical Speciation in Soils*. Berkeley, CA: University of California at Berkeley.

Stumm W and Morgan JJ (1996) *Aquatic Chemistry: Chemical Equilibria and Rates in Natural Waters*. New York: Wiley Interscience.

Truesdell AH and Jones BF (1974) WATEQ, a computer program for calculating chemical equilibria in natural waters. *US Geological Survey Journal of Research* 2: 233–248.

Westall JC, Zachary JL, and Morel FMM (1976) *MINEQL, a Computer Program for the Calculation of Chemical Equilibrium Composition of Aqueous Systems*. Massachusetts Institute of Technology, Technical Note 18. Cambridge, MA: MIT Press.

Whitfield M (1974) Thermodynamic limitations on the use of the platinum electrode in Eh measurements. *Limnology and Oceanography* 19: 857–865.

Wolt JD (1989) SOILSOLN: a computer program for teaching equilibria modeling of soil solution composition. *Journal of Agronomic Education* 18: 40–42.

Chemical Speciation Models *See* **Surface Complexation Modeling**

Chernozems *See* **Grassland Soils**

CHILDS, ERNEST CARR

E G Youngs, Cranfield University, Silsoe, Bedfordshire, UK

© 2005, Elsevier Ltd. All Rights Reserved.

From early times it has been recognized that the physical state of soil is important in the practice of agriculture. Soil physics has thus become a subject studied as part of agriculture, not as a subject in mainstream physics. Progress in soil physics owes much to the few physicists in the first part of the twentieth century who left careers in pure physics to work in agricultural research, and these few laid the foundations of the scientific discipline of soil physics. Ernest Carr Childs was one of the physicists who left a career in pure physics to help pioneer a basic physical approach to soil behavior (Figure 1). Born 6 November 1907 in Forest Gate, London, he attended East Ham Grammar School and then studied for his BSc degree in physics at King's College in the University of London. After graduating with first-class honours in 1927, he continued at King's as a research student under E.V. Appleton and later as a demonstrator, with an interruption of a year with the Cambridge Instrument Company. There he was in charge of small-instrument development and this led to the collaboration with A. Campbell in writing a book, *The Measurement of Inductance, Capacitance and Frequency*, which was published in 1935. He was awarded a PhD by the University of London in 1931 for a thesis on the radiofrequency properties of ionized air. He then became a research student at the Cavendish Laboratory in Cambridge in the 'golden years' under Lord Rutherford. His research on the diffraction of slow-speed electrons in certain metal vapors earned him a second PhD, this time from the University of Cambridge, in 1934.

Figure 1 Ernest Carr Childs 1907–73.

With this physics background, in the forefront of research in one of the most prestigious physics departments in the world, Childs crossed the road to the School of Agriculture in the neighboring museum's site in Downing Street, Cambridge, to join H.H. Nicholson in 1934 to study water movement in heavy clay soils, particularly as it related to the newly awakened interest in land drainage in the UK. Why he left such an eminent band of pure physicists to work largely by himself on a rather obscure (at least in the world of physics) subject was never really discussed by him, although it was said that he found the pressure in a competitive environment for quick publication of results at the expense of scientific thoroughness to be unacceptable. The result of the move was that someone with an uncompromising and critical mind for thoroughness in physics was thrown into the practical world of agriculture. Throughout his life, Childs' research philosophy was to develop a physical understanding of soil-water phenomena and to apply this to practical problems in agriculture, hydrology, and engineering. Two topics dominated Childs' work: one was water movement in unsaturated soils, and the other was groundwater movement to land drains.

In considering water in unsaturated soils, Edgar Buckingham, at the beginning of the century, had postulated that water was held in the interstices between solid particles by means of surface tension forces, and had defined the capillary potential dependent on the water content. A gradient of this capillary potential would produce movement of the water. L.A. Richards, in 1931, argued that Darcy's law would be obeyed in unsaturated soils because the air-filled pores could be regarded as the same as the solid soil particles as far as water conduction was concerned. However, the hydraulic conductivity would decrease with the water content, since the pathways for water movement would be more tortuous, fewer, and narrower, with the larger pores draining first. The work of W.B. Haines at this time on the physics of hysteresis in the relationship between water content and soil-water pressure in unsaturated soils was a complicating factor. Childs saw that these basic physics concepts were of little use to agriculture unless they could be applied to practical situations.

The study of water-table heights in drained agricultural land had been pursued by groundwater hydrologists by applying Darcy's law through the

work of Jules Dupuit, Philipp Forchheimer, and J. Boussinesq. These studies resulted in the now-familiar drainage equations, the first obtained by the Danish engineer L.A. Colding in 1864, and in the 1930s developed for practical drainage engineering by S.B. Houghoudt in the Netherlands. However, these developments were not rigorous enough for Childs, and he pursued a more basic analytical approach, as did Don Kirkham in the USA.

Childs brought his physics discipline to the ongoing work at Cambridge on the mole drainage of heavy clay soils in which unlined channels, formed by a bullet drawn through the soil, act as drains. In his first soil physics papers, he recognized that gravity played a minor role in the movement of soil water when these soils were unsaturated, and gradients of soil-water pressure were all-important. In order to be able to predict the soil-water movement, he hypothesized that water moves in soils according to the diffusion equation that is amenable to mathematical solution for known boundary conditions. He was thus able to demonstrate how the moisture profile reacts to various surface conditions of evaporation and rainfall, predicting soil-water behavior and comparing analytical results with those observed in measurements on the Gault clay soil at the Cambridge University Farm, as well as with other published results. A decade later, Childs introduced the concept of soil-water diffusivity that depends on the soil-water content. Nevertheless, in the series of papers on water movement in heavy clay soils that used a constant diffusion coefficient to analyze soil-water profiles, he concluded "that the theory of diffusion of water through soil accounts qualitatively for the water movement from wet to dry soil, fits the best experiments numerically, and is at present alone in doing this." This still summarizes the situation so long as it is recognized that the diffusion coefficient is dependent on soil-water content and that difficulties associated with hysteresis occur in using diffusion theory, difficulties that were apparent to Childs even in this series of papers.

This analytical work on the movement of water through heavy clay soils was accompanied by an interest in practical problems of mole drainage, the common method of draining heavy clay soils. He developed an automatic recorder to measure the rapid increase and decrease of flow rates from mole drains and employed it in studies of drain deterioration with the annual weather cycle. This deterioration depends on the stability of clay soils to cycles of wetting and drying. Childs used moisture-retention curves, which he termed 'moisture characteristics,' after the characteristics of the thermionic valve with which he was familiar through his earlier work in pure physics, to indicate the pore-size distribution of soil samples. He devised a test to determine the stability of clay soils to wetting cycles, and hence their suitability for mole drainage, by subjecting samples of soil crumbs to different times of wetting and then measuring their moisture characteristics. The stability of a soil was shown by its ability to maintain large pores that drained at small tensions. By plotting the slope of the moisture characteristic against the pressure deficiency, he was thus able to show the stability of the Gault clay and the instability of London clay when left undisturbed in water. Besides this work on soil behavior during mole drainage, Childs discussed the mechanics of the design of the mole plough itself. For this early work, he was awarded the ScD degree of the University of Cambridge in 1945.

Childs' studies on moisture characteristics of soil crumbs in the context of the stability of mole drains led to more fundamental studies connected with the interpretation of such curves in terms of their pore-size distribution that determines the conduction of water through soils. Measurement of the hydraulic conductivity in unsaturated soils was a time-consuming and difficult task. Childs and N. Collis-George led the way in modeling the hydraulic conductivity function from the pore-size distribution obtained from the moisture characteristic. They compared their modeled results with accurately measured values of the hydraulic conductivity of unsaturated sands obtained using the method of flow to a water table down long columns that they devised.

While the Richards equation gives the basis for describing the moisture-profile development in soil profiles under unsaturated conditions, computing such profiles from this equation was practically impossible with the tools available in the middle of the twentieth century. Research workers, if they were lucky, used electromechanical calculators, but most relied on slide rules, logarithmic tables, and pencil and paper to do their calculations, so little progress seemed possible. Childs' recollection of his earlier work, in which he hypothesized that water movement could be described by the diffusion equation, led him to the diffusion form of the Richards equation with a soil-water diffusivity dependent on the water content. This was the beginning of the theoretical calculation of moisture profile developments from the Richards equation.

It is interesting to note that Don Kirkham and C.L. Feng concluded in 1949 that "the differential equation of diffusion theory is not valid" while producing moisture profiles during horizontal infiltration that can be analyzed to give a moisture content-dependent diffusivity. They were clearly of the opinion that the diffusion coefficient needed to be a constant.

Advances were being made in the 1950s in the numerical solution of the nonlinear diffusion equation in other scientific disciplines. Arnold Klute, using

a soil water-dependent diffusivity, used these advances to compute the moisture profiles with the steep wetting fronts that occur during horizontal infiltration. Later, John Philip developed a very efficient method of solving numerically the highly nonlinear diffusion equation, and he extended this to the situation of vertical infiltration, when the effect of gravity has to be considered. Computers have allowed great improvements in numerical techniques of solving the Richards equation in two and three dimensions, as well as the simpler one-dimensional case, making the use of the diffusion approach redundant. Nevertheless, the diffusion approach that Childs introduced produced the progress in soil-water studies that occurred in the middle part of the last century. A particular short-coming of the diffusion analysis of soil-water flow is that it is limited to profiles either being entirely in the wetting state or in the draining state, because at a reversal of trend the soil-water diffusivity is discontinuous. In describing the soil-water distribution during the redistribution of infiltration water, Childs and his colleagues at Cambridge were quick to point out that the diffusion approach was unsuitable in situations where a reversal of wetting or draining takes place.

Childs' introduction to the physics of soil water through the drainage of heavy soils that was ongoing in Cambridge in the 1930s led to an interest in land-drainage theory in general. He was critical of the general acceptance of work based on applying the approximate Dupuit–Forchheimer analysis that properly relates to shallow flow to ditches that penetrate down to a horizontal impermeable floor. He was particularly critical of drainage equations incorporating the concept of radial flow to drains to account for a depth of soil below drain level. He thus turned his attention to a study of land drainage using a fundamental physics approach.

Childs noted that the water flow in soils was analogous to the flow of electricity in conductors, and this led him to pioneer the use of electric analogues in the study of the two-dimensional problem of steady-state drainage of uniform rainfall to pipe drains. To do this he produced a conducting paper that simulated the soil, by carefully spraying filter paper with a graphite suspension in water to produce a paper with uniform conductivity. The marketing of commercial Teledeltos graphited paper was later a great help in this work. In this way, Childs was able to trace the streamlines and equipotentials in drained lands under given rainfall conditions and obtain the position and shape of the water table. 'Rain' was introduced through electrodes along the 'water table' whose position was found by trial and error and cut to shape to give the requirement that the potential there was equal to the height above the 'drain' formed by an electrode cemented to the paper. This work was of great significance in correcting earlier viewpoints regarding land drainage.

It seemed to Childs that there were limits to the insights to be gained by rigid mathematical analysis of problems relating to the flow of fluids in porous materials, especially those relating to the drainage of agricultural lands, and further progress was possible only by model or a full-scale study of porous materials themselves. Childs therefore had constructed a laboratory containing a large sand tank with the aim of furthering the understanding of the physics of land drainage. This was built on Cambridge University Farm. It formed the nucleus of the Agricultural Research Council Unit of Soil Physics, affectionately known as the 'Tank,' which was formed in 1951 and disbanded in 1977. Childs, an Assistant Director of Research at the University of Cambridge and later the Reader in Soil Physics, was the Unit's Director. The tank, which measured $10\,\text{m} \times 10\,\text{m} \times 1.5\,\text{m}$, was filled with uniform Leighton Buzzard sand. There was the facility to provide uniform 'rain' of various intensities on to the sand surface. Portholes in the side of the tank provided outlets for drains that could be installed in the sand. Pressure cells were developed and installed in the sand so that pressure heads could be continuously recorded during experiments.

Experiments in the sand tank were used to develop and test methods of measuring hydraulic conductivity below the water table as well as for its main purpose of advancing our understanding of water-table behavior in drained lands. From work done in the Unit's sand tank, advances were made concerning: the influence of depth of soil below drain level on water-table heights; the non-steady-state problem of the moving water table above drains with changing rainfall intensity, including a reappraisal of the concept of specific yield; the hydraulic behavior around gappy drains; and the water-table behavior in three-dimensional drainage situations that was inspired by mole drains being laid above lateral drain channels, recalling Childs' early work when he joined the School of Agriculture. Present-day modeling of water regimes in drained lands depends much on the basic understanding that emerged from these studies.

Having built the large sand-tank laboratory, Childs left others at Cambridge to use it. Childs' attention was diverted to the use of conformal mapping techniques to solve the problem in potential theory of the difficult and unusual boundary conditions presented by land drainage. This technique of solving two-dimensional groundwater problems had long been used by Russian scientists, but in general their work had gone unnoticed in the West. Childs

learned of the solutions given by the Dutch mathematician J.J. van Deemter in 1950, and the Danish mathematician Frank Engelund in 1951, to the steady-state drainage problem of uniform rainfall and upward artesian water flow to uniformly spaced cylindrical drains installed in an infinitely deep soil. These gave the shape and height of the water table for drains of given diameter. Van Deemter also considered the problem of ditch drainage when a surface of seepage occurs on the ditch wall above ditch-water level. The analysis of the drainage problem using conformal mapping techniques gave a much more rigorous approach to the determination of water-table heights in drained lands than theory developed from the Dupuit–Forchheimer analysis.

Childs made an attempt to extend the analysis to the situation where there was an impermeable barrier at some depth, but perhaps the more significant use of the theory given by van Deemter and by Engelund was its use in the consideration of the nature of the drain channel. It was accepted that drains did not conform to the uniform sinks that theory assumed, because they were mostly of a gappy nature, being constructed initially from tiles butted together and then from slotted plastic pipes with water entry taking up a limited area of the drain surface. It was argued, and confirmed experimentally in the Unit's sand tank, that the converging flow to the gaps in the drain produced almost cylindrical equipotentials within a short radial distance, thus allowing the oncept of 'equivalent drain radius' to be applied. This concept could simply be used in van Deemter's and Engelund's analyses to show the effect on the water-table height due to the resistance effect of the gappy nature of the drain channel. In particular, it was noted that the gappy nature of the drain had a relatively small effect on the water table midway between drain lines, but raised the water table above the drains significantly so that the water table became flatter the greater the drain resistance. All this time, during which Childs was attracted to more theoretical studies, work progressed in the Unit's sand tank.

From early times in his career in soil physics research, Childs was very mindful of the role hysteresis played in the way water redistributed itself in soil profiles. It is impractical to undertake the amount of measurement required to obtain a complete trace of the hysteresis in soil-water relationships, so that bringing hysteresis into a rigorous analysis of soil-water movement seemed impossible. It was with some excitement that Childs' attention was drawn to the independent-domain theory proposed by L. Néel in the context of magnetic hysteresis. This excitement was short-lived, however, when it was realized that pores that could be viewed as the 'domains' behaved far from independently. Further work, particularly by Ed Miller and Alex Poulovassilis, led to the dependent-domain theory of hysteresis that would apply to soil-water relationships. At the time of his death, Childs was working on a unified-domain theory of hysteresis that remained uncompleted. Generally, however, it appears that the amount of experimental measurements required to obtain the necessary data for the dependent-domain theory is practically the same as that required for the direct tracing of the hysteresis paths, so that little is achieved. Despite this, what domain theories may give are insights into the pore structure of soils.

Childs died at the age of 65 on 24 May 1973. His legacy was a fundamental physics approach to soil physics that inspired those who worked with him, as well as those who followed in his footsteps. His work emphasizes the importance of a basic understanding of soil physics phenomena in tackling practical soil problems.

See also: **Capillarity**; **Darcy's Law**; **Diffusion**; **Drainage, Surface and Subsurface**; **Hydrodynamics in Soils**; **Hysteresis**; **Infiltration**; **Macropores and Macropore Flow, Kinematic Wave Approach**; **Porosity and Pore-Size Distribution**

Further Reading

Childs EC (1940) The use of soil moisture characteristics in soil studies. *Soil Science* 53: 239–252.

Childs EC (1942) Stability of clay soils. *Soil Science* 53: 79–92.

Childs EC (1950) The water table, equipotentials, and streamlines in drained land. *Soil Science* 56: 317–330.

Childs EC (1969) *An Introduction to the Physical Basis of Soil Water Phenomena*. London, UK: John Wiley.

Childs EC and Collis-George N (1950) The permeability of porous materials. *Proceedings of the Royal Society of London. Series A: Mathematical and Physical Sciences* 201: 392–405.

Childs EC and Youngs EG (1974) Soil physics: twenty-five years on. *Journal of Soil Science* 25: 408–419.

Gardner WH (1986) Early soil physics into the mid-20th century. *Advances in Soil Science* 4: 1–101.

Nicholson HH (1942) *The Principles of Field Drainage*. London, UK: Cambridge University Press.

Youngs EG (1977) *The Agricultural Research Council Unit of Soil Physics, 1951–1977*. University of Cambridge Department of Applied Biology Memoir, 49, pp. 4–10. Cambridge, UK: Cambridge University Press.

Youngs EG (1983) The contribution of physics to land drainage. *Journal of Soil Science* 34: 1–21.

Youngs EG (1984) Developments in the physics of land drainage. *Journal of Agricultural Engineering Research* 29: 167–175.

Youngs EG, Towner GD, and Poulovassilis A (1974) In memoriam: Ernest Carr Childs. *Soil Science* 117: 241–242.

CIVILIZATION, ROLE OF SOILS

D Hillel, Columbia University, New York, NY, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Human Management of the Soil

Manipulation and modification of the environment was a characteristic of many societies from their very inception. Long before the advent of earth-moving machines and toxic chemicals, even before the advent of agriculture, humans began to affect the land and its biota in ways that tended to destabilize natural ecosystems. In many of the ancient countries, where human exploitation of the land began early in history, we find disturbing examples of once-thriving regions reduced to desolation by human-induced degradation. Some of the early civilizations succeeded all too well at first, only to set the stage for their own eventual demise. The poor condition of the Fertile Crescent today is due not simply to changing climate or to the devastation caused by repeated wars, though both of these may well have had important effects. It is due in large part to the prolonged exploitation of this fragile environment by generations of forest cutters and burners, grazers, cultivators, and irrigators, all diligent and well-intentioned but destructive nonetheless.

An example of soil abuse on a large scale can be seen in the rainfed parts of the Mediterranean region, which has borne the brunt of human activity more intensively and for a longer period than any other region on earth. Visit the hills of Israel, Lebanon, Greece, Cyprus, Crete, Sicily, Tunisia, and southeastern Spain. There, rainfed farming and grazing were practiced for many centuries on sloping terrain, without consistent or fully effective soil conservation. The land has been denuded of its natural vegetative cover, and the original mantle of fertile soil has been raked off by the rains and carried down the valleys toward the sea. That may have been the reason why the Phoenicians, Greeks, Carthaginians, and Romans, each in turn, were compelled to venture away from their own country and to establish far-flung colonies in pursuit of new productive land. The end came for each of these empires when it had become so dependent on distant and unstable sources of supply that it could no longer maintain central control or ward off growing competition from other land-hungry nations.

Consider, for another example, the southern part of Mesopotamia. Aerial and satellite photographs of this area, now part of Iraq, reveal wide stretches of barren, salt-encrusted terrain. Long ago, these were fruitful fields and orchards, tended by enterprising irrigators whose very success inadvertently doomed their own land. The once-prosperous cities of Mesopotamia are now 'tells,' mute time capsules in which the material remnants of a civilization that lived and died there are entombed. Similarly ill-fated was the ancient civilization of the Indus Valley in present-day Pakistan.

There were, on the other hand, some societies that did better than others. The more successful ones were those who were able to develop modes of soil and land management that enabled them to thrive in the long run. Impressive evidence exists regarding the terrace-building farmers of eastern Asia and the Near East, as well as the wetlands-based societies of Meso-America and South America. Remarkably productive wetland management systems have survived in China and other parts of Southeast Asia. In contrast with the irrigation-based civilization of Mesopotamia, the similarly based civilization of Egypt sustained itself for more than five millennia – though it too (owing to its intensified management, impelled by its 20-fold increase of population in the last two centuries) is now beset with problems of waterlogging and salinity.

Historical Attitudes Toward the Soil

Early societies generally revered the earth and tended to deify it. The earth was held sacred as the embodiment of a great spirit, the creative power of the universe, manifest in all phenomena of nature. The earth spirit was believed to give shape to the features of the landscape and to regulate the seasons, the cycles of fertility, and the lives of animals and humans. Rocks, trees, mountains, springs, and caves were recognized as receptacles for this spirit.

The cult of the earth is perhaps the oldest and most universal element in all religions. The Australian aborigines and the African Bushmen, among the last to have maintained the preagricultural hunter-gatherer mode of life, have regarded the earth as the Great Provider, the source of all sustenance. So did – and still do – the native Americans. The ancient Egyptians represented the earth as the god Geb, who mated with the sky goddess Nut. The sexual roles were reversed in the culture of the ancient Canaanites, who worshiped the male Baal as the god of the sky, who provided the rain that fructified the earth goddess Ashera. To the ancient Greeks, the earth was Gaea, the maternal goddess who, impregnated by her son and consort Uranus

(god of the sky), became mother of the Titans and progenitor of all the many gods of the Greek pantheon.

In the Hebrew Bible, there are two very different accounts of creation and the role granted or assigned to humanity in the scheme of life on earth. In the first chapter of Genesis, we read that God (called 'Elohim') decided to "make man in our own image, and let them rule over the fish of the sea, and over all the earth, and over every creeping thing that creepeth upon the earth." And God blessed man and woman and said unto them: "Be fruitful, and multiply, and fill the earth, and conquer it... Here, I have given you every herb yielding seed and every tree with fruit... to you shall it be for food."

But the divine injunction to humanity is defined quite differently in the second chapter of Genesis. In this version, God (called 'Yahweh') "formed man out of the soil of the Earth and blew into his nostrils the breath of life, and man became a living soul." Then God planted a garden in Eden in the east and placed the man therein... to serve and preserve it." (Those words are this author's translation of the Hebrew words 'l'ovdah ul'shomrah' (Genesis 2:15), usually rendered "to dress it and keep it" (King James Version; or "to till it and keep it," Revised Standard Version.) Here, humanity is not given license to rule over the environment for self-gratification, but – quite the contrary – is charged with the responsibility to nurture and protect it.

Thus, latent in one of the main founts of Western Civilization we have two opposite perceptions of humanity's destiny. One is anthropocentric: humans are set above nature, to be its omnipotent masters. They are endowed with the power and the right to dominate all other creatures, toward whom they have no obligations. The other view is more modest. The human earthling is made of soil and is given a "living soul." There is no mention of being made "in the image of God." Humanity's appointment is not an ordination but an assignment, which is to serve as the custodians of God's garden.

Over the generations, unfortunately, it has generally been the arrogant and narcissistic view, implied in the first Biblical account, that has prevailed. It has repeatedly been cited and used as a religious justification or rationale for the unbridled and relentless exploitation of the environment. The imperative now is to accept and realize the long-ignored second view of our proper role in relation to nature.

Readers of the Bible in translation miss much of the imagery and evocative verbal associations in the original language. The indissoluble link between humanity and soil is manifest in the very name 'Adam,' derived from 'adamah,' a Hebrew noun of feminine gender meaning earth or soil. Adam's name encapsulates the notion that his existence is derived from the soil, to which he is tethered throughout life and to which he is fated to return at the end of his days. Likewise, the name assigned to Adam's mate, 'Hava,' rendered Eve in transliteration, literally means 'living.' In the words of the Bible, "Adam called his wife Hava because she was the mother of all living." Together, therefore, Adam and Eve signify "Soil and Life."

The ancient association of humanity with soil is echoed in the Latin name for man, 'homo,' derived from 'humus,' the stuff of the soil. This powerful metaphor suggests an early realization of a profound truth that has since been too often disregarded. Since the words 'humility' and 'humble' also derive from humus, it is rather ironic that we should have assigned our species so arrogant a name as *Homo sapiens sapiens* ('Wise Wise Man'). Perhaps a more appropriate and certainly more modest name would be *Homo sapiens curans*, with the last word denoting caring or caretaking, as in 'curator.'

Human Origins

Our species' birthplace was evidently in the continent of Africa, and its original habitat was probably the subtropical savannas that constitute the transitional areas of sparsely wooded grasslands lying between the zone of the humid and dense tropical forests and the zone of the semiarid steppes. We can infer the warm climate of our place of origin from the fact that we are naturally so scantily clad, or furless; and we can infer the open landscape from the way we are conditioned to walk, run, and gaze over long distances.

For at least 90% of its career, the human animal existed merely as one member of a community of numerous species who shared the same environment. Humans were adapted to subsist within the bounds defined by the natural ecosystem. By and large, our ancestors led a nomadic life, roaming in small bands, foraging wherever they could find food. They were gatherers, scavengers, and opportunistic hunters. Unlike their primate cousins who remained primarily vegetarian, humans diversified their diet to include the flesh of whichever animals they could find or catch, as well as a variety of plant products such as nuts, berries and other fruits, seeds, some succulent leaves, bulbs, tubers, and fleshy roots.

The story of how humans gradually ventured far from their original birthplace to range over a variety of climates and landscapes is a remarkable saga of audacity, ingenuity, perseverance, and adaptability. The mode of human adaptation was not entirely genetic or physical: there was not enough time for that. Rather, their adaptation was in large part behavioral.

Instead of relying on physical prowess, they had to use inventiveness to survive the elements and to compete successfully against stronger animals. The increase in brain size and manual dexterity, as well as the invention of various tools and stratagems, gradually enabled humans to overcome the constraints of their ancestry.

By 1 million years ago, hominids had become taller (approximately 1.5 m in height) and had acquired a larger brain. Some time later, the so-called *Homo erectus* learned to set and use fire, probably at first only for cooking and softening food. That achievement, along with the fashioning of stone tools, was a momentous innovation, celebrated in the Greek myth of Prometheus. Eventually, it had a great effect on the environment. Evidence has been found in southern and eastern Africa of repetitive occurrences of brush fires, whether purposeful or accidental, apparently set by humans nearly a million years ago. This early manifestation of pyrotechnology signifies the beginning of human manipulation of the earth's ecosystems. The use of fire became even more important when humans moved out of the tropics into colder climes, where bonfires and hearths were needed to warm their shelters in winter, as well as to cook their food.

The Paleolithic Transformation

At some point, humans began to use fires deliberately and systematically to flush out game and to modify the vegetation. The resultant suppression of woody plants and the fertilizing effect of ash encouraged the growth of herbaceous plants and improved their nutritional quality. This benefited foraging species and raised the carrying capacity for game animals. It also facilitated travel and hunting by humans. In time, the practice of clearing woodlands and shrublands by repeated firings also set the stage for the advent of agriculture.

As vegetation is affected by fire-setting hunters, so are soils. Following repeated fires and deforestation, soil erosion and landslides often result in the greatly increased transport of silt by streams, and in the deposit of that silt in river valleys and estuaries. The dating of fluvial sediments in river valleys in England, for example, suggests that they were the products of erosion caused by anthropogenic clearings in the originally closed deciduous forest during the Late Paleolithic period.

By approximately 40 000 years ago, modern humans, evidently indistinguishable from us today in physical features and in intelligence, had gained dominance. Clad in sewn garments made of animal skins, able to make and use a variety of implements and weapons, humans were able to range and settle in locations and climes far from their ancestral home. All the while they continued to evolve by natural selection, increasingly aided by cultural and technological development. They also contrived increasingly sophisticated methods of obtaining and storing foods, including the selective gathering, processing, and preservation of biological products, and eventually the domestication of plants and animals.

The described series of changes has been termed the Paleolithic (Early Stone Age) Transformation. It was marked by the development of adaptive mechanisms and modes of social organization suited to exploiting potentialities within the environment. Each modification of the environment entailed additional human responses, which in turn further modified the environment, so that a process of escalating, dual metamorphosis was instigated. Human intelligence and culture were both cause and effect in that fateful interplay. The peculiarly dynamic and progressive evolution of human ecology is the true history of our species.

The Agricultural Transformation

The gradual intensification of land use continued throughout the Paleolithic period, so that by its later stages nearly all the regions of human habitation had experienced some anthropomorphic modification of the floral and faunal communities. At some stage, humans began to delineate sections of the environment that they could control and manage to suit their special needs, and in which they could find secure shelters for habitation.

The process of intensification of land use can be seen as an adaptation to increasing population pressure. Several millennia of occupation by hunter-gatherers, even at a very low density and slow rate of population growth, filled up the terrain and decimated the natural forageable resources to the point where subsistence was difficult. The choice was then between migration and some form of intensification aimed at inducing the same area to yield a greater supply. The selective eradication of undesirable animal species and the encouragement of desirable ones led eventually to domestication and herding. Similarly, selective manipulation of plant communities involved suppressing some species and promoting the growth of others. The entire series of activities quite logically led to plant domestication and propagation, and to purposeful soil management aimed at creating favorable conditions for crop production – that is to say, these activities culminated in the development of agriculture and the agricultural way of life.

The Agricultural Transformation is very probably the most momentous turn in the progress of humankind, and many believe it to be the real beginning of civilization. Often called the Neolithic Revolution, this transformation apparently first took place in

the Near East, approximately 10 000 years ago, and was based on the successful domestication of suitable species of plants and animals. The ability to raise crops and livestock, while resulting in a larger and more secure supply of food, definitely required attachment to controllable sections of land and hence brought about the growth of permanent settlements of larger, coordinated communities. The economic and physical security so gained accelerated the process of population growth and necessitated further expansion and intensification of production. A self-reinforcing and self-perpetuating pattern thus developed, so the transition from the nomadic hunter-gatherer mode to the settled farming mode of life became in effect irreversible.

The Agricultural Transformation radically changed almost every aspect of human life. Food production and storage stimulated specialization of activities and greatly enhanced the division of labor that had already started in hunting-gathering societies. The larger permanent communities based on agriculture required new forms of organization, both social and economic. Domestication affected family structure and the roles and status of men, women, and children. With permanent facilities such as dwellings, storage bins, heavy tools, and agricultural fields came the concept of property. The inevitably uneven allocation of such property resulted in self-perpetuating class differences. Religious myths and rituals, as well as moral and behavioral standards, developed in accordance with the new economic and social constellation and the new relationship between human society and the environment.

The evolution of agriculture left a strong imprint on the land in many regions. The vegetation, animal populations, slopes, valleys, and soil cover of land units were radically altered. The processes of tillage and fallowing, of terracing, of irrigation and drainage have had considerable consequences for such processes as the erosion of slopes and the aggradation of valleys, as well as the formation of deltas in seas and lakes where silt from the land surface naturally comes to rest. Soil lost from deforested and subsequently cultivated slopes is unlikely to be regenerated unless the land is allowed to revert to its forest cover for many scores, perhaps even many hundreds, of years.

Soil Husbandry and Ceramics

An important factor in the evolution of agriculture in the Near East, as elsewhere, was the development of the tools of soil husbandry. Seeds scattered on the ground are often eaten by birds and rodents, or subject to desiccation, so their germination rate is likely to be very low and uneven. Given a limited seed stock, farmers would naturally do whatever they could to promote germination and seedling establishment. The best way to accomplish this is to insert the seeds to some shallow depth, under a protective layer of loosened soil, and to eradicate the weeds that might compete with the crop seedlings for water, nutrients, and light.

The simplest tool developed for the purpose was a paddle-shaped digging stick, by which a farmer could make holes for seeds. The use of this simple device was extremely slow and laborious, however, so at some point the digging stick was modified to form the more convenient spade, which could not only open the ground for seed insertion but also loosen and pulverize the soil and eradicate weeds more efficiently. In time, the spade developed a triangular blade, at first made of wood, but later made of stone, and eventually of metal. Such a spade, initially designed to be used by one person, was later modified so that it could be pulled by a rope so as to open a continuous slit, or furrow, into which the seeds could be sown. A second furrow could then be made alongside the first, to facilitate seed coverage. In some cases, the rows were widely enough separated to permit a person to walk between the rows, weeding the cultivated plot.

The human-pulled traction spade or 'ard' gradually metamorphosed into an animal-drawn plow. The first picture of such a plow, dating to 3000 BCE, was found in Mesopotamia, and numerous later pictures have been found both there and in Egypt, as well as in China. It was not long before these early plows were fitted with a seed funnel, so that the acts of plowing and sowing could be carried out simultaneously. The same ancient implement is still in use today in parts of the Near and Middle East.

While the development of the plow represented a huge advance in terms of convenience and efficiency of operation, it had an important side effect. As with many other innovations, the benefits were immediate, but the full range of consequences took several generations to play out, long after the new practice became entrenched. The major environmental impact was that plowing made the soil surface – now loosened, pulverized, and bared of weeds – much more vulnerable to accelerated erosion. In the history of civilization, contrary to the idealistic vision of the prophet Isaiah, the plowshare may well have been far more destructive than the sword.

As farming induced sedentary living in villages, there also developed an important new industry that depended directly on the soil – pottery, which began in the Near East at approximately 6000 BCE. The shaping and baking of clay to form hardened vessels for grain, for liquid storage and conveyance, and for cooking, represented the first transmutation of material by humans. Such an innovation could not

have been possible, owing to the fragility of the ceramic objects, during the nomadic hunting-gathering phase.

From Rainfed to Irrigated Farming

The Mediterranean-type climate of the Near East is at best semihumid, but more typically semiarid, with a rather high incidence of drought. Hence the practice of rainfed farming could not provide anything like total food security. The early farmers, who depended only on seasonal rainfall to water their crops, were always at the mercy of a capricious and highly unpredictable weather regime. The Hebrew Bible, for instance, is replete with references to the ever-present threat of drought and consequent famine. In time of need, therefore, it was only natural for farmers located near river courses to attempt to augment the water supply to crops by diverting water from the river. It was also reasonable to try to raise crops on riverine flood plains that were naturally inundated, and thereby irrigated, periodically.

Thus, many centuries after its advent, farming was extended from the relatively humid centers of its origin toward the extensive river valleys of the Tigris-Euphrates, the Nile, and the Indus. As the climate of these river valleys is quite arid, a new type of agriculture based primarily or even entirely on irrigation came into being. With a practically assured perennial water supply, an abundance of sunshine, a year-round growing season, deep and fertile soils, and relative security from the hazards of drought and erosion that beset rainfed farming, irrigated farming became a highly productive enterprise. However, behind its success lurked an insidious problem that could not have been foreseen initially: the problem of soil waterlogging and salination.

Silt and Salt in Ancient Mesopotamia

Ancient Mesopotamia owed its prominence to its agricultural productivity. The soils of this alluvial valley are deep and fertile, the topography is level, the climate is warm, and water is provided by the twin rivers Euphrates and Tigris. However, the diversion of river water onto the valley lands led to a series of interrelated problems.

The first problem was sedimentation. Early in history, the upland watersheds were deforested and overgrazed. The resulting erosion was conveyed by the rivers as suspended silt, which settled along the bottoms and sides of the rivers, thus raising their beds and banks above the adjacent plain. During periods of floods, the rivers overflowed their banks, inundated large tracts of land, and tended to change course abruptly. The silt also settled in channels and clogged up the irrigation works.

The second and more severe problem was salt. Seepage from the rivers, the irrigation channels, and the flood-irrigated fields caused the water table to rise throughout southern Mesopotamia. Because all irrigation waters contain some salts, and because crop roots normally exclude salts while extracting soil moisture, the salts tend to accumulate in the soil and groundwater. As the undrained water table rose, it took the salts back into the soil.

The farmers of ancient Mesopotamia attempted to cope with the process of salination by periodically fallowing their land, and by replacing the salt-sensitive wheat with relatively salt-tolerant barley. However, the process proceeded inexorably; so the ancient hydraulic civilizations of Sumer, Akkad, Babylonia, and Assyria each, in turn, rose and then declined, as the center of population and culture shifted over the centuries from the lower to the central to the upper parts of the Tigris-Euphrates valley.

The Sustainability of Egyptian Agriculture

In contrast to Mesopotamia, the civilization of Egypt thrived for several millennia in the same location. What explains the persistence of irrigated farming in Egypt in the face of its demise in southern Mesopotamia? The answer lies in the different soil and water regimes of the two lands. Neither clogging by silt nor poisoning by salt was as severe along the Nile as in the Tigris-Euphrates plain.

The silt of Egypt is brought by the Blue Nile from the volcanic highlands of Ethiopia, and it is mixed with the organic matter brought by the White Nile from its swampy sources. It was not so excessive as to choke the irrigation canals, yet was fertile enough to add nutrients to the fields and nourish their crops. Whereas in Mesopotamia the inundation usually comes in the spring, and summer evaporation tends to make the soil saline, the Nile rises in the late summer and crests in autumn. So in Egypt the inundation comes at a more favorable time: after the summer heat has killed the weeds and aerated the soil, just in time for the prewinter planting of grain.

The ancient Greek name for Egypt was 'Khemia,' from the word 'Khami' signifying black soil, which was what the Egyptians themselves called their land. So fabulously fertile was that dark deposit of the Nile (which contrasted vividly with the yellowish color of the nearby desert sand) that the Greeks considered it the mother lode of all material substances and apparently named the science of materials after it. That name has been transmuted into our term 'chemistry.'

The narrow floodplain of the Nile (except in the Delta) precluded the widespread rise of the water

table. Over most of its length, the Nile lies below the level of the adjacent land. When the river crested and flooded the land, the seepage naturally raised the water table. As the river receded and its water level dropped, it pulled the water table down after it. The all-important annual pulsation of the river and the associated fluctuation of the water table under a free-draining floodplain created an automatically repeating, self-flushing cycle by which the salts were leached from the irrigated land and carried away by the Nile itself.

The basis of Egypt's civilization was the nearly optimal combination of water, soil, nutrients, and organic matter, provided by the regular annual regime of the river, which was more dependable and timely than the relatively capricious floods of Mesopotamia. It enabled Egyptian farmers to produce a surplus that fed the artisans, scribes, priests, merchants, noblemen, and, above all, the Pharaohs, who used their power to order the building of monuments. Those monuments still stand today, less in testimony to the kings who ordered them than to the diligence and organization of a society of labor rooted in the soil.

See also: **Desertification**; **Erosion:** Water-Induced; **Irrigation:** Environmental Effects; **Salination Processes**

Further Reading

Carter VG and Hale T (1974) *Topsoil and Civilization*. Norman, OK: University of Oklahoma Press.

Hillel Daniel (1992) *Out of the Earth: Civilization and the Life of the Soil*. Berkeley, CA: University of California Press.

Hyams Edward (1976) *Soil and Civilization*. New York: Harper & Row.

Jacobsen Thorkild (1982) *Salinity and Irrigation in Antiquity*. Malibu, CA: Udena Publications.

Mitchell John (1975) *The Earth Spirit: Its Ways, Shrines, and Mysteries*. New York: Thames and Hudson.

Redman CL (1978) *The Rise of Civilization: From Early Farmers to Urban Society in the Ancient Near East*. San Francisco, CA: W. H. Freeman.

Rindos David (1984) *The Origins of Agriculture: An Evolutionary Perspective*. San Diego, CA: Academic Press.

Thirgood JV (1981) *Man and the Mediterranean Forest: A History of Resource Depletion*. New York: Academic Press.

Classification of Land Use *See* **Land-Use Classification**

CLASSIFICATION OF SOILS

R W Arnold, Formerly with USDA–Natural Resources Conservation Service, Washington, DC, USA

© 2005, Elsevier Ltd. All Rights Reserved.

A classification is an organized body of knowledge about something of interest. It is intended to show relationships among and between entities, and to help recall important properties of these entities. Three principles deal with the setup of a classification: 'purpose' states the reasons for wanting to organize soil knowledge; 'domain' specifies the universe of objects relevant to the purpose; and 'identity' defines and names the individual members of the domain. Four additional principles deal with the organization of a system: 'differentiation' specifies a protocol-guided hierarchical structure with categories and classes within categories; 'prioritization' is evident by sequencing categories and sequencing classes within categories; 'diagnostics,' whereby selected soil properties and features (diagnostics) are quantified, provide objectivity; and 'membership' is based on quantified class limits and described central tendencies. A final principle of certainty recognizes change as inevitable and the driving force for continual testing of a system.

Soils occur in most terrestrial environments and commonly have observable properties such as color and structure, arranged as layers; that is, soils have morphology. There are patterns of soil morphology throughout the world that are systematic enough to suggest causal relationships with other features of location.

In the late 1800s, 'pedology' originated from 'genetic soil science,' as termed by V.V. Dokuchaev, a Russian. He envisioned soils to be natural bodies of transformed materials at or near the interface of the lithosphere, biosphere, and atmosphere (**Figure 1**). Soils were recognized as cause-and-effect results of processes that were, and are, influenced by natural environmental factors and conditions: the factors of climate, biota, parent materials, and relief (landscape features) interact over time to produce soils. This was the most fundamental change in the concept of

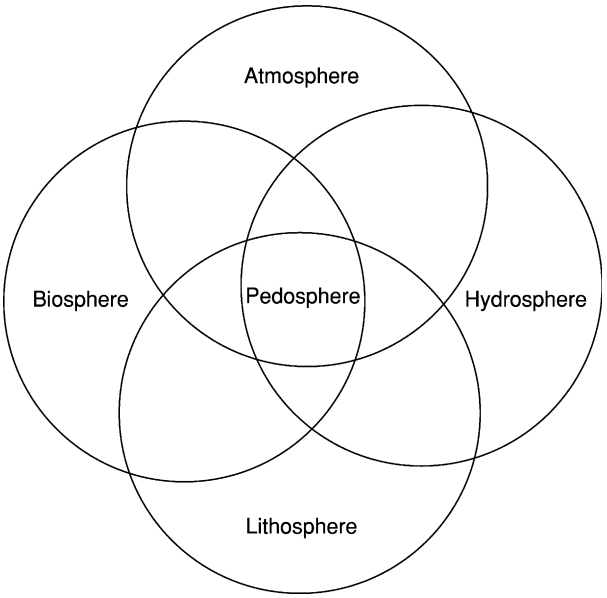


Figure 1 The pedosphere results from the interactions of associated spheres of influence. Reproduced with permission from the international journal cover of *Pedosphere*.

soil in history and significantly altered the manner in which soils were classified.

Although most soil interpretation schemes are also classifications, the more popular soil classification schemes of the twentieth century have been morphogenetic ones based on hypotheses and theories of soil development and pedological transformation of so-called parent materials into the entities called soils. Soil properties were described and genetic hypotheses proposed to explain the presence and spatial occurrence of soils. Names were given to different morphologies, e.g., Podzols and Chernozems, and they served as short-hand identifiers of the natural bodies in the environment.

Soil classification has seldom been static, consequently each system is an abstract of knowledge about soils at a given moment. Over time, technology has provided opportunities to make more precise measurements, enabling a major shift from qualitative definitions to quantitative ones. Field studies and soil survey have revealed the complexity and spatial intricacies of global pediment formation in unconsolidated materials. Such evidence greatly influenced soil genesis research and models of soils as landscapes. The significance of the time factor has become more apparent as details of paleopedology have been unraveled. The realities of polygenetic cycles of soil genesis have challenged prevailing theories of when and how soils form and develop. Not only have some soils developed in preconditioned geologic materials, but others have formed in previously developed soils, thus complicating the issue of inherited properties versus pedogenically formed ones.

Taxa limits in most classifications are direct consequences of theories, models, and experience; however, quantitative definitions of limits and even of central concepts of taxa are coldly factual, and departures of theory from facts become readily apparent. Some cherished concepts have received harsh treatment at the hands of precise definition.

As a source from which a thing proceeds, a principle serves as a foundation for the framework of knowledge about a population being classified. The following principles are fundamental statements upon which a hierarchical system of classification is developed and, as such, provide answers to commonly asked questions about classification. Eight principles are listed: the first three, purpose, domain, and identity, are concerned with the rationale of the system. The next four, differentiation, prioritization, diagnostics, and membership, are concerned with the organization of the information. The last principle, certainty, is concerned with the future viability of a system.

Principle of Purpose

Why do we want to organize soil knowledge? Each classification is dependent on the purpose, the rationale, of arranging current information in a systematic structure. With the new emphasis on genetic soil science, it was desirable to group soils according to their pathway or mode of formation. The groups could be related to each other based on the understanding of how soils obtained their properties, and by assigning names to these groups it helped recall many of their properties. Consequently the principle of purpose for most soil classification schemes is to show order in nature and to remember major features of soils. The purpose of the American Soil Taxonomy is to serve a soil survey program that is conducted to provide information about soils relative to their use and management. Considerable overlap with a purely genetic system is likely, but the selection of many properties is associated with a different purpose.

Principle of Domain

What do we want to include, or exclude, in our classification? In other words, what is the domain of interest? The realm of soils has many connotations, consequently it is necessary to specify what is to be included, and what is to be excluded, in a classification. At one time organic deposits were not considered to be soils and so were excluded. For some people soils must have a vegetative cover, for others they only need to be capable of supporting plant life.

The principle of domain requires that the universe of interest must be specified and, as such, will have to be defined. Instead of the collection of natural soil bodies, a domain might consist of terrestrial entities defined by subjective functional characteristics and associated environmental parameters. Thus, use-oriented groupings, in contrast to genetic ones, are not as concerned with order in nature; rather they generally show rankings of suitabilities of soils for specific uses. The choice of the domain depends on the purpose of the classification.

Principle of Identity

Who are the members of the domain of interest? How are they identified? A domain indicates the universe, or population, that is being considered; however, it does not define the members that will be the source of data for the classification. Where detailed soil survey has been important, the individuals are members of the taxa of the lowest category in a hierarchical system. Identity is a way of providing a name for a nondivisible component, or individual, that would not otherwise be recognizable. A major difference between some systems is the selection of the basic unit; in some systems, members of high-level taxa are the basic units, like the Russian Soil Type, whereas in others they are members of the lowest category, like the American Soil Series.

Examples of individuals of interest include: polypedon, pedon, profile, arbitrary body, soil landscape unit, and segment of a continuum. In use-oriented classifications, the members may be functional units or capability units rather than soil bodies themselves. Arbitrary soil individuals are meaningful reference units to compare classification schemes if their boundaries are independent of soil properties and class limits. Although precise definitions may be difficult and rather cumbersome, it is nearly impossible to proceed without resolving the principle of identity.

Identifying taxa at all categorical levels assists in visualizing relationships and patterns within the structure of the classification framework. A systematic nomenclature that indicates location within the system and also provides mnemonic links to important properties promotes understanding and use of the classification. The use of connotative elements has been successfully demonstrated in global systems, including Soil Taxonomy and the World Reference Base, and in several national schemes.

Principle of Differentiation

What kind of structure is needed? How should the domain be divided into groups, and how many groups are needed? How should the categories be defined? The sheer numbers of soils in the world, or even a small area of it, suggest that a hierarchical structure would be appropriate. There would be few classes in the higher categories and many more in the lower categories. It is possible to conceive of grouping individual soils together, and then combining those groups into more comprehensive groups, until the whole domain has been combined into only a few classes. It is more common to start with the domain and divide it into groups, then subdivide those groups into smaller, less-inclusive classes until the classes at the lowest level contain individuals with many features in common. It is sometimes forgotten that once a population is organized into a hierarchy it is only possible to employ the system from the top down, that is, by differentiating more and more specific classes. At the present time, the 'rules of engagement' for soil classification have only been devised, or clearly stated, for the processes of separation.

A hierarchical system of classification is a separation of a domain into successively more specific classes. It divides the domain into smaller groups of members that are also mutually exclusive classes. Consequently the differentiae (criteria) that separate classes focus on the limits or boundaries of the classes, and subordinate the central concepts of the classes. This is done for operational reasons; nevertheless the essence of a class is its central concept, and its description serves as the focus and image of the class.

The first separation of a domain into classes creates the highest category. The definition of this particular set of classes must be quite abstract to capture the intent of the classification. It may be genetic, geographic, interpretive, or functional, but it must apply to all members of the population (a rule sometimes forgotten). The definition of the category suggests possible indicators that are consistent with, and satisfy, the definition. Such marks or evidence are soil properties assumed to be closely related to the category definition. It is usual to expect classes of the highest category to reveal large areas when displayed on a global map indicating where soil-forming conditions are, or have been, similar enough to produce common tendencies of soil features. The number of classes of the highest category is subjective, that is, it depends on the designers of the system; for example, Soil Taxonomy has 12 and the World Reference Base has approximately 30.

The separation of each class into subclasses at the next-lower level is guided by the definition of this next-lower category. It must be less abstract than the definition of the category above. For example, difference of kind among soils is more abstract than

difference in degree. An important feature of a hierarchy is that successively lower-level classes accumulate the limits and descriptions that apply to the classes above which it is a subdivision. This means there is an accumulation of defining characteristics with each successive subdivision. In addition it is appropriate also to make statements about those properties and features that are correlated with the defining characteristics; thus each class has a set of defining characteristics and a set of associated, or accessory, characteristics which permit many useful statements to be made about the classes in the lower categories.

Two important aspects of a hierarchy are: (1) that all members of the domain being considered are included in each categorical level; and (2) that differentiating criteria accumulate in each lower set of classes.

Defining categories is a difficult challenge. The set of category definitions may refer to morphogenesis of soil profiles, geographic associations of kinds of soils, or even potentials for use, but they must be derived from the purpose intended for the classification. The set of category definitions establishes a priority among the concepts and theories of the paradigm (Table 1).

For example, the definition of the highest category of Soil Taxonomy, the Order, can be stated as 'soils whose properties result from, or reflect, major conditions of soils formation and evolution.' The processes cannot be measured, but soil morphology and some additional properties are assumed to correlate with the formation and presence of particular features. It is easy to produce confusion when the definitions are violated.

Due to the many areas of uncertainty in the measurement of properties and the relevance of different combinations of sets of properties, in addition to the lack of precision of concepts and relationships, it is obvious that classification is not a truth that can be discovered. There is no obvious way to determine the correct details of structure of a hierarchical system. The number of categories required to represent adequately a specific domain of millions of individuals is not known, and whether exclusive classes and those that overlap are equally relevant and proper is still unclear. Evaluating the adequacy of a system relies heavily on understanding the definitions of categories and their classes. This is where most 'pet theories and sacred cows' are led to slaughter.

Table 1 Definitions of categories of Soil Taxonomy

Category	*Definition*	*Abstraction*
Order	Soils whose properties result from major conditions of soil formation and evolution	Genetic pathways (soil evolution)
Suborder	Soils within an order whose additional properties or conditions are major controls of the current set of soil-forming processes	Major controls of current processes
Great group	Soils within a suborder whose additional properties constitute subordinate or additional controls, or reflect such controls, of the current set of soil-forming processes	Subordinate controls of current processes
Subgroup	Soils within a great group whose additional properties result from a blending or overlapping of sets of processes in space and time that cause one kind of soil to develop from, or toward, another kind of soil (intergrades), or whose conditions have not previously been recognized (extragrades)	Merging of processes or has specific conditions
Family	Soils within a subgroup whose additional properties characterize parent material and ambient conditions	Constraints on further change; affect use
Series	Soils within a family whose additional properties reflect relatively narrow ranges of soil-forming factors and localized processes	Smaller range of factors and processes

Principle of Prioritization

How are priorities set? How are categories, and classes within a category, sequenced? It is usually assumed, or clearly stated, that measurable soil properties and features are the source of data, and not the concepts or theories themselves. A category definition is commonly given in abstract terms rather than by naming soil properties; that is, the concepts of soils' development, or use, guide the structure of a classification to show how the groupings are related to each other. In morphogenetic-based classifications, the concepts of genesis are the bases for the definitions of categories; however, the indicators or evidence of those concepts are specific soil properties and features that are thought be to appropriately correlated with the conceptually based definitions. In Soil Taxonomy the suborder category (division of the classes of the order category) can be stated as 'soils whose additional properties result from or reflect major controls of the current soil-forming

processes.' As a concept, 'constraints on present day processes' is thought to be less abstract than 'pathways of development,' which is used to define the order category.

Soil-forming processes are important to an understanding of how and when soil features develop, but most processes cannot be adequately measured to confirm or deny such relationships. Which processes are more important? That is subjective and at the discretion of the designers of systems. For example, some conditions such as coldness and dryness are believed to restrict soil-forming processes. The sequence used in a key indicates the priority given to them. Because many soil schemes are morphogenetic, cause-and-effect relationships and concepts guide the selection of soil properties. Consequently, properties are surrogates for the concepts of genesis or soil behavior as the case may be.

The sequence of classes in the highest category establishes the precedent for the remainder of the framework and serves as the entry into the system. Most hierarchical arrangements rely on concepts of exclusion to reduce the volume of repetitive definitions. Such a technique also retains flexibility in what is being included in a given class. The list of classes at the order level of Soil Taxonomy demonstrates the possibilities:

1. Gelisols
2. Histosols
3. Spodosols
4. Andisols
5. Oxisols
6. Vertisols
7. Aridisols
8. Ultisols
9. Mollisols
10. Alfisols
11. Inceptisols
12. Entisols

Gelisols have permafrost near the surface during most years and, without further specification, it is implied that this group includes both organic and mineral soils. All soils that do not meet the permafrost definition are excluded from this group. Histosols are the remaining organic soils (those without permafrost) that may or may not be saturated with water most of the time. All other mineral soils are excluded from this group. This procedure is also evident in the structure of the classes of the first level in the World Reference Base used to correlate and compare major soil classification systems.

The sequence of classes within each category is also a priority scheme that promotes consistency and facilitates using the system as an identification key. The list of subdivisions of the class of Alfisols (a particular kind of texture-differentiated soil) at the suborder level (Table 2) indicates a preference for the major controls of the current processes.

The Alfisols are divided into Aqualfs, Cryalfs, Ustalfs, Xeralfs, and Udalfs. The presence of an aquic moisture regime is considered to be a stronger control of processes than temperature and is commonly used as a high priority in groups of soils that have well-expressed properties recognized at a higher categorical level. Subdivisions of the Ustalf suborder of Alfisols into great groups (Table 2) are a key to the preferences given to subordinate controls of current soil-forming processes: first is the presence of a duripan; then horizons dominated by plinthite; next, those with a natric horizon; followed by thin and thick kandic horizons; followed by a thick argillic or a petrocalcic horizon; then those having red colors associated with active iron oxides; and finally Ustalfs, which have a common argillic horizon.

Because of prioritization of the categories, and of prioritization of classes that are subdivisions of a class in a higher category, it is critical to evaluate alternatives and to maintain consistency in the development of a classification scheme. A key is a simplified exclusion technique to show relationships and minimize long, complicated definitions of each taxon in a system. The first class in a sequence that accepts a soil being considered is the correct placement. To place a soil in an Alfisol requires that all of the criteria of the first nine orders have been considered and rejected.

Table 2 Suborders of Alfisols and great groups of Ustalfs in Soil Taxonomy

	Description
Suborders	Major control of processes (affects biological activity)
Aqualfs	Aquic moisture regime
Cryalfs	Cryic temperature regime
Ustalfs	Ustic moisture regime
Xeralfs	Xeric moisture regime
Udalfs	Udic moisture regime
Great groups	Subordinate control of processes (affects water movement)
Durustalfs	Has a duripan (Si-cemented)
Plinthustalfs	Plinthite-dominated horizons
Natrustalfs	A natric horizon (columnar, sodium-affected)
Kandiustalfs	A thick kandic horizon (medium-textured, low CEC and ECEC)
Kanhaplustalfs	A thin kandic horizon
Paleustalfs	Petrocalcic (cemented) or thick argillic horizon
Rhodustalfs	Dark, dusky red argillic horizons, reflect active iron oxides
Haplustalfs	Has an argillic horizon

CEC, cation exchage capacity; ECEC, effective cation exchange capacity.

The principle of prioritization allows others to visualize the rationale applied throughout a system. It also provides a check on what kind of properties are appropriate at different categorical levels and promotes consistency in placing soils. Prioritization does not, however, eliminate differences of opinion about the correctness of the selection of properties nor the placement of particular soils.

Principle of Diagnostics

Must all soil properties be quantified? Why are methods of measurement needed? In most classifications, it is assumed or clearly stated that measurable soil properties and features are the data source and not the concepts or theories themselves. The definitions of soil properties and sets of properties to be used as criteria must be based on the methods of measurement. The main reason is objectivity. There are many choices, such as color by Munsell color charts or by spectrometers, and readings may vary according to conditions outside or inside a laboratory. General availability of methods and equipment often restrict widespread acceptance of the parameters selected. If the properties and methods of measurement are specified then other scientists can repeat the procedures and observe the same, or very similar, values. This removes part of the bias of the classification architects.

Any property selected for measurement is obviously subject to the understanding of relationships between the concepts used to explain the development of the property and the data obtained for that soil property. These connections give substance to the framework of the classification and provide meaning to the pattern of order that is displayed.

Often it is thought best to present some information as ratios or percentages, and others as weights or concentrations. Sets of properties may define horizon sets and be given specific names, such as a 'mollic epipedon,' or a 'kandic horizon.' Depths of occurrence of features may be diagnostic for some features. Quantification facilitates consistency, and diagnostics represent a short-hand way to aid recognition and placement of individuals.

The principle of diagnostics refers to quantified soil properties and features and not to the rationale of generalities or abstractions used to define categories. For example, the nomenclature of placement within Soil Taxonomy is built with syllables mostly from Latin and Greek. It is common in most systems to recognize diagnostic properties by mnemonic names to assist in recalling these important features (see Table 2 and Principle of Prioritization, above, for examples). The presence or absence of features may also be diagnostic insofar as the features are defined and are relevant to the differentiae of the classes.

The orders all end with a '-sol' syllable, indicating 'soil.' All suborder classes have two syllables, whereas great group classes have an additional one or two syllables prefixed to the suborder name. The linguistics are also mnemonic to aid in recalling important features of the soils: Alfisols are soils having clay-enriched subsoils (argillic horizons) and modest reserves of basic ions; Ustalfs (two syllables) are Alfisols having a wet–dry (ustic) moisture regime; Durustalfs (three syllables) are Ustalfs with a water-restrictive duripan in the subsoil.

Principle of Membership

What determines membership? What if properties overlap taxa boundaries? Regardless of how the individuals are defined, their membership into taxa is characterized by two concepts of classes. One is the central tendency that is like a mean, mode, or idealized abstract entity. The second is the boundary of a class with other classes. The limits of a class include properties with adjacent classes in the same category and with classes in other categories that share a common property. Although mutually exclusive groups of members are desirable and theoretically possible, the uncertainties of measuring properties and the relevance of precise limits indicate that actual membership acceptance may be more probabilistic than deterministic.

Keying out the placement of an individual relies on the limits of a class more than on the properties of the central concept. It is a process of elimination and, when the properties fall within the first defined class of the priority sequence, that is the proper placement. For example, to classify a soil as a Xeralf, first its placement as an Alfisol must be verified. Then within the Alfisols the criteria of Aqualfs, Cryalfs, and Ustalfs must be considered and rejected to arrive finally at the Xeralfs.

In systems that do not rely on mutually exclusive classes, such as fuzzy c-means classification, membership in a class can be expressed on a continuous scale from 0 to 1, thus partial class membership is common. Membership values are interpolated with kriging techniques, and thresholds are commonly set for class boundary detection. Currently guidelines and applications for using these techniques in morphogenetic soil classifications are not common.

Because of ambiguity in some taxa definitions, improved techniques for decision-making are desirable. The principle of membership is fraught with doubt because there are usually only 'yes' or 'no' answers, with no options for 'almost,' or 'not quite' and other

'near misses.' There appear to be opportunities for innovation.

Principle of Certainty

What happens in the future? Should a classification be viable and flexible? How can we minimize prejudicing the future? It has been found prudent to accept yet another principle to assure that a classification does not stagnate before its time. The principle of certainty suggests that change will occur, new facts and improved knowledge will be available, and continual evaluation for relevancy is a valid construct. Thus provision should be made to test each modification; check for inconsistencies and ambiguities; and decide whether the changes are merely busy work or whether they enhance the agreement with the purpose of the classification. Using the principle of certainty enables a classification to remain flexible and open to the changing world of science.

Summary

Soil classification is a means of organizing knowledge about soils. It is common to use hierarchical schemes because there are many soils and they have numerous physical, chemical, and biological properties. Protocols for designing such systems have been widely used and are summarized in eight principles. Three deal with the rationale of classification: purpose, domain, and identity of entities and their nomenclature. Another four principles guide the organization of the available information: differentiation, prioritization, diagnostics, and membership of individuals into the defined taxa. An additional principle of certainty is concerned with the future viability of a classification scheme.

List of Technical Nomenclature

Diagnostics Quantified soil features and properties used in definitions

Differentiation Separation according to defined criteria

Hierarchy Structured arrangement of information into multiple categories

Key Priority listing of subclasses belonging to the same class

Pedon A small, arbitrary volume of soil (1 m^2 in area) that represents a type of soil; its properties are compared with defined properties for classifying a soil

Polypedon A small segment of landscape dominated by a single type of soil; a landscape unit composed of similar pedons

Taxa Classes of organized information, which can be formed at different categorical levels

See also: **Classification Systems:** Australian; FAO; Russian, Evolution and Examples; USA

Further Reading

Cline MG (1949) Principles of soil classification. *Soil Science* 67: 81–91.

Cline MG (1961) The changing model of soil. *Soil Science Society of America Proceedings* 25(6): 442–446.

Cline MG (1963) Logic of the new system of soil classification. *Soil Science* 96: 17–22.

Cooperative Research Group, CST (2001) *Chinese Soil Taxonomy*. Beijing, China: Science Press.

Deckers JA, Nachtergaele FO, and Sapargaren OC (eds) (1998) *World Reference Base for Soil Resources. Introduction*, 1st edn. ISSS Working Group RB. Leuven, Belgium: Acco.

Food and Agriculture Organization of the United Nations, International Soil Reference and Information Centre, and International Society of Soil Science (1998) *World Reference Base for Soil Resources*. World Soil Resources report 84. Rome: FAO.

Rozanov BG (ed.) (1990) *Soil Classification*. Reports of the International Conference, Alma Ata. USSR State Committee for Environmental Protection. Moscow, Russia: Publication Centre for International Projects.

Shishov LL, Tonkonogov VD, Levedeva II, and Gerasimova MI (2001) *Russian Soil Classification System*. Moscow, Russia: VV Dokuchaev Soil Science Institute.

Smith GD (1963) Objectives and basic assumptions of the new soil classification system. *Soil Science* 96: 6–16.

Smith GD (1986) *The Guy Smith Interviews; Rationale for Concepts in Soil Taxonomy*. Soil Management Support Services, Technical Monograph No. 11. Washington, DC: USDA–SCS.

Soil Survey Staff (1975) *Soil Taxonomy: A Basic System of Soil Classification for Making and Interpreting Soil Surveys*, 1st edn. USDA–NRCS, Agriculture Handbook No. 436. Washington, DC: US Government Printing Office.

Soil Survey Staff (1999) *Soil Taxonomy: A Basic System of Soil Classification for Making and Interpreting Soil Surveys*, 2nd edn. Agricultural Handbook No. 436. Washington, DC: US Government Printing Office.

Stremski M (1975) *Ideas Underlying Soil Systematics*. (Translation of 1971 Polish edn.) Warsaw, Poland: Foreign Scientific Publications Department of the National Center for Scientific, Technical and Economic Information.

Sumner ME (ed.) (2000) *Handbook of Soil Science*, sect. E, Pedology. Washington, DC: CRC Press.

Wilding LP, Smeck NE, and Hall GF (eds) (1983) *Pedogenesis and Soil Taxonomy. I. Concepts and Interactions*. Amsterdam, the Netherlands: Elsevier Science.

CLASSIFICATION SYSTEMS

Contents
Australian
FAO
Russian, Background and Principles
Russian, Evolution and Examples
USA

Australian

R W Fitzpatrick, CSIRO Land and Water
Glen Osmond, SA, Australia

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Classifying soils (for a particular purpose) involves the ordering of soils into groups with similar properties and for potential end uses. The historical evolution of soil-classification systems currently used in Australia that have national, regional, and special-purpose applications can be traced as follows:

- General-purpose soil classifications that have been used in Australia since 1931 to communicate soil information and soil distributions at national scales;
- State and regional 'user-friendly' soil classifications designed both to assist with communication of soil information and to account for the occurrence of soils that impact on existing and future industry development and prosperity;
- Special-purpose and more-technical classification systems for single-purpose applications that involve using detailed soil-assessment criteria with recommendations for soil-management practices have been developed for a range of specific Australian industries.

The wide occurrence in Australia of relict erosional and depositional landscapes has given rise to soil materials not encountered in younger glaciated landscapes of the northern hemisphere. Some of these materials are: ferricretes, silcretes, and companion strongly weathered zones; the alluvial and eolian deposited materials on the plains of the Murray–Darling basin in eastern Australia; and the deeply weathered soils of the erosional terrain of low gradient in northern Queensland. A large proportion of soils in Australia have strong texture contrasts in the profile and their B horizons are susceptible to dispersion and erosion because of sodicity. More than 60% of the 20 million hectares of cropping soils in Australia are sodic, and dryland farming is mainly practiced on these soils. More than 80% of sodic soils in Australia have dense clay subsoils with high sodicity (exchangeable sodium percentage (ESP) more than 6%) and alkalinity (pH > 8.5). There is also high potential for dryland salinity to develop through shallow saline groundwater tables. Thus, compared with soils in other major agricultural regions in the world, many Australian soils present problems for plant growth due to one or more of the following:

- Poor capacity to store water;
- Salinity, sodicity, and alkalinity;
- Low nutrient availability;
- Susceptibility to wind and water erosion;
- Poor physical status of surface and subsoil horizons (e.g., hard, compact, or slowly permeable).

Contemporary objectives of soil and land resource survey programs in Australia now deal with practical issues of land and soil evaluation for a broad range of land-management issues required by landholders, private enterprises, researchers, and government agencies, which is aided by an improved system adopted in 1996. Less emphasis is given to soil genesis. The 1996 *Australian Soil Classification* officially replaced the 1968 *Handbook of Australian Soils* and *The Factual Key for the Recognition of Australian Soils* (1979). Various special-purpose classification systems have also been developed for utilitarian ends. Development of numerical methods for soil classification has also been investigated by researchers.

History of General-Purpose National Soil Classifications

The first classification of Australian soils was devised by Prescott in 1931 and was subsequently revised in 1947. This classification system incorporated concepts of the Russian system of soil classification, an approach that was followed broadly by Stephens. These systems laid emphasis on soil genesis. This was followed by the *Handbook of Australian Soils*

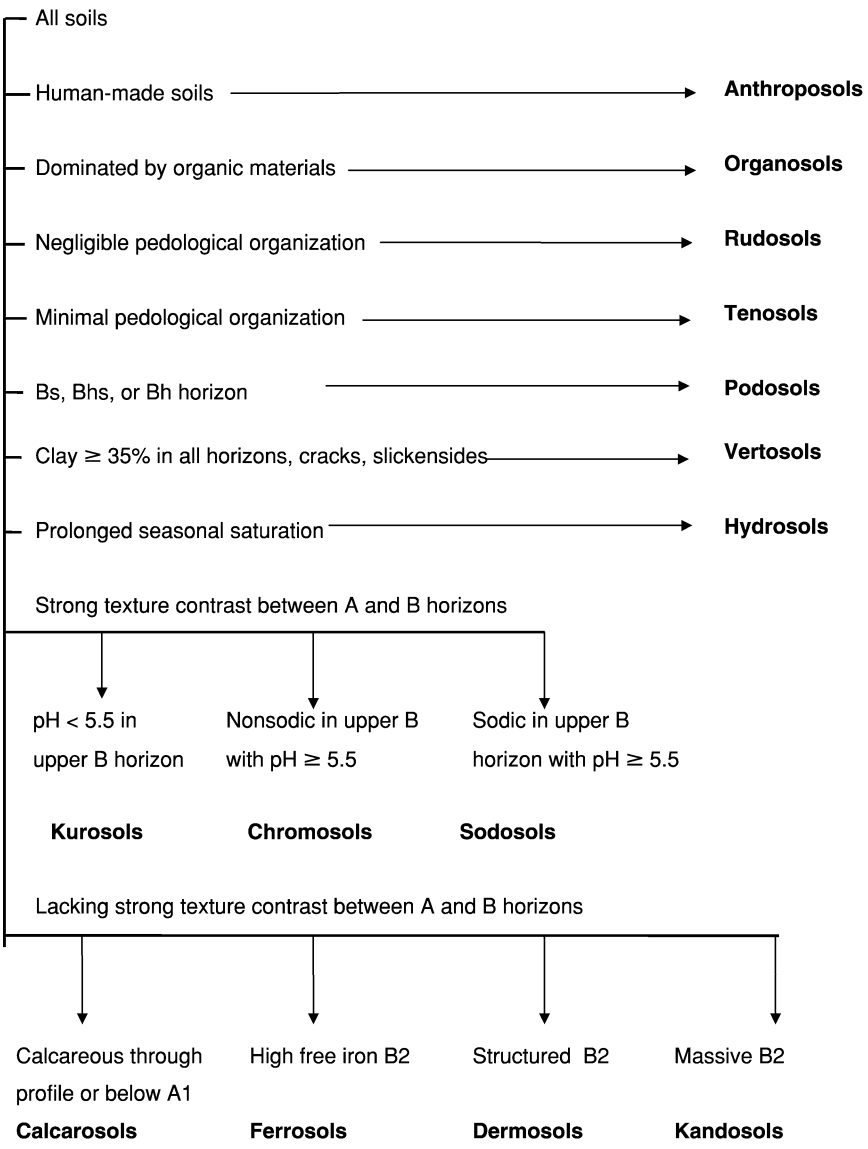


Figure 1 Structure of the Australian Soil Classification. Modified from Isbell RF (1996) *The Australian Soil Classification*. Copyright CSIRO 1996. Reprinted 1998. Published by CSIRO Publishing, Melbourne, Australia.

in 1968, which had many features in common with the 40 Great Groups of the American Great Group system. These genetic systems were criticized because they did not contain morphological information and did not emphasize profile criteria such as marked texture contrast features. Such features were used by Northcote in his *Factual Key for the Recognition of Australian Soils* to develop a more objective system. This classification was the basis for mapping soils of the Australian continent through the *Atlas of Australian Soils*. Forty-three Great Soil Groups were described and supported by representative profile data in *The Handbook of Australian Soils*. However, because separation between soil classes was not always defined clearly, many soils were inconsistently classified. For this reason Isbell devised a new national soil-classification system, which combined concepts from previous Australian systems and overseas classification schemes.

The New National Classification

The new national Australian soil-classification system is broad in its application and its hierarchical structure allows for unambiguous allocation of soils to particular classes. The system has 14 soil orders at the highest level that reflect important features of the soil continuum in Australia (Figure 1). The new system deals with agricultural and rangelands (arid zone) soils, tidal soils, and human-modified soils (Anthroposols, which are poorly understood). For example,

the occurrence of the following two extreme soil orders is reflected by aridity and strong weathering in a continent with an absence of modern glaciation:

- The Sodosol order is defined by a high exchangeable-sodium percentage in the subsoil and by soils that have also been affected by salt during formation, sometimes under past semiarid or arid conditions;
- The Kandosol and Ferrosol orders usually have a history of strong weathering and leaching.

In addition, at the lowest level (family level), properties such as soil depth, thickness, texture, and gravel content of the A horizon and the maximum texture of the B horizon can be used to predict incisively soil- and land-management responses. While the format of this system is suited for pedologists, it is considered inappropriate for teaching and broader public adoption. Consequently, a user-friendly, CD-based interactive key has been developed as a tool to communicate its use to people who lack expertise in soil classification.

Soil Classifications Used on Australian State and Regional Scales

In Australia, there is a strong trend at state and regional levels toward mapping and describing land based on specific soil and landscape attributes. The Australian soil-classification system has proved useful in developing a 'labeling system' for these soil–landscape mapping units, but it is considered too complex for land resource assessors, who lack pedological skills. Consequently, there remains a need for a form of soil classification which distinguishes key soils at the state or regional level. This is considered essential for public communication, and for some geographic information system (GIS)-based modeling applications.

The Basis for State and Regional Soil-Classification Systems

Soil profile classes were defined at various levels of generalization (e.g., series, family, great soil group or phase), depending on the information available, purpose of survey, and scale of mapping. A class of soil profiles is a 'group,' not necessarily contiguous, which may be grouped on morphological similarities and possibly some laboratory-determined properties. For example, Butler in 1980 supported the concept to "develop and use the soil classification that arises from the landscape itself," and this approach takes into account national classification systems in some instances, but also uses soil features and terms that are recognized and used locally. For example, this has been done in various ways in each state by combining a locality descriptor with the taxonomic name, e.g., Dorrigo Red Ferrosols. These systems are easy to use and understand, and have practical implications for managers of agricultural land, while also encapsulating the higher resolution of the national system.

State and Regional Examples

A soil-classification system was developed in 2001 that defined several 'soil groups and soil classes,' which reflected the morphological characteristics having the greatest impact on land use and soil management in the agriculture regions of South Australia – an area covering about 16 million hectares (Table 1). Soil groups and soil classes were defined using the presence of calcium carbonate (i.e., calcareous soils or depth to a calcrete horizon) at the highest level of classification, because these characteristics are a major component of South Australian soils (59%; Table 1) and are particularly important for agricultural management.

In contrast, many agricultural soils in Western Australia have been distinguished by the relative abundance of ironstone gravels, which again affect land use and agricultural management. These soil groups were developed to satisfy a need for a simple, standardized and easy-to-understand way of classifying the most common soils in Western Australia – an area covering approximately 250 million hectares. Soil groups can be allocated by nontechnical people

Table 1 Soil groups for South Australia's agricultural districts, covering 15.7 million hectares

Soil groups for South Australia	*Area (%)*
Calcareous soils	23.9
Shallow soils on calcrete	20.2
Gradational soils with highly calcareous lower subsoils	4.5
Hard, red-brown texture-contrast soils with highly calcareous lower subsoils	10.7
Cracking clay soils	1.4
Deep loamy texture-contrast soils with brown or dark subsoils	2.8
Sand over clay soils	9.0
Deep sands	14.4
Highly leached sands	0.8
Ironstone soils	1.5
Shallow to moderately deep acidic soils on rock	2.4
Shallow soils on rock	4.0
Deep uniform to gradational soils	1.5
Wet soils	2.8
Volcanic ash soils	0.1

Adapted from: PIRSA Land Information (2001) *Soils of South Australia's Agricultural Lands;* and *Atlas of Key Soil and Landscape Attributes – Agricultural Districts of South Australia* (CD-ROMs). Primary Industries and Resources South Australia.

who may have difficulty using the new national Australian Soil Classification. Similar systems have been devised in other states and regions of Australia to assist with the public communication of soils information and are suitable for mapping at broader state and regional scales. Each classification system takes into account the most important features of that state or region. For example, in the western region of New South Wales, the devised soil-classification system uses soil physical properties (inferred from soil texture) and microtopography indices (gilgai) as an indicator of infiltration rate, water-holding capacity, and presence of root-restricting layers.

Special-Purpose Soil-Classification Systems

Special-purpose, technical classification systems have been devised and designed to cover a wide spectrum of practical issues, and are required for finer scales of resolution. These include: matching soils for viticulture and forestry (hardwoods and softwoods), engineering applications (defining best options for installing optical fiber cables), rehabilitation of disturbed mine sites, saline soils, coastal acid sulfate soils (with direct links to policy and jurisdiction), soil tillage (abrasive soils), top-dressing soil for turf, urban planning for infrastructures and mineral exploration (Table 2). These special-purpose classification systems all involve using soil-assessment criteria and also provide recommendations for improving soil-management practices. These classification systems mainly rely on soil attributes but invariably also include relevant landscape features such as geology, terrain, vegetation, hydrology, or soil chemical features. These together provide a more complete understanding of how soils and their properties vary and behave within landscapes, and how this variability needs to be managed satisfactorily. Two contrasting case studies are presented to illustrate how special-purpose soil-classification systems are being developed and utilized by Australian industries.

Viticultural Soils

All the Australian general-purpose or national soil-classification systems were found to be inadequate and could not be adapted for identifying soil profiles and soil properties within vineyards by managers lacking pedological skills. Accordingly, the Australian viticulture industry called for the development of a user-friendly soil key that could be adopted by viticulturists to select and match grapevine rootstocks with appropriate Australian soils. Subsequently, an Australian Viticultural Soil Key was developed in 2002. The key essentially uses nontechnical terms to categorize soils based on attributes important for vine growth and also correlates these attributes with the new national soil classification, great soil groups, and several international soil-classification systems (Soil Taxonomy, World Soil Reference Base, and South African). The soil features used in the key are easily recognized and focus on the following soil diagnostic features: depth to mottling caused by waterlogging; consistency; color; structure; calcareousness in different restrictive layers; cracks; texture trends down profiles (e.g., texture contrast at A/B horizon boundary or duplex character) (Table 2). The key layout is bifurcating, being based on the presence or absence of particular soil-profile features. The key is used for correlating rootstock performance with soil properties and as a vehicle for delivering soil-specific land development and soil-management options to viticulturists.

Engineering Applications: Laying of Telecommunication Optical Fiber Cables

A special-purpose soil-classification system was developed to minimize soil damage to Australia's network of telecommunication optic fiber cables. Buried optical fiber cables can develop transmission faults by soil movements caused from soil shrink–swell properties or by corrosion from saline soil solutions. Such faults are very costly to repair and if avoided can save millions of dollars. Close liaison between soil scientists and engineers ensured that research investigations led to the development of a practical soil-classification system, comprising a 1–10 rating of soil shrink–swell risk. The rating is derived logically by using a series of questions and answers set out in a manual entitled *Soil Assessment Manual: A Practical Guide for Recognition of Soils and Climatic Features with Potential to Cause Faults in Optical Fiber Cables*. The manual, used in planning optical fiber routes, describes practical, surrogate methods to assist engineers estimate shrink–swell indices in soils using either published soil maps in office assessments (*Atlas of Australian Soils*) or by undertaking simple visual observations and chemical measurements of soil properties in the field (Table 2). Guided by this manual, telecommunication engineers have learnt how to integrate pedological, climatic, and soil chemical information. Firstly, shrink–swell and corrosive soils are avoided along optical fiber cable routes. Secondly, soils are matched to appropriate types of cable, thereby circumventing the need to lay expensive, heavy-duty cables along the entire route. Thirdly, problems affecting cables previously installed in troublesome soil types have been rectified.

Table 2 Examples of special-purpose soil-classification systems used in Australia

Industrial issue	*Soil (and other) attributes used in classification*
Viticultural soils: Identification of restrictive soil layers that limit effective root depth. (9 categories; 36 subcategories)	Depth to waterlogging (mottling), hard (nonrippable) or soft rock (rippable), rockiness and stoniness, soil consistence, color, texture and structure, calcareousness in different restrictive layers, cracks; three types of texture change with depth: contrast (duplex character), uniform (little change) or gradational (gradual change)[a]
Engineering: soil damage to telecommunication optic fiber cables. (10 soil shrink–swell risk classes/soil-assessment key to select cable type)	Soil shrink–swell and corrosion risk classes: rock type (geology), cracks, gilgai, soil color, structure (slickensides), texture, soil depth, dispersibility (sodicity), soil salinity. Soil assessment key: shrink–swell and corrosion risk, soil maps, vegetation, climate hydrology[b]
Minesite rehabilitation: soils on waste-rock and spoil dumps. (14 new subgroup classes of Spolic, Anthromorphic Anthroposols)	Rockiness and stoniness, rock type, soil color and mottling, structure, texture, depth, dispersibility (sodicity), soil salinity, pH, acid sulfate soils, impermeable crusts, watertables
Salinity hazard: soil salinity: Linked hydrology and soil chemical hazards. (29 categories or classes of primary, secondary, and transient salinity)	Halitic (sodium chloride dominant); gypsic (gypsum or calcium sulfate dominant); sulfidic (pyrite dominant); sulfuric (sulfuric acid dominant); and sodic (high exchangeable sodium on clay surfaces); hydrology (presence or absence of groundwater); water status (natural or primary as opposed to induced or secondary status)
Soil degradation within catchments: waterlogging, salinity, and sodicity. Linked to options for land use and remediation. (8 soil classes)	Rockiness and stoniness, soil consistence (ease of excavation), color and mottling, structure, texture, depth, dispersibility (sodicity), soil salinity (EC), pH, sulfidic material, topography, watertables, vegetation type[c]
Coastal acid sulfate soils: Identify actual (AASS) and potential acidification hazard (PASS). (2 soil classes)	AASS: soil pH <4, shells, yellow, jarositic horizons, water of pH <5.5, iron stains, scalds PASS: waterlogged, unripe muds, black to blue-gray color, pH >7, positive peroxide test, shells
Forestry soils for future development: site productivity for hardwood and softwood plantations in Tasmania. (4 site productivity classes)	Soil color, texture, depth of each soil layer to a minimum depth of 80 cm or to an impeding layer if shallower, native vegetation type and species, and rock type (geology), elevation, rainfall, soil drainage, tree-rooting conditions, and nutrient availability
Tillage: abrasive wear of cultivation equipment. Abrasive soils. (3 soil classes: highly and moderately abrasive; nonabrasive)	Highly: hard-setting, high bulk density, ploughpan, many rough-surfaced, magnetic ironstone gravels, high silt and sand content. Moderately: few ironstone gravels; moderate organic matter, calcareous gravel, silt, and sand contents. Nonabrasive: friable, no gravels; high clay, fine carbonate and organic matter contents
Topdressing soil for turf. (3 classes: suitable, restricted, and unsuitable)	Soil structure, soil coherence, soil mottling, macrostructure, ped strength, soil texture, gravel and sand content, acidity, salt content, soil color, cutans, other toxic features (sulfides, metals, etc.)
Urban planning. Capabilities and limitations. (5 primary classes and several subclasses)	Soil properties: depth, permeability, shrink–swell potential, Gilgai, bearing strength, drainage properties, erodibility, salinity, and pH Terrain properties: mass movement, watertables, subsidence, and flooding
Mineral exploration: soil sampling medium. (6 soil classes of soil material)	Saline seepages, acid sulfate soils, iron- and aluminum-rich precipitates, sulfidic material, mottles in sulfuric horizons; salinity, pH, geochemical analyses

Adapted from: Fitzpatrick RW, Powell B, McKenzie NJ *et al.* (2002) Demands on soil classification in Australia. In: Eswaran H, Rice T, Ahrens R, and Stewart BA (eds) *Soil Classification: A Global Desk Reference*, pp. 77–100. Boca Raton, FL: CRC Press.

[a]Adapted from Maschmedt DJ, Fitzpatrick RW, and Cass A (2002) *Key for Identifying Categories of Vineyard Soils in Australia.* CSIRO Land and Water, Technical Report No. 30/02.

[b]Adapted from Fitzpatrick RW, Slade PM, and Hazelton P (2001) Soil-related engineering problems: identification and remedial measures. In: Gostin VA (ed.) *Gondwana to Greenhouse: Australian Environmental Geoscience*, pp. 27–36. Geological Society of Australia Special Publication 21. Australia: GSA.

[c]Adapted from Fitzpatrick RW, Cox JW, Munday B, and Bourne J (2003) Development of soil-landscape and vegetation indicators for managing waterlogged and saline catchments. *Australian Journal of Experimental Agriculture* 43: 245–252.

Future Directions

Transfer of soil information in Australia to land managers, regional and urban planners, and researchers has become less 'pedocentric' and is now better tailored to suit client needs (end users). This has involved improvements in the communication of general-purpose or national classification systems and the development of several special-purpose schemes for particular industries. These soil-classification systems continue to be developed and become more dominant within Australia. The new national Australian soil-classification system also continues to evolve and have a role in developing landscape models with attached spatial soil attributes. Finally, the need to develop an integrating system or approach for describing soils in terms of landscape dynamics is being actively researched.

Further Reading

Butler BE (1980) *Soil Classification for Soil Survey*. Oxford, UK: Oxford University Press.

Charman PEV and Murphy BW (eds) (2002) *Soils: Their Properties and Management*. Soil Conservation Commission of New South Wales. Australia: Oxford University Press.

Fitzpatrick RW, Slade PM, and Hazelton P (2001) Soil-related engineering problems: identification and remedial measures. In: Gostin VA (ed.) *Gondwana to Greenhouse: Australian Environmental Geoscience*, pp. 27–36. Geological Society of Australia Special Publication 21. Australia: GSA.

Fitzpatrick RW, Powell B, McKenzie NJ, Maschmedt DJ, Schoknecht N, and Jacquier DW (2002) Demands on soil classification in Australia. In: Eswaran H, Rice T, Ahrens R, and Stewart BA (eds) *Soil Classification: A Global Desk Reference*, pp. 77–100. Boca Raton, FL: CRC Press.

Fitzpatrick RW, Cox JW, Munday B, and Bourne J (2003) Development of soil-landscape and vegetation indicators for managing waterlogged and saline catchments. *Australian Journal of Experimental Agriculture* 43: 245–252.

Isbell RF (1992) A brief history of national soil classification in Australia since the 1920s. *Australian Journal of Soil Science* 30: 825–842.

Isbell RF (1996) *The Australian Soil Classification*. Melbourne, Australia: CSIRO Publishing.

Jacquier DW, McKenzie NJ, Brown KL, Isbell RF, and Paine TA (2001) *The Australian Soil Classification. An Interactive Key*. Melbourne, Australia: CSIRO Publishing.

Maschmedt DJ, Fitzpatrick RW, and Cass A (2002) *Key for Identifying Categories of Vineyard Soils in Australia*. CSIRO Land and Water, Technical Report No. 30/02. Melbourne, Australia: CSIRO Publishing.

McBratney AB (1994) Allocation of new individuals to continuous soil classes. *Australian Journal of Soil Research* 32: 623–633.

Moore AW, Isbell RF, and Northcote KH (1983) Classification of Australian soils. In: *Soil – An Australian Viewpoint*, pp. 253–266. CSIRO Division of Soils. CSIRO: Melbourne, Australia: CSIRO Publishing/London, UK: Academic Press.

Northcote KH (1979) *A Factual Key for the Recognition of Australian Soils*, 4th edn. Adelaide, Australia: Rellim.

PIRSA Land Information (2001) *Soils of South Australia's Agricultural Lands;* and *Atlas of Key Soil and Landscape Attributes – Agricultural Districts of South Australia* (CD-ROMs). Primary Industries and Resources South Australia

Schoknecht NR (ed.) (2001) *Soil Groups of Western Australia*. Resource Management Technical Report 193. Australia: Agriculture Western Australia.

Stace HCT, Hubble GD, Brewer R, Northcote KH, Sleeman JR, Mulcahy MJ, and Hallsworth EG (1968) *A Handbook of Australian Soils*. Glenside, South Australia: Rellim.

FAO

F O Nachtergaele, FAO, Rome, Italy

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

The development of the Food and Agriculture Organization (FAO) soil classification took place in three distinct stages. The first one occurred between 1960 and 1981, when FAO was the lead agency for the preparation of the FAO–UNESCO *Soil Map of the World*, for which the legend was published in 1974. A second stage involved the refinement of the original legend into a more comprehensive system, which took place between 1981 and 1990 and resulted in the *Revised Legend of the Soil Map of the World*. During the last stage, the FAO revised legend was taken as a basis for the development of a universal soil nomenclature by the International Society of Soil Science (ISSS), which had been working since 1981 on such a worldwide reference system. This work, the *World Reference Base for Soil Resources*, was published in 1998 and was subsequently accepted as the preferred international soil nomenclature in a motion by the Sixteenth Congress of the ISSS, in Montpellier, France, in 1998.

The Development of the FAO *Legend* 1960–1981

The Seventh Congress of the ISSS recommended in 1960, in Madison, Wisconsin, USA, that a world soil map be prepared. The project started in 1961, when,

Table 1 Descriptive overview of diagnostic horizons[a]

Horizon	*Description*
Histic surface	Significant amounts of organic matter
Mollic surface	Well-structured, dark, thick nutrient and organic matter-rich topsoil
Umbric surface	Well-structured, dark, thick, base-desaturated topsoil, with moderate to high amounts of organic matter
Ochric surface	Surface horizon which is light-colored, or thin, or has a low organic matter content, or is massive and hard when dry
Argillic	Subsurface horizon showing a clear accumulation of clay
Natric	Subsurface horizon with more clay than the overlying horizon and having a high content of exchangeable sodium. Usually dense with a prismatic or columnar structure
Cambic	Young subsurface horizon showing evidence of alteration such as modified colors, removal of carbonates, or soil structure
Spodic	Dark-colored subsurface horizon with accumulation of amorphous substances composed of organic matter and aluminum with or without iron
Oxic	Strongly weathered; the clay fraction is dominated by low-activity clays
Calcic	Distinct calcium carbonate enrichment
Gypsic	Distinct calcium sulfate enrichment
Sulfuric	Extremely acid subsurface horizon in which sulfuric acid has formed through oxidation of sulfides
Albic	Bleached eluviation horizon with the color of uncoated soil material, usually overlying an accumulation horizon

[a]The actual definitions are fully quantitative and for those the following should be consulted: FAO/UNESCO (1974) *FAO-UNESCO Soil Map of the World*, vol. 1, *Legend*. Paris, France: UNESCO.

in a meeting at FAO, Rome, an advisory panel laid the basis for the preparation of an international legend, the organization of field correlation, and the selection of the scale of the map and its topographic base. The first draft of definitions of soil units was issued in 1968. The legend of the soil map was finalized in 1974. The first draft of the soil map of the world was presented to the Ninth Congress of the ISSS, in Adelaide, Australia, in 1968. The map itself was completed over a span of 20 years. The first sheets, those covering South America, were issued in 1971; the final map sheets for Europe appeared in 1981.

At the global level, the 1:5 million scale *FAO–UNESCO Soil Map of the World* is still, 20 years after his finalization, the only worldwide, consistent, harmonized soil inventory that is readily available in digital format and comes with a set of estimated soil properties for each mapping unit. The FAO legend for the soil map of the world had, as its main characteristic, that it was based to the maximum extent possible on factual information derived from actual soil surveys.

The legend was based on the diagnostic horizon approach adapted under the soil taxonomy classification system developed by the US Department of Agriculture (USDA) during the 1950s and 1960s. Similar measurable and observable diagnostic horizons to those in *Soil Taxonomy* were defined. The diagnostic horizons all have a set of quantitatively defined characteristics. An overview and brief description of the 13 diagnostic horizons are given in Table 1.

A number of soil characteristics that are used to separate soil units cannot be considered as horizons. These are considered as diagnostic features of horizons or of soils which, when used for soil classification purposes, are quantitatively defined. There were approximately 20 of these defined in the FAO legend. The most important ones are summarized in Table 2.

As the legend was the outcome of a vast international collaboration of soil scientists, it was by necessity a compromise. Certain historical soil names were retained to accommodate some national sensitivities. Examples of these at the highest level are Rendzinas, Solonetzes, Solonchaks, and Chernozems. Some of the names had a dubious scientific connotation (such as the Podzoluvisols in which no podzolization in the sense of accumulation of organometal complexes takes place), while others were defined nearly identically to those developed in *Soil Taxonomy*, for example the Vertisols.

In contrast with *Soil Taxonomy*, climatic characteristics were not retained in the FAO legend, with the exception of the Xerosols and Yermosols, which largely coincided with soils developed under an aridic moisture regime and classified as Aridisols in *Soil Taxonomy*.

The FAO legend of 1974 recognized 26 Great Soil Groups, subdivided in 106 Soil units, which were the lowest category recognized on the world soil map. In addition, 12 soil phases were recognized, three general texture classes (fine, medium, and coarse), and three general slope classes (with slopes less than 8%, 8–30%, and more than 30%). Most soil mapping units were in fact soil associations, the composition of which were indicated at the back of each paper map sheet. The dominant soil unit gave its name (and appropriate color) to the mapping unit, followed by a number unique for the associated soils and inclusions.

Table 2 Descriptive overview of the main diagnostic properties in the FAO legend[a]

Property	*Description*
Abrupt textural change	Considerable increase in clay content over a very short distance
Aridic moisture regime	Moisture is not stored in the soil for more than half of the year and less then 3 months consecutively within the year
Dominance of amorphous materials	Combination of low bulk densities and a high cation exchange capacity of the clay
Hydromorphic properties	Visible evidence of prolonged waterlogging either by groundwater or by a perched water table
Plinthite	Iron-rich, humus-poor mixture of kaolinitic clay with quartz which changes irreversibly to a hardpan on exposure to air after repeated wetting and drying
Permafrost	Horizon layer in which the temperature is perennially less than 0°C
Soft powdery lime	Significant amounts of translocated lime soft enough to be readily cut with a finger nail
Sulfidic materials	Waterlogged soil materials which contain sulfur in the form of sulfides. They occur in brackish water and, when the soil is drained, they oxidize to form sulfuric acid
Tonguing	Penetration with greater depth than width of an albic horizon into an argillic one
Vertic properties	Cracks that are more than 1 cm wide and go 50 cm deep in clayey Cambisols and Luvisols

[a]The actual definitions are fully quantitative and for those the following should be consulted: FAO–UNESCO (1974) *FAO/UNESCO Soil Map of the World*, vol. 1, *Legend*. Paris, France: UNESCO.

Texture (1, 2, 3) and slope symbols (a, b, c) were included in the mapping-unit symbol.

The 26 Great Soil Groups can be further aggregated based on the major soil-forming factors that determine their characteristics.

Set 1

Set 1 holds all soils that are characterized by the accumulation of organic matter. These soils, which receive their parent material from the top (from the vegetation), rather than from the bottom (parent rock), have a histic horizon and are called Histosols (from Greek *histos*, tissue).

Set 2

Set 2 contains all mineral soils whose formation and characteristics are determined by the nature of their (mineral) parent material. These include:

1. The Andosols (from Japanese, *An*, dark, and *do*, soil), derived from volcanic ash dominated by amorphous materials;
2. The sandy Arenosols (from Latin, *arena*, sand) of desert areas, beach ridges, inland dunes and areas where soils are derived from sandstone, all characterized by coarse-textured materials;
3. The swelling and shrinking clayey Vertisols (from Latin, *verto*, to turn) of backswamps and river basins in (sub)tropical areas with an expressed dry season and dominated by expanding 2:1 lattice clays.

Set 3

Set 3 accommodates mineral soils whose formation was markedly influenced by their topography and physiography. Particularly soils in low terrain positions suffer form recurrent floods and prolonged wetness. Others on steep slopes in particular are continuously subjected to erosion processes, and remain shallow. The first two are in low-lying positions; the next four, in elevated and eroding areas:

1. Young alluvial Fluvisols (from Latin *fluvius*, river), which show stratification of recent seasonal sedimentation;
2. Nonstratified Gleysols (from Russian *gley*, mucky soil mass) occur in waterlogged areas and show hydromorphic properties linked to the process of oxidation and reduction of iron;
3. The very shallow Lithosols (from Greek *lithos*, stone) occur in the most eroding positons of the landscape and are less than 10 cm thick;
4. The slightly deeper Rendzinas (from Polish *rzedic*, noise) develop on calcareous rocks and have a mollic surface horizon;
5. The Rankers (from Austrian, *Rank*, steep slope) are also shallow, develop on acid rocks, and are characterized by an umbric horizon;
6. The deeper Regosols (from Greek *regos*, blanket) occur in unconsolidated materials, which have only superficial profile development because of prevailing low temperatures, prolonged dryness, or erosion.

Set 4

Set 4 holds soils that are only moderately developed because of their limited pedogenetic age or because of rejuvenation of the soil material. These soils are called Cambisols (from Latin, *cambiare*, change) and may occur in all climates and all landscapes.

Set 5

Set 5 accommodates the typical red and yellow soils of wet tropical and subtropical regions where the high soil temperatures and the relatively abundant rainfall result in a more profound and faster weathering of the parent material and a rapid decay of organic matter. This reults in deep and pedogenetically mature soils grouped in three major Great Soil Groups:

1. The deeply weathered Ferralsols (from Latin, *ferrum* and *aluminium*) have an oxic horizon in which 1:1 lattice clays (kaolinite) dominate;
2. The deep Nitosols (from Latin, *nitidus*, shiny) occur often in more-basic rocks and are more fertile than other soils in the tropics. They are characterized by shiny, nut-like structural elements and high clay content;
3. The strongly leached Acrisols (from Latin, *acris*, very acid) are low in bases and have an expressed clay accumulation (argillic horizon) with depth.

Set 6

Set 6 groups soils of arid and semiarid regions. Redistribution of calcium carbonate, gypsum, and more soluble salts is common under dry climatic conditions:

1. The Solonchaks (from Russian, *sol*, salt) are characterized by their high content of soluble salts;
2. The Solonetz (from Russian, *sol*, salt) contain high levels of sodium on the clay complex and have a natric horizon;
3. The Xerosols (from Greek, *xeros*, dry) are characterized by an aridic (but not cold) climate and moderate contents of organic matter in the topsoil;
4. The Yermosols (from Spanish, *yermo*, desert) are also characterized by an aridic (but not cold) climate and very low contents of organic matter in the topsoil.

Set 7

Set 7 groups soils that are mainly found in steppe regions (pampa in South America, prairie in Northern America), characterized by a vegetation of ephemeral grasses and dry forests, where accumulation of organic matter and less-soluble salts dominates over leaching processes. This set holds four different Great Soil Groups, all characterized by a mollic horizon:

1. The Chernozems (from Russian, *chern*, black, and *zemlja*, earth), characterized by their great depth and dark surface horizon, rich in organic matter, with a pronounced accumulation of calcium carbonate with depth;
2. The Kastanozems (from Latin, *castaneo*, chestnut, and Russian, *zemlja*, earth) that generally occur in the driest part of the steppe zone, with lighter-colored topsoils and stronger accumulations of calcium carbonate or gypsum;
3. The Phaeozems (from Greek, *phaios*, dusky and Russian, *zemlja*, earth) that occur in the wettest part of the steppe and that show no signs of enrichment of calcium carbonate, but remain rich in bases;
4. The Greyzems (from Anglo-Saxon, *grey*, and Russian, *zemlja*, earth) that occur in the colder parts of the steppe and that show gray coatings of silica powder in the topsoil.

Set 8

Set 8 contains the brownish and grayish soils of humid temperate areas. The soils in this set show evidence of redistribution of clay and organic matter. The cool climate and the short pedogenetic history of these soils explain why some are still rich in bases in spite of an expressed leaching process (Luvisols). In more sandy material, eluviation and illuviation of metal–humus compounds result in the grayish colors of the eluvial horizons and the blackish colors where the compounds accumulate (Podzols).

1. Acid Podzols (from Russian, *pod*, under and *zola*, ash), with a bleached eluvial horizon over an accumulation horizon of organic matter with aluminum and iron (spodic horizon);
2. Planosols (from Latin, *planus*, flat, level), with a bleached topsoil overlying abruptly a dense, slowly permeable subsoil;
3. The base-rich Luvisols (from Latin, *luo*, to wash), with a distinct clay accumulation and argillic horizon with a high base saturation;
4. The base-poor Podzoluvisols (from Podzols and Luvisols), with a bleached eluviation horizon tonguing into a clay-enriched subsurface layer.

Although initially developed as a legend for a specific map, not as a soil classification system, the FAO *Legend* found a quick acceptance as an international soil correlation system and was used, for example, as a basis for national soil classifications and regional soil inventories as in the *Soil Map of the European Communities*.

The FAO *Revised Legend* 1981–1990

With the applications of the FAO *Legend* as a soil classification, numerous comments and suggestions

were received to improve the coherence of the system. In fact some combinations of diagnostic horizons could not be classified, while one soil unit identified in the key (gelic Planosols) did not occur in the *Soil Map of the World*. The revision effort undertaken in the 1980s finally resulted in the publication of the revised legend of the *FAO/UNESCO Soil Map of the World*. This revised legend was applied to the world soil resources map at 1:25 000 000 scale, accompanied by a report, and presented at the Fourteenth ISSS Congress in Kyoto in 1990 (Figure 1).

The revised legend retained many features of the original FAO legend, but definitions of diagnostic horizons and properties were much simplified. Where drastic simplifications were made, the diagnostic horizons were given another name; for example the argillic horizon was renamed 'argic,' the oxic horizon was renamed 'ferralic.'

The number of Great Soil Groups increased from 26 to 28. The Rankers and Rendzinas were grouped with the Leptosols (formerly called Lithosols), the 'aridic' Yermosols and Xerosols were not retained, while new Great Soil Groups of Calcisols and Gypsisols, characterized by respectively calcic and gypsic horizons, were created.

The Luvisols and Acrisols, both characterized by clay accumulation but with different base status, were further divided according to the activity of the clay fraction, resulting in four symmetric groups (Luvisols, high base saturation, high-activity clays; Acrisols, low base saturation, low-activity clays; Lixisols, high base saturation, low-activity clays; and Alisols, low base saturation, high-activity clays).

The revised legend created the Anthrosols at the highest level, grouping in this way soils strongly influenced by human activities. The number of soil units increased from 106 to 152. Texture and slope classes remained unchanged but were not represented on the map, for obvious reasons of scale.

A start was made with the development of a set of units at the third level. The latter were elaborated in a comprehensive set of third-level 'qualifiers' and presented at the Fifteenth ISSS Congress in Acapulco, in 1994.

As had happened with the original FAO *Legend*, the *Revised Legend* was used as a basis for the national map legends, e.g., in Botswana. The *Revised Legend* was also successfully used as a basis for university teaching of soil science.

The *World Reference Base for Soil Resources* 1981–1998

In a parallel development, FAO in cooperation with the ISSS had been involved in the development of an internationally acceptable soil classification system. In 1980, an international meeting of soil scientists, held in Sofia, launched a joint project, the International Reference Base for Soil Classification. This initiative was endorsed by the ISSS at its Twelfth Congress in New Delhi, in 1982. A Working Group RB (reference base) was set up in the framework of Commission V of the ISSS. In 1992, at a meeting of this working group, the recommendation was made that, rather than developing a fully new soil classification system, the working group should consider the FAO *Revised Legend* as a base and give it more scientific depth and coherence. This principle was accepted, and a first draft of the *World Reference Base* (WRB) appeared in 1994 that still showed large similarities with the FAO *Revised Legend*.

The WRB aims to satisfy at the same time two very different kind of users of soil information: first, the occasional interested user, who should be able to differentiate the main reference soil groups; and second, the professional soil scientist who needs a universal nomenclature for soils in a simple system that enables communication about the soils covered by a national soil classification.

The first version of the WRB was presented in 1998 at the 16th ISSS congress in Montpellier and was endorsed, in a historical motion, as the official soil correlation system of the International Union of Soil Sciences (IUSS). Major differences from the FAO *Revised Legend* include:

- The development of a two-tier system in which the first level describes and defines, in simple terms, 30 soil reference groups and classifies them with a key. The second level is composed of a list of uniquely defined qualifiers (121 of them) that can be attached to the soil reference groups. For international soil correlation purposes, a specific, preferred priority ranking is given for each soil reference group. This approach results in a very compact system that can be summarized in fewer 20 pages;
- The definition of a number of prefixes that permit the precise description where a soil phenomena occurs (epi-, endo-, bathy-) or how strongly it is expressed (hypo-, hyper-, petri-). This results in the ability to give a very precise characterization of the soil;
- The explicit emphasis put on morphological characterization of soils rather than on analytical procedures. Granted this aim is not always fully realized, but at least it is an attempt to permit soil classification to take place in the field rather than in the laboratory;

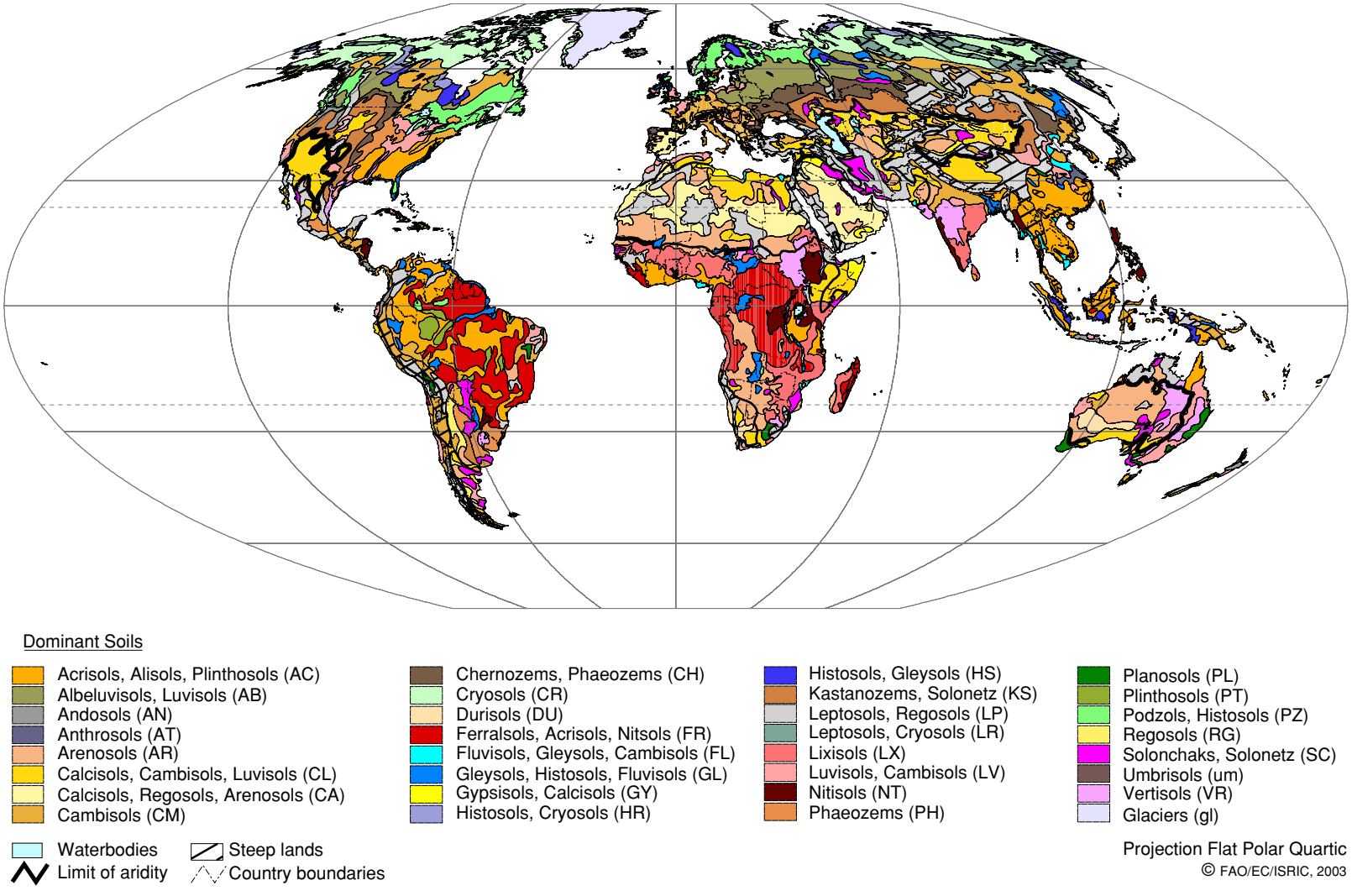


Figure 1 World soil resources map. Original scale 1:25 000 000. Flat polar quartic projection.

- The parallel publication of a WRB topsoil characterization system with approximately 70 different topsoil combinations recognized, permits classification in a much more elaborate way than is possible with six epipedons, the most important portion of the soil for most of its uses.

The WRB system as such has met with considerable success since its official endorsement by the IUSS in 1998. In the three years since its appearance, 3000 WRB books have been distributed, a website established, the system has been translated into 10 languages, and it has been adopted in several regions (European Union, West Africa, global soil and terrain database) and in individual countries (Italy, Lithuania, Georgia, South Africa), for national correlations or classification purposes, which has led to a revival of interest in soil classification and soil nomenclature.

Summary

The developments described above and the association of the FAO with the WRB work indicate that the latter has effectively replaced the FAO *Legend* with an FAO-backed and internationally endorsed world soil reference base for soil resources.

Further Reading

Brammer H, Antoine J, Kassam AH, and van Velthuizen HT (1988) *Land Resources Appraisal of Bangladesh for Agricultural Development*. Report 3. Land Resources Database, vol. II. Soil, Landform and Hydrological Database. Rome, Italy: UNDP/FAO.

Commission of the European Communities (1985) *Soil Map of the European Communities 1:1 000 000*. Luxembourg: Directorate General for Agriculture, Coordination of Agricultural Research.

Deckers J, Driessen P, Nachtergaele F, Spaargaren O, and Berding F (2003) Anticipated developments of the World Reference Base for Soil Resources. In: Eswaran H, Rice T, Ahrens R, and Stewart BA (eds) *Soil Classification: A Global Desk Reference*, pp. 245–256. CRC Press LLC.

Driessen P and Dudal R (1991) *The Major Soils of the World*. Lecture notes on their geography, formation, properties and use. Wageningen, the Netherlands: Wageningen University/Leuven.

Driessen P, Deckers J, Spaargaren O, and Nachtergaele F (2001) *Lecture Notes on the Major Soils of the World*. World Soil Resources Report 94. Rome, Italy: FAO.

Dudal R (1968) *Definitions of Soil Units for the Soil Map of the World*. World Soil Resources Report 30. Rome, Italy: FAO.

Dudal R (1980) Towards an international reference base for soil classification. *Bulletin of the International Society of Soil Science* 57: 19–20.

FAO (1993) *World Soil Resources*. An explanatory note on the FAO world soil resources map at 1:25 000 000 scale, 1991. Rev 1993. World Soil Resources Reports 66. Rome, Italy: FAO.

FAO (1995) *The Digitized Soil Map of the World Including Derived Soil Properties* (version 3.5). FAO Land and Water Digital Media Series 1. Rome, Italy: FAO.

FAO/UNESCO (1971–1981) *The FAO–UNESCO Soil Map of the World, Legend*. Paris, France: UNESCO.

FAO/UNESCO (1974) *FAO–UNESCO Soil Map of the World*, vol. 1, *Legend*. Paris, France: UNESCO.

FAO/UNESCO/ISRIC (1988) *FAO–UNESCO Soil Map of the World. Revised Legend*. World Soil Resources Reports 60. Rome, Italy: FAO.

Nachtergaele FO, Remmelzwaal A, Hof J *et al.* (1994) Guidelines for distinguishing soil subunits. In: Etchevers BJD (ed.) *Transactions of the 15th World Congress for Soil Science*, vol. 6a, pp. 818–833. Commission V: Symposia. Mexico: Instituto Nacional de Estadistica, Geografia e Informatica.

Nachtergaele FO, Spaargaren O, Deckers JA, and Ahrens R (2000) New developments in soil classification. World reference base for soil resources. *Geoderma* 96: 345–357.

Nachtergaele FO, Berding FR, and Deckers J (2001) *Pondering Hierarchical Soil Classification Systems*. Proceedings, Soil Classification 2001. Godollo, Hungary: University of Godollo.

Purnell MF, Nachtergaele FO, Spaargaren OC, and Hebel A (1994) Practical topsoil classification – FAO proposal. In: Etchevers BJD (ed.) *Transactions of the 15th World Congress for Soil Science*, vol. 6b. Commission V: Symposia. Mexico: Instituto Nacional de Estadistica, Geografia e Informatica.

Remmelzwaal A and Verbeek K (1990) *Revised General Soil Legend of Botswana*. AG/BOT/85/011. Field Document 32. Gabarone, Botswana: Ministry of Agriculture.

Soil Survey Staff (1960) *Soil Classification, a Comprehensive System,* 7th approximation. SCS, USDA. Washington, DC: US Government Printing Office.

Soil Survey staff (1999) *Soil Taxonomy. A Basic System of Soil Classification for Making and Interpreting Soil Surveys*, 2nd edn. Agricultural Handbook 436. Washington, DC: NCS, USDA.

Sombroek WG, Braun HMH, and van der Pouw BJA (1982) *Exploratory Soil Map and Agro-climatic Zone Map of Kenya*. Nairobi: Kenya Soil Survey.

Spaargaren O (ed.) (1994) *World Reference Base for Soil Resources. Draft*. Rome, Italy: ISSS/ISRIC/FAO.

Working Group RB (1998) *World Reference Base for Soil Resources. Introduction*. Leuven, Belgium: Acco.

Working Group RB (1998) *World Reference for Soil Resources. Atlas*. Leuven, Belgium: Acco.

Working Group RB (1998) *World Reference Base for Soil Resources*. (ISSS/ISRIC/FAO). World Soil Resources Reports 84. Rome, Italy: FAO.

Russian, Background and Principles

M Gerasimova, Moscow State University, Moscow, Russia

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

The first classification systems in Russia were elaborated at the end of the nineteenth century – in the period when pedology was born as a new branch of natural sciences. They were developed by the great founders of this science in accordance with their perception of soil as an important part, or 'mirror,' of the landscape, and this environmental approach exerted a strong influence on further activities in the area of soil classification. Thus, the majority of Russian classification systems have a 'factor-genetic' background, which presumes that the origin of soils, soil-forming processes, and agents of soil formation serve as criteria to group and subdivide soils at almost all taxonomic levels. This means that soil diagnostics is based more on the analysis of environmental conditions and concepts of soil-forming processes at present or in the past than on ranking soil properties, unlike many western systems.

Despite similar priorities in principles, diverse systems were proposed by a number of individual soil scientists and official bodies in Russia. In terms of conceptual background, the systems may be conventionally subdivided into the groups genetic, evolutionary-genetic, factor-genetic, and substantive-genetic. The most recent system is a member of the latter group, and its substantive principles were partially inspired by the international experience in the area of soil classification.

The hierarchy of soil classification systems in Russia comprised three groups of categories. The low-level categories, or soil systematic, were concerned with texture, intensity of manifestation of processes or properties, erosion phases, etc.

The central level was always regarded as basic: it was presented by soil types, or genetic soil types in present-day versions, which were practically the same in all systems, although their definitions had some differences depending on the classification principles.

The high-level categories, or above-type classes, were specified in accordance with the approach used, thus indicating the concepts of the author(s). Most of the early systems dealt with the upper categories and embraced either the world soils, or soils of Russia/USSR. Unlike American soil classification systems, there were no elementary soil units similar to soil series.

Principles, Categories, and Classes in Major Classification Systems Developed in Russia

Early Systems

The first system was that of V.V. Dokuchaev. It was published as a simple descriptive zonal scheme in 1886 after his famous expedition to Nizhniy Novgorod. The system was in a tabulated form with short comments. All soils, 14 groups, were subdivided into 'normal' soils, 'transitional' (boggy-meadow soils, rendzinas, and solonetzes), and 'abnormal' ones (swamped, alluvial, and aeolian). The system was slightly changed by N.M. Sibirtsev in 1893, who introduced the idea of zonality for grouping the same few soils known in those times. High-category classes were named 'zonal soils, intrazonal soils, incomplete soils, and surface geologic formations,' lower categories were provided for soil texture and parent rock properties. Based on his system, Sibirtsev gave the first description of soils of Russia in 1898. However, in the field of soil classification he always emphasized the priority of Dokuchaev, who issued another version in 1896 with allowance for the remarks of his pupil.

The most popular version (*Soil Classification of Professor Dokuchaev for the Northern Hemisphere*) was published in 1900 in *Pochvovedenie* journal. In this version, normal soils were regarded as synonymous with zonal ones. Seven zones with one or two dominant soils were characterized in terms of pedogenetic processes, parent rocks, climate, vegetation, fauna, and relief. The classes of transitional (between 'normal' and 'abnormal' soils), and abnormal soils were the same as in the first version. 'Surface geologic formations' were mentioned by Dokuchaev as climate-dependent neighbors of 'abnormal' soils gradually merging with them.

Parent rocks and climate as criteria of equal importance to specify classes at the upper taxonomic level were introduced by K.D. Glinka in his first system of 1915 (exo- and endo-dynamomorphic soils, respectively); later (in 1921) it was rearranged in accordance with the idea of soil-forming processes. Five main types of soil formation (lateritic, podzolic, steppe, solonetzic, boggy) corresponded to seven soil groups with soil types as their members.

The idea of pedogenetic processes was supported by S.S. Neustruev, who proposed (in 1924) a more detailed scheme derived from the concept of elementary soil processes. Soil formation was represented there by two classes: automorphic and hydromorphic, with a further subdivision of the former in accordance with the intensity of processes. The

criteria for the third level were a broad set of chemical characteristics (R_2O_3 behavior, humus properties, acidity/alkalinity) of soil constituents. The resulting 17 classes of major processes roughly corresponded to soil types; combinations of processes were provided as well. Information on landscapes was scarce and complementary.

There were other classification systems in the first decades of the last century using different approaches to combine the same few soils at the upper level: zonality (G.N. Vysotskii, 1906), vegetation (A.N. Sabanin, 1909), soil-forming agents (S.A. Zakharov, 1927).

NonOfficial Systems in the Second Half of the Twentieth Century

Many years later, I.P. Gerasimov, who was proud to be Neustruev's pupil, revived the idea of elementary pedogenetic processes – EPP (1973, 1975). Five broad groups of processes, such as 'pedomorphism of organic material, removal or accumulation of substances, pedoturbation', etc., were identified for soils of the USSR; major processes were further subdivided into elementary ones, whose number reached 25. Combinations of EPP produced soil types. These combinations, additionally differentiated in terms of process intensity in a given soil, were named 'process codes'; they produced 'profile codes' – sequences of genetic horizons, or profile formulas. For example, humus-illuvial podzols had the following combination of processes: hydration of primary and secondary minerals, formation of raw humus in an acid medium, podzolization and iron pan formation. This corresponded to the profile formula: At–**A2–Bih**–Cf, where major diagnostic horizons are shown in boldface.

It is worth emphasizing that this was the first system of soil classification to use profile formulas, which means that it combined genetic and substantive principles, namely, soil-forming processes and soil properties. It was simple but not exhaustive.

Soil chemical properties as criteria to group soils at high taxonomic levels were used in the system proposed by M.A. Glazovskaya (1966). Being a geochemist, she subdivded the world soils in accordance with acidity/alkalinity and redox conditions into 'geochemical associations' at the upper level, then into 36 'classes' by major pedogenetic processes. For example, the association of acid fulvic subaerial soils comprised classes of acid humus-enriched, acid metamorphic, acid metamorphic ferruginated, acid cryogenic, lateritic, and acid eluvial-illuvial soils (Podzols).

The criteria for the third category of 'families' presumed chemical properties of soil formation-derived products: humus forms, segregations, and genetic horizons. Families were subdivided into 'soil types' (fourth category), which were traditional in their essence, but were defined in terms of heat regimes within families. Thus, in the example with podzols, two families were specified: humus-illuvial podzols and podzolized iron-clay illuvial soils with different horizons and chemical compounds; the former family embraced types of dwarf podzols of the north, several 'normal' podzols of the temperate climate, tropical giant podzols, and podzols in the mountains.

Despite the emphasis placed on geochemical criteria by Glazovskaya, the system dealt more with geographic units than with soil bodies. Its elements were implemented in the world maps published in 1981 and 1998 in Russia.

V.A. Kovda presented an original evolutionary-genetic approach to soil grouping at the above-type level. The latest and most popular version of his system was used for the legend of the World Soil Map, 1975. Twelve classes of the upper category are ' soil-geochemical formations' differentiated by major trends of weathering, types of humus, clay minerals, acid–base properties, neoformations, and supplemented by characteristics of climate and vegetation. The second category, 'stadial groups of soils', corresponded to hypothetical evolution stages of pedogenesis from the submerged sites to excessively drained ones. The following stadial groups were proposed: hydroaccumulative → hydromorphic → mesohydromorphic → paleohydromorphic → proterohydromorphic → automorphic (including mountain soils) → neoautomorphic. The number of stadial groups, as well as that of soil-geochemical formations, varied in different versions of this system; moreover, climatic zones or belts were introduced in the early ones. Stadial groups of soils comprised soil types and subtypes in final version, whereas the definition of soil type had no importance.

Authors of all these systems were concerned with the upper taxonomic categories; they arranged soil types in different ways in accordance with their ideas and paid less attention to the type level. However, it should be emphasized that the notion, or 'central image' (integrity of properties, occurrence) of the soil type was the same in all systems, whereas its definition was different and depended on the conceptual background of the system. Lower categories, 'soil systematic,' were beyond the scope of these systems.

Official Systems

Unlike the systems of the previous group, official systems are 'complete' hierarchical structures with more emphasis put on soil systematics. The first official system for the purposes of soil mapping,

inventory, and research in the USSR was published as a draft in 1967, then in final version in 1977; its English translation appeared in 1986. Both versions are regarded as based on soil ecology and genesis. The new system is now under preparation; its first version appeared in 1997, and the next one is planned to replace the old official system of 1977. It is assessed as substantive-genetic in its ideology because of soil horizons and soil properties serving as criteria to categorize soils.

Classification and diagnostics of soils of the USSR – 1977 Conceptually, this classification system was based on the 'neo-Dokuchaev triad': factors → processes → properties, and is regarded as adhering to national traditions in genetic soil science and soil terminology. However, among the agents of soil formation, climate, or 'bioclimate', as this combination was termed in the book, had the priority in grouping soils.

The system started with the type category; any higher levels were absent. 'Genetic soil types' were differentiated with respect to bioclimatic conditions or, more precisely, to the place of soils in the system of geographic zones. At the same time, soil type was regarded as a direct result of the main pedogenetic process inherent to the zone; superposition of processes in adjacent zones was responsible for the formation of subzonal subtypes – intergrades. Thus, geographically, subtypes partially corresponded to subzones and/or to climatic facies, which were specified by soil temperature regimes, hence, there were subzonal and facial subtypes. The degree of soil hydromorphism was also taken into account at both taxonomic levels, and along with automorphic (zonal) types, semihydromorphic and hydromorphic types were recognized. This principle was strictly followed.

Genetic soil types were characterized in a descriptive manner, ABC horizons were mentioned in a general way, and some quantitative parameters with flexible and overlapping boundaries were used for both type and subtype definitions.

The third level (genus) was most severely criticized for its inconsistency. The criteria to specify genera were the following: parent rock and groundwater properties, paleopedological events, depth of effervescence, solonetzic features, faunal activity, and base saturation. Quantitative criteria were designed for the fourth level (species); they were elaborated in detail for many properties and many soils. Basically, the category was believed to characterize the degree of development of the major and superimposed pedogenetic processes. For example, chernozems were differentiated at the species level in accordance with the depth of humus horizon and the humus content in the topsoil as indicators of humus-accumulative process development. There are two more categories in the system: varieties for topsoil texture and phases for erosion phenomena. Thus, humus-illuvial podzol is referred to as: (1) the type of podzolic soils under taiga vegetation, having a percolative water regime . . . , acid reaction, profile differentiation . . . ; (2) the subtype of typical (lacking humus accumulation and gley features) podzolic soils; (3) the group of genera on light-textured materials, among which there is a special genus of humus-illuvial soils. At the species level, this podzol is characterized by the depth of podzolic horizon, and content of illuviated humus.

The system was highly appreciated by users in the former Soviet Union for its logical structure, especially at lower levels, and suitability for survey because of easy soil interpretation by analyzing the soil-forming agent patterns. It has been efficiently used in large-scale soil surveys performed for arable soils by regional institutes for land management ('Giprozem') for more than 20 years. The classification system is tightly bound to manifestations of pedogenesis in geographic zones and catenas within them, and allows forecasting soil development in changing environments. One more advantage for users is its adherence to the traditional nomenclature of Russian soils.

The main drawbacks of the classification system of 1977 were the following:

- Its factor-oriented diagnostics could not provide adequate and reliable identification of soils; moreover, it contained some virtual soils
- Many soils of Russia were not included in the system, e.g., most Siberian soils, and those of tundra areas
- The system was not open to newly discovered soils, since all the 'ecological niches' in the matrix with bioclimatic entries were filled
- Human transformed soils were weakly presented.

These and some other reasons promoted the development of the next version of the national classification system.

Russian soil classification system – 1997 The new classification system is based on the principles suggested by V.M. Fridland; it was developed in the Dokuchaev Soil Institute. The system presumes the priority of soil profile morphology as a result of pedogenetic processes. That is why soil properties and soil genesis are regarded as differentiating criteria. Environmental characteristics are no longer involved in the definitions of taxa.

Genetic soil types as central units of the system are preserved, and lower categories remain, although slightly changed in comparison with the former system. However, two above-type categories are introduced, and the system now comprises eight hierarchical levels. At the uppermost level, trunks, soils are grouped in accordance with the ratio of pedogenesis versus lithogenesis (i.e., parent rock formation); hence, there are postlithogenic, synlithogenic, and organogenic soils. Criteria for the next category are soil-forming processes in a broad meaning; recognized by the character of the profile composition. Thus, orders of texturally differentiated soils, humus-accumulative, alluvial, and volcanic soils may serve as examples. The classes of the third category are represented by genetic soil types. Although the essence of soil type central images remains close to the traditional one, the criteria for identifying soil types are different in this system. Genetic soil type is defined as a soil body with a certain sequence of diagnostic horizons; consequently, each soil type has its own formula of the profile. These formulas are complemented by indices corresponding to diagnostic features serving as criteria for the fourth category, subtypes.

Therefore, diagnostic horizons and diagnostic features as manifestations of soil properties acquire great importance, and many diagnostic horizons are specified and defined in terms of morphological, and to a lesser extent, chemical, parameters and/or regimes. The full horizon definition in the English version of the system is the following: "Diagnostic horizons are genetic horizons specified by the integrity of their properties, which derive of soil-forming processes. Among these properties the major ones are regarded as diagnostic; that is, they are inherent to the 'central image' of a horizon and serve for differentiating among horizons."

Soil genesis is regarded as a tool to control the separation of horizons and features along with the choice of differentiating criteria for them. Such an attitude to the contribution of genetic approaches is very close to that in western soil classification systems, although the horizons themselves are similar but not identical.

Diagnostic features are mostly modifications of genetic horizons produced by superposition of several mechanisms responsible for horizon development, specific characteristics of present-day soil regimes, or are inherited from the earlier pedogenesis, or correspond to peculiar human-produced features in horizons that were formerly natural. Consequently, many subtypes are intergrades among types.

There are four lower categories after the subtype one. They retained their names: genus, species, variety, phase. Few changes were introduced at the 'quantitative' species level (concerning gradations, boundaries), still less at the 'textural' variety. Criteria related to parent rock properties, as it was in the former system, were completely removed from the generic level; they were partially accounted for in the classes of types and orders, and referred to at the lowest level – phase. Hence, new criteria were introduced to discriminate among genera: base saturation and salinity parameters (Table 1).

Special attention is paid to human-modified soils, which are considered to be the result of soil evolution under the impact of human activities, basically farming. Several stages of such agrogenic evolution are distinguished; they are identified by the occurrence of human-modified horizons and features, and are reflected by the position of soils in the system. Thus, strongly modified soils form separate orders (second level); they may be correlated with Anthrosols of western systems. Moderately modified soils (with one new horizon and the remainder of the natural profile) are qualified as types within natural soil orders, whereas weakly modified ones (no changes in the sets of horizons) are included as subtypes into natural type classes (Figure 1). Goals and

Table 1 Taxonomic units and criteria for classes in the new Russian Soil Classification System

Category	*Criteria*
Trunk	Pedogenesis-to-lithogenesis (peat formation) ratio
Order	Similar major elements of the profile and similarity of pedogenic processes
Type	System of diagnostic horizons
Subtype	Modifications of diagnostic horizons (recorded by diagnostic features), intergrades between types
Genus	Features of cation exchange capacity or salinity
Species	Degree of diagnostic soil properties development (quantitative parameters)
Variety	Texture, amount of stones
Phase	Parent material, depth of the solum

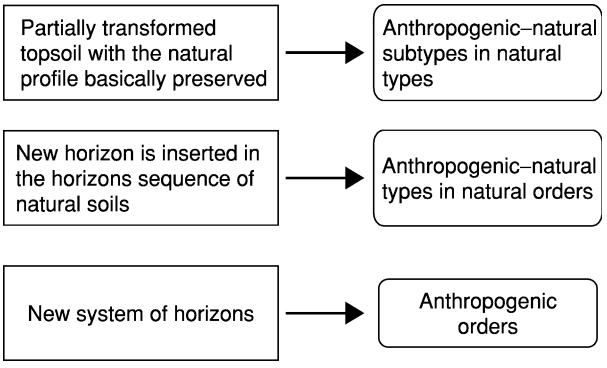


Figure 1 Major elements of profile composition and taxonomic level of human-modified soils.

Table 2 Criteria of genetic soil-type definition in diverse systems as presenting their main principles

	Above-type levels		
Classification system, author	*Number*	*Criteria*	*Genetic soil type*
Early systems	1	Zonality, processes	No strict definitions
Gerasimov			Set of processes
Glazovskaya	3	Geochemical	Regimes within geochemical classes
Kovda	2	Evolutionary	Factors + morphology + regimes
Official – 1977			Factors + processes + profile + regimes
New – 1997	2	Profile properties	Set of diagnostic horizons

character of human impacts on soil and the level of soil fertility are not taken into account.

The new classification retains most of the soil names that have been traditionally used in Russia, with two exceptions. First, all soil names indicating environmental conditions (e.g., forest soils, meadow soils, hydromorphic soils, etc.) are excluded. Second, new names were introduced for a few 'new' soils that were absent in the former system, as well as for human-modified soils. The prefix 'agro' was proposed for the agrogenic soil types: agro-chernozem, agro-solonetz, etc.

Conclusion

Soil classification in Russia started with environmental-genetic ideas, and the systems developed applied the genetic concepts to identify high- and medium-level taxa in different ways. Soils were grouped in accordance with the combinations of soil-forming factors (climate, time), pedogenetic processes, or soil properties controlled by theories on soil genesis. Most of these approaches were implemented at higher taxonomic levels. Since the early 1980s there has been a shift to substantive-genetic principles in classifying soils. After a series of publications in journals, the first version of the *Russian Soil Classification System* was published in Russian (1997, 2000) and in English (2001). The second version is awaited in July 2004. Preserving the basic principles, it presents a broader group of soils (mainly natural); hence, the list of genetic horizons and diagnostic features is considerably enlarged.

The definition of the basic level, soil types, depends on the principles applied (**Table 2**). In the *Russian Soil Classification System* of 1997 genetic soil types are defined by combinations of diagnostic horizons, which have much in common with the diagnostic horizons of western systems. In order to summarize the information on this system, an example of a full soil name is presented.

Example: Soil type – soddy-podzolic soil (Albeluvisol in the WRB system; soil near Moscow). Upper categories: postlithogenic trunk → order of texturally differentiated soils. Lower categories: typical subtype → unsaturated genus → nondeeply podzolic species → sandy loamy variety → on mantle loam phase.

See also: **Classification of Soils**; **Classification Systems:** Russian, Evolution and Examples

Further Reading

Classification and Diagnostics of the Soils of the USSR (1977) Moscow: Kolos. English translation (1986) Washington, DC: USDA and National Science Foundation.

Classification and Diagnostics of the Soils of the USSR (1986) Washington, DC: USDA and National Science Foundation.

Dokuchaev VV (1900) Soil classification of Professor Dokuchaev for the northern hemisphere. *Pochvovedenie*, appendix – special issue.

Russian Soil Classification System (1997) Tonkonogov VD, Lebedeva II, Shishov LL, and Gerasimova MI (eds) Moscow: Dokuchaev Soil Institute (English version edited by RW Arnold (2001)).

Russian, Evolution and Examples

D Konyushkov, V.V. Dokuchaev Soil Science Institute, Moscow, Russia

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

"Every soil classification is a philosophical system of pedology formulated in logical categories and symbols. It reflects the general credo and the current advancement of science. A series of classifications reflects the evolution of science, consecutive stages of its development" (Afanasiev JN (1927) The

classification problem in Russian soil science. In: *Uspekhi Pochvovedeniya* (*Advances in Soil Science*). Moscow Izd, Akad, Nauk SSSR, pp. 49–108 [in Russian]). Classification has to organize our knowledge about things so that "they are thought of in such groups, and those groups in such an order, as will best conduce to the remembrance and to the ascertainment of their laws ..." (Mill JS (1891) *A System of Logic*, 8th edn. New York: Harper). For practical reasons, it is important to subdivide the universe of soils into nonintersecting groups with objective diagnostics.

The revolutionary concept of soil as an independent natural body, the function of soil-forming agents (climate, relief, biota, parent material, and time) was formulated by V.V. Dokuchaev in the 1870s. It provided excellent grounds for developing diverse systems of soil classification. Soil type – a group of soils forming in similar conditions and thus having similar genesis and properties – is the central taxon of Russian soil classifications. The subdivision of soil types into lower taxonomic categories is generally based on substantive soil properties (the humus content, texture, and acidity) and has much in common in all classifications. The difference in approaches is seen at high levels of the taxonomic hierarchy.

In factor–genetic (factor–ecological) classifications, the highest taxonomic categories of soils are distinguished on the basis of the analysis of the factors of soil formation; usually, with emphasis on bioclimatic conditions. This approach was advanced by N.M. Sibirtsev (1895) and realized in full measure in the classification by Ye.N. Ivanova and N.N. Rozov (1967) and in the official Classification and Diagnostics of Soils of the USSR (1977). (All dates are given for the original publication: the dates for translated versions may differ.)

In substantive–genetic (profile–genetic) classifications, the morphology of soil profiles and essential substantive soil properties reflecting the character of pedogenesis and the influence of geodynamic processes (erosion and deposition of sediments) on the soils are used as classification criteria at all taxonomic levels. This approach was clearly stated in the early classifications by V.V. Dokuchaev (1879, 1886) and P.S. Kossovich (1910).

The idea of a polycomponent basic substantive–genetic classification of soils was advanced by I.A. Sokolov (1978, 1991) and V.M. Fridland (1979, 1982). They considered it as a system of three complementary components: (1) the substantive profile–genetic component that describes soil genesis as reflected in the morphology of soil profiles; (2) the lithologic–mineralogical–textural component that classifies the features of parent material inherited by the soil bodies; and (3) the regime component that considers data on soil water and temperature regimes. The new Classification of Russian Soils (1997) combines the substantive profile–genetic component with the lithologic–textural component.

It can be expected that the further development of soil classification in Russia will follow this line. Along with soil classification, the ecological classification of landscapes is being developed as an integral system that takes into account not only climatic and lithological characteristics but also the geomorphic position of soils, the character of vegetation, the presence of geochemical barriers, and other indices important from the viewpoint of predicting soil behavior and elaborating the strategy of sustainable soil management.

Pre-Dokuchaev Period

Agroproductive Soil Groups, Folk Soil Nomenclature, and First Scientific Soil Classifications

The first data on Russian soils were registered in Pistsovye knigi (descriptive books) that appeared in the fifteenth century and were aimed at the evaluation of lands for taxation purposes. They contained qualitative estimates of soil fertility (rich (kind), intermediate, poor, and utterly poor soils) and scarce information on soil properties (sandy, clayey, swampy, water-logged soils). In some areas, yield-based semiquantitative estimates of soil fertility were applied to croplands and hayfields.

Extensive geographical explorations of Russia in the eighteenth century enriched scientific literature in folk soil names (Chernozems, Podzols, Solonchaks, Meadow soils, Birch soils, etc.), first concepts of soil genesis ("Chernozem derives from rotting remains of plants and animals," M.V. Lomonosov), and the notion of nature zones with specific climatic, soil, and vegetation conditions. The notion and the nomenclature of soil types – the main taxa in Russian soil classifications – are derived from local folk names of soils. However, it should be stressed that the same local soil name (e.g., Chernozem or black earth) could mean different soils in different places.

The first cartographic inventories of Russian soils were undertaken in the middle of the nineteenth century. Soil maps were compiled on the basis of questionnaires sent to local administrations. As noted by Dokuchaev, these questionnaires were greatly influenced by the first scientific classification of soils suggested by the German agronomist A.D. Thaer (1810). In the map compiled by V.I. Chaslavskii (1875),

the list of soils included 32 names that can be roughly grouped into several categories: (1) textural (sandy, loamy, silty, clayey, pebbly soils); (2) local soil names (Podzols, Gray Northern soils, Chernozems, Solonchaks, Bogs, Tundra); (3) soil fertility (rich Chernozems, poor Chernozems); (4) petrographic features (Marl soils); and combinations of (1) and (2) (loamy Chernozems) and (2) and (4) (calcareous Chernozems).

The agrogeological approach to soil classification that was advanced by German scientists F.A. Fallou (1862) and F.P. Richthofen (1882, 1886) also influenced the development of soil classification in Russia. These scientists distinguished between the classes of (1) nontransferred (residual, eluvial, original) and (2) transferred (alluvial, eolian, colluvial) soils with further subdivision with respect to their petrographic composition and/or the genetic type of sediment. In 1886, P.A. Kostychev suggested an original classification system on the basis of petrographic (textural) and chemical (the content of mineral oxides and humic substances) properties of soils. Kostychev classified soils into five orders: (1) silty clayey; (2) loess; (3) sandy loamy; (4) sandy; and (5) gravelly (pebbly) soils. They were subdivided into 25 classes (classes of quartz or silicate; clayey; marl, calcareous, and dolomitic; and humus-rich soils (swampy, mucky, and chernozemic) were separated in the first four orders; quartz or silicate and calcareous soils were separated in orders 4 and 5).

Dokuchaev's Period

Genetic Approach to Soil Classification

A critical review of these initial classification systems was made by V.V. Dokuchaev (1886). He noted that they: (1) consider soil from different viewpoints (chemical, physical, geological, agricultural) instead of giving the definition of soil *per se* as an independent natural historical body; (2) do not pay attention to the fact that many soil properties are tightly linked to particular climatic, relief, geological, and vegetative conditions; and (3) give preference to some artificially selected soil properties instead of classifying soils on the basis of the total integrity of their essential properties. Dokuchaev suggested that natural scientific classification of soils as independent natural bodies should be based on their genesis studied in relation to soil-forming factors. He advanced a concept of the soil profile and an A-B-C system of horizon designation. His ideas predetermined the further development of soil classifications in Russia and abroad.

Initially (1879), Dokuchaev's scheme of soil classification looked as follows:

Division A. Normal soils lying in the place of their origin and unaffected by other dynamic processes:
- Class 1. Terrestrial vegetative soils
  - a. Northern gray soils
  - b. Chernozemic soils
  - c. Chestnut soils
  - d. Reddish solonchakous soils
- Class 2. Terrestrial swampy soils

Division B. Abnormal soils, i.e., soils that are strongly transformed by geodynamic processes:
- Class 3. Outwashed (eroded) soils
- Class 4. Inwashed (aggraded, sedimentary) soils

Dokuchaev stressed that the laws of soil formation should be established on the basis of studying normal soils, the main object of pedology. He foresaw that the number of soil types should increase in the future in parallel with the general progress of science, which, however, should not disturb the general structure of his classification. He considered color-based names of soil types as some labels reflecting the thickness of soils, their essential properties, and their relation to vegetation and climate. Finally, he believed that lower soil taxa should be distinguished with respect to the character of parent materials. It was the first substantive (based on soil properties) genetic classification of soils with special emphasis on the effect of proper pedogenic and geomorphic (geodynamic) processes on the character of soils. In 1886, Dokuchaev modified this classification and separated three groups of normal, transitional, and abnormal (sedimentary alluvial) soils by the manner of soil occurrence (soil bedding).

In the 1890s, Dokuchaev advanced the idea of natural horizontal and vertical zonality in the world of soils. It was a brilliant geographical concept. The magic of natural zonal regularities in the pedosphere was very attractive. Dokuchaev's disciple N.M. Sibirtsev (1895, 1900) introduced the zonal perception of soils into the classification scheme. At the highest level, he distinguished between the classes of: (1) zonal soils, or full-profile mature fine-earth soils forming continuous belts (zones) on the Earth's surface and developing under the impact of typical zonal soil-forming agents; (2) intrazonal soils, that appear within soil zones due to the impact of some specific local factors; and (3) azonal soils, or immature soils with not fully developed profile.

Zonal soils included the types of: (1) Lateritic (red earth); (2) Eolian – dust (loess); (3) Desert – steppe; (4) Chernozemic; (5) Gray forest; (6) Soddy – podzolic; and (7) Tundra soils. Intrazonal soils comprised

the types of: (1) Solonetzic; (2) Bog (peat); and (3) Mucky-calcareous (Rendzic) soils. Azonal soils were separated into subclasses of nonriparian soils with the types of skeletal and coarse (dune) soils and flood-plain soils with the type of alluvial soils. In essence, this classification puts the emphasis on soil-forming factors (mainly climate) rather than on soils proper. It can be referred to as the first factor–genetic classification, the main branch of soil classification systems in Russia.

The last classification scheme by Dokuchaev (1900) developed the scheme by Sibirtsev and gave a concise description of typical zonal soil-forming conditions (predominant rocks, climate, vegetation, fauna, and relief) and the character of zonal pedogenetic processes. In this scheme, the concepts of normal, transitional, and abnormal soils were considered as synonyms for the concepts of zonal, intrazonal, and azonal soils, respectively.

Post-Dokuchaev Period

Diversification of Classification Decisions

Some ambiguity in the factor–genetic approach to soil classification was quite obvious for the followers of Dokuchaev. There were several attempts to make the classification 'closer to the soil,' i.e., to base classification decisions on proper soil properties rather than on soil-forming factors.

The concept of the types of pedogenesis (major directions of soil formation that manifest themselves in essential inner soil properties and are governed by some similarity in the character of soil-forming factors) as the highest level of soil classification was advocated by K.D. Glinka, the leader of soil geographical studies in Russia. In his latest (1921–1927) works, Glinka stressed that the concepts of zonal, intrazonal, and azonal soils should be considered geographical rather than classification concepts. He noted that some isolated areas of Chernozem can occur in the forest zone and, in this case, should be placed into the category of intrazonal soils. Thus, the same type of soil (Chernozem) can be found in two different groups of soils (zonal and intrazonal) at a higher level of 'zonal' classification, which is inconsistent with the general principle of classification. Glinka distinguished between the five types of pedogenesis (Lateritic, Podzolic, Steppe, Bog, and Solonetzic) and subdivided them into 25 soil types. In fact, these types of pedogenesis represented ectodynamomorphic soils (pedogenesis governed by external (climatic) conditions) that had been earlier (in 1908) separated by Glinka from endodynamomorphic soils (pedogenesis governed by the specificity of parent rocks). In the system of 1908, Glinka divided ectodynamomorphic soils into six classes on the basis of the degree of climatic moistening.

G.N. Vysotskiy (1906) tried to improve Sibirtsev's classification by introducing into it the factors of relief (orography) and parent materials. He suggested that zonal soils should be defined as the soils of particular climatic zones developing on flat surfaces composed of loamy sediments. Intrazonal soils were subdivided into subclasses of: (1) soils that become zonal in neighboring climatic zones; (2) absolutely intrazonal soils (affected by groundwater or by the surface water stagnation in depressions); and (3) skeletal (sandy and calcareous) soils. Azonal immature soils were subdivided into subclasses of denuded (eroded) and aggraded (accumulative, sedimentary) soils. Vysotskiy distinguished between the soils of flat well-drained warm sites (e.g., soils of southern slopes that become zonal in a more arid climate), and poorly drained cold sites (e.g., soils of northern slopes that become zonal in a more humid climate). This was an orographic–climatic factor–genetic classification.

A.G. Sabanin (1909) suggested the classification system, with special emphasis on the role of vegetation as the main factor that dictates the difference between parent material and soil. He separated all soils into six divisions by the character of vegetation: (1) soils of evergreen deciduous forests; (2) soils of coniferous–deciduous forests; (3) soils of dark forests; (4) soils of meadow forests; (5) soils of wormwood-grass communities (semideserts); and (6) soils of swamps. At the second level (soil classes), characteristic chemical and physical properties of soils were taken into account. This was a vegetation-based factor–genetic classification.

S.A. Zakharov (1927) grouped soil types into suborders and orders by the character of predominant soil-forming factors. He distinguished between the orders of (1) climatogenic soils (zonal soils of Sibirtsev on plains); (2) orogenic (or oroclimatogenic) soils (zonal soils in the mountains); (3) hydrogenic soils (excessive moistening, soils of depressions); (4) halogenic and hydrohalogenic soils (salt-affected soils); (5) fluvigenic soils (alluvial soils); and (6) lithogenic soils (soil properties are governed by the character of soil-forming rocks). The zonal sequence of climatogenic soils included the types of Red-Earth soils, Sierozems, Chestnut soils, Chernozems, Forest-Steppe soils, Podzolic soils, and Tundra soils. The niches for the soils from orders (3–6) were found in every soil zone belonging to orders (1) and (2). Thus, every soil zone was characterized by a series of soils from different orders. These series were called analogous soil series. Zaharov's system was substantive–genetic (morphogenetic) at

lower levels and factor–genetic at higher levels of soil taxonomy.

The idea of analogous soil series in different climatic conditions was also developed by J.N. Afanasiev (1927) and D.G. Vilenskii (1925, 1945). Afanasiev's classification was based on the idea that analogous combinations of parent rocks, topography, and vegetation can occur in different climatic zones, which specifies the development of analogous series of soils in these zones. Climatic zones were differentiated in coordinates of temperature belts (cold, temperate cold, temperate warm, subtropical, and tropical) and sectors of climatic continentality (maritime, moderately continental, continental, and extremely continental). The resulting cells of the classification table were then subdivided into columns with different vegetation types (forest, forest-meadow (herbaceous), and grassy) and rows characterized by different kinds of parent material (silicate, calcareous, saline). In general, it was a climate-oriented factor–genetic classification.

Vilenskii developed several classification schemes using the idea of analogous series of soils. In 1925, he distinguished between thermogenic (soils of hot deserts), phytogenic, hydrogenic, halogenic, and compound soil divisions by the predominant factor of soil formation, within which analogous series of soils were distinguished on the basis of soil moistening and vegetation conditions. Later (1945), soil divisions were distinguished with respect to the type of weathering: lithogenic (weak weathering), H-siallitic, Ca-siallitic, Na-siallitic, Fe-siallitic, and Ferrallitic (for zonal soils); and hydrogenic, halogenic, fluvigenic, and orogenic divisions (for intrazonal and azonal soils). Analogous soil series within these divisions reflected the stages of soil development (from embryonic mineral soils to organomineral, organoaccumulative, and, finally, organoeluvial soils). Soil types were placed into this system of coordinates that can be referred to as an evolutionary factor–genetic soil classification.

S. S. Neustruev (1924) developed an original classification of soil-forming processes characteristic of the particular types of pedogenesis (as defined by Glinka) and soil types. The divisions of automorphic and hydromorphic (affected by groundwater) processes were separated at the highest level. Automorphic processes were subdivided into three classes by the intensity of transformation of the mineral mass of soils (strong, moderate, and weak). At the third level, the fate of residual products was considered. Hydromorphic processes were subdivided into subgroups with (1) the accumulation of precipitates from the groundwater and (2) the stagnation of water and the development of anaerobic reducing conditions. In essence, it was a process-oriented substantive–genetic classification of soils.

The importance of substantive characteristics in soil classification was clearly stated by P. S. Kossovich (1906, 1910). He considered pedogenesis as a combination of the processes of transformation of mineral (the source of bases) and organic (the source of acids) substances that is always accompanied by either the removal or accumulation of these substances. Special attention was paid to the input of substances into the soil with ground or surface waters. The classes of genetically independent (autonomous) and genetically dependent (geochemically conjugated, subordinate, heteronomous) soils were separated at the highest level. At the second level, soils were grouped with respect to: (1) the rate and character of mineral weathering; (2) the character of migration and redistribution of substances within the soil profile; (3) the rate of decomposition of organic substances; and (4) the character of accumulation of humic substances within the soil profile. Kossovich characterized seven types of autonomous pedogenesis (Desert, Semidesert, Chernozemic (or Steppe), Podzolic, Tundra, Peat-moss, and Lateritic) and four types of genetically dependent pedogenesis linked with corresponding autonomous soils. Kossovich showed that the genetically dependent soils accumulate the substances leached from the autonomous soils. His system can be referred to as the first process-oriented geochemical–genetic soil classification.

K.K. Gedroits (1924) tried to substantiate the separation of the main types of pedogenesis on the basis of data on the properties of the soil adsorption complex (SAC). At the highest level, he distinguished between the groups of (1) base-saturated soils (not containing H^+ ions in the SAC) and (2) unsaturated soils (with H^+ ions in the SAC). The major types of pedogenesis (Chernozemic, Solonhakous, Solonetzic, Solodic, Podzolic, and Lateritic) were characterized with respect to the composition of adsorbed cations, the state of the SAC, and the character or redistribution of pedogenetic products in the soil profile. Gedroits believed that the further development of physicochemical studies would make it possible to expand his system, that can be referred to as a physicochemical substantive–genetic classification.

B.B. Polynov (1933) tried to combine the substantive–genetic principle of soil classification with the idea of consecutive stages of soil evolution. At the highest level, he distinguished between the soils of the eluvial (automorphic, zonal) and the lacustrine-bog-solonchakous soil series. The first series was subdivided into the groups of alkaline carbonate-containing soils and acid soils. Then, these groups were subdivided into eight types of pedogenesis that

were believed to be evolutionary-linked (primitive alkaline, prechernozemic, and chernozemic types were separated in the alkaline pedogenesis).

Official Soil Classification and Unofficial (Authors') Systems

In the 1930s, the needs of soil mapping required the development of more detailed soil classification systems encompassing the real geographical diversity of soils in the former Soviet Union. The factor–genetic approach was laid in the basis of classification decisions. The prototype of the new system was suggested by I.P. Gerasimov, A.A. Zavalishin, and Ye.N. Ivanova in 1939. The basic unit of soil classification – soil type – was defined as "a group of soils developing in similar conditions and characterized by common origin and common processes of the transformation and migration of substances." The system of 1939 listed 10 zonal soil types. In the 1950s, this list was expanded to over 100 soil types. In 1956, Ye.N. Ivanova and N.N. Rozov suggested the classification of world soils on the basis of the ecologic–genetic approach.

In 1958, the definition of soil taxa at the below-type level was officially adopted. The work on compiling the unified systematic list of soils of the Soviet Union was entrusted to the Dokuchaev Soil Science Institute and headed by Professor Ye.N. Ivanova. In 1966, after several revisions, this list was officially adopted by the Ministry of Agriculture of the USSR for use in large-scale soil surveys and it included 110 soil types. The principles of their grouping at higher taxonomic levels are shown in Table 1. The position of soil types in the coordinate system of bioclimatic parameters and the degree of soil hydromorphism were very rigid; some deviations (e.g., the appearance of soluble salts in some soils of the humid climate) were 'shifted down' to the soil genus level. It was a very consistent and fully developed system. Virtually all soil niches could be characterized.

Further work on the development of soil diagnostics resulted in publication of the only official Classification and Diagnostics of Soils of the USSR (1977). Soil types were characterized by particular sequences of genetic horizons. The diagnostics of genetic horizons and a system of indices for their designation were suggested. The classification of 1977 considered only the soils of the agricultural area; soils of vast northern and Siberian regions with permafrost were not included.

In parallel with the development of the official classification on the basis of the factor–genetic approach, several prominent Russian pedologists suggested their own original classification schemes of world soils. The most influential classification decisions were suggested by M.A. Glazovskaya (1966, 1972) and by V.A. Kovda, Ye.V. Lobova, and B.G. Rozanov (1967).

Glazovskaya renewed the ideas of Kossovich and Polynov and suggested the substantive geochemical–genetic system of soil classification at the above-type level (Table 2). In the 1990s, Glazovskaya developed a series of applied classifications of soils by their tolerance toward pollutants (heavy metals, acid rain). The soil classification of 1972 proved to be very useful for this purpose, as it was based on the parameters affecting the migration of substances in the soil profile (*Eh*—pH conditions, the character of organic matter and humic substances, mineralogy of the clay fraction, and the presence of geochemical barriers in the soil profile).

Kovda, Lobova, and Rozanov combined the evolutionary and geochemical approaches to soil grouping. The authors tried to arrange a system of soils that would "reflect the history of the balance of substances in the course of pedogenesis taking place under the particular energy potential that depends on the radiation balance and the humidity factor." By the energy potential, soils were grouped into 14 energy orders: humid tropical, arid–humid tropical, arid tropical, humid subtropical, and so on. Provisional evolutionary stages of soil formation were separated on the basis of the hypothesis that soil evolution on great plains proceeds from hydroaccumulative (subaqual) soils (e.g., mangrove soils) to hydromorphic (hydrobioaccumulative), mesohydromorphic, paleohydromorphic (with relic hydromorphic features), protohydromorphic (with weak features of paleohydromorphism), primitive automorphic (bioaccumulative), automorphic, paleoautomorphic, and mountainous (erosional) soils. The authors also considered soil–geochemical formations as the highest taxon of their system. Initially, there were eight formations (acid allitic, acid allitic–kaolinitic, acid kaolinitic, acid siallitic, neutral and slightly alkaline siallitic, neutral and slightly alkaline montmorillonitic, alkaline and saline, and volcanic soils). Later, in the legend to the Soil Map of the World (1:10 M scale), soil formations were considered as particular manifestations of the energy potential of weathering and pedogenesis.

The systems by Glazovskaya and Kovda (who was the main initiator of the work on the evolutionary–geochemical soil classification) were designed to classify soils at high taxonomic levels; they could not be applied to large-scale soil mapping.

The Problem of Basic Substantive–Genetic Soil Classification and the new Russian Classification

The official soil classification of 1977 did not consider soils lying beyond the main agricultural area of

Table 1 Principles of soil types grouping in the factor–genetic soil classification by Rozov and Ivanova (fragment, abridged)

		Biophysicochemical soil orders				
Ecologic–genetic (bioclimatic) soil classes	*Genetic soil orders*[a]	*Unsaturated fulvatic, leached from soluble salts and carbonates*	*Slightly unsaturated fulvate-humatic, leached from soluble salts*	*Saturated Ca-humatic, leached from soluble salts in the upper horizons*	*Ca-humate-fulvatic, with soluble salts in the middle part of the profile*	*Na-humate-fulvatic; soluble salts can occur at any depth*
Groups of soil classes: Arctic and tundra, Boreal, Subboreal, Subtropical						
	A	Taiga frozen (pergelic), soddy taiga pergelic	Palevye Taiga pergelic, sod-calcareous pergelic, meadow-forest pergelic			
	AH	Alluvial soddy pergelic				
Permafrost-affected taiga soils, $\Sigma T^{\circ} > 10^{\circ}C$ 600–800, humid climate, siallitic weathering	SH	Taiga swampy pergelic	Palevye swampy pergelic, muck-calcareous pergelic, taiga solod pergelic	Meadow-chernozemic pergelic		
	H	Bog (high-moor peat) pergelic	Meadow pergelic, meadow-bog pergelic			Meadow solonetzes and solonchaks pergelic
		Low-moor peat pergelic	Low-moor peat pergelic			

[a]Genetic soil orders (series) are distinguished with respect to the character of soil water supply and the degree of soil hydromorphism: A, automorphic soils (atmospheric water supply; deep groundwater); AH, alluvial hydromorphic (short-term inundation by flood water with simultaneous deposition of alluvium); SH, semihydromorphic soils (periodical waterlogging with surface and/or groundwater; groundwater 3–6 m); H, hydromorphic soils (permanent waterlogging; ground water <3 m).

Table 2 High-level taxonomic categories of soil classification

Rank	*Taxonomic category of soils*	*Criteria for separation*
I	Geochemical associations of soils	Soil reaction (acid, acid/alkaline, neutral/alkaline) and redox conditions (oxidative, oxidative/reducing, reducing) as related to the soil water regime in the upper/lower horizons
II	Soil generations (classes) (within geochemical associations of soils)	Manifestation of major soil processes: (1) organic matter accumulation; (2) biochemical weathering and mineral neoformation; (3) translocation of the products of pedogenesis in the soil profile (soil differentiation); (4) gleyzation; and (5) hydrogenic accumulation of substances on oxidative and evaporative barriers
III	Soil families (within soil generations)	Qualitative composition of pedogenetic products: (1) humus and organic matter (fulvate, humate, Ca-humate, muck, peat, etc.); (2) secondary minerals (carbonates, gypsum, soluble salts, allophanes, sesquioxides, etc.); (3) character of eluvial and illuvial horizons (gley-eluvial, Ca-humus argillic, etc.); (4) character of the horizons of ancient or recent hydrogenic accumulation (hydrogenic calcareous, hydrogenic saline, etc.)
IV	Soil types (within soil generations)	The degree of development of characteristic family features as conditioned by the soil temperature regime and the intensity of biological turnover (analogous soil types in different climatic zones, e.g., shallow humus-illuvial podzols of humid boreal climate and deep (tropical) humus-illuvial podzols of humid tropical and subtropical climate)

the former Soviet Union. The principles of this classification were not formulated. Soil taxa above the type level were absent. The dominance of climatic rather than proper soil characteristics was evident, especially at the level of facial (provincial) subtypes of soils. Anthropogenic modifications of soils were not properly considered. To improve this situation, the Classification Commission was organized by the All-Union Soil Science Society.

Preliminary schemes of a polycomponent basic substantive–genetic classification of soils were advanced by I.A. Sokolov (1978, 1991) and V.M. Fridland (1979, 1982). The idea of this classification is grounded in the fact that soil bodies combine stable features acquired in the course of pedogenesis (pedogenic) and inherited from the parent material (lithogenic), as well as dynamic soil characteristics reflecting the current state of soils (first of all, temperature and water regimes). Though all these characteristics are interrelated, the regularities governing them have a different nature, and there is no strict correspondence between stable lithogenic and pedogenic soil properties and dynamic soil regimes. It can be expedient to classify them in a system of three complementary classification components. The substantive profile–genetic component describes soil genesis as reflected in the morphology of soil profiles. The lithologic–mineralogical–textural component classifies the features of parent material inherited by the soil bodies. The regime component considers data on soil water and temperature regimes. The new Classification of Russian Soils (1997, 2001) compiled by L.L. Shishov, V.D. Tonkonogov, and I.I. Lebedeva (in collaboration with M.I. Gerasimova) is a realization of the substantive profile–genetic component combined with the lithologic–textural component. This classification differs significantly from the classification of 1977.

First, soils are separated on the basis of the soil profile morphology, i.e., the system of soil horizons that reflects in its properties the genesis of soils. Factors of soil formation are not taken into account. Soil names, which were indicative of bioclimatic conditions (e.g., desert soils, taiga soils, meadow soils, and so on), are replaced by the new names consistent with the logic of the profile–genetic system.

Second, anthropogenically modified soils are separated from natural soils at a high taxonomic level. Provisional principles and criteria for classifying human-created (technogenic) surface nonsoil formations are also suggested.

Third, the new system of diagnostic genetic horizons and genetic features of natural and human-transformed soils is suggested; a special section of the classification considers the correlation of major genetic horizons with diagnostic horizons of the World Reference Base (1998) system.

Fourth, the taxonomic structure of the classification is modified. At high levels, it came closer to Dokuchaev's classification of 1886. The highest taxon of the new classification is soil trunk reflecting the division of soils by the relationship between proper pedogenic and sedimentation (or organic matter accumulation) processes. The orders of Postlithogenic (soil develops from the already existing parent material, and pedogenesis is not disturbed by the deposition of new

portions of sediments), Synlithogenic (pedogenesis is synchronous to sediment deposition, as in alluvial and volcanic soils), and Organogenic (organic) soils are distinguished. Soil divisions – the second taxonomic level – comprise the soils characterized by the similar trend of pedogenesis and having the same major diagnostic horizon (as a rule, subsurface horizons are taken into account). Considerable changes are made at the level of soil types that are distinguished by the similarity in the system of the main diagnostic horizons, i.e., by the similar horizonation of the soil profile. Thus, the previously single type of Chernozem is differentiated into two types (with and without textural differentiation of the profile). Some new soil types (Gleyzems, Cryozems, Dark Vertic soils) have been added. The types of mountainous soils having the same genetic horizons as the corresponding soils of plains are united with the latter. Overall, 181 types of natural and agronatural soils are distinguished.

Soil types are characterized by the similarity of the system of the main diagnostic horizons, except for the character of the parent material. Soil subtypes are distinguished as the soils having qualitative modifications of the main diagnostic horizons; as a rule, they represent intergrades between soil types. Quantitative criteria are used as soil differentiate at lower taxonomic levels (similar to the classification of 1997). Soil genera are separated by the peculiarities of their exchange complex and the chemistry of salinization. Soil species are distinguished on the basis of the degree of development of soil features taken into account at the type, subtype, or genera levels. Soil varieties take into account soil texture and stoniness. Soil phases are distinguished by the character of soil-forming and underlying rocks and thickness of the fine-earth part of the soil profile.

Further development of this system implies the creation of classification schemes for the mineralogical and textural peculiarities of soils and soil-forming rocks and for soil temperature and water regimes. However, the authors of the new classification argue that it would be difficult to develop appropriate systems for soil regimes because of the lack of adequate data for Russian soils. The development of a separate lithologic–mineralogical component reflecting soil features inherited from the parent rock is hampered by the lack of adequate criteria allowing us to distinguish between proper pedogenic alteration and the initial state of the parent material. At the same time, the need for ecological (environmental) characterization of soils is evident. Special ecological–landscape classifications should be developed, taking into account not only climatic and lithological indices but also the geomorphic position of soils, the character of natural and anthropogenic vegetation, the presence of geochemical barriers in soils, and other indices important from the viewpoint of predicting soil behavior and elaborating the strategy of sustainable soil management.

See also: **Classification of Soils**; **Classification Systems:** Australian; FAO; USA

Further Reading

Afanasiev JN (1927) *The Classification Problem in Russian Soil Science*. Russian Pedological Investigations, commission V. Leningrad: Publishing Office of the Academy of Sciences of the USSR.

Classification and Diagnostics of Soils of the USSR (1986). New Delhi: Oxonian Press.

Gedroits KK (1966) *Genetic Soil Classification Based on the Absorptive Soil Complex and Absorbed Soil Cations*. Jerusalem: Israel Program for Scientific Translation.

Glazovskaya MA (1983) *Soils of the World*, vol. 1. *Soil Families and Soil Types*. New Delhi: Amerind.

Goryachkin SV, Tonkonogov VD, Gerasimova MI *et al.* (2002) Changing concepts of soil and soil classification in Russia. In: Eswaran H, Rice T, Ahrens R, and Stewart BA (eds) *Soil Classification – A Global Desk Reference*, pp. 187–200. Boca Raton, FL: CRC Press.

Ivanova EN and Rozov NN (eds) (1970) *Classification and Determination of Soil Types*. Nos. 1–5. Jerusalem: Israel Program for Scientific Translation.

Neustruev SS (1967) *A Tentative Classification of Soil-Forming Processes as Related to Soil Genesis*. Jerusalem: Israel Program for Scientific Translation.

Shishov LL, Tonkonogov VD, Lebedeva II, and Gerasimova MI (2001) *Russian Soil Classification System*. Moscow: V. V. Dokuchaev Soil Science Institution.

Strzemski M (1975) *Ideas Underlying Soil Systematics*. Warsaw, Poland: Foreign Scientific Publications Department of the National Center for Scientific, Technical and Economic Information.

USA

D J Brown, Montana State University, Bozeman, MT, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

> "People who like sausage and respect the law shouldn't watch either being made" (commonly attributed to Otto von Bismarck, German chancellor, 1871–1890).

The construction of the US Department of Agriculture's Soil Taxonomy system has more in common with the legislative process and sausage-making than

with the development of a carefully reasoned philosophical doctrine. To be sure, there are philosophical justifications and rationalizations for Soil Taxonomy. But the key to understanding Soil Taxonomy, as with most systems of laws or rules, lies in understanding the history and evolution of the contemporary system.

There are many thorough reviews of Soil Taxonomy, with summaries of the soil orders found in virtually every pedology text. US Soil Survey staff have published documents on the details of Soil Taxonomy and how it is supposed to work. This chapter is intended to complement and explain these sources, particularly the Keys to Soil Taxonomy. To the uninitiated, the Keys can appear overwhelmingly complex and difficult to interpret. This article provides a simple, clearly stated introduction for the uninitiated – though, to be fair, many experienced soil scientists struggle to understand Soil Taxonomy. The emphasis is on explanation, not description, with insight valued over completeness. Toward this end, I take a pragmatic stance – with both an appreciation for historical contingency and an irreverent focus on 'where the rubber meets the road,' how Soil Taxonomy is actually employed in practice.

The fundamental sampling unit for Soil Taxonomy provides an illustration of the pragmatic approach. Officially, the pedon is the smallest soil unit, with a 1–3.5 m^2 surface area and approximately 2 m depth. The polypedon, a contiguous set of pedons, is the fundamental unit for Soil Taxonomy. While there are philosophical arguments both for and against the pedon and polypedon concepts, these have little bearing on the way soils are actually classified. With few exceptions, in practice the exposed side of a soil pit (the profile) serves as the basic unit for description, sampling, and classification.

Soil Classification Criteria

The criteria in Soil Taxonomy are designed to classify soils "on the basis of their characteristics and not on the basis of the supposed or partly proved causes which have produced the characteristics," (a mandate from the early survey leader Curtis Fletcher Marbut). Yet while the classification system is superficially based on contemporary soil characteristics, historic soil-forming processes and regional atmospheric climate play a fundamental role in shaping the structure of Soil Taxonomy. The current role of climate and soil-forming processes can be traced to past classification systems. An appreciation of the history of soil classification in the USA greatly clarifies the seemingly obscure contemporary system.

The most important taxonomic requirements for soil materials, and diagnostic horizons and diagnostic characteristics are outlined below. For complete requirements, the reader is referred to the Keys to Soil Taxonomy.

Soil Composition

The amount and type of chemically active surface in a soil are controlled by: (1) secondary clay minerals; (2) organic material; and (3) noncrystalline Fe and Al materials (imogolite, ferrihydrite, Al–humus complexes, and the semicrystalline allophane included by convention). We can subdivide the secondary clay minerals into: (a) high-activity clays (2:1 layer silicates, smectites, and vermiculites); and (b) low-activity clays (kaolinite, Fe-oxyhydroxides, and Al-oxides). Likewise, organic materials can be classified according to the degree of decomposition, from least to most: (a) fibric; (b) hemic; (c) sapric. These differences in soil composition, which greatly affect soil management, are captured through various criteria in Soil Taxonomy and summarized in Figure 1.

All soil materials can be classed into two mutually exclusive categories: (1) mineral soil materials; and (2) organic soil materials. Most accumulations of organic soil materials can be found in locations where the soil is saturated for at least 30 days $year^{-1}$. For these materials to be classed as organic, they must have more than 12–18% organic C, depending on clay content (0–60% respectively). Organic requirements depend on clay content because the contribution of organic matter to overall soil activity becomes more important as clay content declines. Organic soil materials can be further classified according to fiber content as fibric, hemic, or sapric (important for classifying organic soils).

The term 'andic soil properties' refers to materials with significant amounts of noncrystalline Fe and Al materials, formed by the weathering of volcanic glass and usually associated with volcanic ash deposits. Key requirements for the andic designation include low bulk density, high phosphate retention, and significant amounts of noncrystalline Al and Fe (as measured by oxalate extraction). For coarser soils (30% coarse silt and sand), there is a tradeoff between volcanic glass

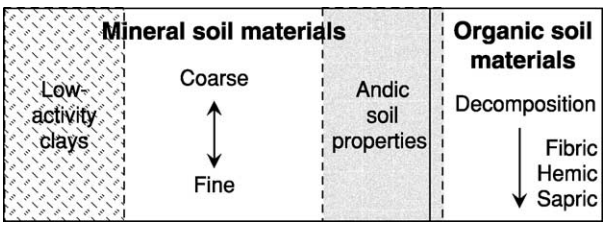


Figure 1 Schematic diagram of important soil constituents, with Soil Taxonomy designations where appropriate.

content and noncrystalline Al and Fe content required. To be termed andic, a soil must also have less than 25% organic C, with the result that both organic and mineral soils can qualify. Because the largely noncrystalline Fe and Al materials are very 'active,' they can be more important to the overall soil management than even large amounts of organic matter. While the laboratory measures required to identify andic properties are demanding, in practice they are rarely undertaken unless a volcanic ash parent material has been identified in the field.

The types and amounts of clay minerals in a soil are vital for soil management. Low-activity clays (1:1 kaolinite, Fe-oxyhydroxides, and Al-oxides) predominate in highly weathered soils, often found in warmer and wetter climates. Within Soil Taxonomy these minerals are not identified with a distinctive class. However, low-activity clays (cation exchange capacity or CEC per kg clay $\leq$16 cmol) are key criteria for two diagnostic horizons, discussed in the next section. Smectitic clays can cause soils to shrink and swell, but they are identified in Soil Taxonomy largely through field observation of shrink–swell features in a profile – not mineralogy. Subsoil clay accumulation is a defining feature for two soil orders, and there are also texture modifiers in many diagnostics, as with the clay content for the determination of organic soil materials.

Genesis and Soil Taxonomy

> "Soil surveys have created a new branch of soil science, that of soil anatomy" (C F Marbut (1921) The contribution of soil survey to soil science. *Society for the Promotion of Agricultural Science Proceedings* 41: 116–142).

Most of the diagnostics for Soil Taxonomy have a basis in soil formation theory. Understanding and using Soil Taxonomy effectively requires at least a rudimentary understanding of soil-forming processes. Soil properties important for land management play a secondary role in this system. For example, surface texture is one of the most important soil properties from a management perspective, yet influences classification primarily at the lowest series level. The emphasis on soil formation in Soil Taxonomy can be traced to the introduction of the biological metaphor in early twentieth-century soil survey and classification.

Curtis Fletcher Marbut, the inspirational leader of the US Soil Survey from 1913 until his death in 1935, modeled the science of soil survey on biology. Due largely to the tremendous success and influence of Darwin's ideas, many natural sciences drew upon a biologic metaphor from the late nineteenth through the early twentieth century, including geomorphology, ecology, entomology, and sociology. The study of profile formation was tied to the biological subfield of morphology – the study of embryonic development. And from this morphology, a hierarchical, 'genetic' soil taxonomy was developed analogous to biological taxonomy. Soils were grouped according to their mode of formation, not necessarily their contemporary properties. "No deviation from strict scientific [genetic] considerations for the sake of the so called practical use of the soil can safely be permitted." For example, texture was not considered important for classification because it was considered "not primarily a product of soil development." Like an old wine in new skins, the current Soil Taxonomy has evolved in many ways from early soil classification attempts, yet still retains the core of Marbut's system.

The genetic basis of Soil Taxonomy can be clearly observed in the requirements for diagnostic subsurface illuvial horizons (**Table 1**). (To present these horizons more clearly the requirements have been summarized: see Soil Taxonomy for precise requirements.) All but a few of these horizons require some evidence of illuviation in addition to compositional requirements. The argillic and natric require clay skins or an increasing fine-to-total clay ratio. The spodic requires an overlying albic or chemical evidence of illuviation. The calcic or gypsic requires observation of secondary (formed in soil) $CaCO_3$ or gypsum, or an increase in $CaCO_3$ relative to the parent material below. Among the nonilluvial diagnostic subsurface horizons (**Table 2**), the oxic horizon requires low CEC clays, and a paucity of weatherable minerals in the fine sand fraction (a genetic requirement). The cambic is a diagnostic horizon with indications of weak soil development. It is not necessary to demonstrate the genesis of these horizons, but the quantitative requirements for most diagnostic horizons are designed to capture genesis as best evidenced by contemporary soil properties.

There is no diagnostic subsurface horizon in Soil Taxonomy to describe soil layers that are periodically saturated and chemically reduced. The term 'aquic conditions' refers to soils which are periodically saturated and reduced. In most cases, redoximorphic features ('mottles' or a 'gleyed' horizon) are used to indicate reducing conditions. An accumulation of organic soil materials can also indicate saturated conditions. Direct measurements of soil hydrology and/or a chemically reduced state can confirm aquic conditions, but due to the cost and time involved are rarely employed. Aquic conditions are deliberately vague, with more detailed requirements specified for each soil order. Confusing matters further, the term 'aquic' is also employed for the aquic soil moisture regime.

Table 1 Illuvial–eluvial diagnostic subsurface horizons

Diagnostic horizon	*Field*	*Genesis*	*Key requirements*
Albic	E	Eluviation/leached	Light-colored horizon below A • High value, low chroma
Argillic	Bt	Clay illuviation	Clay-enriched subsoil • Clay 3–8% (absolute) > A/E horizons • Evidence of illuviation: – clay skins (macro/micro), or – increased fine : total clay ratio
Natric	Btn	Clay dispersion and illuviation	High sodium argillic • Argillic requirements • Columnar or prismatic structure • Sodic soil (high proportion of Na)
Spodic/ortstein	Bs Bh Bhs Bhsm	Chelation and illuviation	Illuvial Al-humus materials, often with Fe • pH $\leq$ 5.9, organic C $\geq$0.6% • Dark and/or red color with overlying albic • If cemented, then ortstein
(Petro) calcic	Bk, Bkm	Illuviation	Secondary $CaCO_3$ accumulation • $\geq$15% $CaCO_3$ (or 5% if sandy) • Evidence of illuviation: – 5% identifiable secondary $CaCO_3$, or – 5% (absolute) > an underlying horizon • If cemented/indurated $\rightarrow$ petrocalcic
(Petro) gypsic	By, Bym	Illuviation	Secondary gypsum accumulation • 5% gypsum ($CaSO_4 \cdot 2H_2O$) • Thickness (cm) $\times$ % gypsum $\geq$150 • Evidence of illuviation: – 1% identifiable secondary gypsum • If cemented/indurated $\rightarrow$ petrogypsic
Duripan	Bqm	Illuviation	Cemented/indurated silica accumulation • Does not slake in weak acid • Does slake in alkali solution

All diagnostic horizons have minimum thickness requirements, 2.5–15 cm.
All pans or cemented horizons require lateral continuity, vertical cracks $\geq$10 cm apart.

The surface diagnostic horizons, called epipedons, also have a genetic basis (Table 3). The histic epipedon, for example, not only requires a surface accumulation of organic materials, but also has a genetic requirement that this organic material accumulated under saturated conditions. The mollic epipedon is designed to capture all surface soils formed under grasslands. Trees generally deposit organic detritus on the soil surface, resulting in surface 'duff' layers and thin accumulations of organic matter in the soil relative to grasslands where annual root turnover adds organic material to the rooting depth. The mollic epipedon criteria (e.g., 0.6% organic C) are designed to capture the lowest prairie organic matter accumulation, such as in Montana, so prairie soils in Iowa usually exceed the requirements many times over.

One important point should be added to this discussion of diagnostic horizons: formal identification of these layers requires extensive, highly specified laboratory characterization. For a calcic horizon, $CaCO_3$ must be greater than 15% as determined by the evolution of CO_2 gas with acid treatment. For argillic horizons, clay films can be observed in the field or through a microscopic analysis of thin sections. For the mollic epipedon, the organic C content must be measured in the laboratory, no matter how thick and dark the surface soil. In fact, there are only a handful of diagnostic horizons that can be determined based on field observations alone. In this way, the contemporary Soil Taxonomy differs greatly from the earlier US Soil Survey classification systems where "the criteria used to classify are those that can be observed or determined rapidly by simple tests in the field."

In practice, however, laboratory characterization is rarely performed. Given the time and expense of laboratory analyses, soil surveyors and others rely on experience, field observations, and a familiarity with similar profiles to make reasonable assumptions as to soil composition.

Climate Zones in Soil Taxonomy

> "without soil climate as a criterion at some level in the taxonomic system, for example, Vertisols from Texas could be in the same class as Vertisols from North

Table 2 Other common diagnostic subsurface horizons or characteristics

Diagnostic horizon	*Field*	*Genesis*	*Key requirements*
Aquic conditions	Bg	Redox	Periodically saturated-reduced soil • Redoximorphic features (gley, 'mottles') • Saturated soil (directly measured) • Chemical reduction (directly measured)
Kandic	Bt	? Clay, mineral weathering	Clay-enriched, weathered subsoil • Clay increase relative to A/E • Low-activity clays (kaolinite, oxides) – CEC $\leq$ 16 cmol kg^{-1} clay
Oxic	Bo	Mineral weathering	Highly weathered soil • Low-activity clays (kaolinite, oxides) – CEC $\leq$ 16 cmol kg^{-1} clay • Weatherable minerals $\leq$10% in fine sand • No clay increase (not Kandic)
Salic	Bz	Multiple pathways	Saline layer • EC $\geq$ 30 dS m^{-1} • EC $\times$ thickness (cm) $\geq$900
Fragipan	Bx	? Unclear	High-bulk-density brittle pan • Dry material slakes in water • Very coarse or weak structure, or massive • At least firm rupture resistance • Brittle, with no roots
Placic	Bsm	Redox	Thin Fe/Mn/organic pan • Cemented with Fe, Mn, or organic matter • Thin: 1 mm minimum, usually less than 25 mm
Cambic		Multiple pathways	Default subsurface horizon • Some soil formation, but not enough to meet any other diagnostic requirements

All diagnostic horizons have minimum thickness requirements, 2.5–15 cm.
All pans or cemented horizons require lateral continuity, vertical cracks $\geq$10 cm apart.
CEC, cation exchange capacity; EC, electrical conductivity.

Table 3 Most common diagnostic epipedons (surface horizons)

Epipedon	*Genesis*	*Key requirements*
Mollic	Prairie grasses	Thick, dark, organic-rich, fertile topsoil • Dark: moist Munsell value and chrome $\leq$3 • Organic C $\geq$0.6% • Base saturation $\geq$50% • Thickness $\geq$18 or 25 cm (see conditions), or meets requirements when mixed to 18 cm depth
Umbric	?	Low base saturation mollic • Meets mollic requirements, except: • Base saturation $\leq$50%
Histic	Wetland	Organic surface horizon • Saturated $\geq$30 days $year^{-1}$ • Organic soil material $\geq$20 cm thick • Not thick enough to be an organic soil (40–60 cm)
Folistic	Litter layer	Nonwetland organic surface horizon • Saturated $<$30 days $year^{-1}$ • Organic soil material $\geq$15 cm thick, or $\geq$20 cm if sphagnum or bulk density $<$0.1
Melanic	Volcanic	Thick, dark volcanic surface horizon • Andic soil properties $\geq$30 cm thick • Organic C $\geq$4% throughout, $\geq$6% average • Dark: moist Munsell value and chroma $\leq$2
Ochric	Forest/shrub	Default surface epipedon • Does not meet the requirements of other epipedons

Dakota" (Soil Survey Staff (1999) Soil Taxonomy. Washington, DC: US Department of Agriculture, Natural Resources Conservation Service, p. 94).

The classification of a dogwood (*cornus*) would not change at any level if we dug up a species in Texas and replanted it in North Dakota. If we dug up a Vertisol in Texas, and 'planted' it in North Dakota, however, the classification of the soil would change by virtue of location, and likely at one of the highest levels of Soil Taxonomy. This geographic–climatic dimension of Soil Taxonomy reflects the origins of soil classification in nineteenth-century Russia.

V.V. Dokuchaev and his student Sibertsev proposed that soils be studied from a 'geologic–geographic' perspective, laying the theoretical groundwork for contemporary pedology, the science behind soil survey. In the nineteenth century, soils were studied from a fertility perspective or assumed to be an extension of the geologic rock below. Dokuchaev argued for the study of soils from a natural history perspective, examining their formation and distribution over the surface of the Earth. From this vantage point, he further outlined five environmental factors that controlled the formation and distribution of soils: "climate, country age, vegetation, topography and parent rock."

Based on these ideas, Sibirtsev published a 'zonal' classification system in 1900 whereby typical profiles were described for the major 'physiographic zones' of Russia. 'Intrazonal types' deviated from the zonal norm due to the influence of one or two soil-forming factors, commonly topography (e.g., bog soils) and/or parent rock (e.g., solonetzes, salt-affected soils). 'Azonal' soils were dominated by parent materials (e.g., fresh alluvium). This classification system was brought into Europe and the USA through the writings of K.D. Glinka, who strongly emphasized the role of climate in soil formation and classification. Determined to free pedology from the discipline of geology, Glinka argued that, unlike geologic strata, soil formation and geography were controlled primarily by climate, which "provides justification for singling out the soils as a particular group of natural bodies, with which a special branch of science should be concerned." The 'zonal' or 'climatic' concept had a strong influence on the early development of pedology in the USA and this influence can still be found in contemporary Soil Taxonomy.

Climate is an important and diagnostic characteristic for both mineral and organic soils. One soil order (Aridisols) and most of the suborders are governed by soil moisture regimes, while soil temperature regimes become important at lower levels of Soil Taxonomy. Another soil order, the Gelisols, is founded on a soil feature (permafrost) that is highly correlated with atmospheric climate.

The requirements for determining soil moisture regimes (Table 4) are excruciatingly complex, with extensive soil monitoring required. In practice these measurements are rarely made because, according to soil taxonomy, "the intent in defining the soil moisture control section is to facilitate the estimation of soil moisture regimes from climatic data." In other words, soil moisture regimes are typically determined from regional atmospheric climate data, without need for soil measurements. For the vast bulk of the Earth's land surface, the soil moisture regime can be mapped in large swathes or 'zones' reminiscent of the Russian system. Local topography and hydrology can also, in some situations, influence the soil moisture regime. In particular, the aquic and peraquic moisture regimes (saturated soils) are usually controlled by local topography and groundwater hydrology and identified by the presence of redoximorphic features or an accumulation of organic soil materials.

Soil temperature regimes (Table 4) are also estimated for the most part from atmospheric climate data, though if necessary, temperatures can be taken at a depth of 50 cm at monthly intervals. Adjustments are available to convert mean annual temperature (MAT) to mean annual soil temperature (MAST). The temperature and moisture regimes for a soil to be classified can also be obtained from other soils in the region. Nonalpine soils in Montana, for example, are generally assumed to have ustic moisture and frigid temperature regimes.

Structure and Nomenclature

Most Soil Taxonomy users will never classify a soil themselves. However, understanding the processes by which soil is classified can greatly enhance the interpretation of soils already classified. Toward this end, the procedure for classifying and naming a soil is described below, followed by a worked example using observations and data from a profile in south Dakota, USA.

How it Works

Modeled on biological taxonomy, Soil Taxonomy has a hierarchical organization (Figure 2). The order, suborder, great group, and subgroup levels are primarily governed by climate and soil genesis, with the family and series levels capturing important physical and chemical properties for crop growth, engineering, and land management. In some cases, soil genesis and management are related: for example, with

Table 4 Brief descriptions of temperature and moisture regimes

Moisture regimes	*Brief description*	*Temperature regimes*	*Brief description*
Aridic/torric	Dry most of the year		
Xeric	Pronounced dry season – summer	Cryic	MAT $< 8°C$ Summer temperature $<$ 6–15°C No permafrost
Ustic	Pronounced dry season – winter	(Iso)frigid	MAT $< 8°C$ Isofrigid; can also be cryic
Udic	Humid climate, no pronounced dry season	(Iso)mesic	$8°C \leq$ MAT $< 15°C$
Perudic	Wet climate, year-round precipitation > evapotranspiration	(Iso)thermic	$15°C \leq$ MAT $< 22°C$
Aquic	Saturated and reduced part of the year	(Iso)hyperthermic	$22°C \leq$ MAT
Peraquic	Saturated and reduced year-round		

MAT, mean annual temperature.
'Iso' prefix indicates that the mean summer and mean winter temperatures differ by less than 6°C.

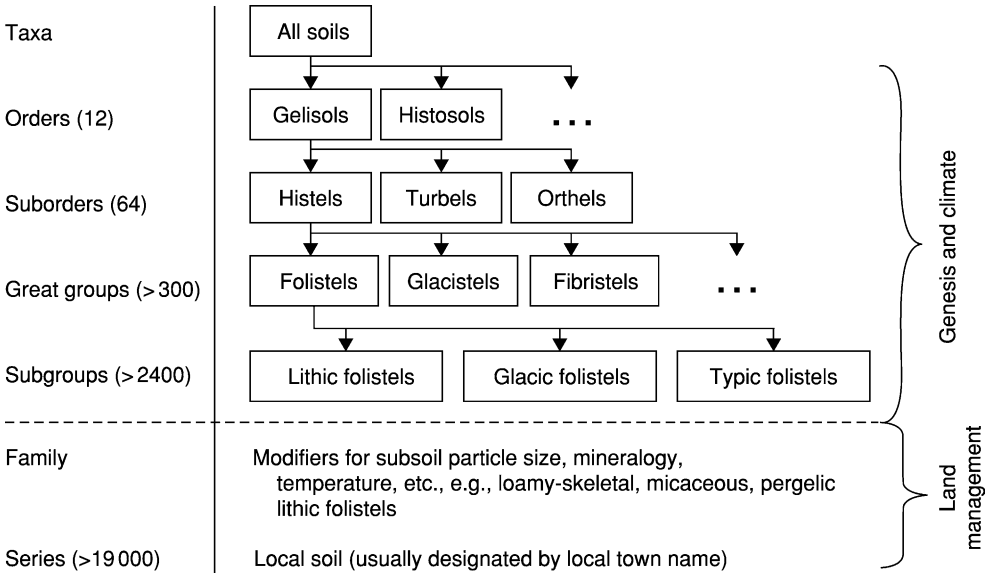


Figure 2 Hierarchical organization of Soil Taxonomy, with climatic–genetic versus management levels identified separately.

Vertisols, where shrink–swell processes are key for both profile genesis and land management. Unfortunately, this not always the case.

The soil orders are summarized in Table 5. Most of the orders can be related closely to genetic factors of soil formation, though most are related to more than one factor. Soils are keyed out in order, starting with Gelisols and ending with Entisols. Implicit in the requirements for each order is the requirement that the soil not meet the requirements for preceding orders. For example, a soil with permafrost at 75 cm, a mollic epipedon, and base saturation greater than 75% for all horizons would be classified as a Gelisol as this order takes precedence over the Mollisol. This sequential approach to 'keying out' a soil profile is used for all levels of soil taxonomy except the lowest, the soil series.

In practice, where Soil Survey activity is largely complete, soils are often classified from the bottom up. At the local level, a soil profile can be described and compared to known soil series. The series that best 'fits' the examined profile can then be applied, and the established taxonomic designation for that series thus derived.

The Soil Taxonomy nomenclature follows the hierarchical structure of the system, and allows users to glean basic soil information from the name alone. Soil names are constructed from right to left, with segments drawn primarily from the diagnostic criteria discussed previously. This nomenclature system can best be illustrated with an example:

Order: Alfisol Suborder: Udalf Great group: Kandiudalf
Subgroup: aquic kandiudalf → Aquic | kandi | ud | alf

Table 5 Summary of soil orders and key factors of soil formation

Order	*Important features and properties*	*Key factors of soil formation*
Gelisols	Permafrost within 1–2 m of surface	Climate: frozen soils
Histosols	Organic soils (usually in wetlands)	Topography: usually in wetlands
Spodosols	Illuvial Fe-Al-humus subsurface horizon, usually sandy	Vegetation and parent material: coniferous trees, sandy deposits
Andisols	Amorphous 'minerals'; volcanic glass	Parent material: volcanic ash
Oxisols	Highly weathered soils; dominated by low-CEC clays (kaolinite and oxides)	Climate and time: mineral weathering
Vertisols	Shrink–swell activity, as evidenced by cracks and 'slickenslides'	? Source of shrink–swell clays
Aridisols	Arid climate	Climate
Ultisols	Clay-enriched subsoil; low base saturation	Climate and time: base leaching, acidification
Mollisols	Thick, dark, organic-rich surface layer; high base saturation	Vegetation: prairie soils
Alfisols	Clay-enriched subsoil, high base saturation	Vegetation, climate, and time: forest soils, moderately leached
Inceptisols	Minimal horizon development	Time and climate: adolescent development
Entisols	Unaltered parent material; usually geomorphic deposits	Time: young

CEC, cation exchange capacity.

Climate, genesis, and key soil properties can be discerned from the four nomenclature segments above. The 'alf' term indicates that this profile has a clay increase in the subsoil, and does not meet the criteria for all of the preceding nine orders. The 'ud' term refers to the udic moisture regime. 'Kandi' is an abbreviation for the kandic diagnostic horizon, a clay-enriched, low-activity (low CEC per kg clay) subsoil. 'Aquic' in the subgroup tag refers to aquic conditions lower in the profile. (By comparison, an 'aqualf' would have aquic conditions higher in the profile.) Given a familiarity with soil moisture regimes, order names, and diagnostic horizons, most Soil Taxonomy names can be similarly interpreted. Default terms are 'orth,' 'hapl,' and 'typic' for the suborder, great group, and subgroup respectively. For example, a 'typic haplorthod' is a spodosol with no additional features to describe in the second to fourth levels of Soil Taxonomy.

Example Profile

The site description, field description, and essential lab characterization data for a Ferney profile in South Dakota are provided in Table 6. Using this information, the key diagnostic horizons are outlined with bold lines in Table 7. As can be seen from this example, diagnostic horizons do not necessarily correspond directly to field horizons. The mollic epipedon encompasses both the Ap and Bt horizons, while the argillic and natric diagnostic horizons are comprised of the Bt and Btkz horizons identified in the field. Furthermore, field designations do not directly correspond with the more precisely defined diagnostics. There are three B horizons with a 'k' designation for illuvial carbonate accumulation, for example, only one of which qualifies as a calcic horizon. There are also three 'z' designations in the subsoil for salt accumulation, though the electrical conductivities of these horizons do not meet the salic horizon requirement.

Both the Ap and Bt horizons meet the color value and chroma requirements as well as the organic carbon requirements for a mollic epipedon. The combined depth is 25 cm, which satisfies the thickness requirement for this epipedon, though the Ap alone would not. Similarly, the Bt and Btkz field horizons both meet the illuvial (observed clay skins) and clay increase requirements ($1.2 \times$ eluvial horizon) for an argillic horizon, with a combined thickness of 51 cm. Horizons below could also meet the illuvial requirement (despite the lack of clay skins) through the examination of thin sections under a microscope, but these data are not available. The calcic horizon has both the overall quantity of $CaCO_3$ required, and a 5% increase over an underlying horizon (evidence of illuviation). The bottom two horizons (C and Bkz2) might also meet the requirements for a gypic horizon, but confirmation of 1% secondary gypsum is lacking. Finally the Bt and Btkz horizons both meet the requirements for a natric horizon: (1) argillic requirements; (2) columnar or prismatic structure; and (3) sodium absorption ratio (SAR) $\geq$13.

Using the diagnostic horizons obtained, together with climatic data, we can readily key out the soil profile (Table 8), indicating the degree to which the designated moisture regime borders on another regime (e.g., ustic bordering on aridic). Care must be taken to distinguish between 'and' and 'or' requirements, but with practice the Keys can be systematically applied to available data, yielding a Leptic Natrustoll in this case. Space constraints do not

Table 6 Ferney profile (based on Soil Survey Pedon no. 86P0004)

Site description

Physiography: glaciated uplands
Slope: 1% concave south-east facing
Parent material: glacial till from mixed material
Temperature regime: frigid
Drainage: moderately well drained
Moisture regime: Ustic
Moisture control section:
dry 3/10 days with soil temperature >5°C

Clay (%)	*Silt (%)*	*Sand (%)*	*Organic C (%)*	*$CaCO_3$ (%)*	*Gypsum (%)*	*Base sat. (%)*	*pH (H_2O)*	*CEC*	*CEC/clay*	*SAR*
								(*$mmol_c\ kg^{-1}$*)	(*$mmol_c\ kg^{-1}$*)	
Ap – 0–13 cm; black (10YR 2/1) interior moist clay loam; moderate fine and medium subangular blocky structure; few fine accumulations of carbonate; strongly effervescent; neutral (pH = 7.0); abrupt smooth boundary										
23.7	47.7	28.6	3.24	–	–	100	6.6	26.3	111	5
Bt – 13–25 cm; very dark brown (10YR 2/2) interior moist clay loam; strong medium columnar structure; continuous faint coats on tops of columns and continuous faint clay films on faces of peds; mildly alkaline (pH = 7.6); clear wavy boundary										
40.3	33.6	26.1	1.31	tr	–	100	8.1	33.2	82	13
Btkz – 25–64 cm; dark grayish-brown (2.5Y 4/2) interior moist clay loam; moderate medium and coarse prismatic structure; continuous faint clay films on faces of peds; few fine carbonate concretions and many fine and medium salt masses; strongly alkaline (pH = 8.8); clear wavy boundary										
40.7	33.7	25.6	0.64	13	2	100	8.2	23.7	58	16
Bkz1 – 64–99 cm; dark grayish-brown (2.5Y 4/2) interior moist clay loam; few fine distinct mottles; weak medium and coarse prismatic structure; common fine carbonate concretions and common fine salt masses; strongly effervescent; strongly alkaline (pH = 8.8); clear wavy boundary										
38.2	35.0	26.8	0.27	17	1	100	8.4	21.2	56	21
Bkz2 – 99–127 cm; olive brown (2.5Y 4/4) interior moist clay loam; common fine prominent and many medium and coarse distinct mottles; weak coarse prismatic structure; common fine carbonate concretions, many coarse salt masses; strongly effervescent; moderately alkaline (pH = 8.2); gradual wavy boundary										
33.3	34.8	31.9	0.22	3	10	100	8.2	19.2	58	19
C – 127–152 cm; olive brown (2.5Y 4/4) interior moist clay loam; common fine prominent and common fine and medium distinct mottles; massive; common fine carbonate concretions, few salt masses; strongly effervescent; strongly alkaline (pH = 8.8)										
33.9	35.9	30.2	0.16	8	7	100	8.2	19.2	55	17

Note: Maximum electrical conductivity for profile is 14.2 dS m^{-1}.

CEC, cation exchange capacity; SAR, sodium absorption ratio.

Table 7 Diagnostic horizons for Ferney profile

Horizon	*Mollic*	*Argillic*	*Calcic*	*Natric*
Ap 13 cm	Value = 2 ≤ 3 Chroma = 1 ≤ 3 OC = 3.2 ≥ 0.6%	Eluvial horizon 23.7% × 1.2 = 28.4% cutoff		
Bt 25 cm	Value = 2 ≤ 3 Chroma = 2 ≤ 3 OC = 1.3 ≥ 0.6%	Clay films 40.3 > 28.4% clay	$CaCO_3$ = trace < 15	Argillic + SAR = 13 ≥ 13 columnar structure
Btkz 64 cm	Value = 4 > 3 XXXXX	Clay films 40.7 > 28.4% clay	$CaCO_3$ = 13 < 15	Argillic + SAR = 16 ≥ 13 prismatic structure
Bkz1 99 cm		No clay films	$CaCO_3$ = 17 < 15 17% > 3% + 5% $CaCO_3$ (3% underlying)	Not argillic
Bkz2 127 cm		No clay films	$CaCO_3$ = 13 < 15%	Not argillic
C 152 cm		No clay films	$CaCO_3$ = 12 < 15%	Not argillic

OC, Organic carbon; SAR, sodium absorption ratio.

Table 8 Keying out Ferney profile

Order	*Yes/no*	*Reason*
A. Gelisol	No	No permafrost or gelic materials
B. Histostol	No	No organic soil materials
C. Spodosol	No	No spodic horizon or ortstein
D. Andisol	No	No volcanic ash or andic materials
E. Oxisol	No	No oxic or kandic horizon
F. Vertisol	No	No surface cracks or slickenslides
G. Aridisol	No	Ustic, not aridic moisture regime, no salic
H. Ultisol	No	Base sat. >35% for all horizons
I. Mollisol	Yes	1a. Mollic epipedon
		2. Base sat. >50% for all horizons
Suborder		
IA. Alboll	No	No albic horizon
IB. Aquoll	No	No aquic conditions at 40–50 cm depth
IC. Rendoll	No	2. Argillic and calic horizons present
ID. Xeroll	No	Have Ustic, not Xeric moisture regime
IE. Ustoll	Yes	Ustic moisture regime
Great group		
IFA. Durustoll	No	No duripan
IFB. Natrustoll	Yes	Natric horizon present
Subgroup (Natrustolls)		
IFBA. Leptic Torretic	No	3a. Dry only 3 days out of 10 in moisture control section with temperature >5°C (not torric)
IFBB. Torrertic	No	2a. Dry only 3 days out of 10 in moisture control section with temperature >5°C (not torric)
IFBC. Leptic Vertic	No	2a. No slickenslides or cracks
		2b. Linear extensibility not 6.0 cm or greater
IFBD. Glossic Vertic	No	No vertic properties (see above)
IFBE. Vertic	No	No vertic properties (see above)
IFBF. Aridic Leptic	No	2a. Dry only 3 days out of 10 in moisture control section with temperature >5°C (not aridic)
IFBG. Leptic	Yes	Visible salt crystals at 25 cm depth <40 cm
Classification of profile to subgroup level: Leptic Natrustoll		

allow classification to the family level, though assuming mineralogical data are available, this can also be accomplished by systematically following the Keys. At the subgroup level, there are often statistically precise climatic requirements. These requirements indicate the degree to which the moisture regime

borders on another regime (e.g., Ustic bordering on Aridic). To simplify the classification for this example, the frequency of dry days during the growing season was provided in the site description, though this is rarely available in practice. As with climate regimes generally, these subgroup climatic modifiers are established regionally and can be obtained from the local soil survey staff.

Closing Thoughts

In this chapter, I explained: (1) how US soil classification took its present form, and (2) how the contemporary system is actually used. Tracing the roots of Soil Taxonomy explains many peculiar features of the contemporary system. Past climatic and genetic classification ideas can be found beneath the utilitarian veneer of the current system. Understanding these historical ideas is vital to understanding contemporary US Soil Taxonomy, and how to use it.

List of Technical Nomenclature

Base sat. Percentage base saturation, exchangeable bases $\times$ $(CEC)^{-1} \times 100$

$CaCO_3$ Percentage of $CaCO_3$ equivalent (measured by CO_2 evoluation) measured relative to fine earth fraction (<2 mm)

CEC Cation exchange capacity ($mmol_c\ kg^{-1}$)

Clay (%) Percentage of fine earth (<2 mm) with particle size less than 2 μm

EC Electrical conductivity ($dS\ m^{-1}$)

Gypsum Percentage of $CaSO_4{\cdot}2H_2O$ equivalent measured relative to fine earth fraction (<2 mm)

Org. C Percentage of organic C measured relative to fine earth fraction (<2 mm)

pH Measure of solution acidity, $\log[H^+]$

SAR Sodium absorption ratio $= Na^+ \cdot \{0.5 \cdot (Ca^{2+} + Mg^{2+})^{0.5}\}$, based on saturated paste extract

Sand (%) Percentage of fine earth (<2 mm) with particle size between 50 μm and 2 mm

Silt (%) Percentage of fine earth (<2 mm) with particle size less than 50 μm

See also: **Classification of Soils**; **Classification Systems:** Australian; Russian, Evolution and Examples

Further Reading

Ahrens RJ and Arnold RW (1999) Soil taxonomy. In: Summer ME (ed.) *Handbook of Soil Science*, pp. E117–136. Boca Raton, FL: CRC Press.

Allen G (1978) *Life Science in the Twentieth Century*. Cambridge: Cambridge University Press.

Baldwin M, Kellogg CE, and Thorp J (1938) Soil classification. In: *Soils and Men*, pp. 979–1001. Washington, DC: United States Government Printing Office.

Bockheim JG and Gennadiyev AN (2000) The role of soil-forming processes in the definition of taxa in Soil Taxonomy and the World Soil Reference Base. *Geoderma* 95: 53–72.

Buol SW, Hole FD, McCraken RJ, and Southard RJ (1997) *Soil Genesis and Classification*. Ames, IA: Iowa State University Press.

Cline MG (1949) Basic principles of soil classification. *Soil Science* 67: 81–91.

Dokuchaev VV (1967) (originally published 1883)*Russian Chernozem. Selected Works of V.V. Dokuchaev 1*. Jerusalem Israel: Israel Program for Scientific Translations.

Glinka KD (1963) (originally published 1931)*Treatise on Soil Science (Pochvovedenie)*. Jerusalem, Israel: Israel Program for Scientific Translations.

Haskett JD (1995) The philosophical basis of soil classification and its evolution. *Soil Science Society of America Journal* 59: 179–184.

Marbut CF (1921) The contribution of soil survey to soil science. *Society for the Promotion of Agricultural Science Proceedings* 41: 116–142.

Marbut CF (1927) *A Scheme for Soil Classification*, pp. 1–31. Washington, DC: First International Congress of Soil Science. American Organizing Committee.

Sibirtsev NM (1966) *Selected Works, 1*. Jerusalem, Israel: Israel Program for Scientific Translations.

Soil Survey Staff (1998) *Keys to Soil Taxonomy*. Washington, DC: US Department of Agriculture, Natural Resources Conservation Service.

Soil Survey Staff (1999) *Soil Taxonomy, A Basic System of Soil Classification for Making and Interpreting Soil Surveys*. Washington, DC: US Department of Agriculture, Natural Resources Conservation Service.

CLAY MINERALS

D G Schulze, Purdue University, West Lafayette, IN, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

The clay-size fraction of soils consists of mineral particles that are less than 2 μm in equivalent diameter. This is the realm of exceedingly small, crystalline particles dominated by planar arrays of SiO_4 structural units and many structural hydroxyls and water. These 'clay minerals' crystallize in the aqueous environment at the Earth's surface from the constituent ions released by dissolving (weathering) 'primary minerals' such as olivines, pyroxenes, feldspars, micas, quartz, and others that were formed under extreme heat and pressure deep within the Earth. Clay minerals are responsible for many of the soil's most important and characteristic physical and chemical properties. Fundamental soil properties such as cation exchange and shrink–swell properties, as well as practical considerations such as how well a particular soil will attenuate a specific pollutant, or how much fertilizer phosphorus will be fixed and unavailable to crops, are all influenced by molecular-scale differences in soil clay minerals.

Clay minerals are distinguished on the basis of their different crystal structures, and there is a close relationship between the crystal structure and the corresponding bulk physical and chemical properties of a particular type of clay. We begin by considering some of the properties of the major chemical elements that make up the clay minerals.

Major Element Composition of Clay Minerals

Most of the mass and volume of the Earth's crust is made up by only a few chemical elements. O and Si alone account for almost 75% of the mass, with most of the remainder, in order of decreasing abundance, consisting of Al, Fe, Ca, Na, K, Mg, Ti, H, P, and Mn. On a volume basis, oxygen alone accounts for more than 90% of the total volume. O, as O^{2-}, is the only abundant anion, while the other abundant elements are all cations. Most of these cations have only one stable oxidation state at the Earth's surface (Al^{3+}, Ca^{2+}, Na^{+}, K^{+}, Mg^{2+}, Ti^{4+}, H^{+}, P^{5+}); Fe (Fe^{2+}, Fe^{3+}) and Mn (Mn^{2+}, Mn^{3+}, Mn^{4+}) are the exceptions. The O^{2-} anions are much larger than most of the positively charged cations. The Earth's crust, therefore, can be characterized as large O atoms in an approximately close-packed arrangement held together by attraction to smaller cations located in the interstitial space.

Most of the elements in the crust and in soils occur in minerals, and the elements listed above are major constituents of the most abundant minerals, including clay minerals. Building on the concept of packing O atoms in space, we will consider atoms as rigid spheres, realizing that this is an oversimplified but convenient model for developing the key structural concepts.

Basic Structural Concepts

Tetrahedra and Octahedra

Two distinct structural features occur within the crystal structures of soil clay minerals as a consequence of packing the large O^{2-} ions together in space. The first consists of four O^{2-} ions packed closely together, and can be described as three O^{2-} ions arranged in a triangle with the fourth O^{2-} occupying the dimple formed by the other three (Figure 1). The centers of the four O^{2-} ions form the apices of a regular tetrahedron, and the small space in the center is called a 'tetrahedral site.' Cations located in tetrahedral sites are in fourfold or tetrahedral coordination, because they are surrounded by and bonded to four O^{2-} ions.

The second structural feature consists of six closely packed O^{2-} ions. Three of them are arranged in a triangle in one plane, and the other three, also in a triangle but rotated 60° relative to the first three, are in a second plane so that the two triangular groups intermesh (Figure 1). The centers of the six O^{2-} ions form the apices of a regular octahedron, and the small space in the center is called an 'octahedral site.' Cations located in the octahedral site are said to be in sixfold or octahedral coordination because they are surrounded by and bonded to six O^{2-} ions.

Tetrahedral and octahedral sites differ in another important way. The space that can be occupied by a cation in a tetrahedral site is smaller than the space that can be occupied in an octahedral site. Since cations vary in size, smaller cations tend to occur in tetrahedral sites, somewhat larger cations tend to occur in octahedral sites, and the largest cations must fit into spaces that are even larger than octahedral sites. Cations with sizes intermediate between the optimum for two sites can occur in either site. The Al^{3+} ion, for example, can occur in either octahedral or tetrahedral sites. Table 1 summarizes the structural sites in which cations tend to occur in clay minerals.

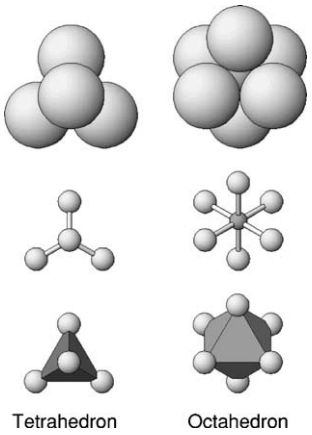


Figure 1 Spheres closely packed to form a tetrahedron and an octahedron. Note the appearance for the three different ways of drawing the model: as a sphere-packing model (top row), ball-and-stick model (middle row), and a polyhedral model (bottom row). (Adapted from Schulze DG (2002) An introduction to soil mineralogy. In: Dixon JB and Schulze DG (eds) *Soil Mineralogy with Environmental Applications*, pp. 1–35. Madison, WI: Soil Science Society of America, with permission.)

Representing Crystal Structures

Octahedra and tetrahedra are commonly represented using different types of models, each of which portrays the same concept, but highlights different structural features. Sphere-packing models give an impression of the space occupied by the atoms, ball-and-stick models highlight the bonds, while polyhedral models emphasize the tetrahedral and octahedral units (Figure 1). There is some ambiguity associated with each representation, and it is important to understand the correspondence between, and limitations of, each type of model. Polyhedral models are used here, along with some sphere-packing models.

Tetrahedral and Octahedral Sheets

The most common and abundant clay minerals belong to a group of minerals called phyllosilicates (from Greek 'phyllon,' meaning 'leaf') or sheet silicates. A common structural theme in all phyllosilicates is the presence of SiO_4 tetrahedra arranged into sheets. Octahedra arranged into sheets are also present in the structures of phyllosilicates and in some hydroxide minerals as well. Different combinations of the two sheets give rise to the different clay mineral structures.

Tetrahedral Sheet

The tetrahedral sheet consists of SiO_4 tetrahedra arranged such that three of the four O^{2-} ions of each tetrahedron are shared with three nearest-neighbor tetrahedra (Figure 2). These shared O^{2-} ions are all in the same plane and are referred to as basal oxygens. Note that adjacent tetrahedra share only one O^{2-} between them (the tetrahedra share apices or corners). The fourth O^{2-} ion of each tetrahedron is not shared with another SiO_4 tetrahedron and is free to bond to other polyhedral elements. These unshared O^{2-} ions are referred to as apical oxygens. Since each basal oxygen contributes a charge of -1 to each Si^{4+} ion, the addition of H^+ ions to the apical oxygens to form hydroxyls should result in an electrically neutral tetrahedral sheet. Such individual tetrahedral sheets do not form stable mineral structures by themselves and only occur in combination with octahedral sheets, as described below.

Figure 2 shows all of the apical oxygens pointing in the same direction, namely, out of the plane of the paper toward the reader. This is the most common arrangement, but structures also occur in which the apical oxygens point alternately in opposite directions.

Table 1 Type of structural sites in which common cations tend to occur in phyllosilicate mineral structures

Type of site	*Cation*
Tetrahedral only	Si^{4+}
Tetrahedral or octahedral	Al^{3+}, Fe^{3+}
Octahedral only	Mg^{2+}, Ti^{4+}, Fe^{2+}, Mn^{2+}
Interlayer sites	Na^+, Ca^{2+}, K^+

Octahedral Sheet

Analogous to the tetrahedral sheet, we can consider the octahedral sheet illustrated in Figure 3 as an assemblage of octahedra in which adjacent octahedra share two oxygens with one another. In other words, adjacent octahedra share edges. For the arrangement of octahedra shown in Figure 3, the octahedral sites are occupied by trivalent cations, typically Al^{3+}, and for charge balance, a proton (H^+) must be associated with each O^{2-}. (The H^+ takes up very little space, and the OH^- ion can be considered a sphere of roughly the same size as an O^{2-} ion.) Each OH^- contributes one-half a negative charge to each cation because each OH^- is shared between two octahedra. Each Al^{3+} cation is therefore effectively surrounded by $6 \times 0.5 = 3$ negative charges and the sheet is electrically neutral. Note the pattern of empty and filled octahedral sites. Two of every three possible octahedral sites are filled when trivalent cations are present in the octahedral sites. This arrangement is called dioctahedral and is the most common in soil clay minerals. If the octahedral sites are filled with divalent cations such as Mg^{2+} then every possible octahedral site must be occupied to produce an electrically

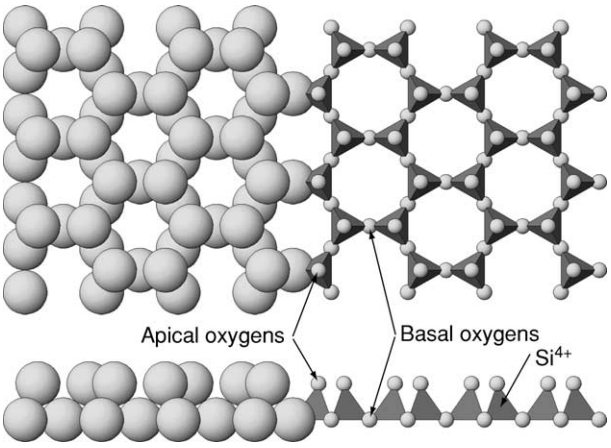


Figure 2 The tetrahedral sheet as a sphere-packing model (left half) and a polyhedral model (right half). (Adapted from Schulze DG (2002) An introduction to soil mineralogy. In: Dixon JB and Schulze DG (eds) *Soil Mineralogy with Environmental Applications*, pp. 1–35. Madison, WI: Soil Science Society of America, with permission.)

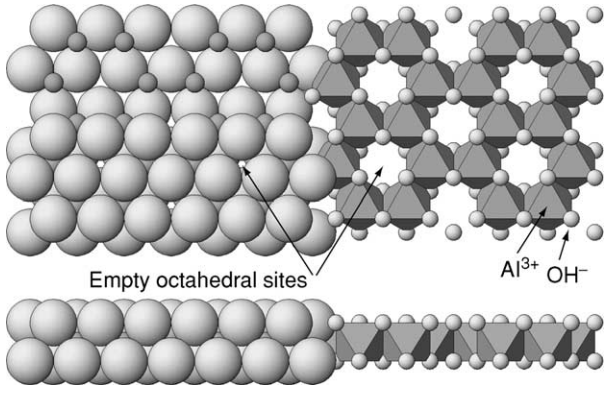


Figure 3 The octahedral sheet as a sphere-packing model (left half) and a polyhedral model (right half). The top three rows of spheres have been omitted from the sphere-packing model so that the underlying cations, represented here as Al^{3+}, can be seen more easily. The dioctahedral arrangement is shown here, in which two-thirds of the possible octahedral sites are filled with cations and one-third are empty. (Adapted from Schulze DG (2002) An introduction to soil mineralogy. In: Dixon JB and Schulze DG (eds) *Soil Mineralogy with Environmental Applications*, pp. 1–35. Madison, WI: Soil Science Society of America, with permission.)

neutral structure. This arrangement is called trioctahedral.

Octahedral sheets, stacked one on top of the other and held together by hydrogen bonds, make up the structure of gibbsite, $Al(OH)_3$, an aluminum hydroxide mineral that occurs in intensively leached soils. The structure of gibbsite is the simplest in a series of structures containing octahedral sheets.

Phyllosilicate Minerals Common In Soil Clays

Phyllosilicates are divided into two groups, 1:1- and 2:1-type minerals, based on the number of tetrahedral and octahedral sheets in the layer structure.

1:1-Type Minerals

1:1 Layer structure The 1:1 layer structure consists of a unit made up of one octahedral and one tetrahedral sheet, with the apical O^{2-} ions of the tetrahedral sheets being shared with (and part of) the octahedral sheet (Figure 4). There are three planes of anions (Figure 4b). One plane consists of the basal O^{2-} ions of the tetrahedral sheet, the second consists of O^{2-} ions common to both the tetrahedral and octahedral sheets (marked 'a' in Figure 4a) plus OH^- belonging to the octahedral sheet ('b' in Figure 4a), and the third consists only of OH^- belonging to the octahedral sheet.

Kaolinite The structure of kaolinite consists of 1:1 layers stacked one above the other. Kaolinite contains Al^{3+} in the octahedral sites and Si^{4+} in the tetrahedral sites (Figure 5). The 1:1 layer is electrically neutral and adjacent layers are held together by hydrogen bonding between the basal oxygens of the tetrahedral sheet and the hydroxyls of the exterior plane of the adjacent octahedral sheet.

Kaolinite is a common mineral in soils and is the most common member of this subgroup. It tends to be particularly abundant in more weathered soils such as Ultisols and Oxisols. Kaolinites have very little isomorphous substitution in either the tetrahedral or octahedral sheets and most kaolinites are close to the ideal formula $Al_2Si_2O_5(OH)_4$. The 1:1 layer has little or no permanent charge because of the low amount of substitution. Consequently, cation exchange capacities and surface areas are typically low. Soils high in kaolinite are generally less fertile than soils in which 2:1 clay minerals dominate.

Kaolinite can form in soils from Al and Si released by the weathering of primary and other secondary minerals. For example, feldspars often weather to kaolinite in soils formed from igneous rocks. Kaolinite can also be inherited from clayey, sedimentary soil parent materials.

Halloysite Halloysite has a 1:1 layer structure similar to kaolinite except that the 1:1 layers are separated by a layer of H_2O molecules when fully hydrated (Figure 5). This water is probably present as hydration shells around a small number of interlayer cations (cations that reside between two adjacent 1:1 or 2:1 layers), although the presence of interlayer

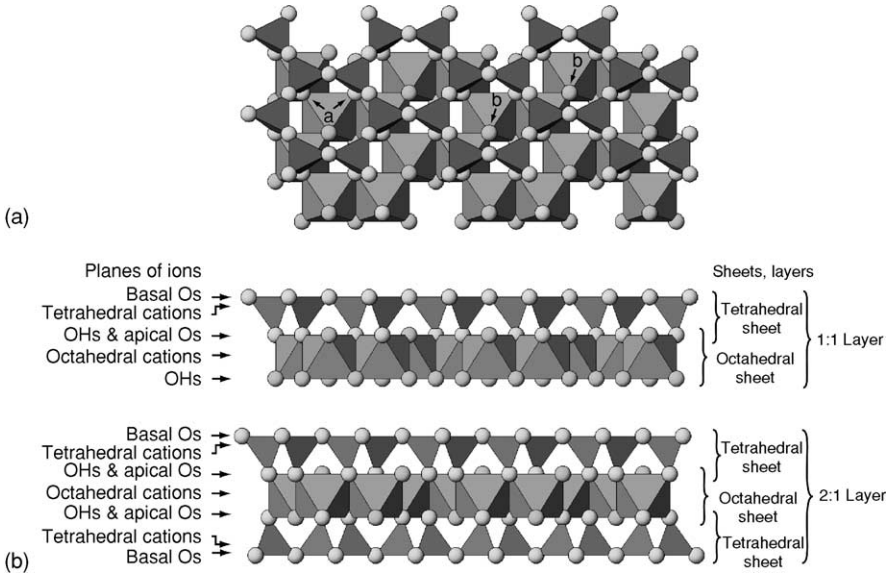


Figure 4 (a) Oblique view of the 1:1 layer structure illustrating the relationship between the tetrahedral and octahedral sheets. Note that adjacent apical oxygens of the tetrahedral sheet (arrows 'a') also define corners of octahedra in the octahedral sheet. Arrows marked 'b' point to OHs that lie directly in the center of the hexagonal rings of tetrahedra, although they appear off-center in this oblique view. (b) Edge view of the 1:1 and 2:1 layer structures, illustrating phyllosilicate nomenclature. (Adapted from Schulze DG (2002) An introduction to soil mineralogy. In: Dixon JB and Schulze DG (eds) *Soil Mineralogy with Environmental Applications*, pp. 1–35. Madison, WI: Soil Science Society of America, with permission.)

cations and the existence of layer charge to attract them has been difficult to confirm. Most clay silicates occur as thin plates, but halloysite often occurs as tubular or spherical particles.

Halloysite is usually found in soils formed from volcanic deposits, particularly volcanic ash and glass. It is a common clay mineral in the Andisol soil order. Halloysite forms early in the weathering process but it is generally less stable than kaolinite and gives way to kaolinite with time.

2:1-Type Minerals

In contrast to the 1:1 minerals, which are represented in soils by only two major minerals, the 2:1 minerals are structurally more diverse and are represented by several mineral species.

2:1 Layer structure The 2:1 layer structure consists of two tetrahedral sheets, with one bound to each side of an octahedral sheet (**Figure 4b**). There are four planes of anions. The outer two consist of the basal oxygens of the two tetrahedral sheets, while the two inner planes consist of oxygens common to the octahedral sheet and the two tetrahedral sheets, plus the hydroxyls belonging to the octahedral sheet.

Pyrophyllite The simple structure of pyrophyllite is a good starting point for discussing 2:1 structures. Pyrophyllite consists of 2:1 layers stacked one above the other. The tetrahedral sheets contain only Si^{4+} and the octahedral sheet contains only Al^{3+}, resulting in the ideal formula $Al_2(Si_4)O_{10}(OH)_2$. The charge is balanced completely within the 2:1 layer, making the layer electrically neutral, and adjacent 2:1 layers are held together only by weak van der Waals forces. Pyrophyllite occurs only rarely in soils, usually only when it is inherited from low-grade metamorphic rocks.

Micas Mica minerals have the 2:1 layer structure described for pyrophyllite but with two important differences. First, instead of having only Si^{4+} in the tetrahedral sites, one-quarter of the tetrahedral sites are occupied by Al^{3+}. Because of this substitution, there is an excess of one negative charge per formula unit in the 2:1 layer. Second, this excess negative charge is balanced by monovalent cations, commonly K^+, that occupy interlayer sites between two 2:1 layers (**Figure 5**). This gives an ideal formula of $KAl_2(AlSi_3)O_{10}(OH)_2$ for a mica mineral with Al in the octahedral sites.

The octahedral sheet can contain either Al^{3+} (the dioctahedral case; **Figure 3**) or Mg^{2+} (the trioctahedral case), and there are several different mica species, because Fe^{2+} and Fe^{3+} can substitute for Mg^{2+} and Al^{3+} in the octahedral sheet and Na^+ and Ca^{2+} can substitute for K^+ in the interlayer.

Mica in soils is usually inherited from the parent rock and is likely to occur in soils derived from various

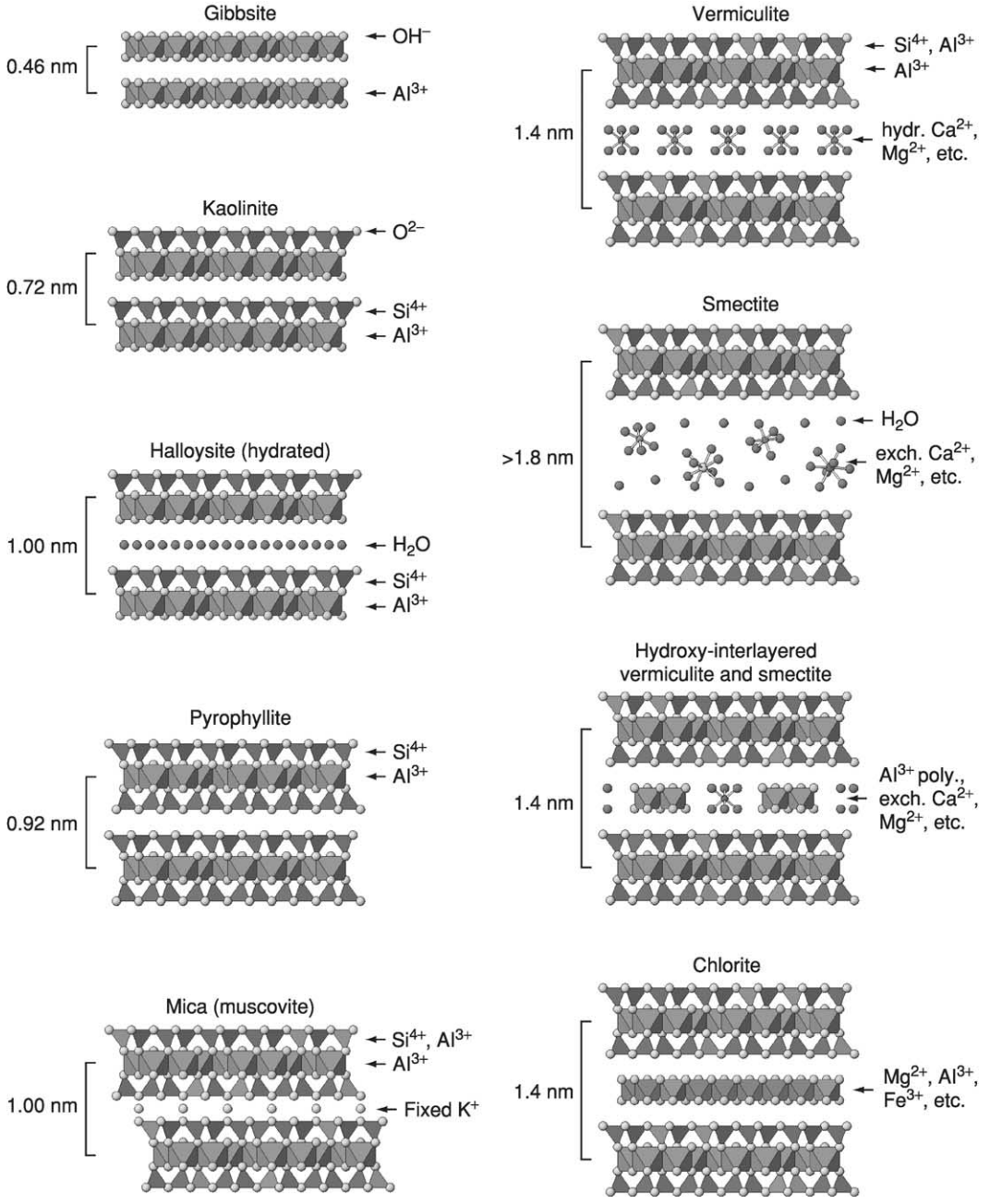


Figure 5 Structural scheme of soil minerals based on tetrahedral and octahedral sheets. (Adapted from Schulze DG (2002) An introduction to soil mineralogy. In: Dixon JB and Schulze DG (eds) *Soil Mineralogy with Environmental Applications*, pp. 1–35. Madison, WI: Soil Science Society of America, with permission.)

igneous and metamorphic rocks, as well as from sediments derived from them. Muscovite, biotite, and phlogopite are the three most common mica group minerals in rocks, and consequently in soils. All three contain K in the interlayer, but they differ in the composition of the octahedral sheet and whether they are di- or trioctahedral. Mica in the clay fraction of soils and sediments differs somewhat from the macroscopic muscovite mica it most closely resembles. This clay-size mica is often referred to as illite. Glauconite is another mica mineral that is similar to illite, but it contains more Fe and less Al in its octahedral sheet than illite.

Micas weather to other minerals, particularly to vermiculites and smectites, and the K^+ released during weathering is an important source of K for plants. As a rule, the dioctahedral micas such as muscovite are more resistant to weathering than trioctahedral

micas. Thus, muscovite tends to be the most common mica mineral in soils.

Vermiculites Vermiculite has a 2:1 layer structure as described for mica, but, instead of having a layer charge of ~1 per formula unit and K^+ in interlayer positions, vermiculite has a layer charge of 0.9–0.6 per formula unit and contains hydrated exchangeable cations, primarily Ca and Mg, in the interlayer (Figure 5). A typical formula for an idealized vermiculite weathered from muscovite is: $M^+_{0.75}Al_2(Si_{3.25}Al_{0.75})O_{10}(OH)_2$, where M^+ represents exchangeable cations. The high charge per formula unit gives vermiculite a high cation exchange capacity and causes vermiculte to have a high affinity for weakly hydrated cations such as K^+, NH_4^+, and Cs^+. Fixation of K^+ by vermiculite can be significant in soils that are high in vermiculite and that have not received large amounts of chemical fertilizers.

Vermiculites in soils are believed to form almost exclusively from the weathering of micas and chlorites. The weathering of micas to vermiculite (or smectite) is believed to occur by replacement of K^+ in the interlayer sites with hydrated exchangeable cations. The integrity of the 2:1 layer is preserved, but there is a reduction in the layer charge. Vermiculite does not swell as extensively as smectite and this is shown in Figure 5 by the presence of only two planes of water molecules surrounding the hydrated cations in the interlayer space.

Smectites The smectite group consists of minerals with the 2:1 structure already discussed for mica and vermiculite, but with a still lower charge per formula weight, namely 0.6–0.2. As in vermiculite, the interlayer contains exchangeable cations (Figure 5). An idealized formula for a common soil smectite, the mineral beidellite, is: $M^+_{0.33}Al_2(Si_{3.67}Al_{0.33})O_{10}(OH)_2$, where M^+ represents exchangeable cations, typically Ca^{2+} and Mg^{2+}.

The most common smectite minerals range in composition between three end-members: montmorillonite, beidellite, and nontronite. All are dioctahedral, but they differ in the composition of the tetrahedral and octahedral sheets. Smectites do not fix K^+ as readily as do vermiculites because smectites have a lower layer charge, but smectites swell more extensively than vermiculite. This is illustrated in Figure 5 by the larger spacing between the 2:1 layers.

Smectites are important minerals in temperate-region soils. Many plant nutrients are held in an available form on the cation exchange sites of soil smectites. Soils rich in smectite tend to be very effective at attenuating many organic and inorganic pollutants because of the high surface area and adsorptive properties of the smectites. Smectites shrink upon drying and swell upon wetting. This shrink–swell behavior is most pronounced in the Vertisol order and in vertic subgroups of other soil orders. The shrink–swell properties lead to cracking and shifting problems when houses, roads, and other structures are built on smectitic soils.

Chlorites Like mica, chlorite minerals have a 2:1 layer structure with an excess of negative charge. In contrast to mica, however, the excess charge is balanced by a positively charged interlayer hydroxide sheet (Figure 5), rather than K^+. The interlayer hydroxide sheet is an octahedral sheet as shown in Figure 3 and can be either di- or trioctahedral. Instead of being electrically neutral as in gibbsite, the hydroxide sheet has a positive charge caused by substitution of higher-valence cations for lower-valence ones, for example, $Mg_2Al(OH)_6^+$. Either octahedral sheet – the one that is part of the 2:1 layer or the interlayer hydroxide sheet – can be di- or trioctahedral, and can contain Mg^{2+}, Fe^{2+}, Mn^{2+}, Ni^{2+}, Al^{3+}, Fe^{3+}, and Cr^{3+}, giving a large number of different mineral species.

Chlorite minerals in soils are often primary minerals inherited from either metamorphic or igneous rocks. They may also be inherited from sedimentary rocks such as shales, or from hydrothermally altered sediments. Chlorites are rather infrequent minerals in soils and when present they generally occur in small amounts. Chlorite weathers to form vermiculite and smectite, and the ease with which chlorites break down makes them sensitive indicators of weathering.

Hydroxy-interlayered vermiculite and smectite Hydroxy-interlayered vermiculite and smectite can be considered a solid solution with vermiculite or smectite as one end-member and chlorite as the other. Hydroxy-interlayered minerals form as Al^{3+} released during weathering hydrolyzes and polymerizes to form large polycations with a postulated formula of $Al_6(OH)_{15}^{3+}$ (or similar) in the interlayers of vermiculite and smectite. These polycations balance some of the charge of the 2:1 layer. The combination of a 2:1 layer with hydroxy Al in the interlayer gives a structure similar to that of chlorite (Figure 5). Thus, these minerals are also called secondary chlorites. The degree of filling of the interlayer with hydroxy Al can vary from none to almost complete, with properties of the clay varying accordingly. The interlayer hydroxy Al is not exchangeable, therefore it lowers the cation exchange capacity of smectite or vermiculite almost linearly as a function of the amount of Al adsorbed in the interlayer.

Interlayer hydroxy Al prevents smectite from shrinking and swelling as it normally would. In

vermiculite, it reduces K^+ fixation by lowering the exchange capacity and by preventing the interlayer from collapsing around the K^+. The positively charged hydroxy interlayers also provide potential sites for anion adsorption. Hydroxy-interlayered vermiculite and smectite are most common in Alfisols and Ultisols. Within a given profile, they tend to be most abundant near the soil surface.

Interstratification in phyllosilicates Because of the structural similarities of all of the phyllosilicate minerals just discussed, phyllosilicates in soils do not always occur as discrete particles of mica, vermiculite, smectite, chlorite, or kaolinite. For example, instead of being made up of a stack of identical 2:1 vermiculite layers, one physically discrete particle may consist of a mixture of both mica and vermiculite layers instead. Such minerals are referred to as 'mixed-layer' or 'interstratified minerals.'

Different types of interstratified minerals have been identified. Two-component systems include: mica-vermiculite, mica-smectite, mica-chlorite, kaolinite-smectite, and others. Three-component mixed layer systems can also occur. The sequence of layers can be either regular or random. A regularly interstratified mineral consisting of two types of layers denoted by A and B could have a sequence such as ABABAB..., or ABBABBABB..., or any other repeating sequence. In a randomly interstratified mineral, the sequence of layers is random – for example, ABBABAABBAAA.... Random interstratification of layer-silicates is more common in soils than regular interstratification, though regular interstratification, especially in weathering micas, is not rare.

Partial removal of interlayer K from micas or of interlayer hydroxide from chlorite is one way that interstratified minerals can form in soils. Other possibilities include (1) fixation of adsorbed K^+ by some vermiculite layers to give mica-like layers, and (2) the formation of hydroxide interlayers to produce chlorite-like layers.

Palygorskite and sepiolite Palygorskite and sepiolite are considered phyllosilicates, but are distinct structurally from the typical 1:1 and 2:1 layer structures. Both minerals have continuous tetrahedral sheets, but adjacent bands of tetrahedra within one tetrahedral sheet point in opposite directions rather than in one direction as in the 1:1 and 2:1 structures. The result is a structure that can be described as ribbons of 2:1 layers joined at their edges, as illustrated in Figure 6. Water molecules occur in the spaces between the ribbons. The 2:1 ribbons are wider in sepiolite than in palygorskite.

Palygorskite and sepiolite are often found in soils of arid and semiarid environments. Both minerals have a fibrous morphology in contrast to the platy morphology of most 1:1 and 2:1 minerals.

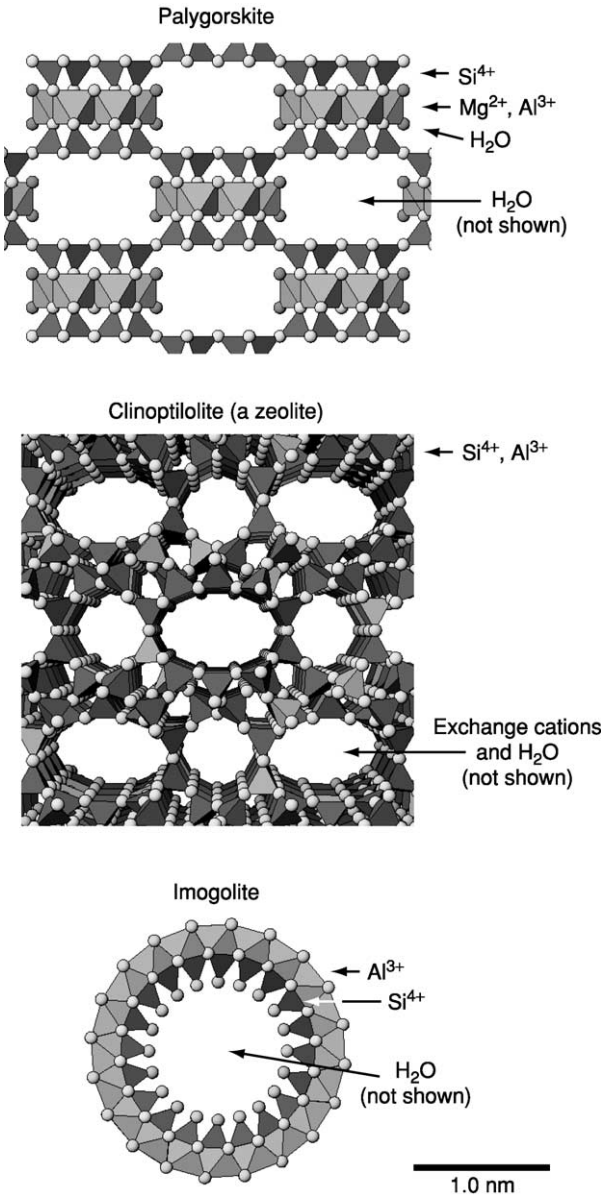


Figure 6 Structural models of representatives of three other aluminosilicate mineral groups that occur frequently in soils. All three are drawn to the same scale. (Adapted from Schulze DG (2002) An introduction to soil mineralogy. In: Dixon JB and Schulze DG (eds) *Soil Mineralogy with Environmental Applications*, pp. 1–35. Madison, WI: Soil Science Society of America, with permission.)

Other Minerals that Occur in Soil Clays

The phyllosilicate clay minerals described above are the most abundant and common in most soils, and they are the minerals that are usually considered to make up the 'clay minerals' group. Several additional minerals or mineral groups require mention as well. Some occur

only in particular soils, others occur in many soils, but usually only at low concentrations.

Zeolites

Zeolites are a large group of aluminosilicate minerals that consist structurally of SiO_4 tetrahedra arranged in ways that result in large amounts of pore space within the crystals (Figure 6). Aluminum substitutes for Si in the tetrahedral sites and, as a result, the $(Si,Al)O_4$ framework has a net negative charge. The charge is balanced by cations that reside in the channels and pores along with water molecules. Because the cations are exchangeable, zeolites have cation exchange properties similar to the phyllosilicates, but, because the tetrahedral framework of the zeolites is rigid and the size of the pores is fixed, small cations can move into and out of the pores freely, while larger cations are excluded. Thus, zeolites are often referred to as 'molecular sieves' because of their very selective cation exchange properties. Zeolites are relatively rare in soils because they weather easily in humid regions, but they occur in some soils in arid regions.

Allophane and Imogolite

The aluminosilicate minerals discussed above have three-dimensional crystal structures, with atoms packed together in a more or less regular manner over relatively long distances (10s of nanometers). They exhibit long-range order. Two other aluminosilicates, allophane and imogolite, exhibit short-range (or local) order. Structures with short-range order exhibit order over several nanometers, but on a larger scale the structure is disordered.

Allophane is a material consisting chemically of variable amounts of O^{2-}, OH^-, Al^{3+}, and Si^{4+}, and characterized by short-range order and a predominance of Si-O-Al bonds. It consists of small (3.5–5.0 nm) spheres, the structure of which has not been determined. The spheres clump together to form irregular aggregates. Imogolite consists of tubes several micrometers long with an outer diameter of 2.3–2.7 nm and an inner diameter of approx. 1.0 nm. The tubes consist of a single dioctahedral sheet with the inner surface OH replaced by SiO_3OH groups (Figure 6). Several individual tubes are arranged in bundles 10–30 nm across to give thread-like particles several micrometers long.

Allophane and imogolite usually occur as weathering products of volcanic ash and are important minerals in the Andisol soil order. Imogolite has also been identified in the Bs horizons of Spodosols. Allophane and imogolite can specifically adsorb many inorganic and organic compounds. Andisols, for example, usually fix large amounts of phosphate, making it unavailable to plants, and the large amounts of organic matter common in Andisols may be due, in part, to adsorption of organic molecules by allophane and imogolite. Soils containing large amounts of allophane and imogolite usually have unique physical properties such as a low bulk density, high water-holding capacity, high liquid and plasticity limits, and a thixotropic consistence.

Aluminum Hydroxide Minerals

Gibbsite, which has already been mentioned as simply a stack of octahedral sheets containing Al^{3+} in the octahedral sites (Figure 5), occurs in situations where weathering and leaching have been intense or long. Gibbsite often occurs in Ultisols and Oxisols and can be important in the fixation of phosphate fertilizers.

Iron Oxide Minerals

Fe is almost as abundant as Al in the Earth's crust, and one or more strongly colored iron oxide minerals are ubiquitous accessory minerals in almost all soil clays. Goethite (FeOOH) is the most common soil iron oxide mineral and accounts for the yellowish-brown colors of many soils. Hematite (Fe_2O_3) is common also and accounts for the red colors of many soils. Iron oxide surfaces are highly reactive and can sorb and fix phosphate, many transition metals, and organic compounds. Redox reactions are an important aspect of soil iron oxide mineralogy.

Manganese Oxide Minerals

Manganese is about 50 times less abundant in the Earth's crust than Fe, so minerals containing large amounts of Mn are proportionally less abundant. Nevertheless, some Mn oxide minerals occur in almost all soils, with birnessite ($(Na,Ca)(Mn^{3+},Mn^{4+})_7O_{14}\cdot 2.8H_2O$) being one of the more common ones. The Mn^{4+} and Mn^{3+} in manganese oxide minerals is reduced to soluble Mn^{2+} under relatively mild reducing conditions, making Mn oxide minerals important in many soil redox reactions.

Titanium Oxide Minerals

Ti is slightly more abundant than Mn in the Earth's crust, and two Ti minerals, anatase and rutile (both TiO_2) occur widely in soil clays. The Ti-oxide minerals, however, are relatively inert chemically and have a negligible impact on most soil properties.

Carbonates, Sulfates, and Soluble Salts

In semiarid to arid climates, calcite ($CaCO_3$), gypsum ($CaSO_4\cdot 2H_2O$), and an array of evaporite minerals (minerals with solubilites greater than or equal to gypsum) often occur coassociated with the

aluminosilicate and oxide minerals discussed above. The impact of these minerals varies greatly, depending on the amount and type of mineral present.

List of Technical Nomenclature

Minerals	gibbsite, kaolinite, halloysite, pyrophyllite, mica, illite, glauconite, muscovite, smectite, montmorillonite, beidellite, nontronite, vermiculite, chlorite, hydroxy-interlayered smectite and vermiculite, interstratified clay minerals, allophane, imogolite, palygorskite, sepiolite, zeolites, iron oxide minerals, goethite, hematite, aluminum hydroxide minerals, manganese oxide minerals, birnessite, titanium oxide minerals, anatase, rutile, calcite, gypsum, evaporite minerals, biotite, phlogopite

See also: **Kinetic Models**; **Thermodynamics of Soil Water**

Further Reading

Brindley GW and Brown G (eds) (1980) *Crystal Structures of Clay Minerals and Their X-ray Identification*. London, UK: Mineralogical Society.

Dixon JB and Schulze DG (2002) *Soil Mineralogy with Environmental Applications*. Madison, WI: Soil Science Society of America.

Dixon JB and Weed SB (1989) *Minerals in Soil Environments*. Madison, WI: Soil Science Society of America.

Klein C and Hurlbut CS Jr (1993) *Manual of Mineralogy* (after James D. Dana), 21st edn. New York: John Wiley.

Moore DM and Reynolds RC Jr (1997) *X-ray Diffraction and Identification and Analysis of Clay Minerals*, 2nd edn. New York: Oxford University Press.

CLIMATE CHANGE IMPACTS

P Bullock, Cranfield University–Silsoe, Silsoe, UK

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Soils form through the interaction of a number of influences, including climate, relief and/or landscape, parent material, organisms (including fauna, flora, and humans) and time. The nature of this interaction varies in different parts of the world, resulting in several thousand types of soil worldwide. It takes thousands of years for a soil to form, and most soils are still evolving as a result of changes in some of these soil-forming factors, particularly climate and vegetation, over the past few millennia. Changes in any of the soil-forming factors, such as climate, will impact directly and indirectly on current soils, with important implications for their development and use.

Unraveling the likely extent and impact of climate change on soils is a complex process and one in which progress has been slow. It is made all the more complicated by the fact that not only can soils be strongly affected directly and indirectly by climate change, but soils themselves can act as both source and sink for greenhouse gases and thus have the potential for either positive or negative feedback to climate change. The lack of specificity of the global circulation models (GCMs) at present, combined with the complexity of the interaction of the various soil-forming processes and the fact that there is still a limited knowledge of many of them, particularly biological ones, makes it difficult to quantify the changes that will ensue. On the basis of current knowledge, it is only possible to describe the likely impacts of climate change on soils in a qualitative or semiquantitative way and highlight the key changes, their direction where there is adequate climate change information, and the implications of them.

Climate Change Predictions

Estimates of global climate change are continually being revised as models are being improved and new data collected. The current main sources of information available on the likely extent of climate change are the Third Assessment Reports of the Intergovernmental Panel on Climate Change (IPCC).

These reports conclude that the globally averaged surface temperature will increase by between 1.4 and 5.8°C over the period 1990 to 2100 and that nearly all land areas will warm by more than this global average, particularly in northern high latitudes in the cold season (Table 1). Increased summer continental drying and associated risk of drought is likely over most midlatitude interiors, the main areas for which the models are consistent with one another.

Table 1 Projected changes in extreme weather during the twenty-first century

Changes in weather phenomenon	*Confidence in projected changes*
Higher maximum temperatures and more hot days over nearly all land areas	Very likely
Higher minimum temperatures, fewer cold days and frost days over nearly all land areas	Very likely
Reduced diurnal temperature over most land areas	Very likely
Increase in heat index[a] over land areas	Very likely, over most areas
More intense precipitation events	Very likely, over many areas
Increased summer continental drying and associated risk of drought	Likely over most midlatitude continental interiors (lack of consistent projections in other areas)
Increase in tropical cyclone wind intensities	Likely, over some areas
Increase in tropical mean and peak precipitation intensities	Likely, over some areas

[a]Heat index: a combination of temperature and humidity that measures effects on human comfort.
Adapted from Houghton JT, Ding Y, Griggs DJ *et al.* (eds) (2001) *Climate Change 2001: The Scientific Basis*, p. 72. Contribution of Working Group I to the Third Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge, UK: Cambridge University Press, with permission.

Global average precipitation is projected to increase during the twenty-first century, with larger year-to-year variations in precipitation and more intense precipitation events very likely. Depending on the regional scale, both increases and decreases in precipitation are likely to be seen. Models predict an increase in precipitation in both summer and winter in high-latitude regions. In winter, increases in precipitation are predicted over northern midlatitudes and tropical Africa, and in summer in south and east Asia. There is likely to be a decrease in rainfall in Australia, Central America, and South Africa. Increased intensity of midlatitude storms, tropical peak wind intensities, and mean and peak precipitation intensities are also likely. Global mean sea level is projected to rise by 0.09–0.88 m between 1990 and 2100, and this will impact on soils of coastal regions. These changes, particularly in combination, are likely to have a major impact on soil formation, soil processes, land use, and rates of land degradation, especially soil erosion, desertification, and salinization. The Third Assessment Reports, although providing much-needed synthesis of model predictions of climate change, give little attention to the actual impacts of these changes on soils.

Timescale for Change

The diverse range of physical, chemical, and biological processes that affect soil formation and modify soil properties will respond to climate change according to varying timescales (Table 2). Parameters such as bulk density, porosity, infiltration rate, permeability, nitrate content, and composition of soil air can change on a daily basis, depending on the weather. At the other end of the timescale, weathering of minerals as part of soil formation and changes in soil texture are more likely to be on millennial timescales.

The effect of climate change will be to modify the rates of these processes and lead to changes in soil properties, with a range of implications for soil formation, soil genesis, and the way in which soils can be used.

Climate Change Impacts on Soil Water and Soil Temperature

Soil Water

The main effects of climate change on soils will be through changes to soil-moisture regimes. Soil moisture is a key driver to most soil processes and is instrumental in the use that can be made of soils. As climate changes, soil-moisture levels will be influenced by direct climatic effects (precipitation, temperature effects on evaporation), climate-induced changes in vegetation, different plant growth rates and different cycles, different rates of soil-water extraction, and the effect of enhanced CO_2 levels on plant transpiration. Changes in soil-water fluxes may also feed back to the climate itself and even contribute to drought conditions by decreasing available moisture, altering circulation patterns, and increasing air temperatures.

Soil water can be influenced in a number of ways by climate change. Changes in precipitation will rapidly affect soil water, since the timescale for response to rainfall in the soil is usually within a few hours. Increasing temperatures will also lead to greater evapotranspiration and hence loss of water from the soil. Much will also depend on land use and vegetation cover, as these will influence water use.

Several soil-forming processes, including organic matter turnover, structure formation, weathering, podzolization, clay translocation, and gleying, are strongly affected by soil-moisture contents. The type

Table 2 Timescale for changes in soils with change in climate

Timescale categories	*Soil parameter*	*Properties and characteristics*	*Horizons and phases*	*Regimes*
$<10^{-1}$ years	Temperature; moisture content; bulk density; total porosity; infiltration rate; permeability; composition of soil air; nitrate content	Compaction; drainage; workability		Aeration; heat regime
10^{-1}–10^{0} years	Total water capacity; field capacity; hydraulic conductivity; pH; nutrient status; composition of soil solution	Microbiota		Microbial activity; human-controlled plant-nutrient regime; erosion
10^{0}–10^{1} years	Wilting percentage; soil acidity; cation exchange capacity; exchangeable cations	Type of soil structure; annual roots biota; mesofauna; litter, fluvic, gleyic, stagnic properties; slickensides	Sulfuric horizon; gelundic, inundic, salic, yermic phases (fine earth properties only)	Moisture; natural fertility; salinity–alkalinity; desertification; permafrost
10^{1}–10^{2} years	Specific surface; clay mineral association; organic matter content	Tree roots; soil biota; salic, calcareous, sodic, vertic properties	Histic (<20 cm), ochric, gypsic, albic, and immature natric and spodic horizons (Podsols); gilgai, placic, sodic, takyric phase	
10^{2}–10^{3} years	Primary mineral composition; chemical composition of mineral part	Tree roots; color (yellowish/reddish); iron concretions; soil depth; cracking; soft powdered lime; indurated subsoil	Histic, mollic, umbric, calcic, albic, natric, cambic, spodic, and nitic horizons; plinthite, placic, yermic phases (stone surfaces)	
$>10^{3}$ years	Texture; particle-size distribution; particle density	Parent material; depth; abrupt textural change	Argic, oxic, petrocalcic, petrogypsic horizons; duripan, fragipan, skeletic, petroferric, lithic, rudic phases	

Adapted from Varallyay GY (1990) Influence of climatic change on soil-moisture regime, texture, structure, and erosion. In: Scharpenseel HW, Schomaker M, and Ayoub A (eds) *Soils on a Warmer Earth*, Development in Soil Science 20, p. 46. Amsterdam, the Netherlands: Elsevier, with permission.

of soil structure that develops under a particular climatic regime is particularly important, because it affects the processes of runoff, infiltration, percolation, and drainage, processes that are vital in the distribution of water across the landscape.

Those areas predicted to have warmer temperatures and less rainfall will have less soil moisture, with potentially large implications for the crops that can be grown and the natural and seminatural ecosystems that can continue to exist. Given these circumstances, arid areas which are already marginal for agriculture may become totally unsuitable for agriculture, new areas will become classed as arid lands, with the associated difficulties of maintaining an agricultural base and survival of its ecosystems, and a larger proportion of the world's land will become unsuitable for agricultural production. However, in view of the many different interacting influences on soil-moisture levels, it is difficult to predict the impact of climate change on soil water at regional or local level.

Soil Temperature

There is a close relationship between air temperature and soil temperature, and an increase in air temperature leads to an increase in soil temperature. The temperature regime of the soil is governed by gains and losses of radiation at the surface, the process of evaporation, heat conduction through the soil profile, and convective transfer via the movement of gas and water. As with soil moisture, soil temperature is a prime factor in most soil processes. Warmer soil temperatures everywhere will accelerate soil processes, leading to more rapid decomposition of organic matter, increased microbiological activity, quicker release of nutrients, increased rates of nitrification, and generally increased chemical weathering of minerals. However, this effect could be minimized or reversed if soils become drier.

Gelisols (*see* **Cold-Region Soils**), one of the two orders of soils largely defined on a climatic basis, are particularly vulnerable to increases in temperature. Gelisols occupy large areas of the northern latitudes where thickening of the seasonally thawed layer above permafrost is predicted. Large areas of permafrost are likely to begin to thaw, with consequent changes in soil drainage, increased mass movement, and thermal erosion likely. There is thus likely to be a significant change in the nature and the distribution of Gelisols. Aridisols, the other group of soils with a strong link to climate, cover more than 12% of the globe. Already such soils are very difficult to manage, are at the margin of cultivation, and are subject to soil erosion, desertification, and salinization. Increasing temperatures, accompanied by increased evaporation, would exacerbate these problems.

Changes in Soil-Forming Processes and Properties

Soil Organic Matter

Soil organic matter is arguably the most important soil component, influencing soil structure, water-holding capacity, soil stability, nutrient storage and turnover, and oxygen-holding capacity, properties that are fundamental in maintaining and improving soil quality. (*see* **Organic Matter:** Principles and Processes.) A decline in organic matter content increases the susceptibility to soil erosion. Organic matter is particularly important as the prime habitat for immense numbers and variety of soil fauna and microflora, which play a critical role in the health and productivity of soils. It is highly susceptible to changes in land use and management and to changes in soil temperature and moisture. In the last decades of the twentieth century, changes in land use and management, particularly conversion of forest and grassland to agriculture, have led to a significant decline in organic matter levels in some parts of the world.

Soil organic matter is one of the major pools of carbon in the biosphere and, unlike most other soil properties, is important both as a driver of climate change (*see* **Carbon Emissions and Sequestration**) and as a response variable to climate change, capable of acting both as a source and sink of carbon. How climate change will impact on soil organic matter is a matter of considerable debate. On the one hand it is recognized that global warming and increasing CO_2 levels in the atmosphere can favor increased plant growth, which in turn would increase organic inputs to the soil. On the other hand, a rise in air temperature and that of the soil are likely to increase decomposition and loss of soil organic matter. Soil organic matter represents a major pool of carbon in the biosphere, estimated at about 1500 Gt of carbon, double that in the atmosphere at present. There is thus significant interest in the fate of such carbon, particularly the extent to which soils and land use may sequester carbon from the atmosphere or lose organic carbon to the atmosphere. The balance of opinion currently is that, in the absence of mitigating action, losses through organic matter decomposition are likely to exceed levels gained from increased plant growth, thus adding to atmospheric CO_2 levels and the greenhouse gas effect, and to lower levels of soil organic matter.

A group of world soils that are particularly vulnerable to climate change are peats (Histosols). (*see*

Organic Soils.) These are soils that are dominantly composed of organic matter throughout their whole depth. Already they have been under threat worldwide because of drainage for use in crop production. Further drying-out of the soils in a warmer, drier climate with concomitant oxidation could lead to losses of this important, highly productive soil type, in addition to releasing a large amount of carbon to the atmosphere as CO_2.

Soil Structure

The structure of the soil, that is, the way in which the soil particles combine together (*see* **Structure**), is an important property of soils, affecting the movement of gases, water, nutrients, soil fauna, and the emergence of crops. The nature and quality of the structure of particular soils are strongly influenced by the amount and quality of organic matter present and also by the inorganic constituents of the soil matrix, cultivation methods, and by natural physical processes such as shrink–swell and freeze–thaw. A decline in soil organic matter levels as could occur under climate change would lead to a decrease in soil aggregate stability, an increase in susceptibility to compaction, lower infiltration rates, increased runoff, and hence an increased susceptibility to erosion.

The structure of the soil is also important for water quality, seepage, and building foundations. Soils with high clay contents, particularly those with smectitic mineralogy, have the potential to shrink when dry, resulting in formation of large cracks and fissures. When the soils rewet again, the cracks close. Drier climatic conditions would be expected to increase the frequency and size of crack formation. The importance of soil structure in determining pathways for the movement of water, pollutants, and contaminants through the soil is now well recognized and future management of soil water, movement of nutrients within the soil and the landscape, and the movement of contaminants once in the soil are important aspects to be considered in a climate change environment. In some areas, as a result of increased cracking and change in structure, there could be an increase in flash-flooding as water moves more rapidly through or over the soil to rivers.

Areas that are likely to experience increased drought may also find that buildings, roads, etc., built to particular specifications relating to current conditions, have foundations which become unstable as soils dry out more. For example, the increase in subsidence to buildings in southeast England alone in the decade to 1997, the warmest recorded, cost the insurance industry several billion euros.

Clay soils with a high shrink–swell potential are termed Vertisols and occupy some 260 million hectares globally. They are important agricultural soils in Africa and Asia but are renowned as some of the problem soils of the world, owing to difficulty in managing them for cultivation. They become very hard and difficult to cultivate when dry and too plastic for trafficking or cultivation when wet. Changes in climate, particularly increased drying of the soil, will lead to increased difficulties in managing these high-clay-content soils and may also lead to other world soils developing the properties of this soil type.

Soil Fauna and Soil Flora

Soil fauna and flora are essential components of all soils. Particularly vital is their role in the retention, breakdown, and incorporation of plant remains, nutrient cycling, and their influence on soil structure and porosity. There are thousands of species in a square meter of most soils but, despite these numbers and their importance, little is known about the roles of the species. Global warming may not have a direct effect on the ecological composition, because soil fauna and flora have a relatively broad temperature optimum. However, changes in ecosystems and migration of vegetation zones are likely in some areas as a result of increased temperature and changes in rainfall. Soil flora and fauna may be seriously affected by such changes, because their migration rates are likely to be too slow.

A further significant impact of climate change on soil fauna and flora is through enhanced CO_2 levels in the atmosphere, leading to enhanced plant growth and in turn increased allocation of carbon belowground. The microbial population and its activity under this regime would increase, potentially leading to higher rates of nitrogen fixation, nitrogen immobilization and denitrification, increased mycorrhizal associations, increased soil aggregation, and increased weathering of minerals. However, as noted above, much will depend on what balance between increased plant growth on the one hand and increased decomposition of soil organic matter on the other under a changing climate will emerge.

Acidification and Nutrient Status

While temperature increases are forecast for most parts of the world, there is less certainty about precipitation changes. Significant increases in rainfall will lead to increases in leaching, loss of nutrients, and increasing acidification, depending on the buffering pools existing in soils. Decreases in rainfall coupled with warmer, more evaporative regimes as are forecast for Australia, Southern Africa, Central America, and southern Europe will increase evaporation, making increased salinization a major risk. The

direction of change toward increased leaching or increased evaporation will depend on the extent to which rainfall and temperature change. In either case the situation could lead to important changes in world soils.

Changes in Land-Degradation Processes

Land degradation is already one of the major problems affecting the world. Currently some 6–7 million hectares are lost annually through soil erosion, desertification affects about one-sixth of the world's population and one-quarter of the world's land, and salinization affects some 20 million hectares of irrigated land. Land degradation through damage to the soil is a serious problem and its causes are often complex and interwoven. Severe damage has already been done to the world's soils, and the impact of climate change needs to be considered in parallel with the effect of the existing pressures on the land. It is difficult to separate the effects of these various impacts and their cumulative impact on soils is often greater than a simple summation.

Soil Erosion

Soil erosion is the movement and transport of soil by various agents, particularly water, wind, and mass movement; hence climate is a key factor. It has been recognized as a major problem since the 1930s and, although there has been some 70 years of research into the causes and processes, it is still increasing and of growing concern. Global rates of soil erosion have been exceeding those of new soil formation by 10- and 20-fold on most continents of the world in the last few decades. The increase in soil erosion to date is strongly linked with the clearance of natural vegetation, to enable land to be used for arable agriculture, and the use of farming practices unsuited to the land on which they are practiced. This, combined with climatic variation and extreme weather events, has created ideal conditions for soil erosion. The main climatic factors influencing soil erosion are rainfall (amount, frequency, duration, and intensity), and wind (direction, strength, and frequency of high-intensity winds), coupled with drying-out of the soil. Land use, soil type, and topography are the other key factors.

Soil erosion by water is more widespread and its impact greater than that by wind. Climate change is likely to affect soil erosion by water through its effect on rainfall intensity, soil erodibility, vegetative cover, and patterns of land use. General circulation models predict for many areas seasonally more intense drying, coupled with increased amounts and intensity of precipitation at other times, conditions that could lead to large increases in rates of erosion by water.

Soil erosion also occurs by wind transport of soil particles by suspension, surface creep, or saltation over distances ranging from a few centimeters to hundreds of kilometers. Wind erosion not only transports soil particles around arid and semiarid landscapes but inputs into ecosystems and may even alter global climatic patterns. Wind erosion is particularly a problem on sandy and organic soils, which are subject to intermittent low-moisture contents and periodic winds. Currently wind erosion is mainly a feature of arid and semiarid conditions. Those areas where climate change is predicted to lead to more droughty soils under increasing temperatures will become increasingly vulnerable to wind erosion. Although general circulation models have in the past been unable to predict changes in wind speed and frequency with any certainty, the latest models are predicting increased summer continental drying and risk of drought in midlatitude areas and an increase in tropical cyclone peak intensities in some areas, both sets of conditions favoring an increase in soil erosion by wind. However, it is important to note that erosion is site-specific, and different permutations of conditions can increase or decrease it.

In the last few decades of the twentieth century, significant advances have been made in modeling erosion risk under different climate change scenarios. Research in the USA, using the EPIC model for two different sets of climatic conditions at 100 sites in the US Corn Belt, has shown that mean water erosion varies approximately linearly with mean precipitation, with approximately a 40% change for a 20% change in mean precipitation (Figure 1). By contrast, with a 20% increase in wind speed, erosion increased fourfold. This suggests that wind erosion is potentially more sensitive to climate change than is water erosion and secondly that, for wind-erosion predictions, it is important to understand and predict wind-speed threshold. These results and others using different models suggest that increased wind speed, rainfall amount and intensity and increased frequency of high-wind events are likely to lead to significant increases in soil erosion, thus exacerbating the already serious situation.

Desertification

Desertification is the process of ecological degradation by which economically productive land becomes less productive, in some cases leading to the development of a desert-like landscape. There is a huge literature on the nature and causes of desertification,

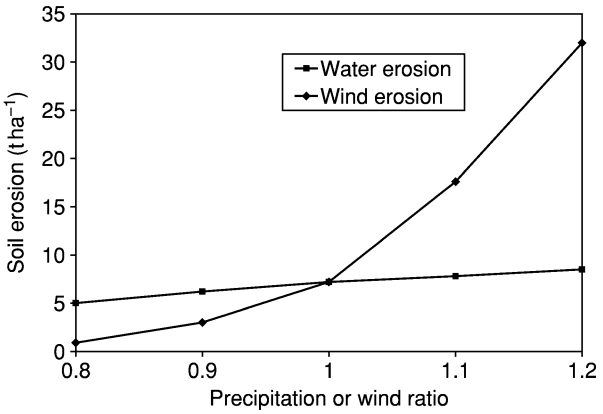


Figure 1 Sensitivity of soil erosion in the US Corn Belt to climate change as estimated using EPIC. For water erosion, temperature, CO_2, and wind speed were held at current values, while precipitation volume (expressed as ratio to current volumes) was varied. For wind erosion, temperature, CO_2, and precipitation volume were held at current values, while wind speed (expressed as ratio to current speeds) was varied. Each point represents the 100-year mean of 100 randomly selected sites. (Reproduced with permission from Ingram J, Lee J, and Valentin C (1996) *Journal of Soil and Water Conservation* 51(5): 378.)

some of which indicates that human impacts arising from overstocking, overcultivation, and deforestation are primarily responsible for the process; some lays the blame on the impact of extended droughts over the last few decades. In reality, desertification is likely to be due to a combination of drought and mismanagement of land, particularly where there is a lack of harmony between land use and management on the one hand and prevailing climate on the other.

Desertification occurs mainly in hyperarid, arid, semiarid, and subhumid climatic zones, ranging from precipitation:potential evapotranspiration indices (P:PET) of less than 0.05 to 0.70. The area of land occupied by these four zones in which true or induced deserts can occur is 47% of the planet. With the increased temperatures and evaporation predicted, this percentage could increase.

Salinization

The accumulation of salts in soils can negatively affect soil properties and processes. It can lead to degradation of soil structure, decline in porosity, and increase in bulk density, and impede nutrient dynamics and nutrient-holding capacity. It leads to a decline in productivity and can cause land to become unsuitable for agricultural production. Soil salinity is already a major global problem in Australia, Africa, Latin America, and the Near and Middle East. It is typically found in areas where evaporation exceeds precipitation. In the USA alone, some 50 million hectares of cropland and pasture is affected by salinity, and the area of land so affected is growing by some 10% annually. Salinization is likely to be prevalent in two situations:

- *Coastal zones*. In these situations salinity is likely to depend primarily on the sea level and its tidal, seasonal, and long-term fluctuations. Given that the global mean sea level is projected to rise by between 9 and 88 cm between 1990 and 2100, there is likely to be a territorial extension of coastal salinity under the direct and indirect effects of saline seawater. Impacts are likely to include flooding of coastal plains by saline seawater, including low-lying coastal fringes, marshlands and swamps, and deltas and estuaries of some of the world's big rivers, increase in storm tides affecting areas several meters in elevation around the coast, with penetration of saline and brackish water inland, and rapid erosion of coastlines.
- *Continental salt transport and salt accumulation processes*. There are three principal mechanisms of this form of salinization: salt accumulation, seepage, and wind deposition. Salinization by salt accumulation happens when leaching is reduced and salt accumulates at or near the surface. It can also occur when salt is leached into a perched water table and the water moves through the landscape to areas of lower elevation, where the water evaporates, leaving the salts. Given suitably exposed deposits of salts, wind erosion can transfer the salts elsewhere in the landscape.

All of these causes of salinity are likely to be enhanced with climate change. Zones subject to rising temperatures during the summer will experience more evapotranspiration and aridity and thus higher concentrations of salt in the soil solution. There may also be enhanced capillary rise from shallow water tables, with the potential to carry salt into the overlying soil horizons. Where climate change leads to increasing wetness as well as temperature, there will be a reduction in the concentration of salts in the salt solution. However, where there is periodic drying also involved, capillary upward transport of salts can take place from shallow groundwater to overlying soil horizons.

Given the predictions for climate change, the balance will be for the area of salt-affected soils to increase, although there may be an improvement to a few of the existing saline areas of the world, albeit a very slow one.

Climate Change Impacts on Soil Functions

The most recognized function of soils is their use for agriculture and the implications this has for feeding the global population. Soil properties, as well as

Table 3 Soil functions and climatic change

Function	*Use of soils*	*Impacts of climate change*
Economic	Food crops, energy crops, timber	Changes in land-use capability
	Sand, gravel, minerals	Changes of productivity
	Foundations for buildings, roads, etc.	Erosion and salinization
		Nitrate leaching and increased use of fertilizers (wetter conditions)
		Need for increased irrigation (drier conditions)
		Changes to foundations of buildings
		Higher building insurance premiums
Ecologic	Habitats for soil fauna and microflora	Loss of some habitats
	Food for ground feeders, e.g., birds	Stress on many habitats
	Nutrient supply and storage	Changes in soil biodiversity
	Cycling of water and air	Likely loss of soil organic matter
		Changes in soil fertility
		Acidification (wetter conditions)
		Loss of peat habitats (drier conditions)
Hydrologic	Water storage, flow control and runoff, absorption, amelioration	Water resources problems (drier conditions)
		Transfer of salts
		Transfer of nutrients
		Increased bypass flow (drier conditions)
		Erosion sediment transfer
Pollutant control	Source and sink for pollutants	Increased bypass flow (drier conditions)
	Waste disposal medium	Erosion and movement of pollutants
		Changes between sources and sinks
Gaseous exchanges	Source and sink of greenhouse gases	Increased decomposition of organic matter (CO_2 release)
		Sequestration of carbon
		Increased/decreased N_2O release (wetter/drier conditions)
		Increased/decreased CH_4 release (wetter/drier conditions)
Heritage protection	Protection of buried archeological sites	Effects of erosion and sedimentation
		Loss of peat
		Cracking

climate, are the main factors governing which crops can be grown, where they can be grown, and their productivity. Changes in the physical, chemical, and biological properties of soil have major implications for agriculture. Changes in soil and air temperature and in rainfall will affect the ability of crops to reach maturity and their potential harvest. Reductions in amounts of soil water may be compensated by irrigation, but scarcity of water may preclude the use of water for irrigation. The increases in land degradation, whether in the form of soil erosion, desertification, salinization, or loss of peat soils, will further impact on the capability of soils to supply the needs of agriculture. If the current predictions of climate change are borne out there is likely to be a shift in agricultural zones, requiring a movement of population, new farming methods, development of new skills, or some combination of these. The rate of change may be sufficiently slow for readjustment to take place in line with the pace of climate change. The main problem areas are likely to be those experiencing increasing aridity where the options for successful change are limited.

There are, however, several other soil functions (Table 3) that have a major influence on the quality of life. Soils are fundamental to land-based ecosystems and, as with agriculture, in association with climate, govern the nature and distribution of the world's natural and seminatural ecosystems, providing water, nutrients, and a growing medium. Although most ecosystems are naturally variable with time and adapt reasonably well to small climatic fluctuations, changes to soils as a result of climate change are likely to be reflected in a change in ecosystems. The form the change will take depends much on the degree of warming and the changes to precipitation, and prediction of the changes will require significantly improved climate change models at regional level.

As discussed earlier, soil is the habitat for an immense number and variety of organisms, important in the decomposition of soil organic matter and the recycling of nutrients on which plant productivity and ultimately life depend.

Soil properties are important in controlling the fate and behavior of water once it reaches the surface of the soil. The soil system influences runoff, flow, storage, and regional distribution of water. There are likely to be major changes to the distribution of water under a changed climate, and the soil will be a

central issue in managing the water regime under these new conditions. The soil also plays an important part as a medium supporting buildings and other infrastructure, and protecting our archeological heritage. Current infrastructure will almost certainly need to change to meet the new climatic conditions. Finally, climate change concerns associated with the levels of greenhouse gases in the atmosphere have focused on the large amounts of carbon stored in the soils, and the ability of the soil to act as a source and sink for carbon. The fact that the soil can act as both a source and sink for several greenhouse gases, e.g., CO_2, N_2O, and CH_4, has already become an important issue in the quest to reduce the impacts of climate change.

See also: **Carbon Emissions and Sequestration**; **Cold-Region Soils**; **Desertification**; **Organic Matter:** Principles and Processes; **Organic Soils**; **Structure**

Further Reading

Houghton JT, Ding Y, Griggs DJ *et al.* (eds) (2001) *Climate Change 2001: The Scientific Basis*. Contribution of Working Group I to the Third Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge, UK: Cambridge University Press.

IPCC (1996) *Climate Change 1995: Impacts, Adaptations and Mitigation of Climate Change: Scientific–Technical Analyses*. Cambridge, UK: Cambridge University Press.

McCarthy JJ, Canziani OF, Leary NA, Dokken DJ, and White KS (eds) (2001) *Climate Change 2001: Impacts, Adaptation and Vulnerability*. Contribution of Working Group II to the Third Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge, UK: Cambridge University Press.

Metz B, Davidson O, Swart R, and Pan J (eds) (2001) *Climate Change 2001: Mitigation*. Contribution of Working Group III to the Third Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge, UK: Cambridge University Press.

Parry ML (2000) *Assessment of Potential Effects and Adaptations for Climate Change in Europe: The Acacia Project*. Jackson Environmental Institute. Norwich, UK: University of East Anglia Press.

Rosenzweig C and Hillel D (1998) *Climate Change and the Global Harvest. Potential Impacts of the Greenhouse Effect on Agriculture*. Oxford, UK: Oxford University Press.

Rounsevell MDA and Loveland PJ (1994) *Soil Responses to Climate Change*. NATO ASI Series. Berlin, Germany: Springer-Verlag.

Rounsevell MDA, Bullock P, and Harris JA (1996) Climate change, soils and sustainability. In: Taylor AG, Gordon JE, and Usher MB (eds) *Soils, Sustainability and the Natural Heritage*. Edinburgh, UK: HMSO.

Rounsevell MDA, Evans SP, and Bullock P (1999) Climate change and agricultural soils: impacts and adaptations. *Climatic Change* 43: 683–709.

Scharpenseel HW, Schomaker M, and Ayoub A (eds) (1990) Soils on a warmer earth. *Developments in Soil Science* 20.

CLIMATE MODELS, ROLE OF SOIL

P Smith, University of Aberdeen, Aberdeen, UK

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Models of climate change usually focus on fluxes of carbon between land (vegetation and soil) and the atmosphere, because carbon dioxide is the most prevalent greenhouse gas. Consequently, the soil components of these models focus on soil organic carbon (SOC) dynamics. Though methane and nitrous oxide are also important biogenic greenhouse gases, and some models include these gases, it is models of SOC dynamics that are described here. Ecosystem models are increasingly used to predict the impacts of climate change on soils and other components of ecosystems, and to investigate the feedbacks between climate change and soil or other ecosystem processes. Most of the modeling of climate change impacts has occurred independently of the modeling of climate change itself, but recently climate models have included fully coupled biospheric carbon models to allow the feedback between biospheric carbon and atmospheric processes to be explored.

Approaches to Modeling SOC Dynamics for the Study of Climate Change

There are a number of approaches to modeling the turnover of SOC, including: (1) process-based, multi-compartment models – this approach is by far the most common; (2) models that consider each fresh addition of plant debris as a separate cohort which decays in a continuous way; and (3) models that account for C and N transfers through various trophic

levels in a soil food web, often termed 'food-web models' or 'organism-oriented models.'

Process-Based, Multicompartment Models of SOC Dynamics

These models focus on the processes mediating the movement and transformations of matter or energy and usually assume first-order rate kinetics. Early models treated SOC as one homogeneous compartment having a single, first-order rate constant. Later two-compartment models were proposed and, as computers became more accessible, multicompartment models were developed. Of the 33 SOC models currently represented within the Global Change and Terrestrial Ecosystems (GCTE)–Soil Organic Matter Network (SOMNET) database, 30 are multicompartment, process-based models. Each compartment or SOC pool within a model is characterized by its position in the model's structure and its decay rate. Decay rates are usually expressed by first-order kinetics with respect to the concentration (C) of the pool:

$$dC/dt = -kC \qquad [1]$$

where t is time. The rate constant (k) of first-order kinetics is related to the time required to reduce by half the concentration of the pool when there is no input. The pool's half-life ($h = (\ln 2)/k$), or its turnover time ($\tau = 1/k$) are sometimes used instead of k to characterize a pool's dynamics: the lower the decay rate constant, the higher the half-life, the turnover time, and the stability of the organic pool.

The flows of C within most models represent a sequence of C moving from plant and animal debris to the microbial biomass, then to soil organic pools of increasing stability. Figure 1 shows the flow of C in the Rothamsted C model.

Some models also use feedback loops to account for links between decomposition (catabolic processes) and the synthesis of new polymers from the breakdown products (anabolic processes) mediated by different groups of microbes. The output flow from any organic pool is usually split, some carbon being directed to a microbial biomass pool, some to other organic pools, and, under aerobic conditions, some to CO_2. This split simulates the simultaneous anabolic and catabolic activities and growth of a microbial population feeding on one substrate. Two parameters are required to quantify the split flow. They are often defined by microbial (utilization) efficiency and stabilization (humification) factors, which control the flow of decayed C to the biomass and humus pools, respectively. The sum of the efficiency and humification factors must be less than 1 to account for the release of CO_2. Unlike food-web models, multicompartment, process-based models generally ignore the separate roles of different organisms involved in organic matter turnover in soil. Instead, they concentrate on the overall processes and quantities of C being transformed.

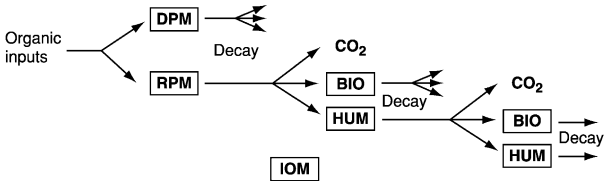


Figure 1 Flows of carbon in the Rothamsted carbon model. BIO, microbial biomass; DPM, decomposable plant material; IOM, inert organic matter; HUM, humified OM; RPM, resistant plant material.

Cohort Models Describing Decomposition as a Continuum

Another approach to modeling SOC turnover is to treat each fresh addition of plant debris into the soil as a cohort. Such models consider one SOC pool that decays with a feedback loop into itself. Some models are represented by a single rate equation with the organic C pool divided into an infinite number of components, each characterized by its 'quality' with respect to degradability, as well as impact on the physiology of the decomposers. The rate equations for such models represent the dynamics of each organic C component of quality q, which is quality-dependent. Exact solutions to the rate equations are obtained analytically.

Food-Web Models

Another type of model simulates C and N transfers through a food web of soil organisms; such models explicitly account for different trophic levels or functional groups of biota in the soil. Some models have been developed which combine an explicit description of the soil biota with a process-based approach. Food-web models require a detailed knowledge of the biology of the system to be simulated and are usually parameterized for application at specific sites.

Factors Affecting the Turnover of Soil Organic Carbon in Models

Rate 'constants' (k), used in all models, are constant for a given set of biotic and abiotic conditions. To adjust a rate constant for nonoptimum environmental circumstances, the simplest way is to modify the maximum value of k using a reduction factor μ, ranging from 0 to 1. Environmental factors considered by SOC models include temperature, water,

pH, nitrogen, oxygen, clay content, cation exchange capacity, type of crop/plant cover, and tillage.

Many studies show the effect of temperature on microbially mediated transformations in soil, either expressed as a reduction factor or the Arrhenius equation. Water and oxygen have a major impact on microbial physiology. While some models simulate O_2 concentrations in soil explicitly, many define the extent of anaerobiosis based on soil pore space filled with water. Soil clay content and total SOC are usually correlated: in general, a soil with a high clay content will contain more organic matter than a soil with a lower clay content if management and climate are similar. This is because clay surfaces have a major role in the stabilization of organic matter through absorption on surfaces. Various schemes simulate the effect of clay on rate equations to obtain SOC accumulation. N is an essential element for microbial growth which will be maximal when enough N is assimilated to maintain the microbial C:N ratio. Table 1 presents an overview of the 33 models represented in the GCTE-SOMNET, including the factors affecting SOC turnover.

Model Performance

There are many reasons for evaluating the performance of an SOC model. Model evaluation shows how well a model can be expected to perform in a given situation. This is important if, for example, a scientist, farmer, or policy-maker wishes to use a model to predict the impact of climate change or a change in land management on SOC content. A model can only be used with confidence in this predictive way if it has been tested previously against real data. In a scientific sense, it is also instructive to evaluate a model against data, because this can help to improve understanding of the system – especially when the model fails, as this may highlight a process, or a factor controlling a rate, that is not properly simulated in the model. Models can be evaluated at a number of different levels. They can be evaluated at the individual process level or at the level of a subset of processes (e.g., net mineralization), or the models' overall outputs (e.g., changes in total SOC over time) can be tested against measured laboratory and field data. Models can also be evaluated for their applicability in different situations, e.g., for scaling-up simulated net C storage from a site specific to a regional level.

Figure 2 shows an example of the simulation of two commonly used SOC models using the same data set. In both cases, the models can be compared with measured data graphically and statistically to indicate the overall error and any evidence of systematic bias.

Models Used to Study the Impacts of Climate Change

Many ecosystem models have been used to examine the potential impact of climate change on ecosystem components. Many of these models include some description of the soil, though the detail with which soil processes are described tends to decrease with the scale of application. Site-specific applications of ecosystem models allow detailed mechanistic descriptions of soil processes to be described, whereas models applied at continental, biome, or global scales (e.g., the Vegetation/Ecosystem Modelling and Analysis Project (VEMAP) exercise) tend to have much more simplistic descriptions of the soil. Models applied at the global scale often have a very simple description of the soil. Many examples exist of the use of models to examine the impacts of climate change, but, in most of these studies, climate change is used to drive the models, with no explicit feedback between the biosphere and the climate system. In the next section, some recent work in which climate models have been linked to biospheric C cycle models to allow the feedback between climate and biosphere to be examined.

Climate Models with Coupled Description of Biospheric Carbon Feedbacks

Until recently, climate (general circulation) models (GCMs) had no soil C component at all, relying instead upon a purely physical description of the Earth's climate system. Recently, however, climate models with a fully coupled biospheric C module (including soil C) have been developed. The group of P. Cox at the UK Hadley Centre and the group of P. Freidlingstein of the Laboratoire des Sciences de l'Environnement, Paris, have independently developed climate models that include biospheric C feedback.

Though the models are very different, both suggest (to different degrees) that the feedback between climate and biospheric carbon leads to accelerated loss of C from the biosphere, which in turn leads to accelerated climate change. One of the key processes leading to this acceleration is the response of soil respiration to increased temperature. The findings from studies using coupled biospheric C and climate models are subject to many uncertainties, but do highlight the need for including soils in assessments of future climate. Future developments require that the process level understanding developed at the site level, using mechanistic process-based models be incorporated in these coupled GCMs so that the full impact of the feedback of soil C change on climate can be quantified.

Table 1 Overview of soil organic matter models represented within the Global Change and Terrestrial Ecosystems–Soil Organic Matter Network (GCTE-SOMNET) database in August 2002

	Inputs					
Model	*Time step*	*Meteorology*[a]	*Soil and plant*[b]	*Management*[c]	*Factors affecting decay rate constants*[d]	*Soil outputs*[e]
ANIMO	Day, week, month	P, AT, Ir, EvW	Des, Lay, Imp, Cl, OM, N[f], pH	Rot, Ti, Fert, Man, Res, Irr, AtN	T, W, pH, N, O	C, N, W, ST, gas
APSIM	Day	P, AT, Ir	Lay, W, C, N, BD, Wi, PG, PS	Rot, Ti, Fert, Irr	T, W, pH, N	C, N, W, ST, gas
Candy	Day	P, AT, Ir	D, Imp, W, N, C, Wi, PD, Nup	Rot, Ti, Fert, Man, Res, Irr, AtN	T, W, N, Cl	C, N, W, ST, gas
CENTURY	Month	P, AT	W, Cl, OM, pH, C, N	Rot, Ti, Fert, Man, Res, Irr, AtN	T, W, N, Cl, pH, Ti	C, BioC, 13C, 14C, N, W, ST, gas
Chenfang Lin Model	Day	ST	OM, BD, W	Man, Res	T, W, F	C, BioC, gas
DAISY	Hour, day	P, AT, Ir, EvG	Lay, Cl, C, N, PG, PS	Rot, Ti, Fert, Man, Res, Irr, AtN	T, W, N, Cl	C, BioC, N, W, ST, gas
DNDC	Hour, day, month	P, AT	Lay, Cl, OM, pH, BD	Rot, Ti, Fert, Man, Res, Irr, AtN	T, W, N, Cl, Ti	C, BioC, N, W, ST, gas
DSSAT	Hour, day, month, year	P, AT, Ir	Des, Lay, Imp, W, Cl, PS, OM, pH, C, N	Rot, Ti, Fert, Man, Res, Irr	T, W, N, Cl, Ti	C, BioC, N, W, ST
D3R	Day	P, AT	Y, PS	Rot, Ti, Res	T, W, N, Cv, Ti	Decomposition of surface and buried residue
Ecosys	Minute, hour	P, AT, Ir, WS, RH	Lay, W, Cl, CEC, PS, OM, pH, N, BD, PG, PS	Rot, Ti, Fert, Man, Res, Irr, AtN	T, W, N, O, Cl, Cv	C, BioC, N, W, ST, pH, Ph, EC, gas, ExCat
EPIC	Day	P, AT	Lay, Imp, W, Cl, OM, pH, C, BD, Wi	Rot, Ti, Fert, Man, Res, Irr, AtN	T, W, N, pH, Cl, Ce, Cv	C, BioC, N, W, ST
FERT	Day	P, AT, WS	Des, Lay, W, Cl, OM, pH, C, N, BD, W, Ph, K, Nup, Y, PS	Rot, Ti, Fert, Man, Res, Irr	T, W, N, pH, Cv	C, N, Ph, K
ForClim-D	Year	P, AT	W, AG	None	T, W	C
GENDEC	Day, month	ST, W	W, InertC, LQ	Can be used; not essential	T, W, N	C, BioC, N, gas, LQ
HPM/EFM	Day	P, AT, Ir, WS	W, Cl, PS	Rot, Fert, Irr, AtN	T, W, N	C, BioC, N, W, gas
ICBM	Day, year	Combination of weather and climate	Many desirable; none essential	C inputs to soil	T, W, Cl	C
KLIMAT-SOIL-YIELD	Day, year	P, AT, ST, Ir, EvG, EvS, VPD, SH	Des, Lay, Imp, W, Cl, PS, OM, pH, C, N	Fert, Man, Res, Irr	T, W, N, Cl	C, BioC, N, W, ST
CNSP Pasture Model	Day	P, AT, Ir	Lay, Imp, W, Cl, CEC, OM, pH, C, N, PS, AS	Fert	T, W, N, pH	C, N, W, ST
Humus Balance	Year	Climate based on P and AT	Des, Lay, PS, OM, pH, C, N	Rot, Fert, Man	N, H, Cl, Cv	C, N
MOTOR	User-specified	P, AT, EvG	Des, OM	Rot, Ti, Fert, Man	T, W, N, Cl, Ti	C, BioC, 13C, 14C, gas
NAM SOM	Year	P, AT	Des, PS, OM, Ero	Man, Res	T, W, Cl, Cv	C, BioC
NCSOIL	Day	ST (P, AT)	W, OM, C, N	Fert, Man, Res	T, W, N, pH, Cl, Ti	C, BioC, 14C, N, 15N, gas

(*Continued*)

Table 1 (*Continued*)

Model	Inputs				Factors affecting decay rate constants[d]	Soil outputs[e]
	Time step	Meteorology[a]	Soil and plant[b]	Management[c]		
NICCE	Hour, day	P, AT, Ir, WS	Imp, OM, C, N, W, TC, PG	Fert, Man, Res, Irr, AtN	T, W, Cl, N	C, BioC, 13C, 14C, N, 15N, W, ST, gas
O'Brien Model	Year	None	Lay, C, 14C	None	None	C, 14C
O'Leary Model	Day	P, AT	Lay, W, Cl, pH, N	Ti, Fert, Res	T, W, N, Cl, Ti	C, BioC, N, W, ST, gas, ResC, ResN
Q-Soil	Year	Optional	C, N	Rot, Fert, Man, Res, AtN	T, W, N	C, BioC, 13C, N
RothC	Month	P, AT, EvW	Cl, C, InertC (can be estimated)	Man, Res, Irr	T, W, Cl, Cv	C, BioC, gas, 14C
SOCRATES	Week	P, AT	CEC, Y	Rot, Fert, Res	T, W, N, Cv, Ce	C, BioC, gas
SOMM	Day	P, ST	OM, N, AshL, NL	Man	T, W, N	C, N, gas
Sundial	Week	P, AT, EvG	Imp, Cl, W, Y	Rot, Fert, Man, Res, Irr, AtN	T, W, N, Cl	C, BioC, N, 15N, W, gas
Verberne	Day	P, AT, Ir, WS, EvS	Des, W, Cl, PS, OM, C, N	Man, AtN	T, W, N, Cl	C, BioC, N, W
VOYONS	Day, week, month	P, ST	Cl, OM, C, N	Fert, Man, Res, Irr, AtN	T, W, Cl	C, BioC, 13C, 14C, N, gas
Wave	Day	P, AT, Ir, EvG	Lay, OM, C, N, W, PG	Rot, Ti, Fert, Man, Res, Irr, AtN	T, W, N	C, N, W, ST, gas

[a]P, precipitation; AT, air temperature; ST, soil temperature; Ir, irradiation; EvW, evaporation over water; EvG, evaporation over grass; EvS, evaporation over bare soil; WS, wind speed; RH, relative humidity; VPD, vapor pressure deficit; SH, sun hours.

[b]Des, soil description; Lay, soil layers; Imp, depth of impermeable layer; Cl, clay content; OM, organic matter content; N, soil nitrogen content/dynamics; C, soil carbon content/dynamics; InertC, soil inert carbon content; pH, pH; W, soil-water characteristics; Wi, wilting point; PD, soil particle-size distribution; CEC, cation exchange capacity; Ero, annual erosion losses; BD, soil bulk density; TC, thermal conductivity; PG, plant growth characteristics; PS, plant species composition; AS, animal species present; AG, animal growth characteristics; Y, yield; Nup, plant nitrogen uptake; LQ, litter quality; AshL, ash content of litter; NL, N content of litter.

[c]Rot, rotation; Ti, tillage practice; Fert, inorganic fertilizer applications; Man, organic manure applications; Res, residue management; Irr, irrigation; AtN, atmospheric nitrogen inputs.

[d]T, temperature; W, water; pH, pH; N, nitrogen; O, oxygen; Cl, clay; Ce, cation exchange capacity; Cv, cover crop; Ti, tillage; F, Fauna.

[e]BioC, biomass carbon; 13C, ^{13}C dynamics; 14C, ^{14}C dynamics; 15N, ^{15}N dynamics; gas, gaseous lossess (e.g., CO_2, N_2O, N_2); ResC, surface residue carbon; ResN, surface residue nitrogen; Ph, phosphorus dynamics; K, potassium dynamics; EC, electrical conductivity; ExCat, exchangeable cations.

[f]N in the soil inputs and outputs section is used to denote all aspects of the N cycle. (*See* **Civilization, Role of Soils**.)

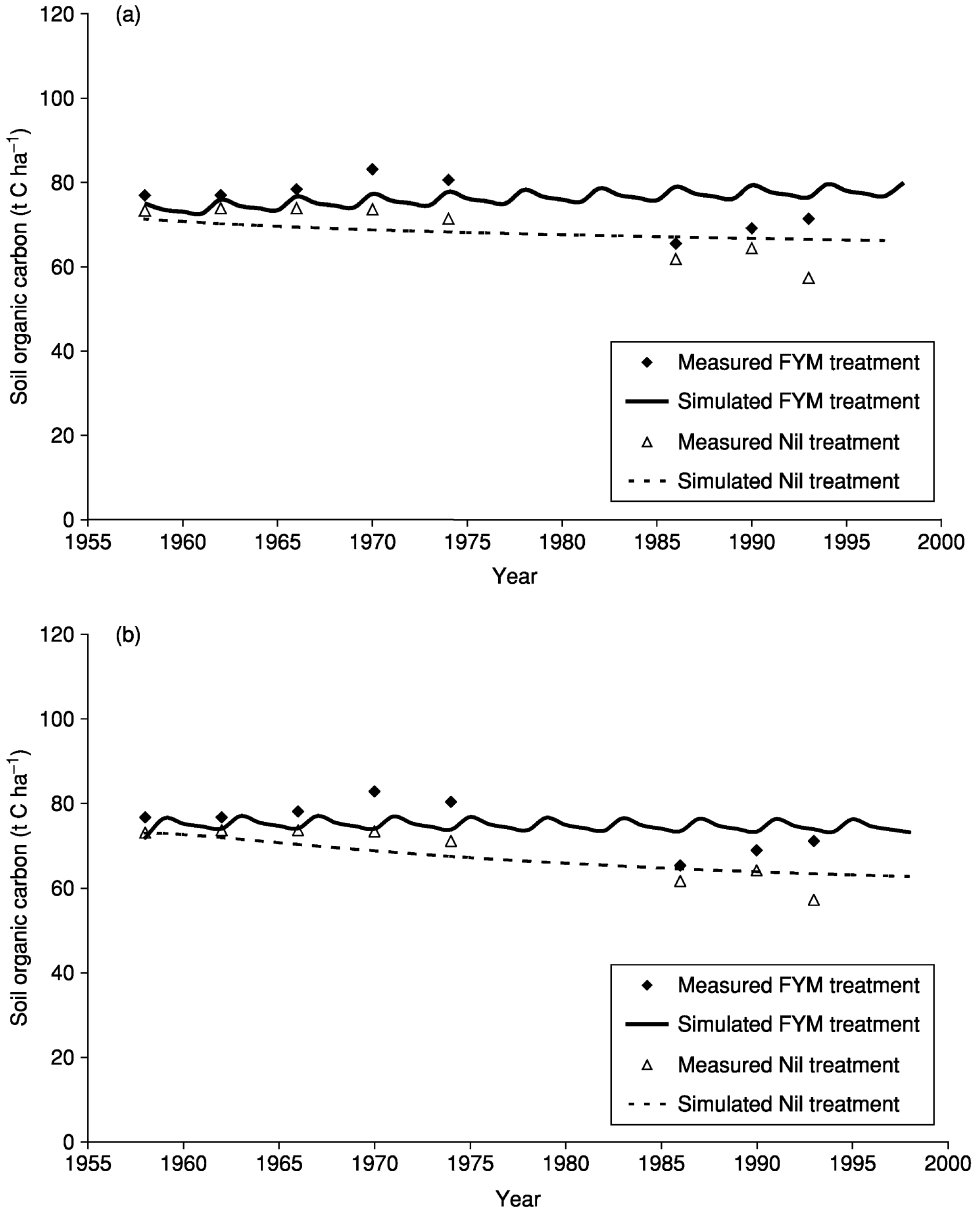


Figure 2 Simulated (lines) and measured (symbols) values of soil organic carbon for two treatments (farmyard manure (FYM) and nil inputs), at the Martonvasar long-term cropping experiment in Hungary, as simulated by two models: Century (a) and RothC (b). Root-mean-square error (RMSE) values for each simulation are Century: FYM = 7.8, Century: nil = 5.3, RothC: FYM = 8.0 and RothC: nil = 7.3.

List of Technical Nomenclature

GCM	general circulation model
GCTE	Global Change and Terrestrial Ecosystems
k	rate constant
RMSE	root-mean-square error
SOC	soil organic carbon
SOMNET	Soil Organic Matter Network

See also: **Carbon Cycle in Soils:** Dynamics and Management; **Carbon Emissions and Sequestration**; **Civilization, Role of Soils**; **Climate Change Impacts**; **Fertility**; **Organic Matter:** Principles and Processes; **Soil–Plant–Atmosphere Continuum**; **Sustainable Soil and Land Management**

Further Reading

Bosatta E and Ågren GI (1995) Theoretical analyses of the interactions between inorganic nitrogen and soil

organic matter. *European Journal of Soil Science* 76: 109–114.

Cox PM, Betts RA, Jones CD, Spall SA, and Totterdell IJ (2000) Acceleration of global warming due to carbon-cycle feedbacks in a coupled climate model. *Nature* 408: 184–187.

Falloon P and Smith P (2002) Simulating SOC dynamics in long-term experiments with RothC and Century: model evaluation for a regional scale application. *Soil Use and Management* 18: 101–111.

Friedlingstein P, Bopp L, Ciais P *et al.* (2001) Positive feedback between future climate change and the carbon cycle. *Geophysical Research Letters* 28: 1543–1546.

Hunt HW, Trlica MJ, Redente EF *et al.* (1991) Simulation model for the effects of climate change on temperate grassland ecosystems. *Ecological Modelling* 53: 205–246.

McGill WB (1996) Review and classification of ten soil organic matter (SOM) models. In: Powlson DS, Smith P, and Smith JU (eds) *Evaluation of Soil Organic Matter Models Using Existing, Long-Term Datasets*, pp. 111–133. NATO-ASI 138. Berlin, Germany: Springer-Verlag.

Melillo JM, Kicklighter DW, McGuire AD, Peterjon WT, and Newkirk KM (1995) Global change and its effect on soil organic carbon stocks. In: Zepp RG and Sonntag CH (eds) *Role of Nonliving Organic Matter in the Earth's Carbon Cycle*, pp. 175–189. New York: John Wiley.

Molina JAE and Smith P (1998) Modeling carbon and nitrogen processes in soils. *Advances in Agronomy* 62: 253–298.

Powlson DS, Smith P, and Smith JU (eds) (1996) *Evaluation of Soil Organic Matter Models Using Existing, Long-Term Datasets*. NATO ASI 138. Berlin, Germany: Springer-Verlag.

Smith P, Powlson DS, Smith JU, and Elliott ET (eds) (1997) Evaluation and comparison of soil organic matter models. *Geoderma* 81: 1–225.

Smith P, Andrén O, Brussaard L *et al.* (1998) Soil biota and global change at the ecosystem level: describing soil biota in mathematical models. *Global Change Biology* 4: 773–784.

Smith P, Falloon P, Powlson DS, and Smith JU (2001) *Soil Organic Matter Network (SOMNET): 2001 Model and Experimental Metadata*, 2nd edn. GCTE Report No. 7. Wallingford, UK: GCTE Focus 3 Office.

VEMAP (1995) Vegetation/Ecosystem Modelling and Analysis Project: comparing biogeography and biogeochemistry models in a continental-scale study of terrestrial responses to climate change and CO_2 doubling. *Global Biogeochemical Cycles* 9: 407–437.

COLD-REGION SOILS

C-L Ping, University of Alaska Fairbanks, Palmer, AK, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Soils of cold regions are those soils formed in areas with a mean soil temperature of generally less than 8°C and having soil formation controlled by cryogenesis (or cryogenic processes). Cryogenic processes involve the formation of ice and subsequent frost heave caused by the increased volume. These soils are distributed throughout high latitudes or areas with high altitudes. The soil temperature strongly influences geochemical and biological processes within the pedosphere. Soil organic matter accumulation is one of the most important features in soils of a cold region. The rate of organic matter accumulation increases with latitudes and altitude, and maximizes at approximately 50–65° N. Above latitude 65°, the rate of accumulation decreases more as a result of lower inputs, but at the same time the rate of decomposition decreases with the increasing latitude. The cold temperatures in high latitudes and altitudes cause freeze–thaw cycles that result in unique cryogenic fabrics and cryoturbated soil horizons. Many of these soils have permafrost that dominates the soil processes in the tundra zone and the northern boreal zone. The polar zone is less affected because it is too dry.

Definition of Cold Soils

Cold soils generally refer to the soils formed in high latitudes and at high altitudes, and they are identified based on vegetation belts by ecologists and on soil temperature by soil scientists. The northernmost region is the polar desert in the High Arctic, where only sparse cushion plants and lichens grow. In these regions, the climate is severe, with mean annual precipitation (MAP) less than 200 mm and air temperatures in the warmest summer month only about 4°C, and mean annual soil temperatures (MAST) at 50 cm being less than −8°C. The region below the High Arctic is the tundra zone where the ericaceous plants and mosses dominate the upland tundra; moss and

sedges dominate the coastal plains. In the tundra zone, the annual precipitation ranges from 200 to 400 mm and the warmest summer air temperatures are less than 10°C, with MAST ranging from −2 to −8°C. The region below the tundra zone is the boreal zone, where the summer air temperature may reach 20°C, the MAST is less than 8°C, and the annual precipitation is more than 200 mm. The vegetation in the boreal zone is dominantly boreal forest, or taiga. Generally, the polar and the arctic zones correspond with the zone of continuous permafrost and the boreal zone corresponds to the zone of discontinuous permafrost. Permafrost is defined as a thickness of soil or other superficial deposit, or even of bedrock, that remains frozen for a consecutive 2 years or more.

Characteristics and Genesis of Cold-Region Soils

What separates the cold-region soils from those of the temperature and tropical regions is their cryogenic nature and the unique properties resulting from freeze–thaw cycles and the formation of ground ice. Due to ice-lens formation during the freeze cycle, a granular structure usually forms in the A horizons of soils associated with earth hummocks, and lenticular and reticulate structures form in subsoils (Figure 1). Commonly an ice-rich layer occurs just above the permafrost table. Thirdly, the presence of permafrost has a profound impact on soil morphology due to its impact on many of the normal soil-forming factors and the fact that it leads to cryoturbation. Cryoturbation is a very common phenomenon not only in permafrost-affected areas, but also in the alpine zone, where the soils are subjected to strong freeze–thaw cycles but lack permafrost. Cryoturbation results in warped soil horizons, mixing of horizons and, most importantly, frost-churning of surface organic matter downward to the upper permafrost and thus sequesters carbon (Figure 2).

Physical Properties

Boreal soils Most of the soils in the northern boreal zone have permafrost, but all soils in this zone are subjected to seasonal freeze–thaw cycle. Those without permafrost are defined as soils with cryic soil temperature regimes in *Soil Taxonomy*. These soils have MAST of less than 8°C with a cool summer. Most of these soils have vegetation dominated by conifer forest or boreal forest. When soil freezes, soil water moves to the freezing front, thus forming a thin layer of ice called an ice lens. The progressive freezing creates a stratified fabric that consists of alternate layers of soil and ice lens. After repeated freeze–thaw cycles, the finer soil particles become orientated along the plate surface. Thus a distinctive, thin platy structure forms in the early stage of this process. With time, the expansion of ice deforms the flat plates into discontinuous, small curved plates referred to as 'lenticular' fabrics by geocryologists (Figure 3). It is also common in the seasonal-frozen as well as some of the permafrost soils; a crusty, or crumb-like structure forms on the surface of the soils. This is caused by needle-ice formation during early winter when water is extracted from belowground and freezes on the soil surface to form needle ice below the very top, thin layer of the soils.

Arctic soils In areas strongly affected by permafrost, the soils have more strongly expressed cryogenic structures. Generally the lenticular structures form in mineral soil horizons in the upper and lower sections of the active layer, directly above the permafrost. The active layer is the portion of the soil profile above the permafrost that is subjected to seasonal freeze and thaw. Soils in the middle section of the active layer have weaker and thicker lenticular or coarse blocky structures. The reason for this is that during freezing the water not only moves to the top due to a descending freezing front, but also moves to the bottom, where the permafrost serves as another freezing front. On top of the permafrost table, there is a zone that has horizontal and vertical cracks, like an ice net (Figure 1). This is caused by the freeze cracks during the freeze-up in early winter. An ice-rich layer is formed by repeated freezing and thawing, and during each thawing more water enters the cracks and later becomes frozen. Thus, in the upper part of this zone, an angular, blocky structure forms and, in the lower part of this zone, the ice-rich layer often contains more than 70% ice. The mineral soil blocks seem to be 'floating' in the ice matrix. When the permafrost table drops and the water drains, a rectangular-shaped lattice forms reticulate structures (Figure 4).

Alpine and high arctic soils Some of the cold-region soils are in the alpine region, where the MAST may be the same as those in the boreal and the arctic regions, but they experience such strong diurnal temperature fluctuations that freeze–thaw cycles are common even in the midst of the summer. In the alpine and the High Arctic regions, rock fragments not only are subject to physical weathering due to a strong freeze–thaw cycle, but also grind against each other during the process, thus producing finer particles, especially silt. This process is referred to as 'underground glaciation' by Russian scientists. Thus many rock fragments in the soil profile are silt-capped.

Figure 1 Tundra soil formed in Yedoma deposit near the Arctic Ocean coast, NE Russia. Note the thin organic horizon (O), cryoturbated A horizon (Ajj), gleyed mineral horizon (Bg), ice-net formation (BC), reticulate structure, ice-rich layer (Cf), and ice wedge (Wfm).

Figure 2 Cryoturbated soil formed in earth hummocks, Northwest Territory, Canada. Note the bare soils at the center of the hummock, thick organic horizon in the trough between the hummocks, and the frost-churned organic matter at 50–70 cm.

Morphological Properties

In soils affected by seasonal frost with or without permafrost, the top horizons usually contain more ice then the underlying horizons, owing to the desiccation process described earlier. During thawing, if the soil drainage is impeded, the topsoil remains saturated for longer then the underlying zone, thus often a reduced or a gleyed layer forms near the top of the profile. However, the permafrost plays an even bigger role in soil morphology. During the growing season, the permafrost table acts as a barrier not only to water movement but also to roots, thus creating a reduced zone, because the water table is positioned on top of the permafrost. Thus the lower active layer is often gleyed (Figures 1 and 3). In areas affected by permafrost, the ground cracks during the winter. The intervening cracks create a unique land pattern called 'polygons.' Each year the melting water enters into the cracks and then freezes in the winter. The ice vein eventually thickens and becomes an ice wedge. The formation of ice wedges increases the internal volume of the soils, and the soils buckle up at the edge of the

Figure 3 Riverbank erosion caused by thawing of ice wedges under boreal forest, Duvany Yar, Lower Kolyma region, NE Russia. The surface horizontal layers belong to the active layer (a), ice wedge at left (b), and distorted frozen soil at right (c).

polygon. Eventually, the center of the polygon raises and forms an earth hummock (Figure 2). The center of the hummock becomes better drained than the surrounding low areas, where it is poorly drained, and thus contains more water, which drives the process even more. More organic matter accumulates and portions of the organic matter get squeezed between two hummocks and eventually are 'frost-churned' into the lower part and mixed with the underlying mineral horizons (Figure 2). In the process, the horizontal soil horizons become warped and distorted by the process called cryoturbation.

Another phenomenon unique to the soils with cryogenic processes is the frost boil. Frost boils appear on the ground surface as areas of the mineral soil material 'boiled' from underneath the surface, forming circles. The vegetation is pushed to the edge of the frost boil. Although there are many theories as to how frost boils form, field morphology suggests that it is due to diaparism; that soils have different viscosity and when the surrounding soils become frozen the unfrozen soils in the center get squeezed out. In a well-developed frost-boil soil profile, it is clear that organic matter has been cryoturbated into the lower soil horizons (Figures 5 and 6). Even though the formation processes involved in earth hummocks and frost boils may not be the same, they all result in similar soil morphology: frost-churned organic matter in the lower horizons; and warped and broken soil horizons.

Soils formed in high mountain, alpine, and high plateau regions have characteristic morphological features. Although these soils also have cryogenic fabrics such as lenticular and reticulate structures, due to the general arid environment, these structures are more abundant in the lower active layers. In the exposed ridges and slopes, sorted circles are very common. In sorted circles, soils form in the center of the circles, whereas the rock fragments form the circles.

Biological and Chemical Properties

Temperature and moisture conditions control soil-profile development, microbial respiration in soil, and hence the character of soil organic matter. Temperature and moisture affect the accumulation and distribution of C stocks in the soil profile through controls on vegetation, cryic soil processes (such as cryoturbation) and the presence of or depth to permafrost. One of the most important features in cold soil formation is the accumulation of organic matter. Based on recent studies, the estimated total carbon stored in the peatlands in the Arctic and the boreal zones of Alaska ranges from 32–80 $kg\,m^{-2}$ to more than 130 $kg\,m^{-2}$, respectively. This agrees with Ovenden's finding that peat accumulation is a function of climate and landscape position and it maximizes in the boreal region. Most of the organic soils are in the boreal regions in the Northern Hemisphere: northern Russia, the Canadian Shield, and interior Alaska. Thus the widest distribution of organic soils is in the cryic zone and the boreal zone. Recent study has indicated that, in the arctic tundra soils, nearly 50% of the total organic carbon is stored in the upper

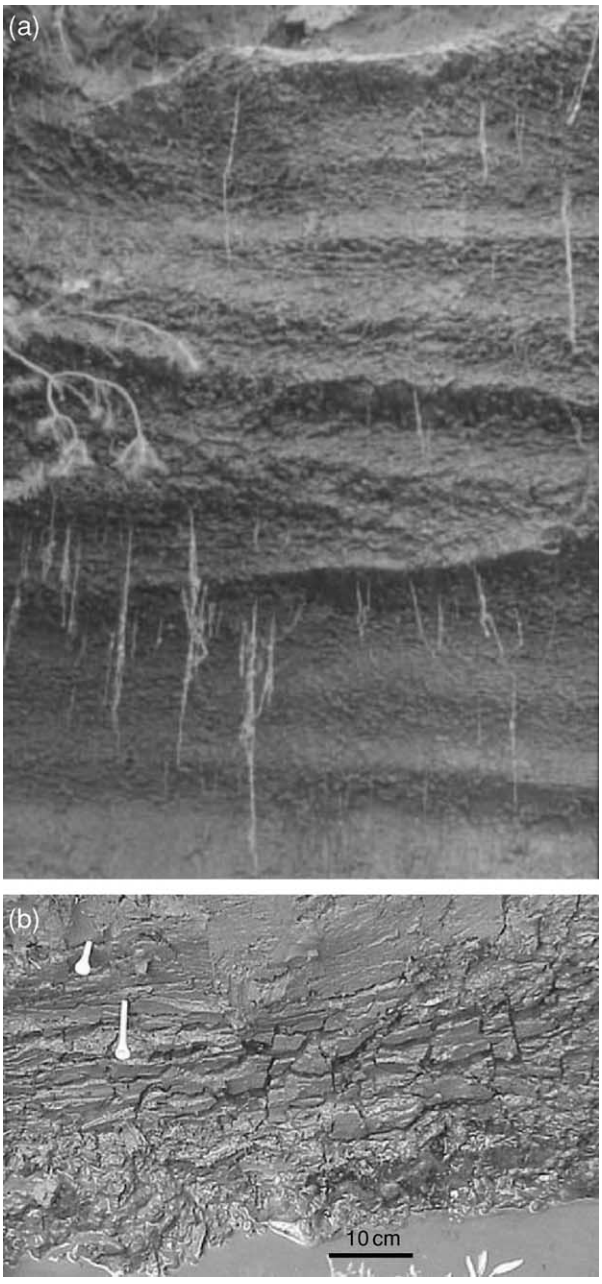


Figure 4 An enlargement of the active layer identified in Figure 3: (a) note the lenticular structure at the top and bottom, and reticulate structure in the middle; (b) the cryogenic structure can be better observed after the frozen soil is partially thawed.

permafrost as a result of cryoturbation. Thus cryoturbation contributes to the sequestration of atmospheric carbon into the upper permafrost. But if there is global climate change and warming of these regions, this portion of carbon could be a major source of CO_2 and methane into the atmosphere and a major perpetrator of more climate change.

Due to the presence of permafrost and seasonal freeze–thaw cycles, leaching is limited in most of the cold soils. In the high alpine regions, the leaching is further retarded by the arid conditions. In the high mountains and plateau, such as the northern Qinghai-Tibet Plateau in west China, the aridity (the ratio of potential evaporation to precipitation) exceeds 10 and is as high as 40 in some areas. Thus, in these soils, soluble salts often accumulate on the soil surface and the pH value ranges from 7.5 to 8.5. As the elevation drops, and with increased precipitation, the soluble salts and carbonates moved to the subsoils, whereas, in the arctic tundra regions, the lithology of the substrates has more influence on soil chemistry. Soils with a higher base status generally form in carbonate-rich loess or glacial drift and support nonacidic tundra vegetation. But with time, the vegetation succession gradually changes the soil chemistry by protonation or acidification of the surface soils, thus the landscape becomes acidic tundra. However, the leaching in these soils is generally weak and the soils are slightly acidic to slightly alkaline. But on the coastal plain and lowlands, base-rich fen forms and the soils are slightly to moderately alkaline.

The character of soil organic matter (SOM) in cold soils is influenced by its cryogenic environment, its position in the soil profile, the surface, subsurface soil active layer, and the presence of permafrost. Generally the SOM in cold soils is less decomposed than that in the temperate and tropical regions. In a recent study, the SOM in arctic soils was separated into extractable (EF) and nonextractable fractions (NEF) based on alkali solubility. The NEF, containing cell-wall and related constituents, such as hemicellulose and soluble fibers, dominate the SOM of cold soils. In a pattern consistent with a process of cold-temperature preservation, these NEF are found in largest proportions in the SOM of the upper permafrost. This fraction has greater potential to influence carbon cycling than EF in these ecosystems with climate warming.

Alpine and Plateau Regions

The high-plateau cold desert exists below the snowline on the high slopes on the mountains rising above the broad plateau in the Qinghai-Tibet Plateau, the Andes, Central Asia, and the High Arctic. The annual evaporation is more than 3000 mm, thus the climate is extremely arid and the aridity index is 23–24. Although the MAST is less than $-3^{\circ}C$, the active layers are generally more than 2 m thick due to strong diurnal temperature fluctuations, dry conditions, and coarse-textured soils. The parent materials consist of

Figure 5 An organic soil formed in the arctic coastal marsh, Prudhoe Bay, Alaska. Note the cryoturbated underlying mineral horizon.

glacial deposits or colluvium. At elevations below 5000 m, ice-cemented permafrost may be present between 1 and 3 m below the surface. This area also experiences strong diurnal and seasonal freeze–thaw cycles that have resulted in granular structures in surface horizons and reticular or platy structures in subsurface horizons. Thus, cryoturbation occurs in these soils.

The alpine meadow zone lies below the high mountain zone and occupies the upland slopes on both sides of the mountains as well as the lowlands in the alpine zone. The climate is cold, semiarid to subhumid; with MAST of −0.2 to −6.0°C, MAP of 300–400 mm, and 85–90% falling between June and September, and annual evaporation of more than 2000 mm; thus the aridity index is more than 10. The active layer is generally more than 1.5 m. In addition to seasonal freeze–thaw, the area also experiences strong diurnal freeze–thaw cycles. The well-drained soils have a dense, black surface horizon, with many grass roots overlying a strong brown subsoil, often with an accumulation of carbonates, Fe, manganese oxides, and humus (to a lesser extent) coatings on soil particles, and well-developed reticular and fine, subangular blocky structures. The strong freeze–thaw cycle results in frost-heaved grass mounds, frost cracks, and gelifluction lobes. In lowlands and depressions of the alpine meadow zone, where drainage is limited, gleization features are prominent in the subsurface horizons. Organic soils form in depressions with restricted drainage.

Vegetation and Fire

The boreal forest is subjected to frequent lightning-ignited fires during the summer. Fires are an important factor in controlling soil properties through altered vegetation succession and permafrost dynamics. Vegetation mosaics and soil morphology in this region strongly reflect the combined effects of fire, permafrost, and slope. It is common to find evidence of past fire events such as accumulations of charcoal particles in soil profiles, especially at the bottom of organic horizons. In areas of soils recently affected by fire of moderate or severe intensity, vegetation is commonly altered to willow shrub and aspen or mixed white spruce–aspen forest types. Combustion and consolidation of the insulating organic mat can result in warming of underlying soils, lowering of the permafrost table resulting in a thicker active layer, and transition to a well-drained and permafrost-free state before eventual return to preburn conditions. Soil reaction in the surface organic horizons ranges from moderately alkaline (pH 7.9–8.4) immediately after fire to slightly acid levels (pH 6.1–6.5) several years later.

Land Use of Cold-Region Soils

The cold soils present a special challenge to land-use managers owing to the effects of frost heave. It has caused highway-maintenance problems, structure-foundation instability, inconvenience to people, and

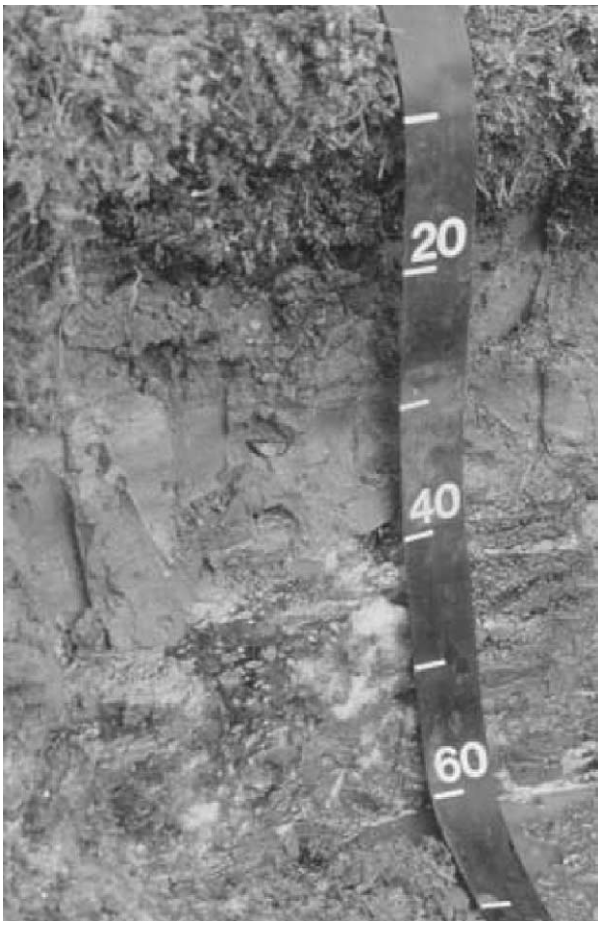


Figure 6 A cryogenic soil formed under moist acidic tundra, Toolik Lake Field Research Station, in the Arctic Foothills, Alaska. Note the frost-churned organic matter in upper permafrost at 45–60 cm.

inhibition of agricultural development. Because of the cold temperature, vegetation takes longer to recover after disturbance, and certain species of wildlife are unique to the cold regions; the ecosystems associated with cold soils are considered fragile. In past decades, natural resources extraction, mainly oil, gas, and minerals, has raised concerns. The optimum solution lies within our understanding of these cold soils and their environment.

Engineering

In the boreal regions, where permafrost is discontinuous and thin, the general practice is either to thaw the permafrost or to insulate the permafrost to minimize the effects of frost damage. But in the arctic region, the best practice is to insulate the permafrost to avoid thawing. Thick beds of gravel are applied to serve as an insulator, and lighter-color pavement is used to decrease solar radiation absorption. For permanent structures and houses, gravel pads or insulated foundations are commonly used in soils containing only ice lenses. But in areas where the soils contain a high content of ice, piling is used for structures and houses. In some cases, the piling or the foundation is refrigerated to avoid heat conduction to the permafrost.

Agriculture Development

Soil affected by permafrost accounts for more then 26% of the global land surface. Most of these areas are, in their natural state, tundra, boreal forest, cold deserts, and bogs. However, some areas have been cultivated for a long time. Cold soils present a great challenge to farm development not because of the low temperatures, but because of the ice content. In the subarctic or boreal regions, farms newly cleared from native forest underlying permafrost often face a subsidence problem. The land surface may become hummocked because of the melting of ice wedges and even thermokarst sinkholes, where large volumes of ice are present. Such problems are more severe on Pleistocene surfaces because of the higher ice content. The Holocene terraces are generally more favorable for land-clearing and farming.

The natives of the circumpolar regions have used the land for reindeer grazing for centuries. These people include the Samis in northern Scandinavian, the Eskimos in northern Alaska and Canada, and the Chukchi people in northern Russia. In the alpine and plateau regions, the Tibetans raise yaks, sheep, and cattle on the Tibetan Plateau; the Indians of the Andes have long raised llamas and other domesticated animals. In recent years there have been concerns of overgrazing in these fragile ecosystems. Once the organic layers or the surface vegetation covers are disturbed or destroyed, the quality of the land deteriorates owing to topsoil erosion.

There is intensive farming as well as grazing. Farming in the circumpolar region is made possible because of the long daylight hours in the summer that enable crops to mature in a shorter time than in the temperate region. In the boreal regions, where the soil conditions are favorable, small grains, especially barley, can mature within 110 days and oats within 130 days. Vegetables, such as carrots, lettuce, and cabbages can be harvested in less than 60 days. Potatoes are grown more than 200 km north of the Arctic Circle, where the permafrost table persists at less than 1 m from the surface. Thus in the boreal regions the potential for agriculture is greater than some of the northern temperate regions, where the summer always remain cool. However, such intensive farming can only exist in high-latitude environments owing

to the prolonged daylight, but not in alpine and plateau environments due to the diurnal temperature changes; the temperature can reach more than 20°C at midday but drops below freezing at night.

It should be noted that the development of these areas leads to an increased active layer and deeper permafrost, and this results in changes in leaching and oxidation and reduction as well as cryogenic processes. Consequently, there will be more methane and CO_2 emission owing to increased wetlands from thawing of the permafrost and thickening of the active layers. Not withstanding all of these problems, these areas can be highly productive because of the long day length and radiation inputs. This also raises the question as to why these soils are classified as high-latitude soils when they have much higher summer temperatures than ones found at high altitudes. One area can be used for crop production and the other cannot.

The greatest concern of farming in soils affected by permafrost is subsidence. Since these soils contain variable amounts of ice, once the soils are stripped of their natural vegetation covers the thermal region changes and the soils will warm. In soils with a high ice content and ice wedges, thermokarst will occur, and the cleared field becomes hummocky and presents difficult management problems. Another concern is the nature of the substratum. When the land is cleared, water from the melted ice has to be drained. Waterlogging or ponding will occur if there is restricted layer or the farm is on a topographic low. Thus good understanding of the soil and the geographic environment is the key to successful farming in cold soils.

Further Reading

Agriculture Canada Expert Committee on Soil Survey (1998) *The Canadian System of Soil Classification*, 3rd edn. Ottawa, ON: NRC Research Press.

Bockheim JG, Walker DA, Everett LR, Nelson FE, and Shiklomanov NI (1998) Soils and cryoturbation in moist nonacidic and acidic tundra in the Kuparuk River Basin, arctic Alaska, USA. *Arctic and Alpine Research* 30: 166–174.

Brown J, Ferrians OJ Jr, Heginbottom JA, and Melnikov ES (1997) *Circum-Arctic Map of Permafrost and Ground-Ice Conditions*. US Geological Survey, Map CP-45, 1:10 000 000. Washington, DC: US Government Printing Office.

Clymo RS (1983) Peat. In: Gore AJP (ed.) *Ecosystems of the World*, vol. 4A, *Mires: Swamp, Bog, Fen, and Moor*, ch. 4, pp. 159–224. Amsterdam, The Netherlands: Elsevier Scientific.

Cryosol Working Group (2004) *Cryosols*. New York: Springer Verlag.

Dai XY, Ping CL, and Michaelson GJ (2002) Characterization of organic matter in Arctic tundra soils by different analytical approaches. *Organic Geochemistry* 33: 407–419.

French HM (1996) *The Periglacial Environment*, 2nd edn. White Plains, NY: Longman.

Gubin SV (1993) Late Pleistocene soil formation on coastal lowlands of northern Yakutia (in Russian). *Pochvovedeniye* 10: 62–70.

Hoefle CM and Ping CL (1996) Properties and soil development of late-Pleistocene paleosols from Seward Peninsula, northwest Alasaka. *Geoderma* 71: 219–243.

Lal R, Kimble JM, and Stewart BA (eds) (2000) *Global Climate Change and Cold Regions Ecosystems*. Boca Raton, FL: Lewis.

Michaelson GJ, Ping CL, and Kimble (1996) Carbon storage and distribution in tundra soils of Arctic Alaska, USA. *Arctic and Alpine Research* 28: 414–424.

Muller SW (1947) *Permafrost or Permanently Frozen Ground and Related Engineering Problems*. Ann Arbor, MI: JW Edwards.

Ovenden L (1990) Peat accumulation in northern wetlands. *Quaternary Research* 33: 377–386.

Péwé TL (1954) *Effect of Permafrost on Cultivated Fields, Fairbanks Area, Alaska*. US Department of Interior, Geological Survey Bulletin 989-F, pp. 315–351. Washington, DC: US Government Printing Office.

Péwé TL (1975) *Quaternary Geology of Alaska*. US Department of Interior, Geological Survey Professional Paper 835. Washington, DC: US Government Printing Office.

Ping CL, Michaelson GJ, and Kimble JM (1997) Carbon storage along a latitudinal transact in Alaska. *Nutrient Cycling in Agroecosystems* 49: 235–242.

Ping CL, Bockheim JG, Kimble JM, Michaelson GJ, and Walker DA (1998) Characteristics of cryogenic soils along a latitudinal transect in arctic Alaska. *Journal of Geophysical Research* 103(D22): 917–928.

Qiu GQ and Cheng GD (1993) Development condition of Cryosols in alpine and plateau regions of western China. In: Gilinchinsky DA (ed.) *Post-Seminar Proceedings. Joint Russian–American Seminar on Cryopedology and Global Change*, pp. 59–75. Pushchino, Russia: Russian Academy of Sciences.

Qiu GQ and Cheng GD (1995) Permafrost in China: past and present. *Permafrost and Periglacial Processes* 6: 3–14.

Reiger S (1983) The genesis and classification of cold soils. New York: Academic Press.

Richardson JL and Vepraskas MJ (2001) *Wetland Soils: Genesis, Hydrology and Classification*. Boca Raton, FL: Lewis.

Samson-Liebig SE, Kimble JM, and Ping CL (1995) Improvements in definition of cryic and pergelic soil temperature regimes in soil taxonomy using daylength/solar radiation. *Soil Survey Horizons* 36(1): 20–25.

Viereck LA (1970) Forest succession and soil development adjacent to the Chena River in Interior Alaska. *Arctic Alpine Research* 2: 1–26.

Viereck LA, Dyrness CT, and Foote MJ (1993) An overview of the vegetation and soils of the floodplain ecosystems of the Tanana River, interior Alaska. *Canadian Journal of Forest Research* 23: 889–898.

Walker DA, Auerbach NA, and Backheim JG (1998) Energy and trace-gas fluxes across a soil pH boundary in the Arctic. *Nature* 394: 469–472.

COLLOID-FACILITATED SORPTION AND TRANSPORT

R Kretzschmar, Institute of Terrestrial Ecology, Zurich, Switzerland

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Colloid transport in soils has attracted the attention of soil scientists since the 1930s, as they have tried to understand the formation of clay skins, argillic horizons, and clay pans in soils. Research on colloidal stability and aggregation of soil clays and reference clay minerals was later stimulated by problems due to soil dispersion in irrigation agriculture, where the release and transport of colloids can lead to drastic reductions in soil permeability. Since the early 1990s, increasing concern has been raised that mobile soil colloids may also serve as carriers for strongly sorbing contaminants such as heavy metals, hydrophobic organic compounds, and radionuclides, thereby facilitating the transport of contaminants which would otherwise be highly immobile. This colloid-facilitated transport mechanism is depicted schematically in Figure 1. A large number of laboratory and field studies on colloid transport in subsurface porous media have been conducted since then, providing a better understanding of the release, transport, and deposition behavior of colloidal particles. Today, many of the observed phenomena are understood qualitatively, but quantitative theories are still lacking and are the subject of ongoing research in environmental soil chemistry and physics.

Four key conditions must be fulfilled for colloid-facilitated contaminant transport to become an environmentally important factor: (1) mobile colloidal particles must be present in sufficiently large concentrations; (2) the contaminant must sorb strongly to the colloidal particles and desorb only slowly; (3) the particles and associated pollutants must be transported over significant distances through less contaminated zones of a soil or subsurface sediment; and (4) the concentrations of contaminants transported via colloids as carriers must exceed the tolerable limits. This chapter discusses the environmental factors controlling the release, transport, and deposition of colloids in soils and their potential influence on the transport of strongly sorbing contaminants.

Nature and Stability of Soil Colloids

Colloids are often defined as solid particles with an equivalent spherical diameter between 1 and 1000 nm dispersed in a liquid phase. The size limits of colloidal particles are to a certain degree arbitrary and should not be considered sharp boundaries between dissolved molecules, colloids, and larger suspended particles. However, colloids have two unique properties: (1) they have a very large specific surface area (more than $10\ m^2\ g^{-1}$) and therefore can be efficient sorbents for inorganic and organic contaminants; and (2) they exhibit extremely low gravitational settling velocities

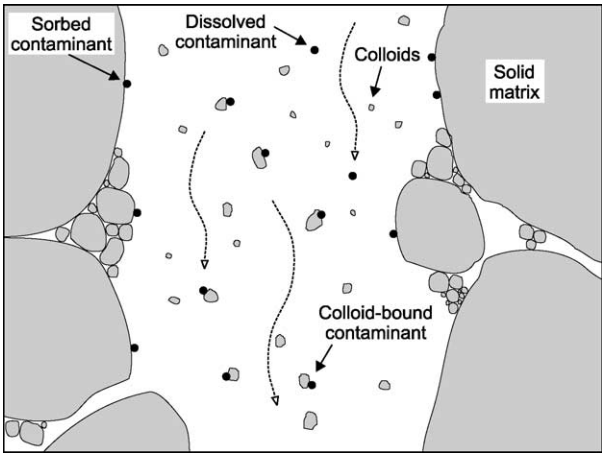


Figure 1 Schematic of colloid-facilitated contaminant transport in porous media. A contaminant (black sphere) can be present as a dissolved species, sorbed to the solid matrix, or bound to colloidal particles that move with the flowing water. For strongly sorbing contaminants, such mobile colloids may serve as carriers and provide a rapid transport pathway. Reproduced from Kretzschmar R, Borkovec M, Grolimund D, and Elimelech M (1999) Mobile subsurface colloids and their role in contaminant transport. *Advances in Agronomy* 66: 121–194 with permission from Elsevier.

in water and therefore remain stable in dispersion over long time periods, unless they coagulate to form larger aggregates or deposit on to liquid–solid or liquid–gas interfaces.

In soils and sediments, the most important source of mobile colloids is the release of small, submicron-sized mineral or organic particles existing in virtually every soil. In fact, the particle size fraction smaller than 1 μm can contribute significantly to the total soil mass and it usually contributes most of the specific surface area, even in sandy soils. As a result, the composition of mobile colloidal particles found in soils and underlying aquifers often reflects the composition of the fine particle size fractions present in the source horizons. Typical components include phyllosilicate minerals (e.g., kaolinite, illite, smectite), oxyhydroxides of Fe and Al (e.g., goethite, hematite, ferrihydrite), colloidal silica, and natural organic matter (NOM, e.g., humic substances). Mostly these components occur in mixtures or as mineral–organic complexes. The size ranges of potentially mobile particles in soils are depicted in Figure 2, along with the conventional particle size classes for comparison.

In some cases, mobile colloids can also be formed by precipitation from oversaturated solutions or from the release of colloids from soil amendments such as sewage sludge. One example where precipitation might play a role is the formation of colloidal calcite particles in the solution, for example, as a result of changes in CO_2 partial pressure. However, little is known about the colloidal stability of calcite particles and their mobility in soils. Viruses and bacteria also exhibit in some respects colloidal properties and are therefore sometimes considered biocolloids (Figure 2).

The stability of colloids in aqueous dispersion strongly depends on their surface charge properties and the ionic composition of the solution. In the vast majority of soils, mobile colloidal particles possess predominantly negative surface charge. Positively charged particles deposit very effectively on to negatively charged matrix grain surfaces due to electrostatic attraction. They are therefore highly immobile, unless the entire soil is dominated by positively charged mineral components. This is only the case in strongly acidic subsoil horizons of highly weathered Oxisols or Ultisols with low organic matter contents. All other soils are dominated by the negatively charged surfaces of clay minerals, feldspars, quartz, and soil organic matter. The surface charge of soil components can have different origins. Firstly, most phyllosilicates in soils have isomorphic substitution of cations either in tetrahedral (e.g., Al^{3+} for Si^{4+}) or octahedral (e.g., Fe^{2+} for Al^{3+}) sheets, resulting in permanent, pH-independent negative surface charge. Secondly, the edge surfaces of clay minerals and oxyhydroxide and oxide minerals have reactive surface hydroxyl groups which can protonate or deprotonate, resulting in pH-dependent negative or positive surface charge. At pH values below the point of zero charge (PZC), the surfaces are positively charged, while at pH values above the PZC the surfaces are negatively charged. Typical PZC values for different soil minerals are given in Table 1.

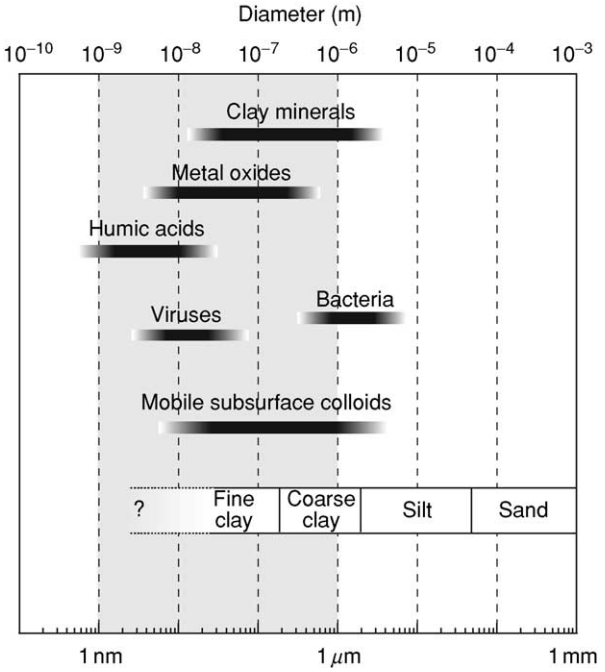


Figure 2 Size range of colloidal particles in soils. The classic size fractions commonly used in soil science are shown for comparison. Reproduced from Kretzschmar R, Borkovec M, Grolimund D, and Elimelech M (1999) Mobile subsurface colloids and their role in contaminant transport. *Advances in Agronomy* 66: 121–194 with permission from Elsevier.

Table 1 Typical points of zero charge (PZC) of various soil minerals in the absence of specifically adsorbed ions. Values reported in the literature vary by up to ±0.5 pH units, depending on mineral source, purity, method, and experimental conditons used

Solid	*PZC*
Calcite	9.5
Amorphous aluminum oxide	9.0
Amorphous iron oxide	9.0
Hematite	9.0
Goethite	8.5
Magnetite	6.5
Kaolinite	5.5
Montmorillonite	2.5
Feldspars	2.2
Quartz	2.0
Amorphous SiO_2	1.8

It is important to note, however, that adsorption of NOM or specifically adsorbing anions such as phosphate can drastically alter the surface charge of oxide and clay surfaces, resulting in a decrease in PZC. Under acidic conditions, one can observe surface charge reversal of oxide particles from positive to negative upon adsorption of NOM. Other ions which significantly reduce the PZC of oxide minerals to varying degrees include adsorbed silica, phosphate, arsenate, and sulfate.

The stability of a colloidal dispersion is determined by the balance between long-ranged electrostatic repulsive forces and short-ranged London–van der Waals attractive forces between colloidal particles. According to the classic Derjaguin–Landau–Verwey–Overbeek (DLVO) theory of colloidal stability, the sum of these forces results in an interaction force profile as a function of separation distance which strongly depends on the ionic composition of the solution and the surface charge density of the particles. At low ionic strength, the colloidal particles must overcome a repulsive energy barrier due to electrostatic repulsion. Therefore, not every particle collision results in coagulation (attachment efficiency $\alpha < 1$) and the process is termed slow or reaction-limited coagulation. With increasing ionic strength, the repulsive energy barrier diminishes and finally disappears due to charge screening and compression of the diffuse double layer. Under such conditions every particle collision results in aggregation (attachment efficiency $\alpha = 1$) and the process is termed rapid or transport-limited coagulation. The attachment efficiency α is the fraction of particle collisions that results in coagulation. In colloid chemistry, the critical coagulation concentration (CCC value) is defined as the minimum electrolyte concentration required for rapid coagulation under given pH and environmental conditions. The relationships between the cation concentration in monovalent, mixed bivalent/monovalent, and bivalent cation solutions and the attachment efficiency α are shown schematically in Figure 3.

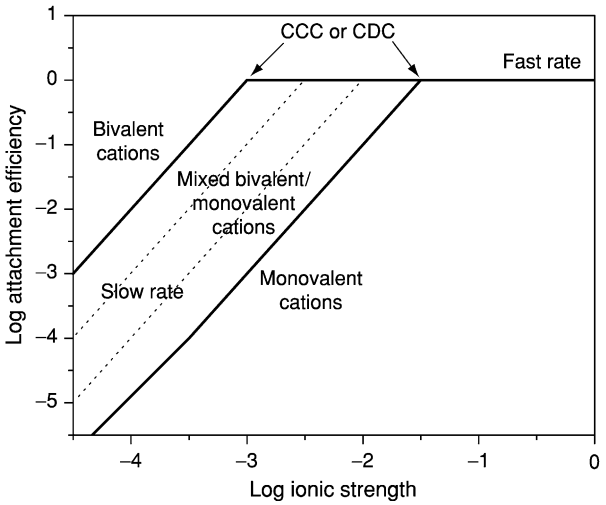


Figure 3 Influence of ionic strength of monovalent, mixed bivalent/monovalent, and bivalent cation solutions on the attachment efficiency (α) of negatively charged colloidal particles during aggregation or deposition on negatively charged surfaces. The transition from slow (log $\alpha < 0$) to fast (log $\alpha = 0$) deposition is called the critical coagulation concentration (CCC) for aggregation and critical deposition concentration (CDC) for deposition, respectively.

The CCC values of soil clays and clay minerals have often been determined by coagulation series tests in a simple batch system, in which the colloidal stability of clay suspensions is evaluated optically after a given time period allowed for coagulation and settling. Such CCC values are strictly operationally defined and one must be aware that they decrease with increasing initial clay concentration and increasing coagulation and settling time. One method to determine CCC values more quantitatively is by direct measurement of the relative coagulation rate using dynamic light scattering (DLS). So far, this method has only been used by a few researchers studying the colloidal stability of soil clays and reference clay minerals, but DLS has great potential.

For several different clay minerals, including illite, kaolinite, and montmorillonite, CCC values of typically less than 40 mol m^{-3} are observed in mixed Na/Ca systems with a sodium adsorption ratio (SAR) up to 60 and for pH values ranging from 6 to 9. The SAR is defined as

$$\mathrm{SAR} = \frac{[\mathrm{Na}^+]}{([\mathrm{Ca}^{2+}] + [\mathrm{Mg}^{2+}])^{0.5}} \qquad [1]$$

where square brackets denote solution concentrations in moles per cubic meter. Substantially higher CCC values than in mixed Na/Ca systems are measured in pure Na systems, where the strong coagulating effect of the divalent Ca^{2+} is absent. Since even sodic and saline-sodic soils contain considerable amounts of Ca^{2+} and Mg^{2+}, CCC values for pure Na systems are of limited practical relevance. Typical CCC values for colloidal dispersions of reference clay minerals in solutions of monovalent and divalent cations are given in Table 2. DLVO theory would predict that the CCC values are roughly 64 times higher for monovalent cations than for divalent cations. Similarly, for trivalent cations the CCC value would be expected to be roughly 700 times lower than for monovalent cations, illustrating the strong coagulating power of Al^{3+}. Experimental data are at least in qualitative agreement with this aspect of the DLVO

Table 2 Typical critical coagulation concentrations (CCC) of monovalent ($CCC_{monovalent}$) and divalent ($CCC_{divalent}$) cations for reference clay minerals

	$CCC_{monovalent}$ ($mmol\,l^{-1}$)	$CCC_{divalent}$ ($mmol\,l^{-1}$)	$CCC_{divalent}/CCC_{monovalent}$
Soil mineral			
Illite	48 ± 11	0.14 ± 0.02	0.003
Kaolinite	10 ± 4	0.3 ± 0.2	0.030
Montmorillonite	8 ± 6	0.12 ± 0.02	0.015
DLVO theory prediction			0.0156

DLVO, Derjaguin–Landau–Verwey–Overbeek.
Adapted from Sposito G (1989) *The Chemistry of Soils.* Oxford, UK: Oxford University Press.

theory, and deviations may partly be explained by specific interactions of cations with the mineral surface. In colloid science, the general dependency of the CCC value on cation valence is also known as the Schulze–Hardy rule, which has been recognized since the end of the nineteenth century.

Soil clays very often exhibit much higher CCC values than corresponding reference clay minerals. This is particularly true for soil clays isolated from surface soil horizons. This higher colloidal stability of soil clays compared to their reference clay counterparts is mainly attributed to the influence of adsorbed NOM such as humic substances. As mentioned earlier, adsorbed humic substances add negative surface charge and can cause charge reversal of clay edges from positive to negative, for example in kaolinite under acidic pH conditions. Adsorbed humic substances therefore lead to increased electrostatic stabilization of soil clay dispersions. In addition, adsorbed humic substances may also cause steric stabilization of clay colloids, which is due to a thermodynamic repulsion that occurs between interpenetrating organic polymer chains attached to the colloid surface. However, steric stabilization by adsorbed NOM may only be relevant if the thickness of the adsorbed organic layer exceeds the thickness of the diffuse double layer. Such conditions are most likely to occur at high ionic strengths or in the presence of multivalent cations.

Due to pH-dependent charge, CCC values of soil colloids are also strongly pH-dependent. For example, oxyhydroxide and oxide mineral colloids coagulate even in deionized water at pH values close to their respective PZC, because electrostatic repulsive forces are absent or too weak to cause electrostatic stabilization. Well above or below the PZC, such dispersions are stabilized by electrostatic repulsive forces. Kaolinite colloids exhibit a different behavior. At pH values less than the PZC of the edge surfaces (near pH 5.5), pure kaolinite particles coagulate rapidly even in deionized water, due to attractive interactions between positively charged edge surfaces and negatively charges face surfaces (edge-to-face coagulation). Above the PZC of the edge surfaces, the entire particles are negatively charged and electrostatic stabilization occurs. Addition of NOM leads to edge charge reversal from positive to negative at pH values below the PZC and strongly increases colloidal stability. Therefore, kaolinite colloids from topsoils often exhibit much higher colloidal stability than kaolinites from subsoils or reference clay sources.

Colloid Transport in Porous Media

The transport of colloidal particles in soils appears to be limited to larger pores, especially to preferential flow paths such as old root channels, earthworm burrows, and other interaggregate pores. Evidence for colloid transport in field soils stems primarily from micromorphological investigations showing oriented clay skins on pore walls and faces of soil peds. The formation of such features is attributed to the translocation and deposition of fine clay material within soil profiles. It has been shown that the composition of clay skin material in subsoil horizons has greatest similarities with the fine clay fraction in the corresponding surface soils, suggesting colloid mobilization near the surface (A and E horizons) and transport through macropores into the subsurface (e.g., Bt horizon).

Laboratory experiments have shown that colloidal particles in natural porous media can be transported faster than a conservative solute tracer such as tritiated water. This effect is termed 'size exclusion effect,' in analogy to size exclusion chromatography. Unlike a conservative solute tracer, colloidal particles are restricted to larger pores, and therefore the colloid peak elutes earlier than the tracer peak. Such effects have been observed both in undisturbed soil and saprolite columns, as well as in repacked soil columns.

The mobility of colloidal particles in soils strongly depends on solution and surface chemistry and on physical factors including water flow velocity, water saturation, and pore size distribution. Colloid transport in soils has been studied mostly in laboratory

column experiments with repacked or undisturbed soil columns. Most experiments were done under water-saturated conditions, which makes it easier to control and vary the flow velocity. Under such conditions, colloid transport can be described by a convective-dispersive transport equation, including terms that account for colloid deposition and release:

$$(\partial C/\partial t) = D_p \frac{\partial^2 C}{\partial x^2} - \nu_p \frac{\partial C}{\partial x} - \frac{\rho_b}{\epsilon}\frac{\partial S}{\partial t} \qquad [2]$$

and

$$((\rho_b/\epsilon)(\partial S/\partial t)) = k_d C - \frac{\rho_b}{\epsilon} k_r S \qquad [3]$$

Along with the appropriate boundary conditions, eqns [2] and [3] describe the colloidal particle concentration in suspension $C(x,t)$ and the amount of deposited particles per unit mass of the porous matrix $S(x,t)$ as a function of travel distance x and time t. Here, D_p is the hydrodynamic dispersion coefficient for colloidal particles, ν_p is the mean interstitial velocity of colloidal particles, ρ_b the solid matrix bulk density, ϵ the porosity, and k_d and k_r the colloid deposition and release rate coefficients, respectively. The deposition of colloids is often observed to follow a first-order kinetic rate law, while the colloid release rate is often extremely low under steady-state flow and constant chemical conditions. Therefore, many researchers have used a simplified equation that neglects particle release to describe colloid transport in laboratory-scale columns:

$$(\partial C/\partial t) = D_p \frac{\partial^2 C}{\partial x^2} - \nu_p \frac{\partial C}{\partial x} - k_d C \qquad [4]$$

The factors influencing the colloid release (k_r) and deposition rates (k_d) are discussed in the following sections.

Factors Influencing Colloid Release

The release of colloidal particles in soils is still poorly understood, and predictive models or theories are currently lacking. However, a qualitative understanding of colloid release has been developed by controlled laboratory column studies, field observations, and theoretical considerations of particle–surface interactions. The release of significant concentrations of colloids in soils is mainly a result of physical or chemical perturbations. Physical perturbations can result, for example, from raindrop impact on bare soil, rapidly infiltrating water, drying and rewetting of soil aggregates, or soil tillage. Chemical perturbations can occur naturally due to atmospheric deposition and leaching of salts, or can be a result of management practices such as fertilizer application or soil irrigation. Possible mechanisms of colloid release due to chemical changes include the dissolution of cementing agents and, much more importantly, the dispersion of soil particles due to electrostatic repulsive forces at low ionic strength and/or a high percentage of exchangeable Na^+ on the cation exchange complex.

Figure 4 shows a conceptual view of the colloid-release process. The kinetics of colloid release in porous media are influenced by particle–surface interactions and the hydrodynamics of the flow field. The release of colloidal particles attached to a surface involves two subsequent steps: (1) detachment of colloidal particles and transport across the interaction boundary layer, and (2) diffusional transport of detached colloidal particles through a stationary water film surrounding the matrix grains, the so-called diffusion boundary layer. The detachment step is strongly influenced by the surface charge of colloids and matrix grains and by solution chemistry. At sufficiently high ionic strength, the particles must escape from a primary energy minimum in order to detach from the surface. Under these conditions, detachment of particles is the rate-limiting step (detachment-limited release) and the release rates are often extremely low. With decreasing ionic strength, the depth of the primary energy minimum decreases. At sufficiently low ionic strength, the interaction forces between colloids and matrix grains can become repulsive at all separation distances, resulting in rapid detachment. The diffusion through the stationary water film is now rate-limiting (transport-limited release). Since the thickness of the diffusion boundary layer decreases with increasing flow velocity, colloid release rates are expected to increase with flow velocity. Additional shear and drag forces can act on colloidal particles and further increase the colloid release rates at very high flow rates, for example, during rapid infiltration of water through macropores.

Chemical changes that can induce colloid release include decreases in soil-solution ionic strength, increases in the percentage of Na^+ on cation exchange sites, increases in soil pH, and adsorption of ions or molecules which alter the mineral surface charge. Decreases in ionic strength can result from rainfall or irrigation with low-ionic-strength water. However, experimental results and field observations show that a decrease in ionic strength alone does not result in significant colloid release if the soil's cation exchange capacity is saturated with multivalent cations such as Ca^{2+} or Al^{3+}. If the soil is partly saturated with monovalent cations, especially Na^+, decreases in

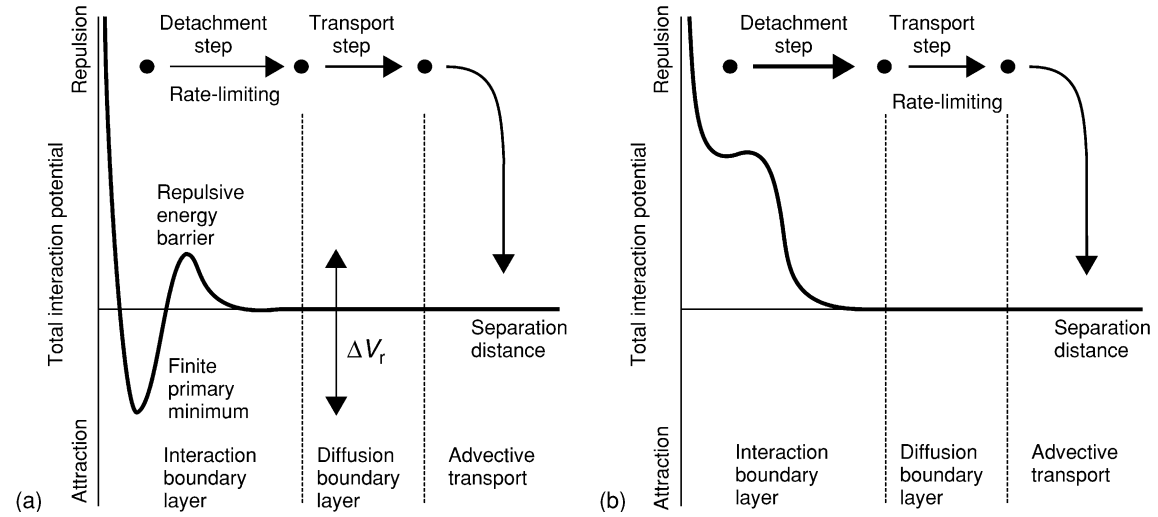


Figure 4 Conceptual view of the colloid-release process consisting of two steps. In the first step, colloidal particles must escape from the primary minimum in which they are attached. The release energy corresponds to ΔV_r. In the second step, the colloidal particles must be transported across the diffusion boundary layer. (a) At medium ionic strength, $\Delta V_r > 0$ and the release kinetics is detachment limited; (b) at very low ionic strength, the barrier height diminishes and the release kinetics becomes transport-limited. Reproduced from Kretzschmar R, Borkovec M, Grolimund D, and Elimelech M (1999) Mobile subsurface colloids and their role in contaminant transport. *Advances in Agronomy* 66: 121–194 with permission from Elsevier.

ionic strength can result in significant colloid release, which usually leads to drastic reductions in soil-water permeability. This problem is rather common in soils where water with a relatively high SAR and low salinity is used for irrigation. In irrigation agriculture, an exchangeable sodium percentage (ESP) of more than 15% of the cation exchange capacity is considered as a critical value. However, also surface soils dominated by kaolinite clay (e.g., in Ultisols) can be highly dispersive and generate mobile soil colloids, which are probably stabilized in suspension by adsorbed NOM.

In strongly acidic mineral soils with high Al^{3+} saturation of the cation exchange capacity, colloid release rates are usually low due to the strong aggregating power of trivalent cations. Exceptions may occur in strongly weathered Ultisols or Oxisols, where positively charged colloidal particles may be mobilized at low ionic strength. However, this has only rarely been observed in practice. Another important factor leading to colloid release is the adsorption of ions or molecules, which make the colloid surfaces more negatively charged. For example, phosphate additions to soils can lead to a decrease in the PZC of oxides and cause charge reversal from positive to negative charge at given pH. As a result, colloidal particles can be more easily released and transported through the soil. Similarly, organic molecules which adsorb to soil particles can add negative surface charge and thereby increase colloid release rates. Examples include humic substances, malonate, citrate, oxalate, and salicylate. Also anionic surfactants and strong complexing ligands such as EDTA can lead to increased colloid release in soils, which has to be taken into consideration when developing in-situ soil remediation technologies. It has been demonstrated that carboxylic acids and other organic constituents released by plant roots can cause kaolinite particle dispersion, suggesting that roots and microorganisms can induce colloid release under natural conditions. On the other hand, many soil organisms are also known to release polysaccharides which tend to stabilize soil aggregates and thereby prevent soil dispersion and release of colloids.

Factors Influencing Colloid Deposition

The most important mechanism of particle removal from solution in porous media is colloid deposition, that is, the collision and attachment of colloidal particles on surfaces of larger grains within the soil matrix. If the chemical conditions are favorable for colloid deposition, granular porous media are extremely efficient particle filters and they are therefore widely used in water treatment. Much research on colloid transport and deposition has been conducted to understand the performance and failures of granular filter beds. Colloid transport and deposition in soils follow basically the the same principles. Particle deposition is commonly assumed to take place in two steps: (1) transport of colloidal particles to matrix

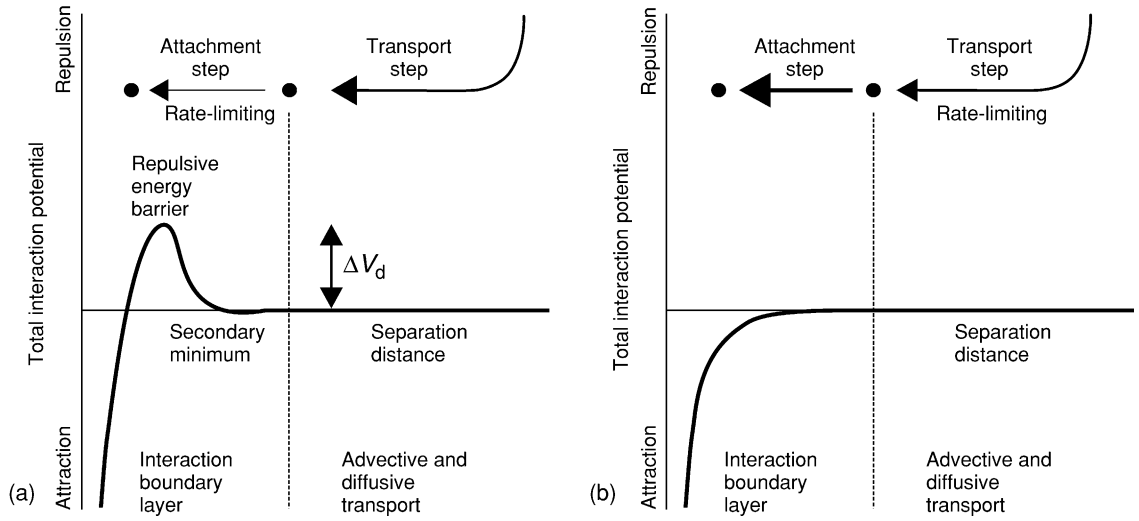


Figure 5 Conceptual view of the colloid-deposition process, consisting of two steps. In the first step, colloidal particles must be transported to the matrix surface by diffusion, interception, or gravitational sedimentation. In the second step, the colloids must be transported across the repulsive energy to be attached in the primary energy minimum. (a) At low ionic strength, a pronounced repulsive energy barrier with height ΔV_d develops between similarly charged colloids and surfaces. As a result, colloid deposition is slow or attachment-limited; (b) at high ionic strength, the repulsive energy barrier disappears and the kinetics of colloid deposition becomes fast or transport-limited. Reproduced from Kretzschmar R, Borkovec M, Grolimund D, and Elimelech M (1999) Mobile subsurface colloids and their role in contaminant transport. *Advances in Agronomy* 66: 121–194 with permission from Elsevier.

surfaces within the porous medium by Brownian diffusion, interception, or gravitational sedimentation, resulting in colloid–matrix collisions, and (2) attachment of colloidal particles to the matrix surfaces. A conceptual view of the particle-deposition process is shown in Figure 5.

The kinetics of the transport step depends on physical factors such as size and density of colloidal particles, accessible surface area for colloid deposition, pore structure, and flow velocity. The particle diffusion velocity decreases with increasing particle size, while interception and gravitational settling velocities increase. Therefore, particles with diameters in the order of 0.1–1 μm are expected to be most mobile in porous media, but this also depends on the specific mass of the colloidal particles.

The kinetics of the attachment step depends strongly on the charge density of colloidal particle and matrix surfaces and on the ionic composition of the solution. In principle, colloid attachment is analogous to heterocoagulation, where the radius and surface charge of two interacting particles are different. If the colloids and matrix surfaces are similarly charged, a repulsive energy barrier develops due to electrostatic repulsive forces and slows down the colloid attachment rate ($\alpha < 1$). With increasing ionic strength, the repulsive energy barrier is diminished and the deposition rate increases until every particle collision results in attachment ($\alpha = 1$). At this point, the deposition kinetics is transport-limited and is largely independent of chemical factors (Figure 3). If the colloidal particles and matrix surfaces are oppositely charged, the deposition rate is always rapid ($\alpha = 1$) due to attractive electrostatic forces.

Experiments with pure hematite colloids have shown that positively charged colloids are highly immobile in soils dominated by negatively charged mineral components, such as 2:1 type clay minerals, quartz, and feldspars. Only if the hematite colloids were coated with humic substances, resulting in charge reversal, could significant colloid transport be observed. This and other research results suggests that NOM coatings play a crucial role in the stabilization and transport of colloidal particles in soils.

In column experiments on colloid transport and deposition through granular porous media, colloid deposition often follows a first-order kinetic rate law, leading to an exponential decrease in suspended colloid concentration with travel distance:

$$\frac{C(x)}{C_0} = \exp(-\lambda x) \qquad [5]$$

where C_0 is the colloid concentration in the column inlet and λ is the so-called filter coefficient. The filter coefficient can be expressed in terms of the macroscopic deposition rate, k_d, by:

$$\lambda = \frac{k_d}{\nu_p} \qquad [6]$$

Colloid deposition rate coefficients k_d can be determined from the colloid breakthrough curves resulting from step-input or pulse-input column experiments.

If large amounts of colloidal particles are introduced into a porous medium the colloid deposition rate can change due to so-called blocking or filter-ripening effects. Again, these terms stem from research in deep-bed filtration, but the phenomena have also been demonstrated in soils and sediments. When colloid–colloid attachment is unfavorable (i.e., stable dispersion) because of interparticle repulsion, the colloid deposition rate tends to decrease with increasing amounts of deposited colloids on matrix surfaces (blocking effect). When colloid–colloid attachment is favorable (i.e., unstable dispersion), the colloid deposition rate tends to increase with increasing amounts of deposited colloids and multiple layers of colloid can form on matrix surfaces, ultimately leading to pore clogging. Pronounced blocking effects have been observed for transport of soil fine clays through undisturbed porous saprolite columns at low ionic strength, where colloid deposition results in a rapid decline in the deposition rate. Filter ripening may occur at high ionic strength, but under such conditions colloid release is usually low.

Another factor which strongly affects colloid deposition rates in porous media is the surface-charge heterogeneity which inherently occurs in soils, sediments, and weathered bedrocks. Surface-charge heterogeneity leads to preferential sites for colloid deposition, that is, sites that are particularly favorable for attachment. For example, a weathered granitic saprolite contains mainly negatively charged mineral components such as quartz, feldspars, and biotite. However, it also contains secondary minerals such as kaolinite, goethite, hematite, and ferrihydrite. Such components can create patches of positively charged surfaces which act as favorable deposition sites for negatively charged colloidal particles. The consequence of surface-charge heterogeneity is that colloid mobility cannot be predicted from streaming-potential or zeta-potential measurements of the matrix and colloidal particles, respectively, because such measurements usually provide only an average zeta potential for the entire sample. Charge heterogeneity may also explain the rapid blocking effects observed in weathered saprolites when small amounts of negatively charged soil colloids are introduced at low ionic strength.

So far, the discussion has been limited to the case of saturated flow. Since soils are often unsaturated, it is also important to consider the influence of a gas phase on colloid transport and deposition. Generally, the presence of air-filled pores leads to a lower connectivity of the water-filled pores and to the presence of water–gas interfaces in addition to water–solid interfaces. Hydrophobic colloidal particles are effectively removed by attachment to water–gas interfaces, while hydrophilic colloidal particles are less affected. Therefore, unsaturated conditions lead to a lower mobility of colloids in porous media, in particular of hydrophobic colloidal particles. For example, it has been shown experimentally that hydrophilic and hydrophobic bacteria have similar mobility in a water-saturated sand, but the hydrophobic bacteria are much less mobile than the hydrophilic bacteria under unsaturated flow conditions.

Colloid-Facilitated Transport of Contaminants

Strongly sorbing contaminants such as heavy metals, oxyanions, radionuclides, certain pesticides, and other hydrophobic organic compounds can be detected much earlier in the subsoil or groundwater than anticipated from traditional solute transport models. This discrepancy can have several different causes, including preferential flow through macropores and colloid-facilitated transport. In many cases, both transport mechanisms probably occur simultaneously, since the mobilization of colloids is most pronounced when dry soils are rewetted during a rainstorm event and dispersed particles are transported into deeper layers of the soil through rapidly infiltrating water along macropores. In contrast, during dry periods or periods with slow unsaturated water flow, the mobility of colloids in soil is small and preferential flow plays a minor role. Thus, colloid-facilitated transport and preferential flow typically occur together and the contributions of both effects are not easily separated.

In agricultural soils, colloidal transport of pesticides and phosphate are the greatest concern. For example, facilitated transport of the herbicides Napropamide (2-(α-naphthoxy-*N*,*N*-diethylpropionamide) and Glyphosate (*N*-(phosphonomethyl)glycine) by dissolved or colloidal NOM and other soil colloids has been demonstrated in column experiments. Drying and wetting of the soil surface and rainstorm events with rapidly infiltrating water enhance colloid mobilization and can lead to colloid-facilitated leaching of pesticides. The result is a rapid appearance of the pesticide in the leachate in association with dispersed colloids or NOM. Colloid-facilitated transport of phosphate is also a major concern on sandy soils that receive regular applications of animal manure, sometimes at excessive levels

from the perspective of plant nutrition. Field lysimeter studies have shown that large percentages of total phosphorus in soil solution can be present in association with colloidal material, probably as colloidal organic matter–metal–orthophosphate complex or occluded in other colloidal particles. Excessive application of animal manure could increase the release of colloids and therefore the risk of groundwater pollution with phosphate and pesticides.

In metal-contaminated soils, colloid-facilitated transport of heavy metals may occur, depending on the vegetation cover, soil management, climate, and soil properties. In soil solutions, some metals such as Cu are often predominantly present as dissolved or colloidal organic complex. Heavy metals such as Pb, Cu, and Zn have also been shown to be associated with clay colloids and biocolloids (bacteria), respectively. Although colloid-facilitated transport of heavy metals has been demonstrated in laboratory experiments, quantitative field studies are lacking and the importance of this process under field conditions is still uncertain.

One important aspect which must be addressed in future research is the role of slow-desorption kinetics of heavy metals or radionuclides from surfaces of colloidal particles. Model calculations show that if sorption–desorption reactions are at local equilibrium (reversible sorption), then colloid-facilitated transport is not important. Contaminants associated with colloidal particles would desorb and be retained by the large excess of binding sites associated with the soil matrix. However, if sorption is irreversible or desorption is at least extremely slow compared with the residence time of the colloidal particles, then colloid-facilitated transport would theoretically be feasible. In soils, most colloids are probably transported during rainstorm events with rapid infiltration of water, and local equilibrium is therefore the exception. However, in the context of safety assessment of geologic disposal of nuclear waste, the flow velocities are extremely slow and local equilibrium may be a valid assumption. In any case, the behavior of mobile colloidal particles and associated contaminants in soils must be further studied in order to assess the importance of colloid-facilitated transport under field conditions.

Further Reading

Buffle J and van Leeuwen HP (1993) *Environmental Particles*, vol. 2. Boca Raton, FL: Lewis Publishers.

Elimelech M, Gregory J, Xia X, and Williams RA (1995) *Particle Deposition and Aggregation. Measurement, Modelling, and Simulation*. London, UK: Butterworth-Heinemann.

Goldberg S, Lebron I, and Suarez DL (2000) Soil colloidal behavior. In: Sumner ME (ed.) *Handbook of Soil Science*, Ch. 6, pp. B195–B240. Boca Raton, FL: CRC Press.

Hiemenz PC and Rajagopalan R (1997) *Principles of Colloid and Surface Chemistry*. New York: Marcel Dekker.

Kretzschmar R, Borkovec M, Grolimund D, and Elimelech M (1999) Mobile subsurface colloids and their role in contaminant transport. *Advances in Agronomy* 66: 121–194.

McCarthy JF and Degueldre C (1993) Sampling and characterization of colloids and particles in groundwater for studying their role in contaminant transport. In: Buffle J and van Leeuwen HP (eds) *Environmental Particles*, vol. 2, pp. 247–315. Boca Raton, FL: Lewis Publishers.

McCarthy JF and Zachara JM (1989) Subsurface transport of contaminants. *Environmental Science and Technology* 23: 496–502.

McDowell-Boyer LM (1992) Particle transport through porous media. *Water Resources Research* 22: 1901–1921.

Ryan JN and Elimelech M (1996) Colloid mobilization and transport in groundwater. *Colloids and Surfaces A, Physicochemical and Engineering Aspects* 107: 1–56.

Sposito G (1989) *The Chemistry of Soils*. Oxford, UK: Oxford University Press.

COMPACTION

J J H van den Akker, Alterra Wageningen UR, Wageningen, The Netherlands
B Soane, Formerly with Scottish Agricultural College, Edinburgh, UK

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Soil compaction is a form of physical degradation in which soil biological activity and soil productivity for agricultural and forest cropping are reduced, resulting in environmental consequences away from the immediate area directly affected. Compaction is a process of densification and distortion in which total and air-filled porosity and permeability are reduced, strength is increased, soil structure partly destroyed, and many changes induced in the soil fabric and in various characteristics. The term 'compaction' is used to identify a process and should be distinguished from the term 'compactness,' which indicates for a given time and position the state of packing of the solid soil constituent.

The compaction process can be initiated by wheels, tracks or rollers, traffic of cultivation machinery, and passage of draft or grazing animals. Soil densification can also be caused by heavy overburdens of ice and snow and by illuviation of clay from the A-horizon into the B-horizon. These processes may take place slowly over long periods of time and result in the B-horizon having much higher bulk density and strength than the A-horizon. Densification also occurs during structural collapse near the soil surface due to the impact of rain, resulting in formation of surface crusts.

In arable land with annual ploughing, both topsoil and subsoil compaction should be considered. We define the subsoil as the soil below the loosened layer (about 20–35 cm thick). This definition of the subsoil includes the panlayer as the upper part of the subsoil. This panlayer is, in many cases, less permeable for roots, water, and oxygen than the soil below it and is the bottleneck for subsoil functions. In contrast to the topsoil, the subsoil is not loosened annually, compaction is cumulative and, in the long run, a more or less homogeneous compacted layer is created. The resilience of the subsoil for compaction is low, and subsoil compaction is at least partly persistent.

Problems of compaction are widely distributed throughout the world, but tend to be most prevalent where heavy machinery is used in agriculture or forestry, in both temperate and tropical areas. Soils that are naturally fragile in structure, such as soils of the humid tropical forest and light-textured soils in areas of low but erosive rainfall, are particularly prone to problems arising from compaction and subsequent high risks of erosion due to reduction of permeability. Compaction is now included in surveys of soil degradation, and preliminary estimates have suggested that the area of degradation attributable to soil compaction may equal or exceed 33 Mha in Europe and 18 Mha in Africa. A project on mapping of Soil and Terrain Vulnerability in Central and Eastern Europe (SOVEUR) by the United Nations Food and Agriculture Organization (FAO) and International Soil and Reference Information Centre (ISRIC) showed that compaction is the most widespread kind of soil physical soil degradation in these countries. About 25 Mha proved to be lightly and about 36 Mha moderately compacted.

Factors and Processes Affecting Distribution and Intensity of Compaction

Compaction under Running Gear

Compaction under wheels and caterpillar tracks is a dynamic process in which a soil volume under the wheel undergoes normal and shearing stresses resulting in normal and shear strains. Neighboring soil volume elements at a certain depth under a wheel are considered in Figure 1. During a wheel pass volume element 1 at a certain depth endures not only compaction but also deformation, as depicted by the volume elements 2–26. The impact of the wheel load on the physical properties of the soil depends on the strength of the soil (*see* **Stress–Strain and Soil Strength**). If the exerted soil stresses exceed the precompression stress and shear strength of the soil, then compaction will be accompanied by large deformations and macropores and structure will be remolded, resulting in degradation of soil physical qualities. Degradation of soil physical qualities will be less severe if only the precompression stress is exceeded, because then the soil will mainly compact with limited deformations, and the remnants of the biomacropores and the intra-aggregate space will still retain a certain continuity.

Loading Characteristics of Individual Wheels or Caterpillar Tracks

For practical purposes a useful parameter is the average ground contact pressure P (kPa), which is defined as the wheel load divided by the ground contact area.

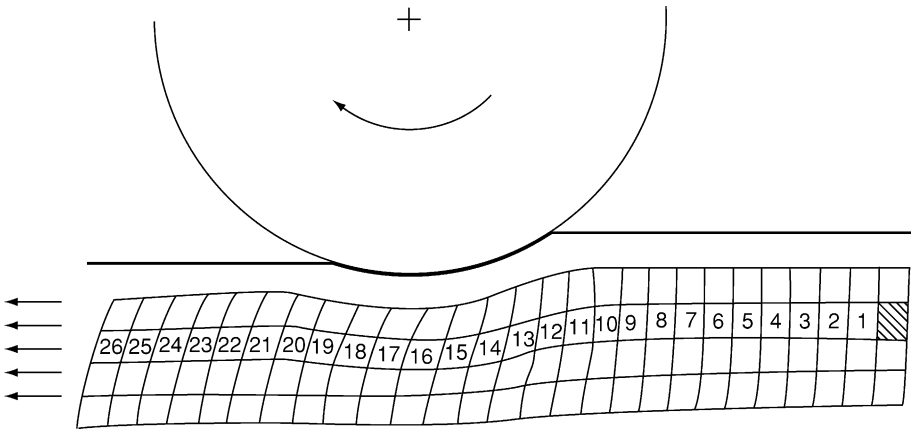


Figure 1 Displacement and deformation in a vertical section of a soil volume element under a wheel. Reproduced with permission from Koolen AJ and Kuipers H (1983) Agricultural soil mechanics. *Advanced Series in Agricultural Sciences* 13, Springer-Verlag, Heidelberg.

The average ground contact pressure can also be calculated as the sum of tire inflation pressure P_i and a pressure P_c for carcass stiffness:

$$P = c\ P_i + P_c \qquad [1]$$

The range for c is 0–1.25 and for P_c is 0–50 kPa. The factor c also depends on the carcass stiffness, and therefore eqn [1] is mostly reduced to $P = c\,P_i$ or $P = P_i + P_c$. At inflation pressures larger than 200–300 kPa, the influence of carcass stiffness diminishes and the factor c can become smaller than 1. The higher the inflation pressure, the more the strength and firmness of the soil determine the ground pressure. The peak stresses in the ground contact area determine the peak stresses in the soil that result in compaction and distortion of the soil structure if the soil strength is exceeded. Peak stresses under lugs can be two to five times higher than the average ground contact pressure. However, the resulting soil stresses decrease rapidly with depth because the ground contact area is small. If soil stresses at a depth of 0.2–0.3 m are considered, then the peak stresses under the lugs can be neglected and a parabolic stress distribution in the ground contact area with a peak stress of one-and-a-half to two times the average ground contact pressure can be assumed. Peak stresses under caterpillar tracks are two to four times the average ground contact pressure and depend strongly on the firmness of the soil and design of the track system.

Examples of load characteristics for various vehicles and animals are shown in Table 1.

Compaction by agricultural machinery extends well into the subsoil (Figure 2). An important cause of subsoil compaction is moldboard plowing in which the furrow-side tractor wheels apply appreciable loads directly to the upper surface of the subsoil, causing a plow pan. Attempts have been made to specify maximum recommended average ground contact pressures in order to minimize compaction, especially in the subsoil. Suggested maximum values range from 80 kPa for wet soils in spring to 200 kPa for dry soils in summer, but progress in gaining official acceptance for such standards has been slow.

Table 1 Maximum loads and ground contact pressures applied by various sources

Source	*Total load (kN)*	*Load per wheel/track/ hoof (kN)*	*Average ground contact pressure (kPa)*
Large-wheeled tractor (120 kW)	100	50	250–350
Small-wheeled tractor (40 kW)	40	20	200–300
Sugarbeet harvester (loaded)	300–600	50–120	200–400
Slurry tanker (loaded)	100–300	25–60	200–500
Track-laying tractor	140	70	40
Horse	8	2–8	75–300
Cow	4–5	1–4	120–480

Field Traffic Intensity and Distribution

The overall incidence of soil compaction within a given field depends upon the distribution of traffic for each field operation and the cumulative value throughout the life of the crop. The weight of the crop to be transported from the field is an important factor in the compaction risk. The traffic intensity and compaction risk for a potato crop and a sugar beet crop are, respectively, more than twice and almost twice that of a winter wheat crop. As the number and width of wheels fitted to vehicles increase, so does the overall proportion of the field area covered

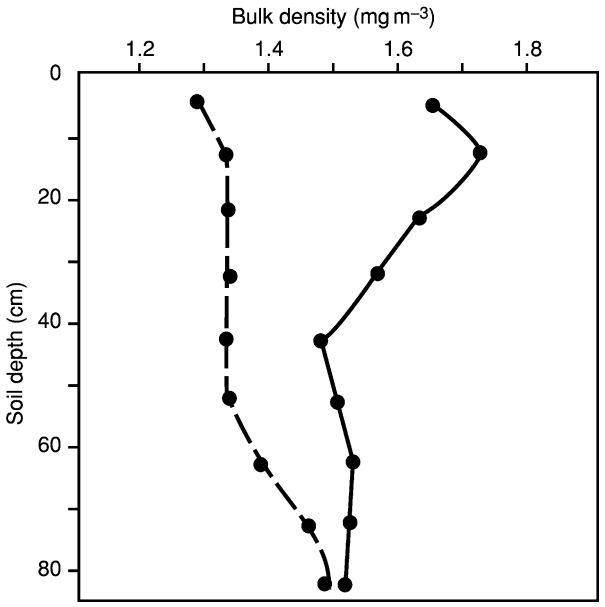


Figure 2 Increase of bulk density to considerable depth when a bulldozer was used for forest clearing in Surinam (---no vehicle traffic; — bulldozer traffic). Reproduced with permission from Van der Weert R (1974) *Tropical Agriculture (Trinidad)* 51(2): 325–331.

by wheels. However, this effect does not counteract the benefits from a reduction in compaction due to reduced ground contact pressure.

Soil Compactibility

Soils can vary from being sufficiently strong to resist all likely applied loads (low compactibility) to being so weak that they are compacted by even low loads (high compactibility). Well-structured soils combine good physical soil properties with high strength. Sandy soils with a single-grain structure and compacted massive soils can be very strong. Rootability and soil physical properties are then often poor. Roots have a binding action and increase the elasticity and resistance of a soil to compaction.

Soil moisture content and soil water suction have a dominant influence on soil compactibility. In soils where capillarity is the main cohesive force, strength increases with drying until it reaches a maximum, and then again decreases upon further dehydration because the cross-sectional area of the menisci decreases more than capillary suction between grains increases. This cohesive force based on soil water suction is called apparent cohesion. Drying also increases the true cohesion. Soil water suction increases the cohesion and so the strength and resistance to compaction of soils vary considerably. In dry structured (aggregated) soils, soil water concentrates in the aggregates, and cohesion and soil strength in the aggregate increases considerably. Dry structured soils shrink and turn into an assembly of individual aggregates that fit together rather neatly. This assembly of aggregates has a very high interaggregate angle of internal friction and moderate interaggregate cohesion. This soil is strong and has a low compactibility. However, if such a soil is overloaded and compacted, the aggregates will be crushed and the interaggregate space will be filled up, resulting in a dense compact soil.

Dry soils resist loads readily. However, extremely dry sandy soils can be deformed and compacted rather easily. As the moisture content increases, soil compactibility increases until the moisture content is approximately at the field capacity point, when a condition known as the optimum moisture content for compaction is reached. At still higher moisture contents, the soil becomes increasingly incompactible as the moisture tends to fill ever more of the total porosity and further loss of air-filled porosity becomes impossible. However, although the compaction may be minimal, the plastic flow of a wet soil results in a complete destruction of soil structure and macropores with accompanying diminishment of soil physical qualities. The reaction of dry and strong soils may be largely elastic, whereas at high moisture content and low strength the reaction may be plastic flow.

Increases in organic matter content tend to reduce soil compactibility and to increase elasticity. For example, peat soils are quite resistant to compaction. The value of the optimum moisture content for compaction increases as the organic matter content increases.

Effects on Soil Physical and Mechanical Properties

Bulk Density, Porosity, and Packing State

Bulk density (the mass of soil solids per unit volume) is the most direct and easy-to-measure indicator of changes in compactness, but changes in packing state can be better quantified by total porosity or void ratio. The relation between porosity (especially macroporosity) and soil physical properties such as saturated and unsaturated hydraulic conductivity and gas diffusivity is much closer than the relation of these soil physical properties with bulk density. Compaction causes major reductions in macroporosity ($>50\,\mu$m), reduction of total porosity, and often an increase in microporosity, resulting in a major impact on soil physical properties. Relative terms for packing state (e.g., relative density, degree of compactness) enable the same threshold values to be found for overcompaction for soils with different texture.

Hydraulic Properties

Saturated hydraulic conductivity is very sensitive to the compaction process by wheels that includes

shearing and kneading of the soil. Figure 3 shows that wet soils are easier to compact and that the effect on the saturated hydraulic conductivity is more than proportional. The destruction of soil structure under wheels explains the greater degradation of soil qualities than observed in uniaxial tests.

In the process of compaction, the macroporosity decreases, whereas the microporosity often increases. This results in larger water contents for a wide range of matric potentials in compacted versus uncompacted soil. This results in a higher unsaturated hydraulic conductivity of a compacted soil versus uncompacted soil. However, near saturation meso- and macropores are also filled and contributes to the hydraulic conductivity, and thus the hydraulic conductivity of an uncompacted soil becomes higher than that of a compacted soil.

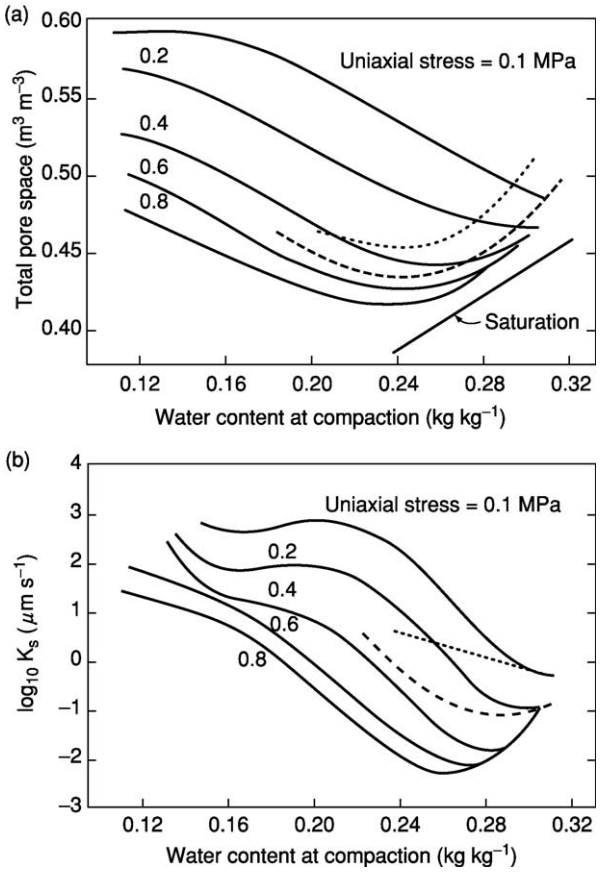


Figure 3 Changes in (a) total porosity and (b) saturated hydraulic conductivity of a sandy clay loam subjected to field traffic with low ground pressure (..... LGP, tire inflation pressure 80 kPa) or high ground pressure (- - - HGP, tire inflation pressure 240 kPa) or subjected to uniaxial stresses of 0.1, 0.2, 0.4, 0.6, and 0.8 MPa at various water contents. Solid lines refer to compression of aggregated mixtures. (After Dawidowski and Lerink, 1990.) Reproduced with permission from Horton R, Ankeny MD, and Allmaras RR (1994) *Soil Compaction in Crop Production.* Amsterdam: Elsevier.

Aeration Characteristics

An essential feature of compaction is reduction in the air-filled pore space. This property is an indication of the aeration status of the soil for plant growth and microbial activity. At air-filled porosity values less than 10%, oxygen deficiency is likely, especially in warm weather, while values less than 5% probably indicate incipient anaerobiosis. Analogous to saturated hydraulic conductivity, aeration status depends strongly on structure and continuous macropores. In a poorly structured compact soil, the threshold values of 10% and 5% should be doubled. In a well-structured soil, these threshold values may be halved. Other indices of aeration status, such as oxygen diffusion rate, oxygen content of soil air, and air permeability, provide more precise indicators of the aeration level.

Strength Characteristics

In most cases, the strength of a soil is increased by compaction. However, a well-structured soil is stronger than a poorly structured soil and a wet soil is weaker than a dry soil. In periods with wet-weather conditions, compacted poorly structured soils tend to become wetter than uncompacted well-structured soils due to limited infiltration capacity, lower saturated hydraulic conductivity, and high microporosity of compacted soils. In these circumstances, the strength and trafficability of compacted soils become lower than that of well-structured soils. After a wet period, a compacted soil will stay wet longer with limited workability. After a dry period, the strength of a compacted soil will increase considerably and result in high trafficability. Soil tillage then requires more powerful equipment and a higher energy input to loosen the soil and reduce clods to acceptable sizes.

Compaction increases the penetration resistance for roots, resulting in limited rooting depth and reduced crop growth. At penetration resistances (measured with a cone with a diameter of 12.7 mm and top angle of 30 degrees) of 1.5 MPa and 3.0 MPa, root growth rates are reduced to 50% and 0% respectively. However, in well-structured soils, roots can make use of continuous macropores to penetrate deeply into the soil. In drying soils, strength and penetration resistance increase.

Compaction in Crop Production

Compacted soils usually have unsatisfactory physical conditions for plant growth, but the extent of loss of productivity depends on soil type, plant species, and weather conditions.

Effects on Germination and Establishment

Compact soils result in cloddy seedbeds after primary tillage, poor soil/seed contact, and reduced germination. Soils may develop a compact surface layer due to crusting after heavy rainfall, which has sufficient strength to restrict or even to inhibit seedling emergence, especially of dicotyledonous species.

Effects on Root Growth and Distribution

Macropores (>50 μm diameter), through which roots can generally proliferate readily, are much reduced in compacted soil. As a result, root growth is restricted or even inhibited. In Figure 4 the limitations to root growth are conceptualized as relations between soil porosity and soil water potential at which soil aeration and mechanical resistance meet specified root requirements. In a soil with a certain pore volume, the soil water suction determines whether root growth is limited by too high mechanical resistance, by aeration problems, or is not limited. In Figure 4 the two thin lines show the situation in the case of a poorly structured soil. A poorly structured soil with few continuous macropores needs more air-filled pores for aeration than a well-structured soil. Also much lower mechanical resistances are allowed in a poorly structured soil than in a well-structured soil because roots can follow the macropores in the well-structured soil.

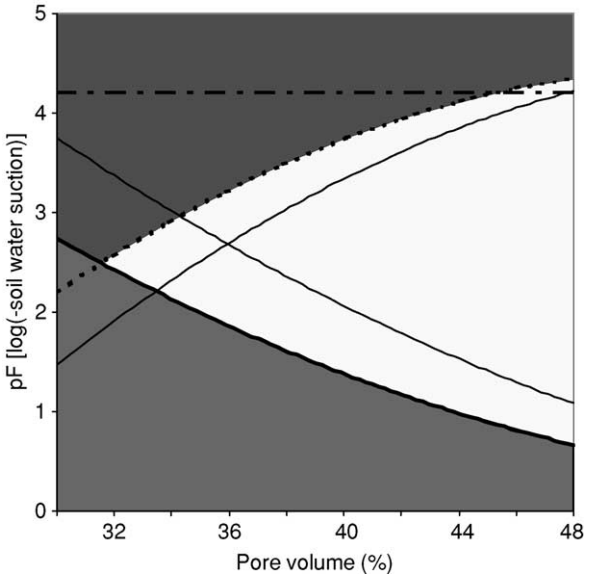


Figure 4 A conceptual relationship between soil porosity and soil water potential in which soil aeration and mechanical resistance (PR, penetration resistance) meet specified root requirements. Root growth is insufficient in the shaded areas and impossible beyond a soil water potential of 1600 kPa. Added to the figure are the same relationships if the structure is deteriorated resulting in a strong reduction of macropores. PR too high; too wet, aeration too low; rootable; PR limiting; poorly structured soil; aeration limiting; too dry. (After Boone, 1988.) Reproduced with permission from Lindstrom MJ and Voorhees MB (1994) *Soil Compaction in Crop Production*. Amsterdam: Elsevier.

Effects on Plant Growth and Yield

The level of compaction (often expressed in dry bulk density) influences the growth, yield, and quality of crops, depending on crop species, soil type, and weather conditions (Figure 5). Optimum level of compaction tends to be higher for sandy soils, in dry seasons, and for monocotyledonous species. At a compaction level less than optimum, crops suffer from reduced soil/root contact, reducing germination and nutrient transfer, while at a compaction level higher than optimum, root growth and aeration are restricted and denitrification can lead to N losses.

Soil biota and biological processes are influenced by soil compactness. This is partly because of the influence of pore size distribution on spatial habitats for bacteria and fungi and partly because compacted soils may suffer from anaerobiosis, which in turn affects microbial metabolism markedly. The burrowing abilities of soil fauna, particularly earthworms, are reduced in compacted soils.

Interactive Crop Responses to Compaction and Fertilizer Application

Where farmers perceive the growth of crops to be adversely affected on compacted soils, a common action is to apply additional N fertilizer. However, the growth responses obtained may be less significant than those that would accrue at lower levels of compactness (Figure 6), and the additional nitrogen applications may be inefficient economically and detrimental to the environment.

Crop Responses to Subsoil Compaction

Amelioration of subsoil compaction is much more expensive and less effective than loosening compacted topsoils. Avoidance of compaction in the subsoil is therefore a much greater need than in the topsoil. Seventeen years after a single full-field compaction action, crop and nitrogen yield losses can still be significant on the compacted sites (Figure 7). After this compaction experiment only moderate axle loads were allowed on the fields. In practice, most subsoil in agricultural fields are partly compacted every year, and the subsoil quality is often worse than in these long-term experiments. Notice that the compaction effect was more pronounced in all years for harvested nitrogen than for grain dry matter. The strong decline of the effect in the first few years was mainly due to

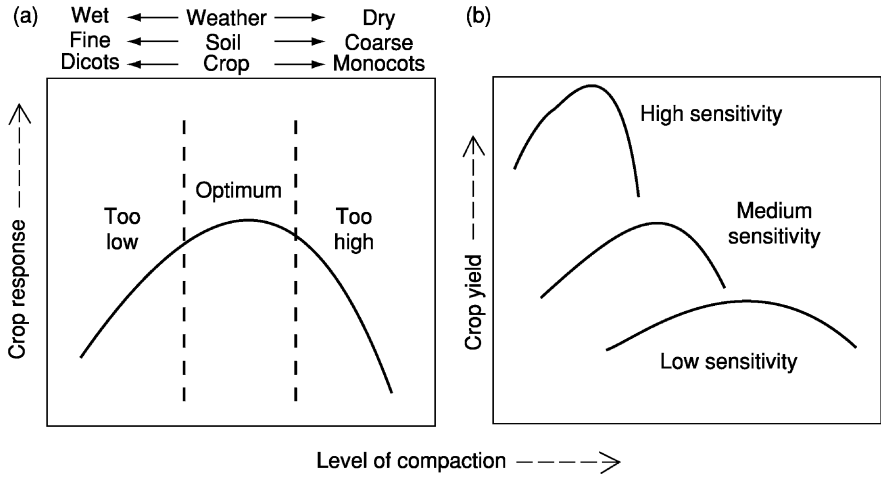


Figure 5 Conceptual relationship of crop response to level of soil compaction in relation to weather, soil texture and (a) crop type and (b) crop sensitivity. Reproduced with permission from Lindstrom MJ and Voorhees WB (1994) *Soil Compaction in Crop Production*. Amsterdam: Elsevier.

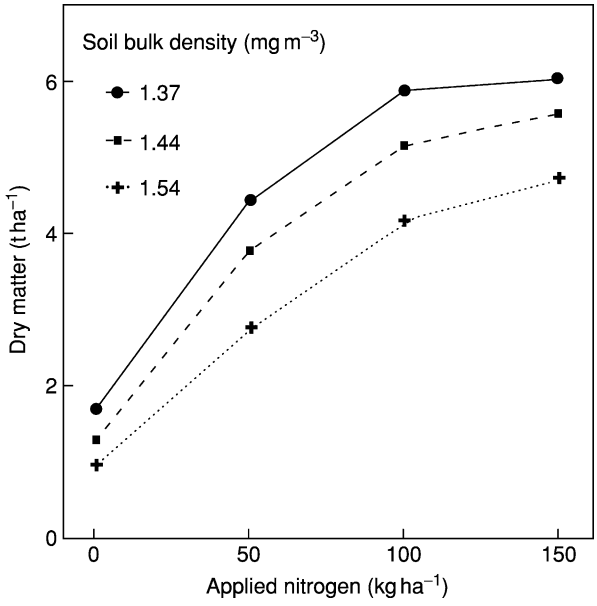


Figure 6 Dry-matter yields of ryegrass (first cut, mean of 3 years) in response to applied nitrogen fertilizer at three levels of bulk density (3–12 cm depth). Adapted from data in Douglas JT and Crawford CE (1993) *Grass Forage Science*. Oxford: Blackwell Science.

the recovery of the topsoil. The effect was influenced by rainfall during the subsequent growing seasons. In dry seasons a moderately compacted subsoil may result in yield advantages due to reduced loss of soil water percolating below and out of the reach of the root zone, whereas in wet seasons, the reduced permeability of the subsoil can lead to anaerobic conditions within the topsoil, resulting in direct damage to the crop and loss of nitrogen by denitrification.

Modeling of Crop Responses to Soil Compaction

Crop growth is less than potential when the uptake of water, oxygen, or nutrients is less than the demand of the crop. Potential crop growth is determined considering the prevailing weather conditions. Reduced crop growth may be caused by reduction of the length of the growing period, low temperature, limited supply from the soil of water, oxygen, and nutrients to the root system, and a limited activity of the root system. Soil water plays a central role in these limiting factors, and effects of soil compaction on crop growth and biological functioning should be modeled in relation to water. In Figure 8 a scheme is presented of the interrelationships among soil tillage, field traffic, soil structure and soil physical, chemical, and biological properties. In Figure 9 a part of this scheme is considered in more detail. To simulate crop responses to soil compaction all aspects presented in Figures 8 and 9 should be included in the model. However, up to now no model exists that includes all aspects, such as limitations of root growth, the role of macropores, reduced availability of nutrients (e.g., loss of nitrogen by denitrification), and effects of reduced biological activity. The existing models in general underestimate the impact of compaction on crop growth.

Effects on Environmental Components

Soil compaction influences a number of environmental parameters even at considerable distance from the original location at which the compaction occurred (Figure 10). Compaction may change the fluxes of greenhouse gases from the soil to the atmosphere

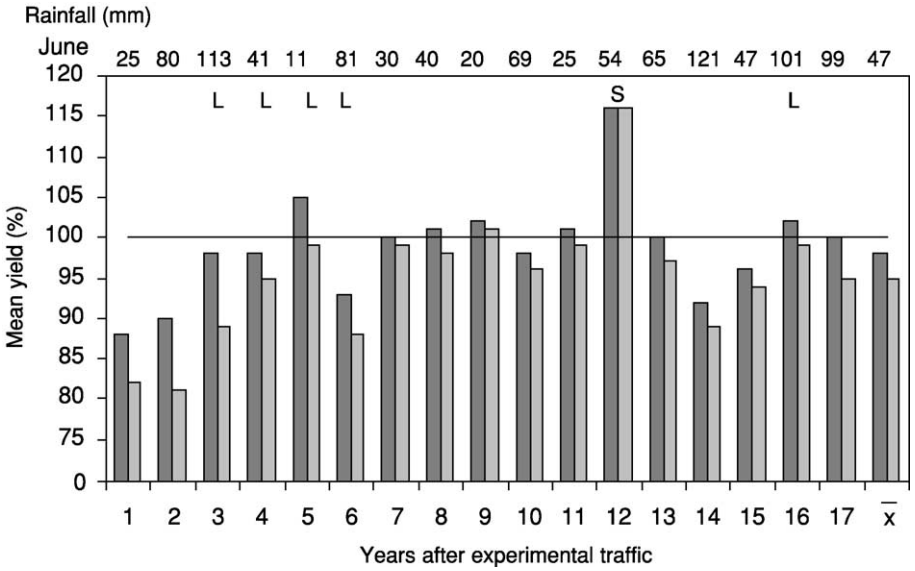


Figure 7 Mean grain and nitrogen yields of annual crops in control treatment (=100%), and relative to the control in loading treatment of four passes with a 50 kN axle load in 1981 on clay soil, for 17 successive years after the loading. L, lodging; S, sprouting; ■ relative grain yield (%); □ relative nitrogen yield (%). Reproduced with permission from Alakukku L (2000) *Advances in GeoEcology* 32. Reiskirchen: Catena Verlag.

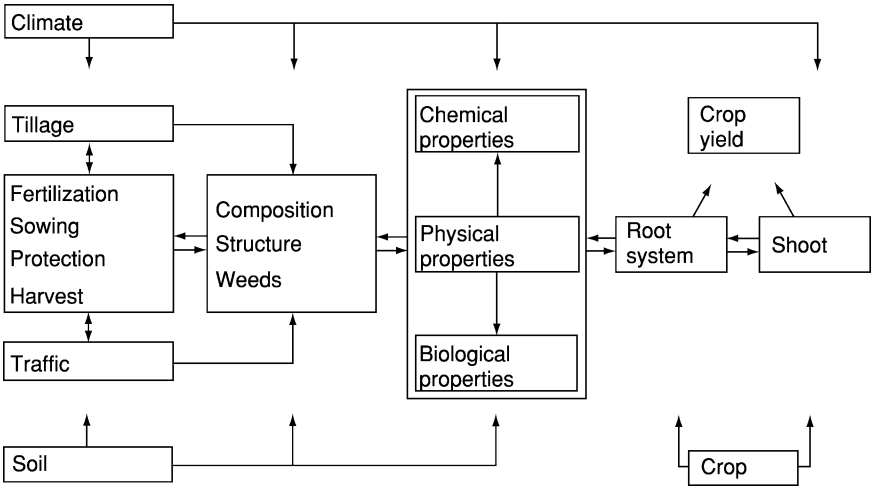


Figure 8 Interrelationships among climate, soil management, soil properties, and crop growth. Reproduced with permission from Boone FR (1988) *Soil and Tillage Research*. Amsterdam: Elsevier.

through mechanisms associated with effects on soil permeability, aeration, and crop development. Compaction increases CO_2 emissions because cultivation of compacted soils requires appreciably more energy than cultivation of uncompacted soils. Approximately 90% of the global N_2O emissions to the atmosphere comes from soils. Compacted soils tend to be wetter than noncompacted soils and denitrification is enhanced. The flux of N_2O increases rapidly as air-filled porosity declines (Figure 11). Compaction from vehicle traffic prior to the establishment of cereal crops can cause marked increases in the N_2O flux during the early growth period in spring (Table 2).

Due to reduced permeability, compacted soils usually show greater runoff and hence greater erosion than noncompacted soils. Surface rills and even gullies are sometimes directly associated with wheel tracks, particularly over seedbeds following periods of highly erosive rainfall. Surface waters may thus carry additional burdens of clay and silt, fertilizer, and pesticides when runoff occurs from areas of compacted soil.

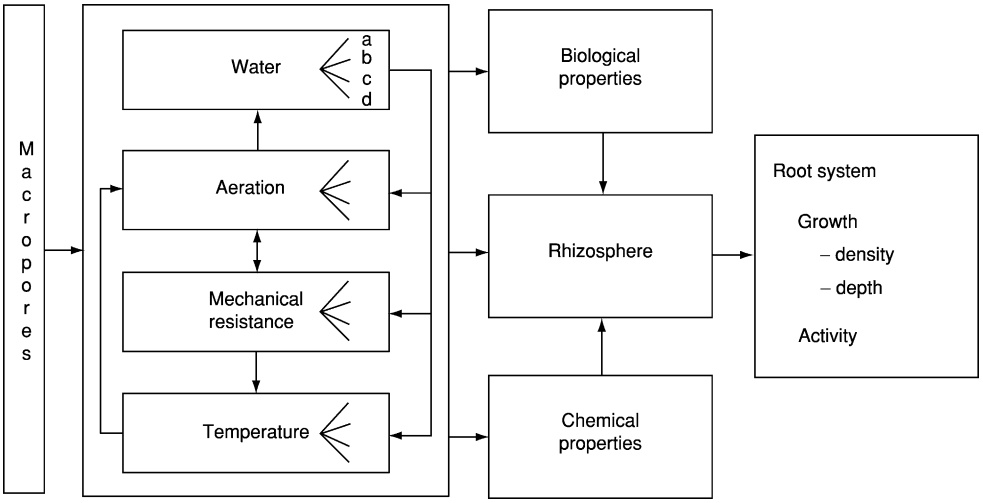


Figure 9 Relationships among soil macropores, soil physical, chemical, and biological properties, rhizosphere and root system. a, surface boundary; b, storage; c, transport; d, sink or source aspects. Reproduced with permission from Boone FR (1988) *Soil and Tillage Research*. Amsterdam: Elsevier.

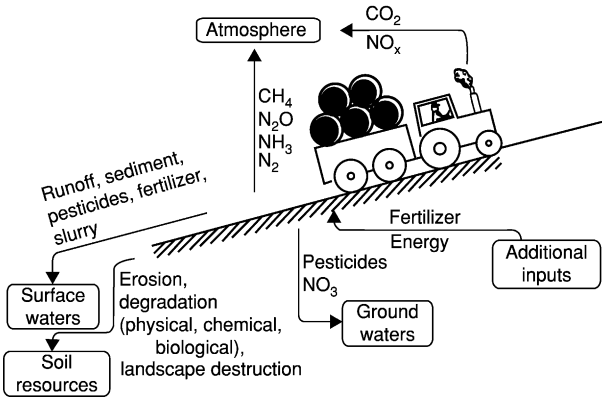


Figure 10 A conceptual diagram showing the major pathways whereby compacted soil conditions may influence components of the environment. Reproduced with permission from Soane BD and Van Ouwerkerk (1995) *Soil and Tillage Research*. Amsterdam: Elsevier.

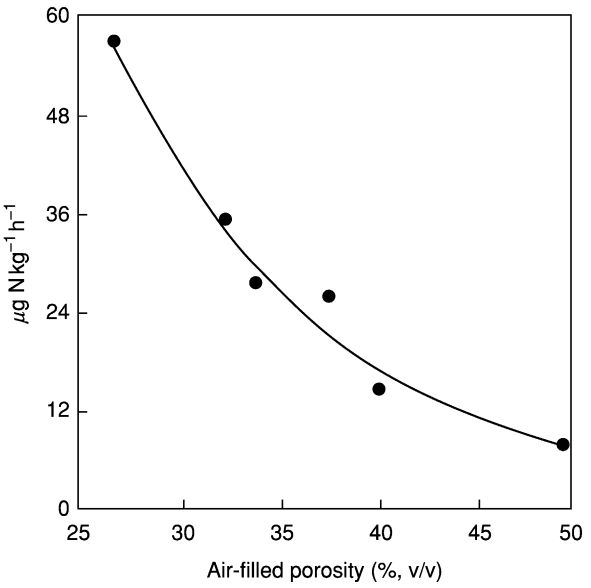


Figure 11 Increased emission of nitrogen by denitrification as air-filled porosity decreases. Reproduced with permission from Sextone AJ, Parkin TB, and Tiedje JM (1988) *Soil Biology and Biochemistry*. Amsterdam: Elsevier.

Table 2 Influence of vehicle traffic (zero, light, heavy) prior to the establishment of wheat and spring barley on N_2O flux during the early spring growth period.

Crop	Period (dates and days)	Cumulative N_2O flux ($g\ N_2O\text{-}N\ ha^{-1}$)		
		Zero	*Light*	*Heavy*
Spring barley	16 May–14 July = 60 days	320	310	401
Winter wheat	8 March–8 May = 62 days	245	210	578

Adapted from data in Ball BC, Parker JP, and Scott A (1999) *Soil and Tillage Research*. Amsterdam: Elsevier.

Techniques for the Reduction of Compaction

The compactive capability of vehicles can be minimized by reductions in overall mass, increases in ground contact area, and reduction of ground contact pressure. Tires of greater width and diameter will increase ground contact area, as will reduction in inflation pressure. Dual wheels, multiple axle systems, especially tandem axles for trailers and tankers with low-pressure tires, provide extra contact area.

Vehicle traffic can be reduced by combining different operations into a single-pass operation. Toolcarriers (gantries), up to 12 m width, have been found capable of providing traffic-free zones in which, because of the lack of compaction, tillage requirement is reduced and crop yield and quality may be improved.

Trafficking and working soil when it is too wet should be avoided. Tillage should be reduced as much as possible to optimize biological and physical processes that improve soil structure, in particular macroporosity.

Improving drainage will result in drier and stronger soils. Increases in soil organic matter, either as a surface mulch or incorporated, will increase soil elasticity and reduce compaction.

Amelioration of Compacted Soils

Natural weathering due to freezing and thawing is usually restricted to the top few centimeters of the surface soil and rarely penetrates to the subsoil. Swelling and shrinking arising from changes in water content can cause the loosening of compacted soils. Roots of certain species can penetrate compacted layers and soil biota can slowly increase macroporosity. In this way, gradual improvements can be made in the soil physical quality of compacted layers. However, highly compacted parts, in particular layers that cannot be penetrated by roots, will not or only slowly recover.

Compacted topsoils can be loosened by tillage. However, loosened compacted soils are cloddy and require additional secondary cultivation to achieve suitable seedbed tilth. Loosening subsoils requires high-draft special equipment and high traction, destroys still existing continuous macropores, and weakens soil structure and strength. Subsoiling should be done when the subsoil is dry with a minimum of loosening and with the aim to form cracks (*see* **Subsoiling**). Many loosened subsoils recompact within a few years with worsened soil physical properties and rootability. The result is that subsoils that have been loosened once must be loosened regularly every 4–5 years. Therefore the subsoil should be inspected beforehand to determine rootability, aeration status, and drainability, to avoid unnecessary subsoiling.

See also: **Conservation Tillage**; **Cultivation and Tillage**; **Stress–Strain and Soil Strength**; **Structure**; **Subsoiling**

Further Reading

Håkansson I (ed.) (1994) Subsoil compaction by high axle load traffic. *Special Issue of Soil Tillage Research* 29: 105–306.

Horn R, van den Akker JJH, and Arvidsson J (eds) (2000) Subsoil compaction: distribution, processes and consequences. *Advances in GeoEcology* 32.

Koolen AJ and Kuipers H (1983) Agricultural soil mechanics. *Advanced Series in Agricultural Sciences* 13.

Larson WE, Blake G, Allmaras RR, Voorhees WB, and Gupta S (eds) (1989) Mechanics and related processes in structured agricultural soils. *NATO ASI Series E: Applied Sciences* 172.

Monnier G and Goss MJ (eds) (1987) *Soil Compaction and Regeneration.* Rotterdam, Netherlands: A. A. Balkema.

Oldeman LR, Hakkeling RTA, and Sombroek WG (1991) *World Map of the Status of Human-induced Soil Degradation, an Explanatory Note*. Wageningen, Netherlands: ISRIC.

Pagliai M and Jones RJA (eds) (2002) Sustainable land management–environmental protection: a soil physical approach. *Advances in GeoEcology* 35.

Soane BD and Van Ouwerkerk C (eds) (1994) *Soil Compaction in Crop Production.* Amsterdam: Elsevier.

van den Akker JJH, Arvidsson J, and Horn R (eds) (2003) Experiences with the impact and prevention of subsoil compaction in the European Community. *Special Issue of Soil Tillage Research* 73: 1–186.

Van Ouwerkerk C (ed.) (1988) Tillage and traffic in crop production. *Special Issue of Soil Tillage Research* 11: 197–372.

Van Ouwerkerk C and Soane BD (eds) (1995) Soil compaction and the environment. *Special Issue of Soil Tillage Research* 35: 1–113.

COMPOST

T L Richard, Iowa State University, Ames, IA, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Compost has been used to improve agricultural soils for hundreds of years, but only in the last few decades have we begun to understand the science behind this practice. This increase in scientific research parallels a dramatic growth in compost use, motivated both by growing demand from organic and conventional farmers, landscapers, and home gardeners, and also by a burgeoning supply of compost from municipal, industrial, and agricultural waste treatment. The resulting diversity of compost feedstocks, processing technologies, and marketing and utilization strategies results in a wide range of compost qualities and characteristics. Each compost has its own attributes, so that generalizations about compost behavior must be fine-tuned with knowledge of the particular compost and application.

When a compost is applied to soil, it initiates a cascade of changes in soil physical, chemical, and biological properties. While the direction of many of these changes is predictable based on soil and compost characteristics, their magnitude varies as a result of the dynamic interactions between these two diverse and biologically active porous media. Understanding these complex interactions is important for both scientific advances and practical management of compost in agronomic and horticultural systems.

Compost Production

Just as soil genesis impacts the structure and function of soil, the composting process plays a key role in the subsequent structural and functional attributes of compost. Compost can be produced at scales ranging from small backyard piles to large industrial factories, with management ranging from benign neglect to intensive engineering. Most composting attempts to encourage aerobic decomposition, and includes an initial thermophilic period that lasts a few days to a few weeks. However, there are exceptions to even this basic characterization of composting as an aerobic, thermophilic process. Anaerobic composting, a form of high-solids anaerobic digestion, is used to produce methane gas as well as stabilize various organic wastes. And ambient temperature processes include both sheet composting, where thin layers of organic material are applied to the soil surface, and vermicomposting, where worms are involved in the stabilization process.

The fundamental requirements for composting are relatively simple: fresh organic substrate, a diverse microbial population, adequate aeration and moisture, and a strategy for managing excess heat. The substrate provides readily biodegradable energy and nutrient sources for the microorganisms, which in the presence of oxygen and moisture will catabolize the substrate to sustain their populations and support active biomass growth. Enzymatic hydrolysis decomposes simple and complex carbohydrates, crude protein, and lipid molecules, some of which are used for microbial growth, while others are mineralized. If oxygen is available the mineralized carbon will be released almost entirely as carbon dioxide, with the hydrogen and oxygen released as water.

Aerobic decomposition releases large amounts of energy, about $14\,kJ\,g^{-1}$ O_2 consumed, and this energy heats the pile to well above ambient temperature. Higher temperatures increase the rate of microbial decomposition, with the maximum rate typically observed around 60°C. At this temperature degradation rates can be five or more times as fast as at 20°C. This thermophilic phase of composting destroys most weed seeds and pathogens, with pathogen reduction requirements for regulated feedstocks or products typically ranging from 3 to 15 days at 55°C or above. These time–temperature combinations have been shown to reduce many types of human, animal, and plant pathogens to below infectious levels and, when combined with adequate stabilization of the feedstocks, prevent pathogen regrowth in the finished product. Although higher temperatures destroy pathogens more quickly, temperatures above 65–70°C cause a steep decline in reaction kinetics and a dramatic reduction in microbial diversity. The combination of advantages and risks associated with operating in the optimum thermophilic range make temperature management a critical factor in controlling the composting process.

Temperature management and oxygen supply are coupled because most heat is removed by the convective flow of oxygen-supplying air, and because the airflow rates needed for adequate heat removal normally provide sufficient oxygen to maintain aerobic conditions. Most composting systems rely on passive aeration by wind, diffusion, and natural convection. Other systems use mechanical aeration, often with temperature-feedback controllers operating a system of blowers and fans. In either case adequate porosity

and permeability of the pile are necessary for uniform distribution of air, as excess moisture or compaction leads to anaerobic pockets where odorous gases can form. This requirement places practical limits on pile dimensions and moisture content, with systems greater than 2–3 m in height or less than 40% solids potentially at risk.

Degradation follows first-order kinetics after a short initial lag period, which can be minimized by recycling mature compost as a microbial inoculant. Oxygen demand and heat production are greatest during the first few days and weeks, when degradation rates are at their maximum. If oxygen becomes limiting during this period anaerobic fermentations form organic acids, which accumulate and result in a measurable but usually ephemeral reduction in pH. These acids normally degrade aerobically as oxygen becomes available, but if they persist their phytotoxicity can affect the compost product. During the thermophilic phase, compost can lose considerable quantities of nitrogen, with overall losses ranging from less than 10% to greater than 60% of the initial mass of N. These losses include volatilization of NH_3, N_2, and N_2O, and in outdoor sites can also include runoff and leaching of NH_4^+ and NO_3. Total N losses are inversely correlated with C:N ratio, and are minimized for C:N ratios greater than 30:1. The ammonia volatilization component of total N loss has been positively correlated with increasing airflow rates, temperatures, and pH.

After a thermophilic period ranging from a few weeks to months, temperatures decline to the mesophilic range as readily degradable compounds become increasingly scarce. The compost then begins a curing or maturation phase, which is characterized by continuing decomposition of hemicellulose, cellulose, and other slowly degradable fractions, reductions in particle size, slowing respiration rates, and increasing humic acid content. Lower temperatures can facilitate nitrification, which increases the plant-available N but can also denitrify and increase N losses. Over extended curing periods of 6 months to 2 years, nitrate can accumulate to a significant fraction of the total compost N. Figure 1 shows these N transformations over time during the composting process.

Compost will continue to stabilize indefinitely, but is often marketed and applied to soil when cumulative carbon mineralization is in the range of 60–70%. Depending on the original bulk density and ash content, as well as factors such as soil or inert matter incorporated during composting, the total volume reduction may be in this same range or somewhat higher, while dry-matter reduction may be somewhat less than 60%. P, K, and other mineral elements

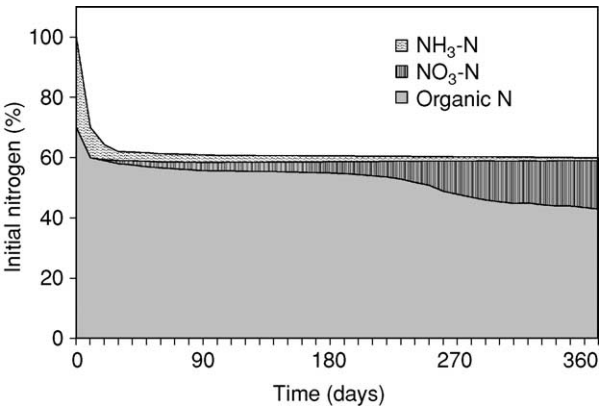


Figure 1 Progressive changes in N partitioning during the composting process. In this example total N loss during composting was 40%.

are normally largely conserved, although K and other soluble elements can leach from systems in outdoor piles. For all the reasons mentioned above, macronutrient contents vary widely in different compost products, as shown in Table 1. Moisture content can also vary from product to product, although the final moisture content from industrial composting systems is often around 40% (wet basis), which allows easy application without excessive dust.

Compost Application and Incorporation

For any particular application, appropriate compost application strategies are a function of both a particular compost's characteristics and specific soil-management goals. Compost application rates vary widely, from as little as 1 Mg ha^{-1} for a nutrient-rich compost in agronomic applications, to more than 60 Mg ha^{-1} for remediation of disturbed sites or severely depleted soils. Compost can also be used to manufacture artificial soils, for landscaping, for ecologic restoration of mines, brownfield sites, and wetlands, or in high-value commercial media such as nursery or greenhouse potting soils. The compost fraction used in these blends is often in the 10–50% range, with higher ratios possible in artificial wetlands and some other specialized situations. Although most of the benefits of compost increase with application rate, some caution must be exercised to insure that both annual rates and cumulative applications are not excessive with respect to nutrients, trace metals, problematic biomolecules, and potential phytotoxic effects. Heavy metal and nutrient loading rates are sometimes regulated to prevent negative environmental impacts, while phytotoxicity concerns can be addressed by insuring adequate compost maturity, sometimes combined with intentional leaching to remove soluble salts.

Table 1 Ranges of nutrient composition in compost products (multisite means in parentheses)

Compost type	*N (g kg^{-1} (dry weight))*	*P (g kg^{-1} (dry weight))*	*K (g kg^{-1} (dry weight))*
Livestock manure	11–37 (20–31)	3–20 (9–20)	2–27 (15–26)
Yard trimmings	5–42 (6–21)	0.5–12 (2–3)	0.4–27 (3–10)
Biosolids	7–38 (18–35)	5–41 (7–21)	0.7–10 (4–5)
Municipal solid waste	4–17 (7–13)	0.8–6 (2–3)	2–10 (3–5)

Compost is normally applied to soil using equipment and techniques designed to handle solid manure, although specialized equipment is available for targeted applications. With drier, more friable composts, spinning and rotating brush spreaders can be used on cropland and turf, while blower hoses can be used for landscaping and erosion control. As with manure, the application rate and degree of incorporation affect the potential for short-term environmental impacts through volatilization, leaching, and runoff, as well as long-term dynamics as the compost interacts with the soil. Surface application of compost is less problematic than with manures, as nutrients are in more stable forms, and a surface layer of compost increases infiltration rates, reduces erosion, and in thicker layers can suppress weed growth. However, incorporation is also common in both agricultural and horticultural applications to distribute the compost product throughout the surface horizons.

Compost–Soil Interactions

The application of compost to soil initiates a series of physical, chemical, and biological transformations that affect both soil properties and processes. Some of these effects are strong initially but decrease over time, others are more persistent, while some are latent and may only manifest under certain soil, crop, or climatic conditions. Table 2 lists some of the most important impacts of compost on soil and indicates the typical direction of change.

Table 2 Results of compost–soil interactions

	Typical change
Physical	
Infiltration	Increases
Erosion	Decreases
Aggregate stability	Increases
Water-holding capacity	Increases
Total porosity	Increases
Permeability	Increases
Bulk density	Decreases
Chemical	
pH	Buffers near neutral
Cation exchange capacity	Increases
Electrical conductivity	Increases
Nutrient concentration	Increases
Nutrient availability	Varies
Trace elements and metals	Varies
Biological	
Microbial activity	Increases
Microbial biomass	Increases
Microbial diversity	Increases
Enzymatic activity	Increases
Phytotoxicity	Varies
Phytostimulation	Increases
Plant disease suppression	Varies

Physical Effects

Compost affects several critical soil physical functions, including water transport and storage, gas exchange, and heat transfer. At the soil surface, compost provides a barrier to raindrop impact, reducing surface sealing and allowing rapid infiltration of rainfall or irrigation water. These factors can delay the onset of runoff dramatically, with 5-cm-thick blanket applications entirely eliminating runoff and thus erosion in all but the most severe storms. When extreme events do occur, compost blankets can reduce erosion by one to three orders of magnitude. These benefits occur immediately after application and thus protect the soil even prior to vegetation establishment.

Deeper in the soil profile, the structural effects of compost particles typically increase both soil porosity and permeability, enhancing gas transfer as well as saturated water flow. Total porosity can increase by 50%, while permeability can increase by up to two orders of magnitude at high compost application rates. However, these potential porosity and permeability gains can be compromised by excessive compaction under wet conditions, when the mechanical strength of these particles is low. Depending on its organic matter content, compost particle density can range from 1600 to 2000 kg m^3, and if porosity has not been compromised this relatively low particle density will reduce the dry bulk density of the soil.

Over time, organic compounds introduced with the compost and produced by subsequent biological activity work to promote soil aggregation and enhance aggregate stability. These soil aggregates, as well as the compost itself, increase water-holding capacity, which can improve irrigation efficiency and mitigate drought stress. These effects occur not just in the surface layers where compost is typically

incorporated, but also in subsurface layers where soluble organic carbon can penetrate by leaching from above. As with other organic amendments, the compost encourages earthworm activity, generating biological macropores, which facilitate deep percolation and drainage of excess moisture.

Many of these physical improvements associated with compost application peak within a few weeks or months after compost is applied and then slowly attenuate over time. While a single high-rate compost application ($>20\,\mathrm{Mg\,ha^{-1}}$) can have significant, lasting effects on soil physical properties, these are more commonly achieved by repeated compost applications over a period of years.

Chemical Effects

When compost is incorporated in soil, there are immediate, calculable changes in concentrations of nutrients, trace metals, and other chemical compounds that result from the application rate and composition of the two materials. However, the organic matter in compost also stimulates changes in biological activity of the soil that have longer-term impacts on soil chemical properties. These compost–soil dynamics are complex, and few studies have characterized the precise mechanisms or kinetics of the resulting chemical transformations and fate. Nonetheless, there are several areas where current understanding can provide important insights and practical guidance.

Carbon is the dominant element in composts, primarily occurring as the structural backbone of organic biomolecules. These biomolecules have open ion exchange sites that can react with other compounds in the soil system, increasing the overall cation exchange capacity (CEC) and chelating minerals and heavy metals in the soil. Some composts also include significant amounts of mineral carbonates, either from feedstocks or accidentally blended in from the surface of compost pads.

Both carbonates and organic carbon in compost can affect the soil pH. Carbonates raise pH through their affinity for hydrogen ions, while organic carbon tends to buffer the system somewhere near neutral. Addition of an immature or anaerobic compost can temporarily lower pH by introducing organic acids, although these degrade rapidly if the soil system is aerobic. In general, mature composts will modify soil pH toward neutral to slightly alkaline levels (pH 6.8–7.8), with larger increases typically associated with composts made with biosolids or other wastewater sludges that have had lime added during dewatering or other treatment processes.

One of the potential negative impacts of compost on soil is the addition of soluble salts. This is a particular concern with manure and biosolids-derived composts but is also possible with composts that include particular food-processing feedstocks and even leaves from municipalities where salt is used for roadway deicing. Electrical conductivity (EC) provides a measure of salt concentration, with levels of less than $4\,\mathrm{dS\,m^{-1}}$ in composts or $1.2\,\mathrm{dS\,m^{-1}}$ in soil–compost mixtures tolerated by all but the most sensitive plants. In situations where high application rates of a high-EC compost may cause problems, leaching by either irrigation or natural rainfall can reduce levels significantly. However, some of the soluble ions that register as EC are also beneficial minerals and nutrients that will be removed by excess leaching.

The CEC of soils facilitates retention of positively charged minerals and nutrients (Ca^{+}, K^{2+}, NH_4^{+}, etc.) for subsequent use by microorganisms and plants. Organic matter, along with clay minerals, is known to have a positive impact on CEC, with humic and fulvic acids playing a particularly important role. While the compost application rates common in agronomic situations are generally not high enough to result in measurable increases in soil CEC, significant increases are observed at high application rates. These increases complement the increased water-holding capacity of compost-amended soils, reducing the potential for minerals and nutrients to leach out of the plant root zone.

Nitrogen remains the critical nutrient in most agricultural and horticultural systems, so its availability from compost is of particular importance. Nitrogen dynamics in compost-amended soils can result in a wide range of outcomes, from net mineralization and a significant fertility boost to net immobilization resulting in crop nitrogen stress. The primary factor influencing this result is the compost C:N ratio, as excess carbon can stimulate microbial growth that immobilizes mineral N from both compost and soil sources. Other compost-specific factors include C and N availability, which can modify the impact of a particular C:N ratio, and particle size, as smaller particles will mineralize more rapidly. Soil conditions also have a strong influence on nitrogen mineralization, especially soil organic matter status, temperature, moisture, oxygen, salinity and pH, with optimum conditions similar to those in soil without compost. The higher biological activity associated with compost application can, however, cause negative perturbations. For example, unstable composts with high respiration rates can deplete oxygen levels in soil microenvironments, increasing denitrification rates and volatile nitrogen loss. The risks of potential negative impacts are minimized by use of stable compost products with a C:N ratio of less than 15:1.

When soil conditions are supportive of microbial growth, compost can prime the decomposition of

existing soil organic matter and stimulate additional N mineralization, providing fertility benefits beyond those associated with the added compost itself. Similarly, synergistic effects can occur when both compost and synthetic fertilizers are applied to soil, with higher net N mineralization rates observed than with either amendment alone. A promising area of current research is in the synchronization of organic N mineralization with crop nutrient demand, which may demonstrate additional advantages for the use of compost as a component of soil-fertility management.

The interactions between these soil and compost factors result in variable mineralization rates and outcomes for nitrogen as well as other macronutrients. Figure 2 shows the range of mineralization rates for N, P, and K, reported in several experiments using different types of compost. Although there will be widespread exceptions to any generalizations from these specific compost–soil experimental results, some of the patterns represented in Figure 2 can be explained by mechanistic considerations. Many manure–crop residue composts are limited by available crop residue in the initial feedstocks, and thus tend to have high concentrations of nutrients (Table 1) that facilitate the relatively high mineralization rates for N and P (no data were found for K mineralization of manure composts). Municipal solid waste (MSW) has a relatively low N mineralization rate, reflecting the relatively high C:N ratio in typical MSW feedstocks. Biosolids, while they can have a high P concentration, are often processed in ways that bind the P to iron and calcium during wastewater treatment, resulting in reduced availability of this element. For any particular compost feedstock, transformations during the composting process (illustrated for N in Figure 1) will also affect subsequent mineralization rates. P availability may be enhanced by organic acids formed during the early stages of the composting process, and then later decline as the compost matures. If compost piles are exposed to precipitation, soluble K may leach out during the process, resulting in reduced availability for the K that remains. The wide range of observed mineralization rates, often varying by a factor of 3 or more, illustrates the need for additional tools and techniques to facilitate accurate and reliable compost nutrient management.

The difficulty of predicting initial compost mineralization rates is compounded as nutrients accumulate from repeated applications over a period of years. Although mineralization rates typically decline over time, the buildup of this pool of moderately available nutrients should also be considered in nutrient-management plans. Over time, mineralization from this 'nutrient bank' will increasingly substitute for current-year application requirements, as shown in the example presented in Figure 3. In this example, 20% of compost N mineralizes in the initial year, and 10% of the remainder mineralizes in each subsequent year. Initial applications must be heavy to supply the 150 kg N ha^{-1} required by a hypothetical crop, but over time these applications can drop off to a steady-state replacement rate as accumulated compost nitrogen from prior applications increases to greater than 1300 kg N ha^{-1}. Heavy initial compost requirements, followed by a gradual decline to steady state, are a common practice in agricultural and horticultural applications.

Environmentally sound compost nutrient management requires consideration of the balance among nutrients also. For many manure and biosolids composts, application at an N-based rate results in an overloading of P relative to crop demand. This imbalance is exacerbated relative to the original manure or sewage sludge feedstocks, because significant amounts of N are lost during composting, while P

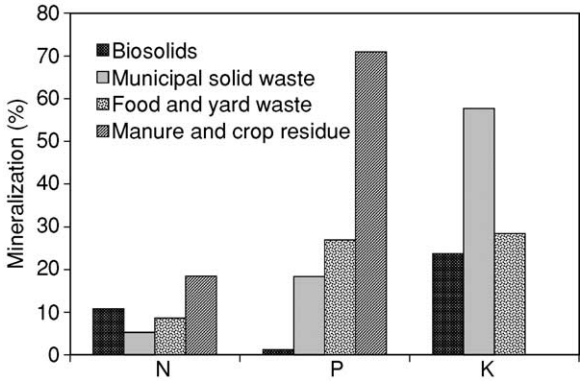


Figure 2 N, P, and K mineralization rates reported for various compost types.

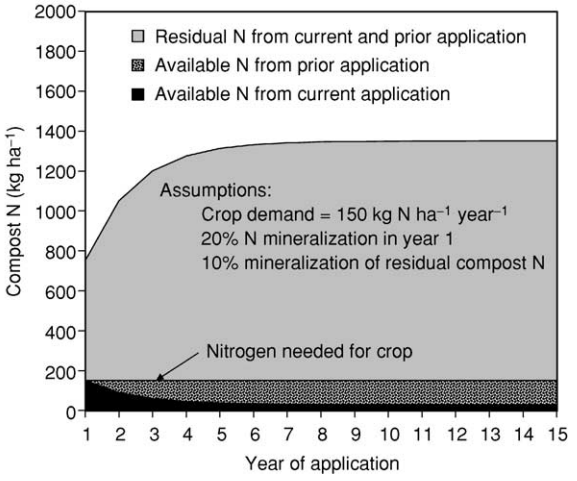


Figure 3 Changes in annual N supply and cumulative N reserve when compost is used to supply crop N requirements.

is largely conserved. This imbalance is particularly problematic on sites where the soil phosphorus levels are already elevated, either from prior manure use or synthetic fertilizer applications. Integration of nitrogen-fixing legumes or other alternative N sources in a crop rotation can be an excellent way to address this issue, while use of synthetic N fertilizers to supplement compost N can also work well. With either approach, compost application rates would then need to be reduced to match P demand, through either lower annual rates or alternate-year applications.

Heavy metals are present in all composts, but only sometimes at levels of toxicological concern. Governmental regulatory agencies have set concentration and cumulative loading limits for metals, particularly for composts whose feedstocks include biosolids and/or MSW, where concentrations of these elements are often higher than in source-separated organic feedstocks. Different metals pose varying types of ecosystem risks, from phytotoxic effects on sensitive plants to bioaccumulation in human and animal food chains. As with macronutrients, the bioavailability of these trace elements is strongly influenced by interactions with the compost and soil matrix. Bioavailability tends to decrease with time, due to chelation of metals and ligand formation. These binding and chelation capabilities of compost have been used to remediate soils with high heavy metal concentrations, and are particularly effective at lower pH. Compost can also be used to remediate toxic organic chemicals, both by enhancing biodegradation and by reducing their mobility and bioavailability.

Biological Effects

Several biological processes have already been mentioned, because they affect the physical and chemical characteristics of compost–soil systems. Organic carbon added through compost provides a major energy resource for microbial activity and growth, while the added nutrients and minerals cycle through both microorganisms and plants. These processes and others operate in multiple feedback loops to regulate nutrient availability, the microbial ecosystem, and plant growth and decay.

One of the dominant features of compost-amended soils is a significant increase in soil microbial activity, the extent of which depends on properties of both the compost and the soil. The microbial biodegradation that fuels this activity is indicated by the carbon mineralization rate, often measured in soil by CO_2 respiration studies. Several similar methods have been developed to measure and classify the degree of compost stability, and results are often available for commercial compost products. Although increased stability reduces the potential for odors and various phytotoxic compounds, it also reduces the carbon available for soil microbial processes. Thus while fairly stable composts may be required for residential markets, greenhouse applications, seedlings, and sensitive plants, agronomic applications may be better served by less-stable composts that can better stimulate microbial activity in the soil. The optimum degree of stability for any particular situation will depend on the application rate, soil system, and plants intended to be grown.

In addition to increasing overall microbial activity, compost also increases the activity of specific enzymes, the amount of active microbial biomass, and the overall diversity of the microbial ecology in soil–compost systems. These effects result from increases in both the amount and diversity of organic constituents introduced with the compost, including both partially decomposed organic matter and living microbial biomass.

Although most of the biological effects of compost are entirely positive, some composts can contain constituents that are detrimental to plants. Potentially phytotoxic compounds include short-chain organic acids, alcohols, and other organic compounds, ammonia, and the soluble salts and heavy metals previously described. Ammonia and the organic compounds are formed by protein degradation and anaerobic fermentations, respectively. Because these conditions are most likely during the early stages of composting, phytotoxicity is most commonly associated with immature composts, but can also occur with older composts that have been stored anaerobically, or that have high EC or heavy metal concentrations. These compounds have the strongest impact during the early stages in a plant's life cycle, with different plant species varying in their response. In general, smaller-seeded plants are more sensitive to the organic acids and phytotoxic biomolecules, while grasses are more sensitive to soluble salts. Laboratory bioassays have been developed to test for reduced seedling germination and root elongation, using species sensitive to specific types of effects. These bioassays can provide assurance to compost producers and users that a compost is safe to use on high-value, sensitive crops.

In weed control applications, the phytotoxicity associated with immature composts can potentially be put to positive use. Because many weeds are small-seeded, they are generally more sensitive to these phytotoxic compounds than large-seeded crops. Blanket surface applications of compost have also been shown to reduce weed emergence, primarily through the physical mulching effect, but biochemical interactions may contribute as well. However, once weeds germinate, the effect of compost on crop and weed competition becomes less well defined. While compost applications appear to reduce competition

between crops and certain weed species, other weed species are luxury feeders on nutrients, and compost may provide these weeds a competitive advantage against certain crops. The use of compost in weed management remains an important and intriguing area for additional research.

The interaction between compost and plants is further complicated by the potential of composts to have phytostimulatory effects, beyond those caused by improvements in soil physical properties, moisture, or nutrient effects. Several biomolecules in compost have been shown to promote plant growth, including indole-3-acetic acid, humic and fulvic acids, and tend to increase in concentration as the compost matures. Humic substances have been linked with the following positive results: increased membrane permeability and nutrient uptake, improved photosynthesis and protein synthesis, enhanced enzyme activity, and other biochemical changes similar to those induced by plant growth regulators. While additional research is needed to confirm these results and explore possible mechanisms, such synergistic effects may help explain increases in plant growth and yield from compost treatments, at levels beyond those predicted by measured physical and chemical effects.

Another important interaction between compost, soils, and plants is in the area of plant disease suppression. Certain composts have been known for decades to suppress particular plant pathogens, including many of the soil-borne pathogens that plague nursery and greenhouse growers. The efficacy of different composts varies, with some having no effect while others perform even better than chemical treatments, depending on the compost, pathogen, and specific mechanisms of suppression involved. In some cases the mechanisms of disease suppression are narrow, such as when composts include specific microorganisms that are antagonistic to or parasitize particular plant pathogens, or somewhat broader when pathogens are suppressed by fungal antibiotic inhibitors or become less competitive when compost changes the nutrient status of the soil. Compost feedstock, processing strategy, and maturity can all contribute to these suppressive mechanisms, with an adequate but not excessive level of maturity particularly important in promoting biocontrol. In some cases biocontrol organisms are inoculated into compost, often just after the thermophilic period, when ecologic niches are relatively open. Certain composts can also induce systemic acquired resistance, stimulating the plant's immune system to ward off pathogens more effectively than it could before. This general mechanism of plant disease suppression is highly effective, suppressing even foliar diseases that have no direct contact with the compost or the soil. Relatively few composts produced today consistently induce systemic acquired resistance, making this a promising area for future research and development.

Compost can have a range of positive effects on soils, microbial communities, and plants. Although the complexity of these interactions and the diversity of compost types and behaviors pose continuing challenges for both scientific understanding and practical management, considerable progress has been made in recent years. With greater emphasis on ecologic management of both organic and conventional crop production, research on the interactions of compost with soil and crops is expected to expand. As more is understood about these physical, chemical, and biological mechanisms and relationships, compost is likely to play an increasing role in improving soil tilth, soil fertility, and managing plant growth and disease.

List of Technical Nomenclature

EC	Electrical conductivity ($dS\,m^{-1}$)
MSW	Municipal solid waste

See also: **Organic Farming**; **Organic Matter:** Genesis and Formation; **Organic Residues, Decomposition**

Further Reading

Avnimelech Y, Kochva M, Yotal Y, and Shkedy D (1992) The use of compost as a soil amendment. *Acta Horticulturae* 302: 217–236.

Dick WA and McCoy EL (1993) Enhancing soil fertility by addition of compost. In: Hoitink HAJ and Keener HM (eds) *Science and Engineering of Composting*, pp. 622–644. Worthington, OH: Renaissance Publications.

Epstein E (1997) *The Science of Composting*. Lancaster, PA: Technomic Publishing.

Fraser DG, Doran JW, Sahs WW, and Lesoing GW (1988) Soil microbial populations and activities under conventional and organic management. *Journal of Environmental Quality* 17: 585–590.

Hoitink HAJ, Krause MS, and Han DY (2001) Spectrum and mechanisms of plant disease control with composts. In: Stoffella PJ and Kahn BA (eds) *Compost Utilization in Horticultural Cropping Systems*, pp. 263–273. Boca Raton, FL: Lewis Publishers.

Inbar Y, Chen Y, and Hadar Y (1990) Humic substances formed during the composting of organic matter. *Soil Science Society of America Journal* 54: 1316–1323.

Kashmanian RM, Kluchinski D, Richard TL, and Walker JM (2000) Quantities, characteristics, barriers, and incentives for use of organic municipal by-products. In: Power JF and Dick WA (eds) *Land Application of Agricultural, Industrial, and Municipal By-Products*, pp. 127–167. SSSA Book Series No. 6. Madison, WI: Soil Science Society of America.

Kuter GA, Blackwood KR, Diaz LF *et al.* (1995) *Biosolids Composting*. Alexandria, VA: Water Environment Federation.

Liebman M and Mohler CL (2001) Weeds and the soil environment. In: Liebman M, Mohler CL, and Staver CP (eds) *Ecological Management of Agricultural Weeds*, pp. 210–268. Cambridge, UK: Cambridge University Press.

Miller FC (1991) Biodegradation of solid wastes by composting. In: Martin AM (ed.) *Biological Degradation of Wastes*, pp. 1–31. London, UK: Elsevier Applied Science.

Ozores-Hampton M, Obreza TA, and Stofella PJ (2001) Weed control in vegetable crops with composted organic mulches. In: Stoffella PJ and Kahn BA (eds) *Compost Utilization in Horticultural Cropping Systems*, pp. 275–286. Boca Raton, FL: Lewis Publishers.

Rynk R and Richard TL (2001) Commercial compost production systems. In: Stoffella PJ and Kahn BA (eds) *Compost Utilization in Horticultural Cropping Systems*, pp. 51–93. Boca Raton, FL: Lewis Publishers.

Sikora LJ and Szmidt RAK (2001) Nitrogen sources, mineralization rates, and nitrogen nutrition benefits to plants from composts. In: Stoffella PJ and Kahn BA (eds) *Compost Utilization in Horticultural Cropping Systems*, pp. 287–305. Boca Raton, FL: Lewis Publishers.

Zucconi F, Monaco A, Forte M, and De Bertoldi M (1985) Phytotoxins during the stabilization of organic matter. In: Gasser JKR (ed.) *Composting of Agricultural and Other Wastes*, pp. 73–86. London, UK: Elsevier Applied Publishers.

CONDITIONERS

R E Sojka and J A Entry, USDA Agricultural Research Service, Kimberly, ID, USA
W J Orts, USDA Agricultural Research Service, Albany, CA, USA

Published by Elsevier Ltd.

Introduction

The use of naturally occurring materials as soil-stabilizing conditioners has been part of agriculture and general land management for millennia. Some of the most familiar conditioners in use since ancient times include animal and green manures, peat, crop residues, organic composts, and lime. These early uses of conditioners resulted from knowledge gained from trial and error long before there was scientific understanding of how efficacy was derived. Other conditioners in use for centuries or decades include composted manures, various organic debris, including sawdust or other milling residues, food, textile, and paper-processing wastes and other organic industrial wastes, as well as mineral materials such as rock phosphates, gypsum, coal dust, rock flour, and sand.

Soil conditioner use and technology, since ancient times, has, in great part, been a marriage of convenience between the agricultural necessity for chemical and physical maintenance or enhancement of the land, and for the disposal or management of waste materials from the full spectrum of human activities. However, since about the early nineteenth century, as modern physics and chemistry emerged and were applied systematically to agriculture, soil conditioner identification, development, and use became more creative and deliberate.

With the development of soil science as a specific discipline, the terminology and concept of soil amendments and conditioners was gradually assigned primarily a physical-conditioning connotation. Chemical conditioning, *vis-à-vis* supplying plant nutrients to soil, has been largely ascribed to materials termed fertilizers. Clearly, however, there is substantial overlap. Many fertilizers affect soil physical properties, both directly and indirectly, and many soil conditioners affect soil fertility both directly and indirectly. The overlap of physical and chemical effects occurs because of the intimate association of all soil physicochemical process and their coupling, as well, to soil-supported biotic processes, cycles, and functions. The designation of fertilizer versus conditioner is often based on the dominant effect intended. Categories are often assigned by law, based on the chemical analysis and/or the proof of claims for the materials.

Early Use of Mineral and Organic Materials

This article provides a brief history of early and traditional conditioner technologies and then focuses on recent developments in inexpensive and highly effective synthetic conditioner materials and use strategies. Organic conditioners have generally been applied to increase infiltration and soil water retention, promote aggregation, provide substrate for soil biological activity, improve aeration, reduce soil strength, and resist compaction, crusting, and surface sealing. The effects of organic conditioners often occur bimodally. That is, some effects, such as improved infiltration

and water retention, are evident immediately upon soil incorporation, whereas other effects, such as improved aggregation, depend on chemical and biological processes over time.

Mineral conditioners are often used to affect soil chemical processes as well as soil physical processes. Lime, for example, raises soil pH. Gypsum or lime is often used to increase base saturation, or reduce the exchangeable sodium percentage (ESP) of retained cations. Because the divalent calcium ion has a compact hydrated radius, it also promotes flocculation of clays and increases aggregate stability. These effects help to reduce particle dispersion and detachment – which reduce erosion and surface sealing. Similarly, the calcium ion promotes flocculation and aggregation. These effects can be particularly important in arid soils with low soil organic matter (SOM) contents. The physical properties of such soils are often impaired when the exchange complex is dominated by the sodium ion, which has a much larger hydrated radius, and thus impedes flocculation and aggregation and favors dispersive phenomena. The physical benefits of calcium addition in low SOM saline soils provide for improved leaching of salts and removal of sodium, especially under irrigated conditions.

Mineral conditioners are especially important for the management of arid or tropical soils where high temperatures promote rapid bio-oxidation of incorporated organic material. A variety of other strategies are used with mineral conditioners to exploit soil physicochemical processes, directly or indirectly improving soil physical and/or chemical status. While the uses of lime and gypsum have ancient origins, another interesting approach in recent decades has been the use of various oxides of iron to promote aggregation in low-organic-matter soils. In the 1970s researchers added iron oxides to increase aggregate stability of soils and found peak aggregation at a 2% addition rate, with aggregation favored by acidic conditions. Others in the 1980s found promising results for addition of ferrihydrite compounds to calcareous soils, with the formation of weak quasicrystalline structures. Recent work shows potential for adding ferric hydrides to low-organic-matter soils for structure improvement and wind erosion resistance. Ferric hydrides are common water-treatment and industrial process waste products.

Soil conditioner research to the present has explored the use of many naturally occurring organic and mineral materials, agricultural and industrial waste products, or by-products of other processes. Materials that have been used as conditioners have included crushed rock, ground coal, gypsum (mined or from ground plasterboard), wood chips, bark, sawdust, food-processing wastes, cheese whey, various manures, composts of manures and/or other organic materials, and, as discussed more fully below, a wide range of synthetic polymer materials, including copolymers of synthetic and naturally occurring substances. All these materials have shown varying capacities to modify soil conditions or soil processes.

Use of Waste Materials as Conditioners

The extent of soil conditioner use has often been limited by economics. The cost of conditioner use has commonly been due more to transportation and application expenses of bulky materials than to the price of the materials *per se*. In many instances conditioner material is available gratis from waste streams of various processes where disposal is an expense. Use of waste materials as soil conditioners eliminates the disposal expense and in some cases creates profitable products. Because of material bulk, transportation, application, and related costs, the widespread use of traditional soil conditioners in mainstream production agriculture has been limited to only a few very highly efficacious materials such as lime, gypsum, and manure. Exceptions have occurred in high-value nursery, cash crops, turf and landscape applications, or in proximity to sources, where transport costs have been minimal.

Advent of Synthetic Conditioners

Since the early 1950s soil scientists have explored using synthetic polymeric conditioners to alter drastically soil physical and, in some cases, chemical properties. During World War II water-soluble polymers were used to stabilize soils in order to hasten the construction of roads and runways. The use of polymeric soil-conditioning chemicals was introduced to agricultural research and the farming community following World War II. In 1949 an industrial process for polymerizing acrylamide molecules was patented. This ultimately enabled a vast new array of water-soluble polymer compounds with thousands of industrial and environmental uses. Sixteen scientific reports of water-soluble polymer soil conditioning appeared by 1952, and 99 reports by 1955.

By enhancing the formation of soil aggregates and prolonging their longevity, water-soluble polymeric conditioners improve soil physical properties, including root penetration, erosion resistance, infiltration, aeration, and drainage. These direct physical improvements usually promote rooting and plant interception of nutrients and water, indirectly improving plant nutrition. The synthetic materials perform

immediate conditioning and structural stabilization that would ordinarily require weeks, months, or years to achieve via a program of organic matter incorporation. Furthermore, synthetic materials can effectively condition soil to the depth of tillage with one to two orders of magnitude less material application than required with traditional conditioners, and can be zone- or spot-applied for even more efficient, targeted application and efficacy. Despite these performance advantages, however, in the early years of synthetic soil conditioner use, cost usually restricted use to high-value crops or specialty applications.

The most common strategy for water-soluble polymeric soil conditioner use from the early 1950s until the early 1990s was the application of sufficient conditioner material to affect significant physical modification of the soil to the depth of tillage. Depending on the nature of the polymer conditioner material, these application amounts often reached hundreds of kilograms per hectare. Generally, this mode of treatment entails multiple application operations, either as bulk solid materials, or as sprayed liquids or slurries. Each application usually requires tillage to incorporate the material to a desired depth. Because the mass of soil in a typical hectare-plow-layer is great (typically 2 000 000 kg per hectare 15-cm slice), many tons per hectare of traditional physical amendments and hundreds of kilograms per hectare of water-soluble polymeric soil amendments are usually necessary to overcome the physical or chemical buffering effect of the large mass of soil being treated.

Some of the most commonly used water-soluble synthetic soil-conditioning polymers since the 1950s have been: hydrolyzed polyacrylonitrile (HPAN), isobutylene maleic acid (IBM), polyacrylamide (PAM), polyvinyl alcohol (PVA), sodium polyacrylate (SPA), and vinylacetate maleic acid (VAMA). Commercial formulations of these compounds sometimes combined polymers and extenders or solubility-enhancing agents. Perhaps the most successful water-soluble soil-conditioning polymer marketed commercially before the 1990s was the Monsanto product Krilium which combined VAMA with a clay extender for improved application uniformity. Krilium and similar products were marketed in the 1950s at costs of \$4–5 kg^{-1}. Then-current application techniques required the application of tens to hundreds of kilograms per hectare, depending on the depth and extent of the soil zone to be treated. This precluded use on all but high-value crops or in specialty situations. After initial enthusiasm for these conditioners, most products have been withdrawn from general marketing to mainstream agriculture because of economic realities.

Hydrogels and Super Water-Absorbent Polymers

There was also interest over the years in super water-absorbent polymers for use in soils. These polymers are not water-soluble, but instead are strongly hydrophilic gel-forming materials that easily absorb hundreds or even 1000–2000 times their weight in water (Figure 1). Hydrolyzed starch-polyacrylonitrile graft polymers (H-SPANs), patented by the US Department of Agriculture in 1975 under the market name Super Slurper, and cross-linked polyacrylamides (gel-forming PAMs) have been the most common polymers for this application. They are used to improve the water retention of soils with low-water-retention properties, or that experience prolonged and untimely drought, especially immediately after planting. Spot placement of gel polymers in proximity to seeds, seedlings, or transplants prolongs the opportunity for emergence and seedling establishment without having to irrigate the entire soil profile. Again, because of cost and application amounts required, it is usually not economically feasible or logistically practical to attempt to modify an entire profile or even tillage zone, even when conditioner cost is as little as \$2 per kilogram.

Polymer chemistry, prior to about the 1980s, generally limited available soil-conditioning polymers to molecules with chain links of a few thousand monomer units. In addition, the purity of the preparations was not always good, sometimes carrying safety or environmental risks from reaction by-products or incompletely reacted base chemicals. Since the 1980s and early 1990s polymer purity and molecular size have increased, greatly improving the efficacy, safety, and affordability of environmental polymers. These changes, coupled with new application strategies that only target critical portions of the soil for treatment,

Figure 1 Superabsorbent cross-linked polyacrylamide in the dry state (left) and hydrated state (right); scale is in centimeters.

and that do not depend on expensive field operations for chemical application, have produced a sustained renewal of interest in environmental polymers for a growing number of uses. Perhaps the best example of this advancement has been the use of PAM for erosion control in irrigated agriculture.

Recent Advances Using Polyacrylamide

Isolated reports in the 1970s and 1980s provided a hint that very small amounts of PAM in irrigation water, flowing over soil in irrigation furrows, virtually eliminated detachment and transport of soil particles. These reports, however, were either anecdotal with respect to erosion or did not adequately identify the polymer used. Thus, the potential importance of the PAM-treatment erosion effects went unnoticed for several years. The foundation was laid for the practical use of PAM to halt erosion in furrow irrigation in a series of studies through the 1990s. The success of this new research came from the realization that the best way to treat soil structure to prevent erosion was to use the eroding water to deliver the soil conditioner. Irrigation is perfectly suited to this mode of application. In this mode of application only 1–2 kg ha^{-1} of PAM was needed to halt an average of 94% of erosion from irrigation furrows (Figure 2). The treated soil was restricted to about 25% of the field surface area and was only treated to a depth of a few millimeters. Inflows only needed to be dosed as water crossed the field, shutting off applicators when runoff began.

This strategy relies on the use of a highly specific class of food-grade PAM to ensure both efficacy and human and environmental safety. These PAMs are anionic, with a charge density of typically 18%; they are what are today regarded as moderately large molecules, having over 150 000 chained monomer segments per molecule for molecular weights of 12–15 million g mol^{-1}. The molecules are manufactured to a high purity, and are actually identical to PAMs used for food processing and drinking water treatment, with residual unreacted actylamide monomer (AMD) contents of <0.05%. The low AMD content and anionic nature of the molecule ensures safety for humans handling the PAM and for aquatic species in the event PAM is lost via runoff to surface

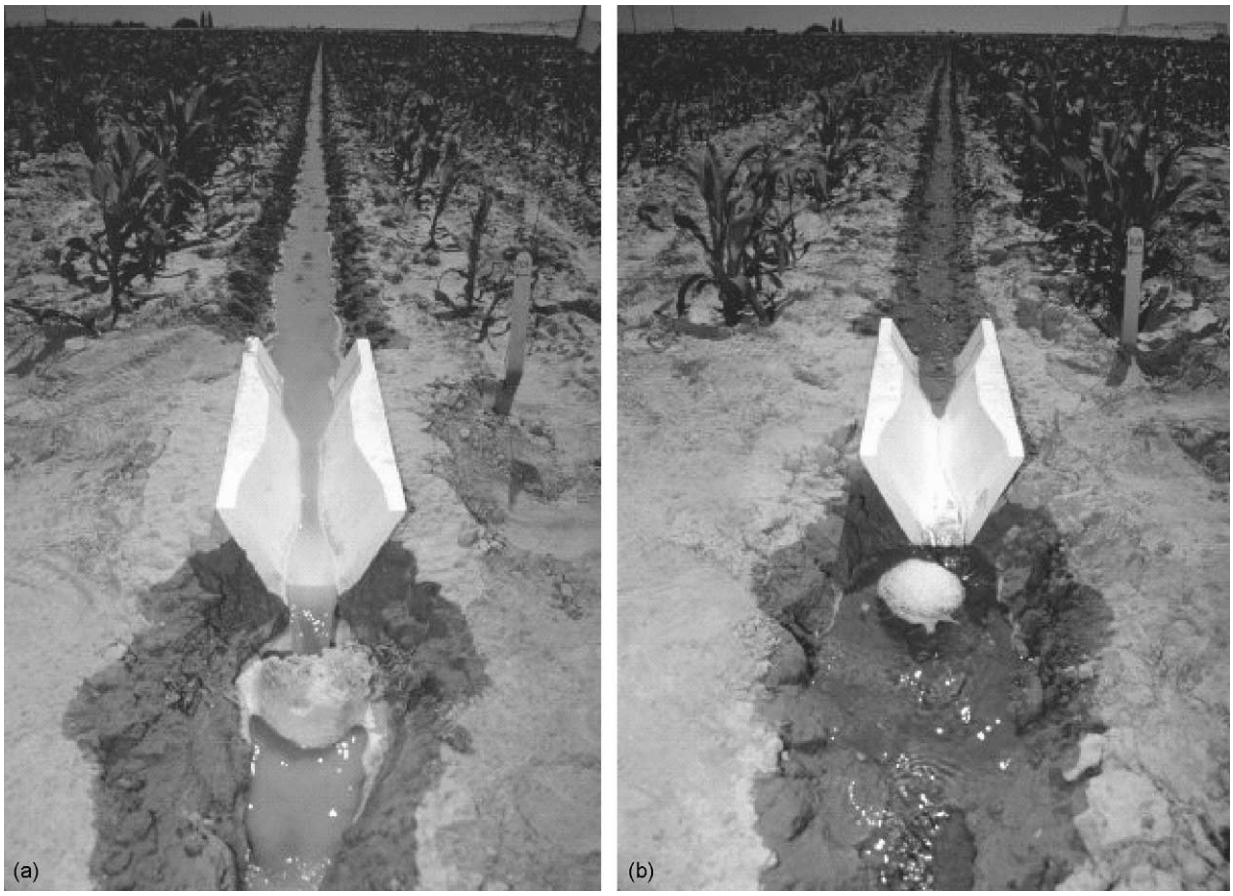


Figure 2 Runoff from irrigation furrows where (a) water is untreated and (b) water is treated with polyacrylamide; note the lack of turbidity, and thus absence of erosion from the polyacrylamide-treated furrow.

waters. However, the anionic charge imparts the need for bridging cations in the solvating water to link the anionic polymer to the anionic surfaces of soil minerals. Waters and soils containing dissolved calcium enable better efficacy than low-electrolyte (pure) water, and efficacy is best when there is little or no sodium present.

PAM is so effective at stabilizing surface structure, even at these small application amounts, that, in most fine- to medium-textured soils, infiltration is increased compared to untreated water, which induces surface sealing. While initial uses of PAM were focused mainly on erosion control, farmers are equally interested in using PAM for infiltration improvement. This is especially true as the technological barriers to use of PAM in sprinkler irrigation are overcome. With proper application strategies, PAM can be used both to increase infiltration amounts or rates as well as to improve infiltration uniformity. Since, with PAM in the water, soil structure is improved and surface sealing is reduced, water droplets enter the ground where they land, rather than causing seals and inducing runoff and ponding.

PAM use with irrigation for erosion control benefits water quality in a number of ways. By preventing erosion it also reduces the desorption opportunity for sorbed nutrients and pesticides, and limits dissolution of soil organic matter in runoff that elevates dissolved organic carbon (DOC) and biological oxygen demand (BOD). PAM-treated irrigation water has also proven highly effective at reducing movement off-site of soil-borne microorganisms and weed seed, greatly reducing the likelihood of downstream inoculation and, ultimately, reducing the need for pesticides.

The prospects for the future development of PAM technology remain good. Because PAM increases the viscosity of water flowing through soil pores, the effects on infiltration are a balance of seal prevention allowing greater infiltration and viscosity slowing the passage of water. Experiments are currently underway to use the viscosity effects with other management strategies for canal sealing, improved infiltration uniformity along long irrigation furrows, and better water retention in soils where infiltration is not a problem.

Natural gas is the cheap abundant raw material from which PAM is currently made. However, current supplies and economics may not reflect the future. Work to develop new copolymers of PAM using chitin and starch as building blocks has proven promising, although results are yet to match those achievable currently with PAMs. Use of these materials as building blocks for effective flocculents and soil stabilizers carries the added benefit of using another agricultural waste stream to produce value-added products. In this case products may eventually be provided that can add to our inventory of environment-protecting and production-improving agricultural tools.

The field of water-soluble polymers for environmental protection and agricultural management is growing rapidly. These polymers are inexpensive, effective, and safe. They can be easily used in many settings and provide nearly instantaneous results in most instances. They can be used effectively in combination with more traditional land management and water quality protection techniques, such as reduced tillage or riparian buffer strips, either enhancing the effectiveness of these more familiar approaches or providing additional 'insurance' for situations when alone they are less reliable. The work of the last decade has emphasized that agricultural and environmental polymers cannot be regarded as 'silver bullets,' but when used in a well-planned approach to agricultural land and water management or environmental protection, offer a significant new capacity for better resource utilization and environmental protection.

Extensive additional information on the current use of PAM as an environmental polymer for erosion and pollution prevention and for irrigation water management improvement can be found at the website www://kimberly.ars.usda.gov/pampage.shtml.

See also: **Aggregation:** Physical Aspects; **Crusts:** Structural; **Structure**

Further Reading

Aase JK, Bjorneberg DL, and Sojka RE (1998) Sprinkler irrigation runoff and erosion control with polyacrylamide – laboratory tests. *Soil Science Society of America Journal* 62: 1681–1687.

Agassi M, Letey J, Farmer WJ, and Clark P (1995) Soil erosion contribution to pesticide transport by furrow irrigation. *Journal of Environmental Quality* 24: 892–895.

Bjorneberg DL and Aase JK (2000) Multiple polyacrylamide applications for controlling sprinkler irrigation runoff and erosion. *Applied Engineering in Agriculture* 16: 501–504.

Entry JA and Sojka RE (2000) The efficacy of polyacrylamide and related compounds to remove microorganisms and nutrients from animal wastewater. *Journal of Environmental Quality* 29: 1905–1914.

Lentz RD and Sojka RE (1994) Field results using polyacrylamide to manage furrow erosion and infiltration. *Soil Science* 158: 274–282.

Lentz RD, Sojka RE, and Robbins CW (1998) Reducing phosphorus losses from irrigated fields. *Journal of Environmental Quality* 27: 305–312.

Malik M and Letey J (1992) Pore-sized-dependent apparent viscosity for organic solutes in saturated porous media. *Soil Science Society of America Journal* 56: 1032–1035.

Orts WJ, Sojka RE, Glenn GM, and Gross RA (2001) Biopolymer additives for the reduction of soil erosion

losses during irrigation. In: Gross RA and Scholz C (eds) *Biopolymers from Polysaccharides and Agroproteins*. ACS Symposium 786, pp. 102–116. Washington, DC: American Chemical Society.

Rhoton FE, Romkens MJM, Bigham JM, Zobeck TM, and Upchurch DR (2003) Ferrihydrite influence on infiltration, runoff and soil loss. *Soil Science Society of America Journal* 67: 1220–1226.

Sojka RE and Entry JA (2000) Influence of polyacrylamide application to soil on movement of microorganisms in runoff water. *Environmental Pollution* 108: 405–412.

Sojka RE, Lentz RD, Trout TJ *et al.* (1998) Polyacrylamide effects on infiltration in irrigated agriculture. *Journal of Soil Water Conservation* 53: 325–331.

Sojka RE, Lentz RD, and Westermann DT (1998) Water and erosion management with multiple applications of polyacrylamide in furrow irrigation. *Soil Science Society of America Journal* 62: 1672–1680.

Stewart BA (ed.) (1975) *Soil Conditioners*. Special Publication 7. Madison, WI: Soil Science Society of America.

Trout TJ, Sojka RE, and Lentz RD (1995) Polyacrylamide effect on furrow erosion and infiltration. *Transactions of the ASAE* 38: 761–765.

Wallace A (ed.) (1995) *Soil Conditioner and Amendment Technologies,* vol. 1. El Segundo, CA: Wallace Laboratories.

Wallace A (ed.) (1997) *Soil Conditioner and Amendment Technologies,* vol. 2. El Segundo, CA: Wallace Laboratories.

Wallace A and Terry RE (eds) (1998) *Handbook of Soil Conditioners: Substances That Enhance the Physical Properties of Soil*. New York: Marcel Dekker.

Conservation *See* **Erosion:** Water-Induced; Wind-Induced; **Sustainable Soil and Land Management**; **Terraces and Terracing**

CONSERVATION TILLAGE

M R Carter, Agriculture and Agri-Food Canada, Charlottetown, PE, Canada

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Conservation tillage (CT) is an umbrella or generic term used to describe tillage systems that have the potential to conserve soil and water by reducing their loss relative to some form of conventional tillage. Precise definitions of conservation tillage are only possible within the context of known crop species, soil types and conditions, and climates. A well-accepted operational definition of CT is a tillage or tillage and planting combination that retains a 30% or greater cover of crop residue on the soil surface. Generally, there are four main types of CT: mulch tillage, ridge tillage, zone tillage, and no-tillage. A main variant of the latter is direct drilling (sometimes termed zero-tillage), while other variants of CT are reduced tillage and minimum tillage. Conservation tillage can provide several benefits for agricultural systems such as soil conservation, economic advantages associated with reductions in crop establishment time and energy use, reduction in soil sheet erosion and nonpoint pollution, and enhanced storage or retention of soil organic matter and improvement of soil quality at the soil surface.

Evolution of Conservation Tillage Systems

Tillage involves the mechanical manipulation of the soil. In an agricultural, horticultural, or forestry context, it involves manipulation of the soil profile to modify soil conditions and to manipulate plant residues, and to control or remove unwanted plant growth. In agricultural systems, tillage functions as a subsystem that influences crop production mainly through crop establishment, modification of soil structure, incorporation of fertilizer and soil amendments (e.g., lime and manure), and weed control. Tillage is also used to alleviate both climatic and soil constraints.

The evolution of conservation tillage is a complex phenomenon with many varied themes. First, excessive soil tillage is associated with soil degradation processes such as compaction, a decrease in soil stability and structure, and increased soil erosion. Thus, one component of CT is a trend towards reducing or minimizing tillage events to address concerns with tillage-induced soil degradation. Second, most arable farming systems developed in climates where

potential evapotranspiration is relatively high and precipitation is moderate to low. Under such conditions, a need arose to conserve soil water and reduce the propensity for soil erosion. Third, the advent of herbicides allowed the weed control aspect of tillage to be drastically curtailed and, in some cases, the need for tillage itself could be eliminated (i.e., adoption of no-tillage). Fourth, the cost of tillage in regard to tillage implement depreciation, tractor fuel use, and labor input is a major aspect of crop production costs. Thus, there is an economic incentive to reduce tillage events.

Within any one specific region, the speed and degree of CT adoption are related to a multitude of factors such as profitability and costs, socioeconomic concerns, the risk involved with adopting a new practice, and farmer skills and awareness of the need for soil conservation, and control of soil erosion. In the majority of reported surveys, concerns about conservation of the soil resource or pollution of the environment are not the foremost reasons for adoption of CT. The main reasons are economic feasibility and cultural compatibility. In many cases the potential economic advantages serve as the main incentive. Table 1 outlines the main reasons for adoption of CT in North America.

Defining Conservation Tillage

An acceptable all-embracing definition of CT is difficult to draft due to the varied conditions that underlie its development and the constraints presented by different soil types, cropping rotations, and climate. However, a common characteristic of CT is the potential to reduce soil erosion and water loss relative to some form of conventional tillage. A general operational indicator used to define CT, mainly in subhumid to semiarid climates, is the maintenance of a 30% soil surface cover with crop residue after seeding or planting. A well-accepted definition for CT, which reflects the above, is as follows: "Any tillage sequence, the object of which is to minimize or reduce loss of soil and water; operationally, a tillage or tillage and planting combination that leaves 30% or greater cover of crop residue on the surface."

Table 1 Reasons, in order of importance, for the adoption of conservation tillage in North America

Reason for adoption of conservation tillage	*Major variables associated with adoption of conservation tillage*
Profitability	Low financial risk; fuel savings; no new equipment needed or savings on equipment cost/depreciation; potential for increase in crop yield and quality
Social and cultural benefits	Level of farmer education; form of land tenure, land owned rather than rented; greater farm size; full-time farmer
Control of soil erosion or degradation	Identification of soil degradation or soil erosion risk on farm
Conservation of soil water	Need to conserve soil water in semiarid warm areas
Timeliness	Need to improve time efficiency for crop establishment due to climate restrictions or labor shortage

Variants of Conservation Tillage

In many cropping situations and climates, innovations in tillage practices lead to soil and water conservation, although the operational definition for CT (i.e., crop residue indicator at 30% cover) may not be met or readily applicable. This has led to the development of various forms of tillage systems closely related to CT (Table 2). For example, in some agricultural systems the adoption of 'reduced tillage' allows a reduction in depth, degree, and frequency of tillage. Another variant of CT is 'minimum tillage,' which mainly refers to the approach or aim of achieving the minimum soil manipulation necessary for crop production, or that required to meet tillage requirements under specific soil and climatic conditions. A third variant is 'shallow tillage' or 'non-inversion tillage,' where the tillage is mainly restricted to a shallow (<15 cm) depth and the soil not turned

Table 2 Forms of tillage systems that meet some of the requirements of conservation tillage

Form of tillage system	*Characteristics of the tillage system*
Reduced tillage	Reduction in total number of primary[a] and secondary[b] tillage operations usually used in conventional[c] tillage, to prepare a soil for crop establishment
Minimum tillage	The minimum use of primary and secondary tillage necessary to meet crop production requirements under existing soil and climatic conditions, usually resulting in fewer tillage operations than for conventional tillage
Shallow or non-inversion tillage	Primary tillage confined to shallow soil depth (<15 cm). Absence of soil inversion[d] tillage

[a]Primary tillage is the initial major soil manipulation usually employed to loosen soil and bury or incorporate crop residue.
[b]Secondary tillage is any sequence of tillage operations that follow primary tillage generally used to prepare the soil for seeding/planting operations.
[c]Conventional tillage, a relative term, are those tillage operations normally used for crop production in a specific geographic area.
[d]Inversion tillage involves primary tillage that 'inverts' or 'turns over' soil causing much soil mixing. The main tillage implement used for inversion tillage is the moldboard plow.

over or inverted. This latter tillage system is sometimes termed 'plowless tillage.' In most cases, tillage systems that utilize moldboard plowing (tillage depth normally >15 cm; soil inverted) can approach the operational definition of CT if supported by use of cover crops and mulches to provide soil surface cover. Other variants use relatively deep tillage, but restrict the area or width of soil disturbed to a small proportion of the field.

Overall, the above variants of CT may not provide a 30% surface cover of crop residue, although most of the residue is still retained near the soil surface.

Common Types of Conservation Tillage Practices

The most common CT systems are 'mulch tillage,' 'ridge tillage,' 'zone tillage,' and 'no-tillage' (Table 3). The latter corresponds to 'direct drilling' and 'zero tillage.' Mulch tillage generally involves disturbance of the whole soil surface, while ridge tillage and zone tillage (sometimes termed 'strip tillage') often only disturb one-third or less of the soil surface. No-tillage restricts disturbance of the soil to that involved with crop seeding or planting. Depending on the type of seeder or planter, the proportion of soil surface disturbed can vary from 30 to 100%.

Constraints to the Adoption of Conservation Tillage Practices

The form of CT adopted in any one area is influenced by soil tillage requirement, climate, and type of farming system and crop rotation. Surveys (conducted in the 1990s) indicate that CT adoption in North America varies according to the above factors. Less than 25% of the coastal plains, delta areas, and humid northeastern regions of the USA utilize CT, while >40% of the southern corn (i.e., maize, *Zea mays* L.) belt, eastern uplands, Piedmont, and central Great Plains have adopted CT. The utilization of extreme forms of CT, such as no-tillage, in the above regions varies from 1 to 23%. In Canada, the adoption of CT ranges from 35% in the prairies to 13% in eastern Canada, with an overall adoption level of 24%.

Economic and sociocultural factors are the main reasons that influence the adoption of CT (Table 1). Generally, agronomic reasons are less important but still provide important incentives for the adoption of CT. Timeliness in crop establishment (especially in areas that use double cropping), the absence of soil compaction risk, and the range or diversity of crops conducive to the use of reduced-tillage practices are the main incentives. The great diversity of CT systems, as expressed by the various types of tillage (Tables 2 and 3), ensures that CT can be applicable to all soil types and farming systems; however, not every variant of CT is suitable for every situation. Several major constraints are associated with climate and soil type, high levels of crop residue, and mixed cropping systems.

Climate and Soil

No-tillage is generally difficult to use in cool, wet soils, or in soils with a relatively high tillage requirement. The latter refers to a soil's need for tillage based on its aggregate stability, potential for shrink-swell, and ability for self-structure. Tillage requirement is also influenced by climate and extrinsic factors such as site drainage and soil moisture class (Table 4). Generally, soils with an imbalance in particle size distribution (i.e., high clay or sand content) and/or those with poor permeability tend to have a high tillage requirement and are more difficult to manage when tillage is reduced.

High Levels of Crop Residue

Excessive amounts of crop residue may restrict the use of CT practices, especially no-tillage. Burning of

Table 3 The dominant types of conservation tillage used in agricultural systems[a]

Form of tillage system	*Some other terminology used*	*Main characteristics of the tillage system*
Mulch tillage	Stubble mulching, trash farming, sod farming, live mulch system	Some form of full-width[b] shallow primary tillage used prior to crop planting. High percentage of crop residues retained at the soil surface
Ridge tillage		Primary tillage confined to formation of raised ridges or beds in rows often on a contour. Planting occurs on the ridges
Zone tillage		Primary partial-width[b] tillage for row crops confined to bands, separated by bands of undisturbed soil, to form a seedbed for each row
No-tillage	Direct seeding, zero tillage	Soil undisturbed with no primary or secondary tillage. Crop seeded/planted directly into the soil

[a]Forms of conservation tillage mainly dominant in subhumid to semiarid climatic regions.
[b]Full-width tillage, tillage conducted over the entire soil surface; partial-width tillage, tillage conducted over a proportion of the soil surface.

Table 4 Soil wetness and physical constraints for adoption of conservation tillage

Constraint	*Unsuitable characteristic*	*Principal constraint*
Soil wetness problems	Imperfectly drained soils	Low soil strength
	Surface waterlogging	Reduced trafficability
	Slow subsoil permeability	Excess soil compaction
Imbalance in soil particle size	Sandy soils with low organic matter	Excess soil compaction and poor structure
	Soils with >32% clay	Excess soil compaction
	Soils with high amounts of silt and fine sand	Excess soil compaction and poor structure

crop residue may not be an option because of environmental concerns, while removal of residue may not be feasible. The main concerns with high residue levels are mechanical interference with seeding operations, delayed drying and warming of the soil surface, potential for allelopathic or residue-derived toxins, increased potential for crop disease and pests, and reduction in the efficiency of fertilizers and pesticides. Generally, some of these residue constraints can be overcome by adoption of technologies that improve seed drilling, harvesting techniques, chaff and straw distribution, and use of banding or placement techniques for fertilizer and pesticides. A switch from no-tillage to forms of mulch and noninversion tillage can be successful under conditions of high crop residue.

Mixed Cropping Systems

Crop rotation is a basic strategy for sustainable cropping systems, but use of varied crops can pose a difficulty for CT as the tillage needs of each crop may differ. Root crops, such as potato (*Solanum tuberosum* L.), have a greater tillage need than cereals. Under these situations, CT needs to be integrated into the rotation by using variations of tillage (i.e., rotational tillage) suited for specific crops.

Beneficial Influence of Conservation Tillage

Many CT systems can improve soil quality over time and reduce the soil erosion risk, improve soil properties, and reduce tillage costs. Soil health may also be improved, as indicated by increased microbial activity and competition and, in some cases, the potential for amelioration of plant pathogen activity and survival. Such microbial antagonism in the root zone can lead to the formation of disease-suppressive soils. Beneficial improvements in soil quality and health are usually associated with organic matter increase and concomitant improvements in the soil physical condition, especially soil aggregation and tilth. Use of CT can also influence storage of organic matter in the soil, which has implications for agricultural greenhouse gas emissions. Overall, these potential effects have important implications for environmental health.

Improving Soil Quality and Health

Adoption of CT can cause an increase in organic matter at the soil surface, relative to a conventional tillage, as a result of maintaining a greater proportion of crop residue near the soil surface. This stratification with depth of soil organic matter has important implications for soil quality and the environment because the soil surface receives much of the fertilizer and organic amendments and rainfall impact, and serves as an interface for gaseous exchange. Enhanced levels of organic matter result in concomitant increases in soil biological activity. This phenomenon, in combination with a reduction in soil disturbance, often results in an increase in the size and stability of soil aggregates, and an increase in the continuity (but not always the size) of the soil pore space. This can have important implications for the way in which soil holds and partitions water, and holds and releases nutrients. For example, most surface soils under CT can better resist the harmful impact of raindrops and have a greater capacity to accept infiltrating water, than their conventional-tilled counterparts. One important factor, as a result of the above, is the development of a continuous pore network down the soil profile in loam to clay soils under CT, which can have implications for rapid movement or leaching of nutrients and pesticides.

Enhancing Soil Organic Matter Storage

The organic matter storage capacity of a soil is dependent upon factors that affect organic matter inputs and outputs such as climate, soil type, landscape position, and management. The latter influences the store of organic matter in the soil by type of cropping system and soil tillage. Combination of cropping practice and reduced tillage can lead to soil organic matter gains and thereby provide a net sink for atmospheric CO_2, although such gains may be of finite magnitude and duration.

Generally, CT itself does not cause an increase in stored soil organic matter when compared to conventional tillage, under conditions of similar carbon inputs (i.e., most studies indicate that crop yield and

crop residue are similar between CT and conventional tillage). Various scenarios are evident for tillage-induced organic matter storage in soil. CT can enhance the organic matter content of surface soils (0–15 cm) in semiarid climates, in comparison to conventional tillage; however, such gains are usually due to labile forms of organic matter (e.g., particulate organic matter). In some climates, extensive tillage can place crop residues at a depth in the soil where decomposition proceeds at a slower rate than that observed for surface soils. Intensive tillage may also enhance organic matter association with soil clay and silt particles and thus encourage aggregation and consequently organic matter storage. Thus, differential placement and incorporation of crop residue and the degree of tillage intensity between tillage systems could result in differences in amounts of stored soil organic matter over time.

The beneficial influence of CT on reducing CO_2 emissions can involve more than increases in soil organic matter content. Full carbon cycle analysis takes into consideration carbon used or emitted from the manufacture and use of agricultural inputs such as machinery, fuel, and other inputs. Most full carbon cycle studies to date suggest that CT provides overall reduction in net CO_2 emissions, compared to intensive forms of tillage.

Reducing Soil Erosion and Environmental Risk

Nonpoint or diffuse pollution is the water pollution associated with land-use activities. In North America, CT is considered to be a beneficial practice to control nonpoint pollution. In general, due to the presence of surface crop residue, CT can slow the rate of water runoff, increase the rate of water infiltration, and reduce soil movement or erosion. Runoff volume (but not always runoff concentration), sediment losses, and losses of sediment-associated materials (e.g., adsorbed phosphorus and pesticides) can be decreased under CT. The leaching potential under CT, however, can be increased, which has implications for groundwater contamination by nitrogen and the limited number of pesticides that are highly mobile. In addition, runoff volume may not decrease in the presence of traffic/tillage-induced compaction, poor internal drainage, or the presence of fragipans and subsurface clay horizons. Reduced runoff can decrease sedimentation losses but not always the amount (i.e., increased concentration) of soluble phosphorus and pesticides in the runoff, resulting in no overall beneficial effect.

Table 5 summarizes the results of many CT studies in regard to range of runoff and leaching parameters. Conservation tillage can markedly reduce sedimentation but the effects on runoff volume are less clear, while the reduction in the nitrogen, phosphorus, and pesticide content of the runoff is variable. Leaching volume is greater under CT. Generally, most of the studies that measured pesticide loss in runoff or leachate indicated that only minor amounts were lost. Overall, CT is a useful management practice for control or reduction of sedimentation, but less effective for decrease of phosphorus movement. Generally, P resides in both inorganic and organic forms in the soil, and the latter is related to increasing levels of organic matter. Organically bound phosphorus is more soluble and therefore more mobile than phosphorus adsorbed on to soil minerals. Generally, CT needs to be integrated with both nutrient and pest management techniques to reduce nitrogen, phosphorus, and pesticide applications to the soil.

Table 5 Influence of CT on sedimentation, runoff and leaching compared to conventional tillage[a]

Measurement	*Decrease (%)*	*No effect (%)*	*Increase (%)*
Runoff volume	38	37	25
Sediment in runoff	100	0	0
Nutrients[b] in runoff	70	11	19
Leaching volume	20	0	80
Nitrogen leached	36	27	36

[a]Comparison of 40 studies (1991–1996) in North America comparing CT and some form of conventional tillage on a range of soil types.
[b]Nitrogen and phosphorus.

Tillage Costs

Use of fuel is generally lower under CT because of the reduction in tillage operations. Since fuel can account for 10% or more of farm energy use, CT clearly presents an economic and environmental benefit. However, general estimates of full production costs are problematic since CT often has higher fertilizer and pesticide requirements than the conventional tillage comparison. Overall, general principles in regard to tillage costs are not possible. Due to the wide variation in type of CT systems, in regard to tillage implements and crop production systems, conclusions on full tillage costs can only be site-specific.

See also: **Aggregation:** Physical Aspects; **Biocontrol of Soil-Borne Plant Diseases**; **Carbon Emissions and Sequestration**; **Crop-Residue Management**; **Degradation**; **Plant–Water Relations**; **Pollution:** Groundwater; **Sustainable Soil and Land Management**; **Zone Tillage**

Further Reading

Allmaras RR, Langdale GW, Unger PW, Dowdy RH, and Van Doren DM (1991) Adoption of conservation tillage and associated planting systems. In: Lal R and Pierce FJ

(eds) *Soil Management for Sustainability*, pp. 53–83. Ankeny, Iowa: Soil and Water Conservation Society.
Cannell RQ and Hawes JD (1994) Trends in tillage practices in relation to sustainable crop production with special reference to temperate climates. *Soil Tillage Research* 30: 245–282.
Carter MR (ed.) (1994) *Conservation Tillage in Temperate Agroecosystems*. Boca Raton, Florida: CRC Press.
Carter MR (1994) A review of conservation tillage strategies for humid temperate regions. *Soil Tillage Research* 31: 289–301.
Carter MR (1998) Conservation tillage practices and diffuse pollution. In: Petchy AM, D'Arcy BJ, and Frost CA (eds) *Diffuse Pollution and Agriculture 11*, pp. 51–60. Aberdeen: Scottish Agricultural College.
Lal R (1989) Conservation tillage for sustainable agriculture: tropics versus temperate environments. *Advances in Agronomy* 42: 85–197.
Langdale GW, Alberts EE, Bruce RR, Edwards WM, and McGregor KC (1994) Concepts of residue management: infiltration, runoff, and erosion. In: Hatfield JL and Stewart BA (eds) *Crop Residue Management*, pp. 109–123. Boca Raton, Florida: CRC Press.
Logan TJ, Davidson JM, Baker JL, and Overcash MR (eds) (1987) *Effects of Conservation Tillage on Groundwater Quality*. Chelsea, Michigan: Lewis Publishers.
Mannering JV and Fenster CR (1983) What is conservation tillage? *Journal of Soil and Water Conservation* 38: 140–143.
Paustian K, Elliott ET, and Carter MR (eds) (1998) Tillage and crop management impacts on soil carbon storage. *Soil Tillage Research* 47: 181–349.
Pierce FJ (1985) A systems approach to conservation tillage: introduction. In: D'Itri FM (ed.) *A Systems Approach to Conservation Tillage*, pp. 3–14. Chelsea, Michigan: Lewis Publishers.
Rasmussen KJ (1999) Impact of ploughless soil tillage on yield and soil quality: a Scandinavian review. *Soil Tillage Research* 53: 3–14.
Uri ND (1998) Trends in the use of conservation tillage in US agriculture. *Soil Use and Management* 14: 111–116.

COVER CROPS

L Edwards, Agriculture and Agri-Food Canada, Charlottetown, PE, Canada
J Burney, Dalhousie University, Halifax, NS, Canada

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

In a global climate of unease over the loss of ecological capital due to accelerated soil erosion and other forms of soil degradation under traditional cropping systems, attention is increasingly turning towards sustainable approaches that will protect the soil surface year-round and enrich it in the long term. Although the principal crop in most annual systems affords adequate ground protection at full canopy, interim cover cropping is a necessity, particularly where the soil is subject to erosion by water or wind. Water erosion takes its toll on sloping land under the influence of rain or snowmelt (which makes humid areas particularly prone), while wind erosion mostly affects areas that are flat, exposed, and dry. The gravity of this process is as great as it is ancient, and, likewise, are the steps taken to correct it. The need for its correction is a pressing urgency as a globally expanding green consciousness demands measures to promote: (1) a soil environment where the surface (most fertile) layer stays intact and builds rather than loses fertility; and (2) an aquatic environment that is free of imported sediment (and chemical pollutants), mostly from unprotected farmland. The judicious use of cover crops (sometimes known as green covers) is central in attaining these achievements.

Cover crops are the most practical means of providing ground cover, and thus the principal agronomic approach to erosion control and fertility building on crop land. Mulching also has a historical place, but can be limited where the material is unsuitable for soil enrichment, or where it is easily dislodged by wind or frequent runoff. The merits of cover crops are, however, well established and are mainly attributed to their relative ease of agronomic control. There is, for example, a wide choice of species and planting techniques for ensuring adequate plant establishment in most temperate or tropical settings, thus making cover cropping feasible under most soil, climatic, and cultural conditions.

The goals of cover cropping have historically varied between single-purpose and multipurpose, and may thus range from pure stands (**Figure 1**) principally for ground protection, to mixed stands of food (**Figure 2**) or feed species. Whereas these latter are, in large part, established towards economic ends (e.g., grain or hay in north-temperate climates and pulses in tropical climates), eventually all systems manifest in soil quality improvement that may take the form of: (1) physical conditioning; (2) nitrogen building; or (3) organic matter enhancements to the benefit of any arable terrain, be it sloping or flat.

Figure 1 Pure stand of red clover (*Trifolium pratense*) as a rotation crop with potatoes (*Solanum tuberosum*).

Figure 3 Winter rye cover (*Secale cereale* L.) following main-crop potatoes (*Solanum tuberosum*).

Figure 2 Leucaena cover (*Leucaena* spp.) with alternating rows of maize (*Zea mays*). Courtesy of P.K.R. Nair.

The conservation and nutrient-enrichment roles of cover crops in orchards and other perennial cultures are recognized, but are not dealt with specifically in this article.

This article examines the agronomy of cover crops, thus addressing species selection and husbandry, and describing their function in soil quality improvement. It also considers the role of cover crops in soil erosion control because of the body of evidence that puts this relationship into focus.

Agronomy

Cover crops are found among both monocotyledonous and dicotyledonous species. The former botanical group, which comprises mostly cereals and grasses, are doubly useful in soil erosion control because, in addition to the direct role of their vegetation in shielding the soil surface against aggregate breakdown by raindrop impact, their fibrous root systems bind the soil, thus increasing its resistance to entrainment by overland flow. On the other hand, the soil-protection attributes of a dicotyledonous cover are due, overwhelmingly, to the broad-leaf nature of this group of plants – specifically their characteristically overlapping system of leaves geared to intercepting rainfall and minimizing splash.

Success in cover-crop establishment and function depends greatly on species choice; and, accordingly, there is wide variation in both botanical groups. Thus, species suitability is hardly ever limiting for the range of cover cropping needs that exist.

Species Selection

Within the limits of adaptability of any cover crop to prevailing climate and soil, foremost among the desired characteristics is establishment vigor, persistence to the end of the required period, and ease of eradication to make way for the succeeding crop – usually the main crop. For instance, in the humid Atlantic region of Canada where cover cropping is practiced against cool-season erosion, cold temperature is the major limitation. Thus, winter rye (Figure 3) is the appropriate cover-crop choice under these circumstances because of its superior autumn-establishment vigor which allows it to fit into the narrow window between autumn potato (*Solanum tuberosum* L.) harvest and the onset of cold temperatures. Its cool-period persistence and excellent regrowth in early spring give this species a place in farming tradition as a cover crop in this region as in parts of the USA where it is planted after soybeans (*Gycine soja* L.) or after cotton (*Gossypium hirsutum* L.) to increase infiltration and reduce erosion on compacted soil. Seeding rates can even be manipulated to compensate for late-autumn seeding, which is its fate where potato harvesting is delayed to increase profitability. Generally, the later the potato harvest, the larger and more suitable are the tubers for both the fresh and processed (mainly french fries) markets.

In sharp climatic contrast to temperate situations, in many tropical areas where dry impoverished soils are all that exist for cover cropping after main crops of grain, cover-crop choice is restricted to hardy, low-growing species of legumes (like *Leucaena* spp.) because of their ease of establishment and spread, deep roots, and prospects of soil enrichment. Forage legumes are generally much hardier and better-adapted to these conditions than are pulses. Drought-tolerant grasses may be cultivated where longer-term cover is appropriate or where erosion is a severe threat. For this purpose, species such as *Chloris* spp. and *Cynodon* spp. have proven valuable.

Not always are cover crops confined to the period of most critical need, but might be deliberately taken to maturity and harvested for their economic products, usually within rotation systems. This is normally a business decision in which species versatility and adaptability to market opportunity play an important part. For example, many eastern Canadian farmers seed winter cereals as cover crops and cultivate them to full maturity if grain market prospects are good, even choosing to plant winter wheat (*Triticum aestivum*) where winter rye is preferable as cover. Under these conditions, winter rye is only an occasional grain crop because of the danger to livestock of fungus-infected grain.

Where the farmer wishes to hedge against cover-performance uncertainty and opts for mixed instead of pure stands, species compatibility is of prime importance in exercising choice, whether for herbage mixtures or herbage-grain companion cropping (as seen in Figure 2).

Seeding and Husbandry

Where the object of cover cropping is solely or primarily vegetative production, seeding method and spacing are far less exacting than where the object is an economic product like grain or pulses. Grass species, for example, are usually broadcast-seeded (surface-spread) for cover while legumes and cereals might be row-seeded (by drilling). This latter is a relatively demanding operation, but is an insurance against loss to pests or poor weather. However, where the system calls for companion cropping with grasses, this component might be interrow-seeded either simultaneously or in relay. Within this system, grass–legume mixtures are common, as are herbage–cereal mixtures, and usually require two operations to accommodate differences in seed size and planting-depth. Broadcast seeding may be accompanied by light harrowing to cover the seed.

Seedbed preparation is traditional for cover-crop seeding, but not always necessary depending on soil moisture conditions, terrain, seed availability, and the availability of planting equipment. Cultural circumstance might also play a part. Some US dairy farmers have been known to seed winter rye successfully in silage corn stubble and attain good winter cover with only light disking. Tea producers in India have successfully used a system of contour-staggered trenching (to conserve water) and leguminous cover crops, and have had yield increases exceeding 30% from the companion tea crop.

A variant of interseeding for cover was realized where winter cereal was successfully broadcast into standing crops of potatoes just days before harvesting in the autumn. Good winter cover was achieved without seedbed preparation. In this system, which was meant to achieve a measure of cover-crop growth (Figure 4) by the time the host crop was harvested, the winter cereal had germinated on undisturbed, relatively compact ground surface (but in a favorable microclimate), and was buried during potato harvesting with negligible damage to the emerging seedling. In this circumstance, soil moisture was adequate for growth – typical of this period in humid temperate climates.

Pregermination techniques to save establishment time are widely used, but are only lucrative for high-value crops (especially vegetables), enabling them to take early advantage of good field conditions in spring and then make a profitable entry to the market upon maturity. However, this technique is less practical for a low-value commodity like ground cover – whether it be in a temperate or tropical setting. Furthermore, short-term cover crops, upon germination, normally face a less-than-ideal period for establishment and growth since they usually follow the main crop and develop under progressively limiting conditions (i.e., cold or dry).

The most practical procedure to pretreat and seed cover crops for quick establishment, particularly

Figure 4 Early stand of winter rye (*Secale cereale* L.) broadcast-seeded into a standing crop of potatoes (*Solanum tuberosum*) just before harvest in the autumn.

under less-than-ideal soil-surface conditions, is the hydroseeding procedure used by highway departments which broadcast a liquid slurry of seeds, lime, fertilizer, and a fine, fibrous material that serves as a mulch, once applied. Straw or hay may also be used as a carrier if finely chopped, but is less capable of uniform spread. In the Atlantic region of Canada, this approach was tested to autumn-seed winter rye as a ground cover, but was constrained by uneven spread due mainly to equipment limitations. For example, standard on-farm equipment like manure spreaders proved inadequate, thus bringing into question: (1) the practicality (for a given farming enterprise) of even attempting to establish cover crops under poor ground-surface conditions; or, more directly, (2) the wisdom of investing in specialized equipment to facilitate cover cropping.

Cover-crop maintenance management naturally depends upon crop type and may, basically, be carried out as for any other crop husbandry to optimize growth. Where crop mixtures or companion cropping is called for, management becomes tricky and must be geared to the more sensitive of the crop components. However, in practical terms, since cover crops are usually relegated to secondary-crop status, postseeding inputs are often minimal. For example, fertilizers might not be used at all where cover cropping is short-term or if legumes are in the system. In any case, cover crops that follow heavily fertilized main crops (e.g., grain in Brazil and central or western Canada, or potatoes in the Atlantic region of Canada) could benefit adequately from nutrient carry-over.

Where cover crops are short-term and sandwiched by main crops into periods of unfavorable growth, husbandry attempts have, nevertheless, been made to boost performance through targeted fertilizer use, particularly with potash, with the object of enhancing winter- or drought-hardiness. However, success has been limited since the plants are unlikely to respond significantly during difficult periods of growth or when following a crop like potatoes that are normally overfertilized with K to improve eating or processing quality. Thus, it was found that increasing K dosages had no effect on autumn-seeded winter rye cover after main-crop potatoes in the Atlantic region of Canada, either in terms of leaf weight or the content of K in the leaves (Tables 1 and 2) at any time between cover establishment and termination in spring.

Termination of the cover-crop phase, in preparation for the main crop, is usually a tillage procedure that incorporates the plant material into the soil. This may range from near-surface harrowing of a short-duration cover of easily manageable plant material to plowing-under the vegetation, which may find itself below the root zone – as is the case with cereal stubble or swards. In either case, the procedure invariably aims to facilitate seedbed preparation and minimize interference with seeding machinery. It also aims to maximize contact between the incorporated material and the soil.

Termination of the cover-crop phase, in preparation for the main crop, may also be achieved through herbicide use, which is more common in large-scale operations or where soil-surface conditions are difficult. Undoubtedly, operational ease is the object of this approach, but it does little or nothing for soil improvement.

In situations of mixed farming, which characterizes many small to medium enterprises in the tropics, it is a common practice to harvest the cover crop periodically as cattle feed, although it may mean weakening the stand to the point where it contributes little to soil conservation or enrichment – even to the point of decimation.

Role in Soil Quality Improvement

Cover crops contribute to soil quality improvement principally through their decomposition by soil microbes. The products of decomposition, while

Table 1 Late autumn response (dry weight) of winter rye to applied K

	Trial					
	Year 1			*Year 2*		
K level[a]	*Plant count (tillers m^{-1})*	*Shoot ($g\,m^{-1}$)*	*Root ($g\,m^{-1}$)*	*Plant count (tillers m^{-1})*	*Shoot ($g\,m^{-1}$)*	*Root ($g\,m^{-1}$)*
K_0	127	4.46	0.93	145	4.46	1.27
K_1	140	4.82	1.01	162	4.36	1.57
K_2	141	4.92	0.85	148	4.11	1.45
LSD ($P=0.05$)	33	1.23	0.23	37	1.44	0.43
($P=0.01$)	46	1.70	0.31	51	1.99	0.59

[a]Year 1: K_0, control, K_1, 50 kg ha^{-1}, K_2, 75 kg ha^{-1}; year 2: K_0, control, K_1, 100 kg ha^{-1}, K_2, 150 kg ha^{-1}.
LSD, least significant difference.
Reproduced with permission from Edwards LM (1986) *Canadian Journal of Soil Science* 66: 31–35.

Table 2 Growth recovery (dry weight) of winter rye in spring with autumn-banded K

	Trial			
	Year 1		Year 2	
K level[a]	Shoot ($g\,m^{-1}$)	Root ($g\,m^{-1}$)	Shoot ($g\,m^{-1}$)	Root ($g\,m^{-1}$)
K_0	60.9	19.9	10.4	6.0
K_1	65.0	17.1	10.3	6.1
K_2	57.6	21.3	11.7	7.2
LSD ($P=0.05$)	17.2	5.9	2.2	1.4
($P=0.01$)	23.8	8.2	3.0	1.9

[a]Year 1: K_0, control, K_1, 50 kg ha^{-1}, K_2, 75 kg ha^{-1}; Year 2: K_0, control, K_1, 100 kg ha^{-1}, K_2, 150 kg ha^{-1}.
LSD, least significant difference.
Reproduced with permission from Edwards LM (1986) *Canadian Journal of Soil Science* 66: 31–35.

Figure 5 Stand of *Pueraria phaseoloides* under a crop of cassava (*Manihot* spp.). Courtesy of IITA, Nigeria.

generally adding to the soil organic matter (SOM) reservoir, benefit the soil in two specific ways, i.e., through soil physical conditioning and through fertility building. The degree of enrichment depends on the quantity and quality of cover-crop biomass. Cellulose-rich plants or plant parts degrade far more rapidly than if they were ligneous – as is the nature of mature grasses. Hence, leafy portions of the shoot system degrade far more rapidly than stemmy portions. In any case, with decomposition, the resulting increases in SOM benefit the soil by improving its structure, thus enhancing gaseous exchange, internal drainage, nutrient exchange and water retention. Increased SOM also means greater benefits to soil biodiversity, which is at the center of all SOM dynamics and increased aggregate stability which, in turn, means a reduced tendency to erode.

From the standpoint of soil nutrient enrichment, cover crops contribute to the nutrient pool through the mineralization of decaying biomass under the influence of soil biota, which release the full range of plant nutrients in absorbable form. Where the decaying material is leguminous, nitrogen enrichment is predominant because of the nitrogenous nature of this family of plants and, particularly, the capacity of their roots to fix atmospheric nitrogen in symbiotic association with bacteria. Thus, by root rupture or necrosis, the soil eventually benefits from this nitrogen-fixing activity. *Pueraria phaseoloides* L. (Figure 5), a leguminous cover crop popular in the tropics, accumulated 150 and 250 kg N ha^{-1} within 4–18 months in parts of West Africa. In the end, the resulting soil nutrient enrichment following cover-crop fallows generally means increased crop production and savings on fertilizer costs, particularly under the poor-fertility conditions long-experienced and reported for Latin America and Africa. Tropical leguminous cover crops are also known for their foliar abundance. Some species contribute over 95% to total litter fall and are particularly favored in orchards and tree plantations if they are shade-tolerant and noncompetitive.

Managing cover crops for soil physical improvement is best achieved with gramineous species. For example, grass cover crops, by virtue of the sheer volume and mass-distribution of their fibrous root systems within the root zone, bind the soil in a manner that gives it excellent structural attributes, the most outstanding of which are internal drainage and mechanical resistance. The vast root mass of which grasses are capable also contributes substantially to soil structure improvement via the SOM reservoir upon decay.

The ultimate cover-crop manipulation towards building soil fertility is the age-old practice of green manuring, the previously mentioned tillage activity that incorporates the plant material with the soil and stands out as the most useful form of crop destruction so far. Since most cover species are herbaceous, this material is usually not difficult to handle and biomass decomposition can be rapid. Because the speed of decomposition depends on the degree of contact between the plant material and the soil, and further depends upon the size of the incorporated material and the thoroughness of its incorporation, green-manuring tillage may require attention to be fully effective.

Role in Soil Erosion Control

Nature of Soil Erosion

There will never be any doubt of the beneficial role of cover crops in the control of soil erosion, which can be devastating to farmland (its customary point of origin) and neighboring environments. Its impact, however, is far more easily seen or felt than understood by either the land-user or the user of the

Table 3 Effect of winter rye (*Secale cereale* L.) cover on soil erosion using a laboratory rainfall simulator

	Mean value			
	Runoff		Splash	
Cell	Volume (ml)	Sediment (g)	Volume (ml)	Sediment (g)
Crop				
Bare	4191*a*	60.2*a*	434*a*	1.50*a*
Early-seeded	3917*a*	10.1*b*	390*a*	0.54*b*
Late-seeded	4943*a*	17.0*b*	437*a*	0.86*b*

a–b Means followed by the same letter in any column for any treatment are not significantly different at $P = 0.05$.
Reproduced from Edwards LM and Burney JR (1987) *Canadian Agricultural Engineering* 29: 109–115.

Table 4 Comparison of treatments: relative soil loss under outdoor simulated rainfall

Treatment	No. of runs	Calculated relative soil loss
Fallow	24	1.00*a*
Potato 0% cover	6	1.00*a*
Potato 5% cover	6	1.00*a*
Potato 70% cover	6	0.50*b*
Potato 90% cover	6	0.04*c*
Clover	12	0.06*c*

a–c Means in any column for any treatment followed by the same letter are not significantly ($P \leq 0.05$) different.
Reproduced from Parsons TS, Burney JR, and Edwards LM (1994) *Canadian Agricultural Engineering* 36: 127–133.

affected off-farm site or resource, be it a culvert, road, stream, river, lake, or estuary.

Fundamentally, the magnitude of soil erosion is contingent on a causative effect, a modulating effect, and a resistance effect. In the context of conserving topsoil on sloping farmland, the resistance effect can be attributed in large measure to live cover.

Undoubtedly, cover must be adequate, and becomes particularly crucial during periods of vulnerability to erosion on the basis that the occurrence, in combination, of high-intensity rainfall and weak soil resistance means that soil erosion on any given land surface would be otherwise high. Since contributing physical factors like weather and the degree of slope are uncontrollable, cover-crop usage becomes invaluable – particularly in tropical cultures where mulching material has to meet a more pressing need as fuel or building material than as ground cover.

Increased foliar mass of the cover crop reduces splash erosion predictably with increasing ground coverage, and this effect is incorporated as a crop-management factor in calculating soil loss. Cover crops work by reducing splash detachment and by intercepting the splashed soil, thus limiting its trajectory movement.

In work done in the Atlantic region of Canada to compare the effectiveness of winter rye (*Secale cereale* L.) cover in splash control using a laboratory rainfall simulator and splash equipment, a 47% reduction (compared to bare soil) was recorded for a highly erodible fine sandy loam (Table 3).

Runoff and Erosion Control

Over longer distances, the progressive build-up of runoff generates an increasing shear stress on the soil surface. As the effects of runoff increase relative to the constant effects of splash downslope, the root structure of the cover crop becomes increasingly important for keeping the soil in place.

A comparison of cover-management effectiveness under varying degrees of canopy, using a field rainfall simulator, showed that soil loss was not affected by a canopy of potatoes of up to 5%. It took a 70% canopy before the crop cover could keep the soil reasonably in place, and a 90% canopy before it was as effective as a mature stand of red clover (*Trifolium aestivum* L.) in the same comparison (Table 4).

Runoff comparisons, using a laboratory rainfall simulator on a fine sandy loam subjected to freezing and thawing cycles, showed that a cover of winter rye decreased runoff sediment to 23% of that from the bare soil, provoking significant deposition, particularly of the coarser-size grades. In taking a close look at the cover parameters and in assessing relative contribution to variations in sediment concentration, this study also showed that leaf area was by far the most important. Leaf area index, together with long-term climatic data, is a key component of many models that are used to determine safe planting dates, especially for winter cover crops.

Perspective

Present approaches to managing cover crops within a farm production system for soil enrichment or to minimize soil-surface destabilization are, for the most part, inadequate as farm management schedules mostly allow only: (1) short-term cover sandwiched between main crops; or, at best, (2) a growing season, where longer-term cover is required to bring the soil to full health. However, anything more than short-term cover could mean economic loss even if the cover is commercializable (e.g., as hay, cereal, or pulse).

The question still arises as to why cover cropping is ignored in the face of its agronomic and ecological merits, and why the pressure of commerce (minimum expenditure and maximum profit) overtakes most other considerations. For instance, monocropping

grain crops or potatoes is common commercial practice in parts of North America and South America, with dire soil-health consequences in the long term. On the other extreme where, instead of cover cropping, bare fallowing is done to rest the soil from commercial production – a common practice with cereals in the Canadian Prairies – the result is a noxious salinization of the surface due to the upward movement of surplus soil nutrients – ultimately wasted. The use of cover crops to minimize nutrient wastage, whether through surface salinization or leaching, especially after intensive cropping, is widely recommended husbandry.

The practice of bare fallow in temperate or tropical regions is a missed opportunity to maintain (even enhance) the soil biological activity built up by the preceding crop. Thus, judicious selection of fallow cover crops can do much to enrich the soil microbial, earthworm and microarthropod populations – with benefits particularly needed under marginal conditions.

The danger to soil environmental health of ignoring cover cropping has struck a chord with many ecologists who press for its adoption at all levels of farming, and are ready to link its neglect, more widely, to global environmental decay. This linkage may, therefore, place cover cropping squarely in the realm of biological agriculture – remedial or sustainable – particularly based on its potential to sequester substantial amounts of soil organic carbon – estimated at nearly 30 millions of tonnes of carbon per year in the USA. Thus, it could only be to the advantage of cover cropping that it is identified with a 'green culture' that seeks a balance between agricultural enterprise and environmental capital.

Cover cropping is seen as green culture, given its role in SOM building, soil structure improvement, soil fertility increase, and disease suppression with minimal or no chemical inputs. Thus, it stands to gain expanding popularity and increasing political importance, considering that sustainable agriculture is successfully moving, as a recognized farming system, from temperate regions (Europe and North America) to tropical regions (Asia, South America, and Africa).

In view of the exigencies of cover cropping and the associated sacrifices, the question further arises of its relative place as a fertility builder at this time when the livestock industry produces, on average, a near excess of excrement whose benefit to soil health is unquestionable – having a greater capacity than cover crops to build SOM and enrich the soil rapidly. Furthermore, its effects are longer-term under intense cropping. In comparison, benefits derived from cover cropping (e.g., in terms of structure improvement) are only fully realized after several years of continual occupancy, and can disappear after a single year of commercial cropping under intense tillage. Therein lies a debate that could become particularly crucial where cropland is scarce.

But beyond any debate on the economy of cover cropping for soil-fertility building and erosion control on farmland is the importance of this practice in the wider context of terrestrial sustainability and ecological balance. Thus cover-crop species find important use in stabilizing fragile sand dunes that serve as habitat for rare coastal birds, in helping to restore surface-mine wastes to pre-mining productivity, and concealing dumps of solid industrial wastes that may pose a danger to animal life where easily accessible. Even the aesthetic value of cover crops is high considering their role in rendering unsightly or near-barren landscapes visually pleasing, even without mentioning their contribution to wildlife habitat and still broader biodiversity in temperate and tropical regions alike.

Having entered into the realm of green culture, cover cropping stands to gain status as a bona fide environmental ethic. It is, therefore, anticipated that the environmentalist movement will increasingly put pressure on farm operators and other land-users to recognize its ecological value and adopt it, at least in the name of environmental stewardship.

See also: **Carbon Emissions and Sequestration**; **Crop-Residue Management**; **Mulches**; **Nitrogen in Soils:** Cycle

Further Reading

Beasley RP, Gregory JM, and McCarty TR (1984) *Erosion and Sediment Pollution Control*. Ames, Iowa: Iowa State University Press.

Edwards LM (1986) Late-fall application of potash to winter rye to improve ground cover effectiveness. *Canadian Journal of Soil Science* 66: 31–35.

Edwards LM (1989) Dry matter growth performance of red clover and Italian ryegrass as cover crops spring-seeded into fall-seeded winter rye in relation to soil physical characteristics. *Journal of Soil and Water Conservation* 44: 243–247.

Edwards LM (1998) Comparison of two spring seeding methods to establish forage crops cover in relay with winter cereals. *Soil Tillage Research* 45: 227–235.

Edwards LM and Hergert G (1990) Establishing winter rye as a cover crop after potatoes. *Canadian Agricultural Engineering* 32: 183–187.

Foster GR (1982) Modelling the erosion process. In: Haan CT, Johnson HP, and Brakensiek DL (eds) *Hydrological Modelling of Small Watersheds*, pp. 297–370. ASAE Monograph Number 5. St Joseph, MI: ASAE.

Govers G (1991) Spatial and temporal variations in splash detachment: A field study. *Catena* Supplement 20: 15–24.

Kay BD (1990) Rates of change in soil structure under different cropping systems. *Advances in Soil Science* 12: 1–52.

Morgan RPC (1985) Establishment of plant cover parameters for modelling splash detachment. In: El-Swaify SA, Moldenhauer WC, and Lo A (eds) *Soil Erosion and Water Conservation Engineering*, pp. 377–383. Ankeny, IA: Soil Conservation Society of America.

Schwab GO, Fangmeier DD, Elliot WJ, and Frevert RK (1993) *Soil and Water Conservation Engineering*, 4th edn. New York: John Wiley.

Siegrist S, Schaub D, Phiffner L, and Mäder P (1998) Does organic agriculture reduce soil erodibility? The results of a long-term field study on loess in Switzerland. *Agriculture, Ecosystems and Environment* 69: 253–264.

Ward AD and Elliot WJ (1995) *Environmental Hydrology.* Boca Raton: CRC Press.

CROP ROTATIONS

C A Francis, University of Nebraska, Lincoln, NE, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Since early in the history of agriculture, farmers have recognized the value of crop rotations. Sequences of different crops, in contrast to continuous cultivation of the same crop in the same field, were observed to have higher yields and were used thousands of years ago in China. In prehistoric Europe, when human population was small, land-clearing for several years of planting crops and pastures was often followed by allowing regrowth of natural vegetation and a repetition of this cycle of different species, but little is known about precise sequences. It is likely that there were rotations of annual crops with pastured areas, although this is speculation. The system is similar to the more familiar slash-and-burn or swidden agriculture still practiced in some parts of the tropics.

There is written evidence of the recognition of importance of crop rotations as well as putting specific crops on unique soil types during the Roman times. Although Cato the Censor (second century BC) dedicated most of his writing on farming to culture of grapes and olives, he also discussed the importance of applying manure and of fallowing land before planting annual or perennial crops. His writings describe the importance of vetch, beans, and lupines in helping to build the soil or fertilize the land, and that cereal crops following these legumes benefited and produced higher yields. Varro and Virgil (first century BC) recognized the importance of alfalfa, a deep-rooted perennial legume, in providing improved fertility for crops that followed, and they recommended rotations, fallow periods, and specifically a bean–spelt wheat rotation. Pliny (first century AD) wrote extensively on farming, including crop rotations in the context of Italy and the Mediterranean region. Likewise, there are written records of early effects of crop rotations and advice against the continuous planting of cereals in the literature from the Muslim era in the Iberian Peninsula.

In the Middle Ages before the Renaissance, there was apparently a reversion to crop–fallow sequences. There are limited reports of perennial pasture-grass plantings in rotation with cultivated cereal crops. Use of animal manure was prevalent. Perhaps the effects of application of manure were more apparent in a given year, if there were side-by-side areas with and without manure, compared with the more subtle differences that were the result of rotation, effects not easily observed. In the seventeenth and eighteenth centuries, more organized legume cereal rotations began to appear in reports from England. They often consisted of 2–4 years of vegetable or fodder crops, 1 year in legumes, and 1 or 2 years in cereals. Mixed farming with crops and livestock was prevalent, and the pastures and animal manure undoubtedly contributed to cereal yields. These rotation systems were taken to North America by early settlers from northern Europe.

Cropping diversity in many early systems included multiple species as well as crop rotations. In Europe the small grains were often mixed cultures of cereal species that were used for grain and for fodder. The well-known maize/bean/squash systems of Central America and Mexico as well as the intercrop and relay crop systems of maize/bean–potato in the Andean Zone are examples of diverse mixtures that evolved over centuries as people selected these New World crop species.

With an emerging understanding of soil chemistry and the responses of plants to applied manures and other soil amendments, crop rotations in the first half of the twentieth century focused again on cereal–legume sequences. In long-term experiments in

Missouri, USA, increased crop yields were found in rotations even when the cereals were supplied with adequate nutrients from animal manure application. According to various state experiment station reports, yields of maize were increased from 20 to 60% in rotations with legumes; in experiments in Ohio, there were 25% higher maize yields in rotation even when nitrogen, potassium, and phosphorus were applied to continuous maize. These popular rotation systems were phased out of most farming operations after World War II when large supplies of synthetic nitrogen and other multiple-nutrient fertilizers became popular. The specialization in field crops, especially cereals, and disappearance of livestock from most farms further pushed the systems into continuous cropping or simplified 2-year cereal–legume rotations. Only in the past two decades, with the interest and growth of organic farming and more integrated systems, are longer and more complex rotations coming back into popularity in the USA and Europe.

A summary of general types of crop rotations is presented in Figure 1, showing the array of different levels of genetic diversity that have been exploited in farmers' systems across the centuries. Rotating different varieties of the same cereal, rotating cereals, and rotating cereals with legumes are all strategies to provide some of the soil-fertility and plant-protection benefits that come from diverse cropping sequences. More diverse are the sequences of summer–winter crops and rotations of pastures with cropping fields, and most diverse of all are the annual–perennial sequences and combinations such as alley cropping.

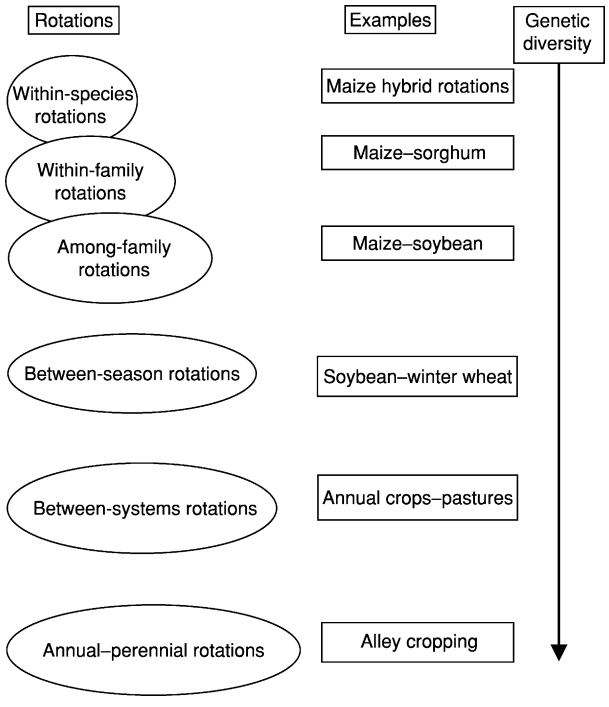


Figure 1 Different types of crop rotation, crop–pasture rotation, and long-term perennial alternatives, showing the increasing genetic diversity of specific systems (from top to bottom).

It is important to note that crop rotations are influenced by a wide range of factors, including biological interactions and contributions to yield on the farm, relative prices of differently adapted crops, and potentials for production and marketing from crop/animal systems. Farmers' potential to use rotations may depend on the agreements in place with landowners, since, for example, about half of the land currently farmed in the US Midwest is not owned by the people farming that land. Further, the political decisions in the EU and USA on crop commodity support payments and promotion for global markets for some selected crops narrow the farmers' options. Political decisions in the Baltic countries to position themselves for entry into the EU caused major changes in cropping systems during the 1990s, as farmers began to compete in a regional and global market and abandoned the guaranteed markets that were common under the former Soviet system. It is within this complex economic, biological, and political framework that crop rotations must be discussed.

Using current improved hybrids and varieties of major field crops, yields are generally improved by an average of 10% or more owing to crop rotation alone. This is a consistent observation, even when all apparent nutrient needs of crops are satisfied and there are no obvious limiting pest problems. If there are serious limiting factors that cannot be remedied by other means, for example a severe insect infestation or plant pathogen infection, crop rotations may increase yields much more. Thus the mechanisms of crop rotation contributions to yields and economic return are many and complex, and still poorly understood; those factors that have been researched relate to soil fertility, crop protection against pests, and the economic and social dimensions of farming.

Soil Fertility and Crop Yields

Cereal–legume rotations are among the most frequently used systems, dating back to before Roman times, and they still provide the foundation for successful design of crop sequences today. Crop rotations among crops from different families are among the most frequently used today in the temperate zone. In areas of adequate rainfall, above 600 mm year^{-1}, a maize–soybean rotation is prevalent in the Corn Belt of the US Midwest. In areas with lower rainfall, irrigation is used to assure a crop, and rotations are less common in these systems. Crop rotation effects are

Table 1 Examples and mechanisms for crop rotation contributions to soil fertility and nutrient cycling

Examples or mechanisms	*Contribution to soil fertility*
Cereal–legume rotation	Improves soil structure, texture, tilth, quality Increases water percolation, aeration, nitrification Promotes more efficient nutrient and water use
Different root structures	Shallow–deep species exploit more nutrients
Legume–grass sequence	Increases soil organic matter, reduces pests
Biennial and perennial legumes	Increases soil organic matter, manages weeds
Longer rotations and diversity	Improves soil quality, increases soil organic matter
Pasture–crops sequence	Increases organic matter, improves nutrient efficiency
Manure application	Stimulates soil mycorrhizae
Unlike crop sequences	Stimulates root exudates, increases mycorrhizal spores
No-tillage	Keeps earthworm burrows intact, improves soil mixing

often attributed to improved soil fertility that accrues from legumes in the sequence and the increased soil organic matter that improves soil quality. A number of specific effects of crop rotations are summarized in Table 1.

Maize in a 2-year rotation with soybean in the US Midwest consistently yields 5–20% more than continuous maize; yield increases in the same experiment in the same field often vary from 5 to 15% across years, depending on different growing conditions and what factors are limiting yields in the weather and soil situation in that array of years. Maize yields after a perennial crop such as alfalfa not only have higher yields than continuous maize, but the need for additional nitrogen is minimal in the 1st year. There are inconsistent results for maize yields in longer crop rotations with other annual crops. Similar yield increases are observed in other cereals grown in rotation with legumes, for example the grain sorghum–soybean rotation. Soybean in these rotations also shows approximately a 10% increase in yields, in spite of the fact that soybean fixes much of its needed nitrogen and fertilizer is most often not applied to this crop.

In areas of less rainfall, wheat is grown in rotation with a fallow year to accumulate enough soil moisture to make the crop profitable. An average of 25% of the fallow-year moisture is stored for a succeeding crop, with the rest lost to evaporation, yet this is enough in traditional systems to make the wheat–fallow rotation consistently profitable. Recently such systems as ecofallow, with minimal tillage and a sequence of two crops in 3 years (e.g., wheat–proso millet, wheat–sunflower, or wheat–maize) with short fallow periods, have intensified cropping and made better use of available rainfall.

Rotation of summer and winter crops, annual and perennial crops, and crops with pastures for grazing all reduce the need for applied fertilizers. This is due to rotations in crop root systems, nutrient-uptake patterns, timing of nutrient needs, diversity in crop residues, and different nutrient needs.

Cultural Management

Alternative soil-management methods, choices of crop species, and harvesting techniques can contribute to soil fertility, pest management, and amount of water available to crops in rotations. Reduced tillage maintains more residue on the soil surface, reduces erosion, and conserves soil moisture compared with plowing and other intensive land preparation and cultivation methods. The additional stored moisture opens up a farmer's options for crop rotation and increasing cropping intensity through the growing season. Catch crops of short-cycle legumes may thus be planted following a principal cereal grain crop, or a green manure crop planted with a cereal, in combinations not possible without this additional moisture.

Reduced tillage increases earthworm populations and soil quality, and slows oxidation and breakdown of organic matter to increase long-term nutrient supply to crops. Rotations of tillage practices that reach different soil depths can reduce the potential for soil compaction that leads to formation of a plow or disk pan. Rotating unalike crops may also require use of different equipment at different times of the year, and this diversity should promote a more diverse and reduced pest population and healthier farm ecosystem.

Pest Management with Rotations

Planting different crops in successive seasons or years can disrupt the reproductive cycles of many weed, insect, plant pathogen, or other pest species. This in turn can lower their potential for reducing crop yields or quality and diminish pest-control costs. Organic crop production systems rely almost entirely on rotations plus genetic resistance to manage crop pests without chemical pesticide applications. Problems with soil-borne pathogens such as bacteria, fungi, and nematodes can be reduced by crop rotations of different species. For some soil biotic pests, rotation is the only economical way to control them. A number

Table 2 Mechanisms of crop rotation contributions to plant protection against weeds, insects, pathogens, nematodes, and other biotic problems

Mechanism or process	*Contribution to plant protection*
Cereal–legume sequences	Favors different weed spp., facilitates mechanical control
Unlike crop sequences	Reduces soil-borne pathogens
Maize–soybean sequence	Controls maize stem borer, maize rootworm
Vegetable–cereal sequence	Controls nematodes
Potato–legume sequence	Controls wireworms
Continuous plant cover	Provides competition with weeds, reduces erosion
Legume cover crops	Reduces soil erosion, competition with undesirable weeds
Alfalfa in crop sequence	Control of annual weeds, increased predators and parasites
Strip-cutting alfalfa	Increases insect movement, increases natural enemies
Diverse crops in sequence	Soil biology diversity and suppression of soil pathogens
Cropping diversity over time	Promotes soil microorganism biodiversity

Table 3 Mechanisms of crop rotation contributions to economic and social consequences of farming systems

Mechanism or process	*Contribution to economics or social viability*
Rotation of fields	Places uniform areas in all crops each year
Diverse crops in rotations	Provides forages, cereals, cash crops, soil building
Sequences of pesticides	Avoids buildup of resistance of insects, pathogens, weeds
Different tillage depths	Avoids buildup of plow-pan or compaction layer
Range in crop-planting dates	Spreads labor needs, reduces equipment costs
Range in harvest dates	Reduces peak labor demand, reduces equipment costs
Diverse array of crops/products	Buffers price changes in marketplace, gives direct sales
Different skills, labor needed	Increases interaction with local community for labor

of mechanisms or methods for pest control are listed in Table 2.

Longer rotations that include annuals and perennials, or crops and pastures, provide both soil-fertility and pest-protection advantages. Annual crops allow frequent cultivation that will often suppress perennial weeds. Perennial crops may provide year-long competition that shades out annual weed species. Crop/animal mixed farming and a sequence of several years of annual crops followed by perennial grass/legume pasture mixes and grazing help to diversify income sources, as well as control many weeds and other pests that become problems in continuous cultivation of annual crops. Rotations of different animal species in the pastures – ruminants, pigs, poultry – make use of their different preferences and make more complete use of available forage, as well as reducing animal diseases that become problems in confined raising of large animal numbers in a small space. Goats will eat many weed species that are passed by other grazing ruminants, while chickens eat insect larvae from the manure deposits of larger grazers. The 4-year annual crop–4-year pasture rotation is a common system in Argentina.

Economic Diversity

In addition to these biological advantages of crop rotations, there are positive economic and social aspects that result from diverse plant combinations on the farm. Table 3 lists a number of direct economic benefits or those that will develop as a result of biological and other soil changes in fields over time. Either rotation of large fields or within-field diversity in strip-cropping annual systems will diversify the income stream and protect against a weak market for a single commodity. Planting summer and winter crops, or annual and perennial crops, will spread the demand for labor and equipment and allow farming of the same area with smaller equipment. Such rotations provide income at different times of the year. The diversification into more intensive crops and products provides the opportunity for using labor from other farms or nearby communities, perhaps opening new markets for direct sale as well.

Economic diversity can be enhanced by both crop rotations and the diverse products that come from the farm. If value-added activities on the farm or in the local community can make effective use of available labor when field operations are less intense, this strategy can increase farm income and return to investment. Crop rotations and diversity of products open new economic opportunities for on-farm sale or cooperation with neighbors in other direct sale options. A diverse farm is also attractive to visitors for educational programs, farm stays, hunting, and other creative ways to add value to the rural landscape and its natural resources.

Future Perspectives

Although economic rewards, consolidation, and specialization drove many farmers away from rotations in the latter years of the twentieth century, it is

likely that cropping diversity and rotations will become more important in the future. Federal and regional subsidy programs that link payments to conservation of soil, water, and other natural resources often stipulate field and farm diversity as conditions for farmers to continue to qualify for payments. Organic farming regulations and subsidies in some countries specify the types of rotations that must be used for farmers to qualify. There is often a rule against continuous cropping of the same crop species and a requisite to include legumes in at least half of the years in a crop sequence.

With increased costs of energy, the nutrients from green manure and animal wastes become more attractive and cost-effective. When animal manure from concentrated confinement operations becomes an excessive liability, due to water-quality dangers, there will be greater pressure from society and from regulators to rescue this resource and cycle it back into the crop production process. Diverse crop rotations provide multiple opportunities through the year to apply manure and compost back in production fields. A growing appreciation of how agroecosystems can be designed to mimic the diversity and function of natural ecosystems provides the motivation to seek innovative and diverse crop rotations that maximize productivity and minimize negative impacts on the environment. Crop rotations provide multiple benefits to farmers and to society, and will be increasingly important in tomorrow's agriculture.

Further Reading

Altieri MA (1994) *Biodiversity and Pest Management in Agroecosystems*. New York: Food Products Press, Haworth Press.

Francis CA and Clegg MD (1990) Crop rotations in sustainable production systems. In: Edwards CA, Lal R, Madden P, Miller RH, and House G (eds) *Sustainable Agricultural Systems*, pp. 107–122. Ankeney, Iowa: Soil & Water Conservation Society.

Francis CA, Flora CB, and King LD (eds) (1990) *Sustainable Agriculture in Temperate Zones*. New York: John Wiley.

Gustafson AF (1941) *Soils and Soil Management*. New York: McGraw-Hill.

Karlen DL, Varvel GE, Bullock DG, and Cruse RM (1994) Crop rotations for the 21st century. *Advances in Agronomy* 53: 1–45.

Lal R and Pierce FJ (1991) *Soil Management for Sustainability*. Madison, WI: Soil Science Society of America.

Olson R, Francis C, and Kaffka S (eds) (1995) *Exploring the Role of Diversity in Sustainable Agriculture*. Madison, WI: American Society of Agronomy.

Thurston HD (1992) *Sustainable Practices for Plant Disease Management in Traditional Farming Systems*. Boulder, CO: Westview Press.

CROP WATER REQUIREMENTS

L S Pereira and I Alves, International Commission on Agricultural Engineering (CIGR), Lisbon, Portugal

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Crop water requirements (CWR) are defined as the depth of water (millimeters) needed to meet the water consumed through evapotranspiration (ET_c) by a disease-free crop, growing in large fields under non-restricting soil conditions, including soil water and fertility, and achieving full production potential under the given growing environment. Defining 'crop evapotranspiration' (ET_c) as the rate of evapotranspiration (millimeters per day) of a given crop as influenced by its growth stages, environmental conditions, and crop management to achieve the potential crop production, then the CWR is the sum of ET_c for the entire crop growth period. When management or environmental conditions deviate from the optimal, then that rate of evapotranspiration has to be adjusted to the prevailing conditions and is called actual crop evapotranspiration (ET_a). Both CWR and ET_c concepts apply to either irrigated or rainfed crops.

For irrigated crops, the concept of CWR has to be complemented by that of irrigation water requirement (IWR), which is the net depth of water (millimeters) that is required to be applied to a crop to satisfy fully its specific crop water requirement. The IWR is the fraction of CWR not satisfied by rainfall, soil-water storage, and groundwater contribution. When it is necessary to add a leaching fraction to assure appropriate leaching of salts in the soil profile, this depth of water is also included in IWR. In practice, IWR has to be converted into gross irrigation requirements to take into consideration the efficiency of the irrigation systems utilized.

Crop Evapotranspiration

The rate of evapotranspiration (ET) can be computed with the Penman–Monteith (PM) equation:

$$\mathrm{ET} = \frac{1}{\lambda}\frac{\Delta(R_n - G) + \rho c_p(e_s - e_a)/r_a}{\Delta + \gamma(1 + r_s/r_a)} \quad [1]$$

where λ is the latent heat of vaporization (kilojoules per kilogram), $R_n - G$ is the net balance of energy available at the surface (kilojoules per square meter per second), $(e_s - e_a)$ represents the vapor pressure deficit (VPD) of air at the reference (weather measurement) height (kilopascals), ρ represents mean air density (kilograms per cubic meter), c_p represents specific heat of air at constant pressure (kilojoules per kilogram per degrees Celsius), Δ represents the slope of the saturation vapor pressure–temperature relationship at mean air temperature (kilopascals per degrees Celsius), γ is the psychometric constant (kilopascals per degrees Celsius), r_s is the bulk surface resistance (seconds per meter), and r_a is the aerodynamic resistance (seconds per meter).

The PM eqn [1] can be utilized for the direct calculation of $\mathrm{ET_c}$ because the surface and aerodynamic resistances are crop-specific. However, data for these crop characteristics are scarce for most crops.

The transfer of heat and vapor from the evaporative surface into the air in the turbulent layer above a canopy is determined by the aerodynamic resistance r_a (seconds per meter) between the surface and the reference level above the canopy:

$$r_a = \frac{\ln\left(\frac{z_m - d}{z_{om}}\right)\ln\left(\frac{z_h - d}{z_{oh}}\right)}{k^2 u_z} \quad [2]$$

where z_m is the height of wind velocity measurements (meters), z_h is the height of air temperature and humidity measurements (meters), d is the zero-plane displacement height (meters), z_{om} is the roughness length relative to momentum transfer (meters), z_{oh} is the roughness length relative to heat and vapor transfer (meters), u_z is the wind velocity at height z_m (meters per second), and k is the von Karman constant (0.41).

Equation [2] assumes that the evaporative surface may be represented as a 'big leaf' inside the canopy. However, exchanges in the top layer of the canopy between heights $d + z_{om}$ and the crop height h (meters) are important as sources of vapor fluxes. Adopting $d + z_{om}$ as the level of the evaporative surface can lead to overestimation of r_a and underestimation of r_s. Thus, an alternative r_a can be computed from the top of the canopy:

$$r_a = \frac{\ln\left(\frac{z_m - d}{z_{om}}\right)\ln\left(\frac{z_h - d}{h - d}\right)}{k^2 u_z} \quad [3]$$

Both parameters d and z_{om} depend upon the crop height, h, and canopy architecture. Information exists relating d and z_{om} to h. Most of these relationships are crop-specific. More general functions also consider the leaf area index (LAI) or the plant area index (Table 1).

The height z_{oh} is estimated as a fraction of z_{om}, commonly $z_{oh} = 0.1\ z_{om}$ for short and fully developed canopies. The factor 0.2 is often preferred for tall and partial-cover crops. However, there is a relatively small impact on ET calculations from selecting a z_{oh}:z_{om} ratio between 0.1 and 0.2.

The surface resistance, r_s (seconds per meter), for full-cover canopies is often expressed by:

$$r_s = r_l/\mathrm{LAI_{eff}} \quad [4]$$

where r_l is the bulk stomatal resistance of a well-illuminated leaf (seconds per meter), and $\mathrm{LAI_{eff}}$ is the effective leaf area index (dimensionless), usually taken as 0.5 LAI. r_l usually increases as a crop matures and begins to ripen. Typical values for r_l and r_s are listed in Table 2. The use of these equations for prediction of crop water requirements is difficult due to differences among varieties and crop-management practices. Information on stomatal conductance or stomatal resistance available in the literature is mainly

Table 1 Relationships of d and z_{om} with crop height (h) and leaf area index (LAI)

d	z_{om}	*Comment*
0.7 h	0.29 h	Theoretical study
0.6–0.7 h	0.12–0.14 h	
0.7 h	0.1 h	Complete cover crops
0.79 h		Cotton
(0.78 ± 0.04) h	(0.041 ± 0.002) h	Potato
0.75 h	0.065 h	Beans
0.644 h	0.034–0.092 h	Peas
(0.66 ± 0.04) h	(0.077 ± 0.04) h	Beans
0.4–0.63 h	0.03–0.13 h	Soybean
0.54–0.604 h	0.076–0.0964 h	Sorghum
0.56 h	0.06 h	Soybean
0.56 h	0.05 h	Wheat
0.56 h	0.11 h	Grass
1.04 $h^{0.88}$	0.062 $h^{1.108}$	
1.04 $h^{0.88}$	0.062 $h^{1.08}$	Rice, maize
	0.025 $h^{1.1}$	Maize, sugarcane
	0.105 h	Barley, potato, alfalfa
0.70 $h^{0.979}$	0.131 $h^{0.997}$	
$h[1 - 2/\mathrm{LAI}(1 - e^{-\mathrm{LAI}/2})]$	$he^{-\mathrm{LAI}/2}(1 - e^{-\mathrm{LAI}/2})$	

Adapted from Alves I (1995) *Modelação da Evapotranspiração Cultural. Resistências Aerodinâmica e do Coberto.* [Modeling crop evapotranspiration. Aerodynamic and surface resistances.] PhD thesis. Lisbon: instituto Superior de Agronomia.

Table 2 Typical values of the leaf resistance per unit leaf area (r_l) and bulk stomatal resistance (r_s) for several canopy types. Parameters $r_{l_{min}}$ and $r_{s_{min}}$ are minimum daytime values when all environmental variables are optimum

Canopy type	r_l (s m^{-1})	$r_{l_{min}}$ (s m^{-1})	r_s (s m^{-1})	$r_{s_{min}}$ (s m^{-1})
Tropical forest	–	–	125–150	50
Deciduous forest	–	45–150	70–160	50–60
Aspen	400	–	–	–
Eucalyptus	200–400	–	–	–
Maple	400–700	250	–	–
Coniferous	–	–	70–150	30–60
Crops, general	50–320	–	30–130	20–150
Grain sorghum	200	150–200	100–140[a]	–
Snapbeans	–	130	–	–
Soybean	120	–	–	50
Maize	160	70	80	25–40
Barley	150–250	–	45–70	40
Wheat	–	50	–	30
Alfalfa	80	50	40	–
Sugar beet	100	50	–	–
Clipped grass (0.15 m)	100–150	–	80–120	70
Clipped and irrigated grass (0.10–0.12 m)	75	40	40–60	25–30

[a]LAI = 2.

Adapted from Allen RG, Pereira LS, Raes D, and Smith M (1998) *Crop Evapotranspiration. Guidelines for Computing Crop Water Requirements*. FAO Irrigation and Drainage Paper 56. Rome, Italy: Food and Agriculture Organization of the UN.

oriented to physiological or ecophysiological studies, rather than practical agricultural management. Information on bulk stomatal resistances is scarce.

Resistances r_l and r_s are influenced by climate and water availability: r_s increases when soil water availability limits ET, the VPD increases, and r_a is high. r_s decreases when energy available at the surface increases. In general, r_s varies according to:

$$r_s = r_a\left(\frac{\Delta}{\gamma}\beta - 1\right) + (1+\beta)\frac{\rho c_p \mathrm{VPD}}{\gamma(R_n - G)} \qquad [5]$$

where β is the Bowen ratio (the ratio between the sensible and latent heat fluxes). In eqn [5], β plays the role of a water-stress indicator. This equation shows that weather variables interact and their influences are interdependent, thus adding to the difficulties in appropriately selecting r_s.

These difficulties create challenges in applying the PM equation or other 'multilayer' resistance equations to estimate ET from agricultural crop canopies. Current research work is focused on improving our ability to apply the PM equation or multilayer ET models to specific agricultural crops; this work often utilizes relatively complex computer models. Meanwhile, the PM equation is used to compute the reference evapotranspiration and to determine ET_c with crop coefficients.

Crop Coefficients

ET_c can be calculated by multiplying the reference evapotranspiration, ET_o (millimeters per day), by a dimensionless crop coefficient, K_c:

$$ET_c = K_c ET_o$$

The reference crop is a hypothetical crop with an assumed height of 0.12 m, having a surface resistance of 70 s m^{-1} and an albedo of 0.23, closely resembling an extensive surface of green grass of uniform height, actively growing and adequately watered. ET_o can then be computed easily with the PM eqn [1], since the aerodynamic and surface-resistance terms can be parameterized, resulting in the United Nations Food and Agriculture Organization (FAO)–Penman–Monteith (FAO-PM) equation:

$$ET_o = \frac{0.408\Delta(R_n - G) + \gamma\dfrac{900}{T+273}u_2(e_s - e_a)}{\Delta + \gamma(1 + 0.34u_2)} \qquad [7]$$

where, in addition to variables defined for eqn [1], T is mean daily air temperature (degrees Celsius) and u_2 is wind speed (meters per second), both at 2 m height. In this equation ET_o is in mm per day and $R_n - G$ in mJ m^{-2} day^{-1}.

The reference crop – corresponding to a living, agricultural crop (i.e., a cold-season clipped grass) – incorporates the majority of the weather effects into ET_o estimates. Therefore, since ET_o represents an index of climatic demand on evaporation, the K_c varies predominantly with the specific crop characteristics and little with climate. This enables the transfer of standard values for K_c between locations and climates.

K_c represents an integration of the effects of three primary characteristics that distinguish the crop from the reference: crop height (affecting roughness and aerodynamic resistance); crop–soil surface resistance

(related to leaf area, fraction of ground covered by vegetation, leaf age and condition, degree of stomatal control, and soil surface wetness); and albedo of the crop–soil surface (influenced by the fraction of ground covered by vegetation and soil surface wetness).

Two K_c approaches are considered. The first uses a time-averaged K_c to include multiday effects of evaporation from the soil. The second concerns the basal crop coefficient and a separate calculation of evaporation from the soil.

The crop coefficient curve represents the changes in K_c over the length of the growing season (Figure 1). Its shape relates to changes in the vegetation and ground cover during plant development and maturation that affect the ratio ET_c:ET_o. Shortly after planting of annuals, or after the initiation of new leaves for perennials, the value for K_c is often small. The K_c increases from that initial value, $K_{c_{ini}}$, at the beginning of rapid plant development and reaches a maximum, $K_{c_{mid}}$, at the time of maximum or near-maximum plant development, the midseason period. During the late-season period, as leaves begin to senesce, the K_c begins to decrease until it reaches a lower value, $K_{c_{end}}$, at the end of the growing period.

The form for the equation used in the dual K_c approach is:

$$K_c = K_s K_{cb} + K_e \quad [8]$$

where K_s is the stress-reduction coefficient (0–1), K_{cb} is the basal crop coefficient (0–~1.4), and K_e is the soil-water evaporation coefficient (0–~1.4). K_{cb} represents the ratio ET_c:ET_o when the soil surface layer is dry but the average soil-water content of the root zone is adequate to sustain full plant transpiration, thus representing the baseline potential K_c in the absence of evaporation from the soil (Figure 2). K_s reduces the value of K_{cb} when the soil-water content is not adequate.

Because eqn [8] requires the calculation of a daily soil-water balance for the surface soil layer, a simplification is required for routine application. The time-averaged K_c is then adopted:

$$K_c = (K_{cb} + K_e) \quad [9]$$

where $(K_{cb} + K_e)$ represents the sum of the basal K_{cb} and time-averaged effects of evaporation from the soil, K_e. Typical shapes for the K_{cb}, K_e, and $K_{cb} + K_e$ curves are shown in Figure 2. When summed, the values for K_{cb} and for K_e represent the total crop coefficient, K_c. The time-averaged K_c is used for planning, irrigation system design, and typical irrigation management. The dual K_c is best where effects of day-to-day variation in soil surface wetness are important to estimate the resulting impacts on daily ET_c, soil moisture profile, and deep percolation.

The Single-Crop Coefficient Approach

A simple procedure may be used to construct the K_c curve (Figure 3):

1. Divide the growing period into four general growth stages that describe crop phenology or development, and determine the lengths (days) of these stages. The four crop growth periods are:
   a. Initial: for annual crops, duration is from planting date to approximately 10% ground

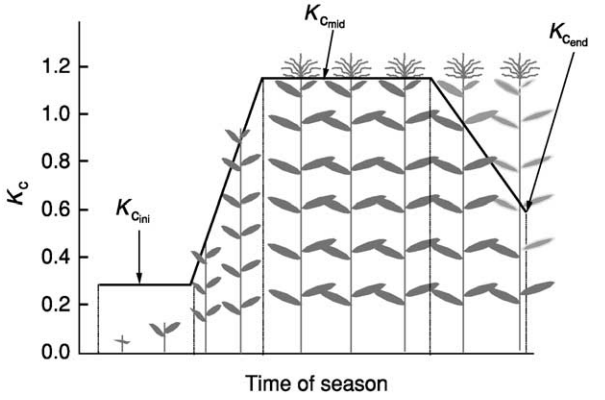


Figure 1 Generalized crop coefficient curve (K_c) during a growing season. Reproduced from Allen RG, Pereira LS, Raes D, and Smith M (1998) *Crop Evapotranspiration. Guidelines for Computing Crop Water Requirements*. FAO Irrigation and Drainage Paper 56. Rome, Italy: Food and Agriculture Organization of the UN.

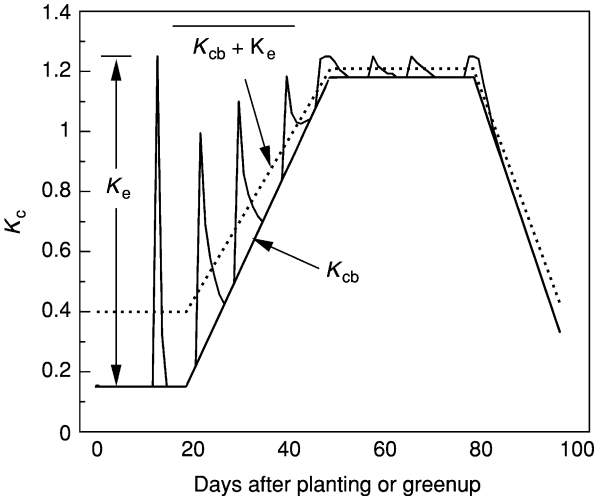


Figure 2 Crop coefficient (K_c) definitions showing the basal K_{cb}, soil evaporation K_c, and time-averaged $K_{cb} + K_e$ values. Reproduced from Allen RG, Pereira LS, Raes D, and Smith M (1998) *Crop Evapotranspiration. Guidelines for Computing Crop Water Requirements*. FAO Irrigation and Drainage Paper 56. Rome, Italy: Food and Agriculture Organization of the UN.

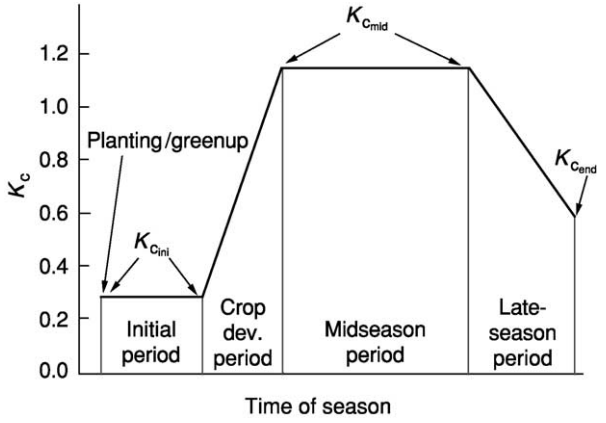


Figure 3 Crop coefficient curve (K_c) and stage definitions. Reproduced from Allen RG, Pereira LS, Raes D, and Smith M (1998) *Crop Evapotranspiration. Guidelines for Computing Crop Water Requirements*. FAO Irrigation and Drainage Paper 56. Rome, Italy: Food and Agriculture Organization of the UN.

cover. For perennials, the planting date is replaced by the 'greenup' date, when initiation of new leaves occurs;

b. Crop development: from 10% ground cover to effective full cover, which often occurs at the initiation of flowering or when LAI reaches 3;
c. Midseason: from effective cover to start of maturity, which is often indicated by the beginning of the aging, yellowing or senescence of leaves, leaf drop, or the browning of fruit;
d. Late season: from start of maturity to harvest or full senescence. For some perennial vegetation in frost-free climates, crops may grow year-round so that the date of termination may be taken as the same as the date of 'planting';

2. Identify the three K_c values that correspond to $K_{c_{ini}}$, $K_{c_{mid}}$, and $K_{c_{end}}$;
3. Connect straight-line segments through each of the four growth-stage periods.

The length of crop growth stages is crop-specific and changes duration with crop variety, planting date, cultivation practices, and weather conditions, mainly air temperature. The length of crop growth stages may be predicted using cumulative, degree-based equations or plant growth models.

The lengths of the initial and development periods may be relatively short for deciduous trees and shrubs that develop new leaves in the spring at relatively fast rates. The $K_{c_{ini}}$ should then reflect the ground condition prior to leaf initiation, including the amount of grass or weed cover, soil wetness, tree density, and mulch density. The length of the late-season period may be relatively short for vegetation killed by frost or for crops harvested before senescence. The value for $K_{c_{end}}$ should reflect the soil surface condition and that of the vegetation following plant death or harvest. Indicative lengths of growth stages are given in FAO guides. However, local observations or information should be used to incorporate effects of plant variety, climate, and cultural practices.

Values for $K_{c_{ini}}$, $K_{c_{mid}}$, and $K_{c_{end}}$ are listed in Table 3 for various agricultural crops. Usually there is close similarity in K_c within the same crop group, since the plant height, leaf area, ground coverage, and water management are usually similar.

The K_c values in Table 3 represent potential water use by healthy, disease-free, and densely planted stands of vegetation, with adequate levels of soil water. When stand density, height, and leaf area are less than that attained under perfect or normal conditions, K_c should be reduced by as much as 0.3–0.5 for poor stands, according to the amount of effective leaf area relative to healthy vegetation with normal planting densities.

The $K_{c_{ini}}$ values in Table 3 are only approximate, because they vary widely with soil wetting conditions and because ET during the initial stage for annual crops is predominantly in the form of evaporation from the soil. Therefore, estimates for $K_{c_{ini}}$ must consider the frequency of irrigation and rainfall that wet the soil surface.

Evaporation from bare soil, E_s (millimeters per day), can be characterized as occurring in two stages (Figure 4). During stage 1, termed the 'energy-limited' stage and having a duration t_1 (days), moisture is transported to the soil surface at a rate sufficient to supply the potential rate of evaporation, E_{so} (millimeters per day), which is governed by energy availability at the soil surface. E_{so} can be estimated from:

$$E_{so} = 1.15 ET_o \qquad [10]$$

where ET_o is averaged for the initial period.

Stage 2 is termed the 'soil water-limited' stage, where hydraulic transport of subsurface water to the soil surface is smaller, thus making $E_s < E_{so}$. Some of the evaporation occurs from below the soil surface, and energy is supplied by transport of heat into the soil profile. E_s decreases as soil moisture decreases and can be assumed to be linearly proportional to the depth of water remaining in the evaporation layer.

When the time interval (days) between two successive wettings is $t_w > t_1$, $K_{c_{ini}}$ is approached as:

$$K_{c_{ini}} = \frac{TEW - (TEW - REW)\exp\left(\frac{-(t_w - t_1)E_{so}\left[1 + \frac{REW}{TEW-REW}\right]}{TEW}\right)}{t_w ET_o} \qquad [11]$$

Table 3 Single (time-averaged) crop coefficients (K_c) and basal crop coefficient (K_{cb}), for nonstressed, well-managed crops in subhumid climates ($RH_{min} \approx 45\%$, $u_2 \approx 2\ m\ s^{-1}$) for use with the FAO Penman–Monteith ET_o, and indicative mean maximum plant heights (h), maximum root depths ($z_{r_{max}}$), and depletion fractions for no stress (p)

Crop	$K_{c_{ini}}$ [a]	$K_{c_{mid}}$	$K_{c_{end}}$	$K_{cb_{ini}}$	$K_{cb_{mid}}$	$K_{cb_{end}}$	*Maximum crop height* h *(m)*	*Maximum root depth*[b] $Z_{r_{max}}$ *(m)*	*Depletion fraction*[c] *(for ET 5 mm day*$^{-1}$*)*p
Vegetables	0.7	1.05	0.95	0.15	0.95	0.85			
Cabbage		1.05	0.95		0.95	0.85	0.4	0.5–0.8	0.45
Carrots		1.05	0.95		0.95	0.85	0.3	0.5–1.0	0.35
Lettuce		1.00	0.95		0.90	0.90	0.3	0.3–0.5	0.30
Onions, dry		1.05	0.75		0.95	0.65	0.4	0.3–0.6	0.30
Onions, green		1.00	1.00		0.90	0.90	0.3	0.3–0.6	0.30
Tomato		1.15[d]	0.70–0.90		1.10[d]	0.60–0.80	0.6	0.7–1.5	0.40
Cucumber, fresh market	0.6	1.00[d]	0.75		0.95[d]	0.70	0.3	0.7–1.2	0.50
Pumpkin, winter squash		1.00	0.80		0.95	0.70	0.4	1.0–1.5	0.35
Watermelon	0.4	1.00	0.75		0.95	0.70	0.4	0.8–1.5	0.40
Roots and tubers	0.5	1.10	0.95	0.15	1.00	0.85			
Cassava, year 1	0.3	0.80	0.30		0.70	0.20	1.0	0.5–0.8	0.35
Cassava, year 2	0.3	1.10	0.50		1.00	0.45	1.5	0.7–1.0	0.40
Potato		1.15	0.75–0.40		1.10	0.65–0.30	0.6	0.4–0.6	0.35
Sugar beet	0.35	1.20	0.70[e]		1.15	0.50[e]	0.5	0.7–1.2	0.55
Legumes (Leguminosae)	0.4	1.15	0.55	0.15	1.10	0.50			
Beans, green	0.5	1.05[d]	0.90		1.00[d]	0.80	0.4	0.5–0.7	0.45
Beans, dry and pulses	0.4	1.15[d]	0.35		1.10[d]	0.25	0.4	0.6–0.9	0.45
Garbanzo	0.4	1.15	0.35		1.05	0.25	0.8	0.6–1.0	0.45
Soybeans		1.15	0.50		1.10	0.30	0.5–1.0	0.6–1.3	0.50
Perennial vegetables (with winter dormancy and initially bare or mulched soil)	0.5	1.00	0.80						
Asparagus	0.5	0.95	0.30	0.15	0.90	0.20	0.2–0.8	1.2–1.8	0.45
Strawberries	0.40	0.85	0.75	0.30	0.80	0.70	0.2	0.2–0.3	0.20
Fiber crops	0.35			0.15					
Cotton		1.15–1.20	0.70–0.50		1.10–1.15	0.50–0.40	1.2–1.5	1.0–1.7	0.65
Oil crops	0.35	1.15	0.35	0.15	1.10	0.25			
Safflower		1.0–1.15[f]	0.25		0.95–1.10[f]	0.20	0.8	1.0–2.0	0.60
Sunflower		1.0–1.15[f]	0.35		0.95–1.10[f]	0.25	2.0	0.8–1.5	0.45
Cereals	0.3	1.15	0.4	0.15	1.10	0.25			
Spring wheat		1.15	0.25–0.40		1.10	0.15–0.30	1	1.0–1.5	0.55
Winter wheat, frozen soils	0.4	1.15	0.25–0.40	0.15–0.50	1.10	0.15–0.30	1	1.5–1.8	0.55
Winter wheat, nonfrozen soils	0.7	1.15	0.25–0.40						
Maize, field (grain; field corn)		1.20	0.60,0.35	0.15	1.15	0.50,0.15	2	1.0–1.7	0.55
Rice	1.05	1.20	0.90–0.60	1.00	1.15	0.70–0.45	1	0.5–1.0	0.20[g]

(*Continued*)

Table 3 (*Continued*)

Crop	$K_{c_{ini}}$ [a]	$K_{c_{mid}}$	$K_{c_{end}}$	$K_{cb_{ini}}$	$K_{cb_{mid}}$	$K_{cb_{end}}$	*Maximum crop height* h *(m)*	*Maximum root depth*[b] $Z_{r_{max}}$ *(m)*	*Depletion fraction*[c] *(for ET 5 mm day*$^{-1}$*)*p
Forages									
Alfalfa hay, averaged cutting effects	0.40	0.95	0.90				0.7		
Alfalfa hay, individual cutting periods	0.40[h]	1.20[h]	1.15[h]	0.30[h]	1.15[h]	1.10[h]	0.7	1.0–2.0	0.55
Grazing pasture, rotated grazing	0.40	0.85–1.05	0.85	0.30	0.80–1.00	0.80	0.15–0.30	0.5–1.5	0.60
Grazing pasture, extensive grazing	0.30	0.75	0.75	0.30	0.70	0.70	0.10	0.5–1.5	0.60
Turf grass, cool season	0.90	0.95	0.95	0.85	0.90	0.90	0.10	0.5–1.0	0.40
Turf grass, warm season	0.80	0.85	0.85	0.75	0.80	0.80	0.10	0.5–1.0	0.50
Sugar cane	0.40	1.25	0.75	0.15	1.20	0.70	3	1.2–2.0	0.65
Tropical fruits and trees									
Banana, first year	0.50	1.10	1.00	0.15	1.05	0.90	3	0.5–0.9	0.35
Banana, second year	1.00	1.20	1.10	0.60	1.10	1.05	4	0.5–0.9	0.35
Date palms	0.90	0.95	0.95	0.80	0.85	0.85	8	1.5–2.5	0.50
Pineapple, bare soil	0.50	0.30	0.30	0.15	0.25	0.25	0.6–1.2	0.3–0.6	0.50
Pineapple, with grass cover	0.50	0.50	0.50	0.30	0.45	0.45	0.6–1.2		
Grapes and berries									
Berries (bushes)	0.30	1.05	0.50	0.20	1.00	0.40	1.5	0.6–1.2	0.50
Grapes, table or raisin	0.30	0.85	0.45	0.15	0.80	0.40	2	1.0–2.0	0.35
Grapes, wine	0.30	0.70	0.45	0.15	0.65	0.40	1.5–2	1.0–2.0	0.45
Fruit trees									
Apples, cherries, pears								1.0–2.0	0.50
No ground cover, no frosts	0.60	0.95	0.75[i]	0.50	0.90	0.70[i]	4		
Active ground cover, no frosts	0.80	1.20	0.85[i]	0.75	1.15	0.80[i]	4		
Apricots, peaches, stone fruit								1.0–2.0	0.50
No ground cover, no frosts	0.55	0.90	0.65[i]	0.45	0.85	0.60[i]	3		
Active ground cover, no frosts	0.80	1.15	0.85[i]	0.75	1.10	0.80[i]	3		
Citrus, no ground cover									
70% canopy	0.70	0.65	0.70	0.65	0.60	0.65	4	1.2–1.5	0.50
50% canopy	0.65	0.60	0.65	0.60	0.55	0.60	3	1.1–1.5	0.50
20% canopy	0.50	0.45	0.55	0.45	0.40	0.50	2	0.8–1.1	0.50
Olives (40–60% ground coverage by canopy)[j]	0.65	0.70–0.45	0.70–0.65	0.55	0.65–0.40	0.65–0.60	3–5	1.2–1.7	0.65

Reproduced from Allen RG, Pereira LS, Raes D, and Smith M (1998) *Crop Evapotranspiration. Guidelines for Computing Crop Water Requirements.* FAO Irrigation and Drainage Paper 56. Rome, Italy: Food and Agriculture Organization of the UN.

[a] $K_{c_{ini}}$ in the table is only indicative. The procedure relative to eqns [11] and [12] should be used.

[b] The larger values for $z_{r_{max}}$ correspond to rainfed conditions.

[c] The tabulated value for p must be corrected when $ET_c \neq 5\,mm\,day^{-1}$ using $p = p_{table} + 0.04\,(5 - ET_c)$.

[d] For crops grown on stalks, $K_{c_{mid}}$ and $K_{cb_{mid}}$ have to be increased by 0.05 to 0.10 and h should be increased too.

[e] The $K_{c_{end}}$ and $K_{cb_{end}}$ are increased by 0.1 to 0.3 when irrigation or significant rain occurs during the last month.

[f] The smaller value is for rainfed conditions when plant density is smaller than under irrigation.

[g] This p-value represents the depletion fraction below soil moisture at saturation.

[h] These K_c coefficients for hay crops represent immediately after cutting, at full cover, and immediately before cutting, respectively. The growing season is described as a series of individual cutting periods.

[i] These $K_{c_{end}}$ (and $K_{cb_{end}}$) values represent K_c prior to leaf drop. After leaf drop, $K_{c_{end}} \approx 0.20$ (0.15) for bare, dry soil or dead ground cover and $K_{c_{end}} \approx 0.50$–0.80 (0.45–0.75) for actively growing ground cover.

[j] The larger values correspond to well-watered conditions and the smaller concern controlled, stressed conditions.

where REW is the readily evaporable water, corresponding to the depth of evaporation when stage 1 drying is complete (millimeters), TEW is the total evaporable water, i.e., the maximum evaporation depth when soil evaporation effectively ceases. From the concept of stage 1 drying results $t_1 = \mathrm{REW}/E_{so}$.

When $t_w < t_1$, the entire process resides within stage 1, then:

$$K_{c_{ini}} = E_{so}/\mathrm{ET_o} \qquad [12]$$

Where furrow or trickle irrigation is practiced, and only a portion of the soil surface is wetted, $K_{c_{ini}}$ in eqns [11] and [12] should be reduced in proportion to the average fraction of wetted soil surface, f_w (ranging from 0, when no rain or irrigation occurs, to 1). Indicative values for f_w are shown in Table 4. The infiltration depth from irrigation, I_w (millimeters), should also be adjusted:

$$I_w = I/f_w \qquad [13]$$

where I is the total irrigation depth (millimeters).

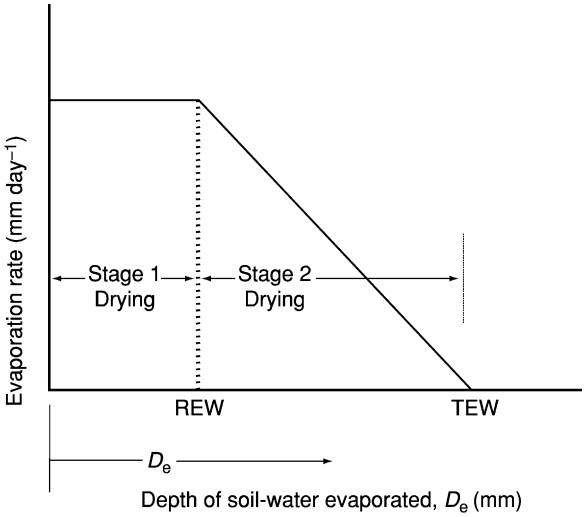


Figure 4 Two-stage model for soil evaporation. REW, readily evaporable water; TEW, total evaporable water. Reproduced from Allen RG, Pereira LS, Raes D, and Smith M (1998) *Crop Evapotranspiration. Guidelines for Computing Crop Water Requirements*. FAO Irrigation and Drainage Paper 56. Rome, Italy: Food and Agriculture Organization of the UN.

Table 4 Indicative values of average fraction of wetted soil surface, f_w

Irrigation method	f_w
Rain, sprinkling, basin, and border irrigation	1.0
Furrow irrigation	0.4–0.6
Irrigation with alternate furrows	0.3–0.4
Trickle irrigation	0.2–0.5

REW is higher for medium-textured soils and is lower for coarse soils. Maximum values for REW ($\mathrm{REW_{max}}$) may be predicted according to soil texture:

$$\mathrm{REW_{max}} = 20 - 0.15(\mathrm{Sa}) \quad \text{for Sa} > 80\%$$

$$\mathrm{REW_{max}} = 11 - 0.06(\mathrm{Cl}) \quad \text{for Cl} > 50\%$$

$$\mathrm{REW_{max}} = 8 + 0.08(\mathrm{Cl}) \quad \text{for Sa} < 80\%, \mathrm{Cl} > 50\% \qquad [14]$$

where Sa and Cl are the fractions of sand and clay in the soil (percentage).

The TEW value is governed by the depth of soil contributing to evaporation, z_e (100–150 mm). The soil water-holding properties within this evaporative layer, the presence of a hydraulically limiting layer beneath it, the unsaturated hydraulic conductivity, the conduction of sensible heat into the soil, and any root extraction of water from the evaporative layer all influence TEW. An approximation to $\mathrm{TEW_{max}}$ is:

$$\mathrm{TEW_{max}} = z_e(\theta_{FC} - 0.5\theta_{WP}) \quad (\mathrm{ET_o} \geq 5\ \mathrm{mm\ day^{-1}}) \qquad [15]$$

$$\mathrm{TEW_{max}} = z_e(\theta_{FC} - 0.5\theta_{WP})\sqrt{\frac{\mathrm{ET_o}}{5}} \quad (\mathrm{ET_o} < 5\ \mathrm{mm\ day^{-1}}) \qquad [16]$$

where θ_{FC} and θ_{WP} are the soil-water content at field capacity and wilting point (millimeters per millimeter). Typical values for θ_{FC} and θ_{WP} are given in Table 5.

The average total water available for evaporation, D_a (millimeters), during each drying cycle is computed from the average depth added to the evaporative layer at each wetting:

$$D_a = P_{mean} + W_{ini}/n_w \qquad [17]$$

where W_{ini} is the available soil water (millimeters) in the evaporation layer at the time of planting, n_w is the number of wetting events, and P_{mean} is the average depth (millimeters) of water added to the evaporating layer at each wetting event. P_{mean} can be obtained with:

$$P_{mean} = \left(\sum P_n + \sum I_w\right)/n_w \qquad [18]$$

but where each value of P_n and I_w must be limited to $P_n \leq \mathrm{TEW_{max}}$ and $I_w \leq \mathrm{TEW_{max}}$. The values for TEW and REW in eqn [11] are calculated from $\mathrm{TEW_{max}}$ and $\mathrm{REW_{max}}$ as:

$$\mathrm{TEW} = \min(\mathrm{TEW_{max}}, D_a) \qquad [19]$$

and:

Table 5 Typical soil water characteristics for different soil types

Soil type (USA soil texture classification)	Soil water characteristics			Amount of water that can be depleted by evaporation for ze = 0.10 m	
	θ_{FC} m^3/m^3	θ_{WP} m^3/m^3	$(\theta_{FC} - \theta_{WP})$ m^3/m^3	Stage 1 REW mm	Stages 1 and 2 TEW mm
Sand	0.07–0.17	0.02–0.07	0.05–0.11	2–7	6–12
Loamy sand	0.11–0.19	0.03–0.10	0.06–0.12	4–8	9–14
Sandy loam	0.18–0.28	0.06–0.16	0.11–0.15	6–10	15–20
Loam	0.20–0.30	0.07–0.17	0.13–0.18	8–10	16–22
Silt loam	0.22–0.36	0.09–0.21	0.13–0.19	8–11	18–25
Silt	0.28–0.36	0.12–0.22	0.16–0.20	8–11	22–26
Silt clay loam	0.30–0.37	0.17–0.24	0.13–0.18	8–11	22–27
Silty clay	0.30–0.42	0.17–0.29	0.13–0.19	8–12	22–28
Clay	0.32–0.40	0.20–0.24	0.12–0.20	8–12	22–29

(TEW = qFC − 0.5 qWP) ze.

Reproduced from Allen RG, Pereira LS, Raes D, and Smith M (1998) *Crop Evapotranspiration. Guidelines for Computing Crop Water Requirements*. FAO Irrigation and Drainage Paper 56. Rome, Italy: Food and Agriculture Organization of the UN.

$$\mathrm{REW} = \mathrm{REW}_{\max}\left[\min\left(\frac{D_a}{\mathrm{TEW}_{\max}}, 1\right)\right] \qquad [20]$$

When eqns 14–16 are used, appropriate checking of results is required.

K_c Adjustment for Climate

The $K_{c_{mid}}$ and $K_{c_{end}}$ values in Table 3 represent $K_{cb} + K_e$ for irrigation management and precipitation frequencies typical of a subhumid climate where $\mathrm{RH}_{\min} = 45\%$ and $u_2 = 2\,\mathrm{m\,s^{-1}}$.

Under humid and calm conditions, the K_c for 'full-cover' agricultural crops generally do not exceed 1.0 by more than about 0.05, because 'full-cover' agricultural crops and the reference crop behave similarly regarding absorption of short-wave radiation, the primary energy source for evaporation under humid and calm conditions. Because the VPD is small under humid conditions, differences in ET caused by differences in r_a between the agricultural and the reference crop are also small, especially with low-to-moderate wind speeds. Thus the values of K_c are less dependent on differences between the aerodynamic components of $\mathrm{ET_c}$ and $\mathrm{ET_o}$. On the contrary, under arid conditions, the effect of differences in r_a between the agricultural and the reference crop on $\mathrm{ET_c}$ become more pronounced because the VPD is often large. Hence, K_c will be larger under arid conditions, mainly for tall crops. Because the $K_{c_{mid}}$ and $K_{c_{end}}$ in Table 3 represent conditions where $\mathrm{RH}_{\min} \approx 45\%$ and $u_2 \approx 2\,\mathrm{m\,s^{-1}}$, when climatic conditions deviate from these values, the tabled values need to be adjusted:

$$K_c = (K_c)_{tab} + [0.04(u_2 - 2) - 0.004(\mathrm{RH}_{\min} - 45)]\left(\frac{h}{3}\right)^{0.3} \qquad [21]$$

where $(K_c)_{tab}$ represents the $K_{c_{mid}}$ or $K_{c_{end}}$ taken from Table 3, u_2 is the average daily wind speed at 2 m height (meters per second), $\mathrm{RH}_{\min}$ is the average daily minimum relative humidity (percentage), and h is the average plant height (meters), all averages referring to the midseason or the late-season period. Indicative values for h are listed in Table 3, although data from field observations are more accurate. When crops are allowed to senesce and dry in the field ($K_{c_{end}} < 0.45$), no adjustment is necessary.

K_c Adjustment for Nonpristine Conditions

The values of K_c in Table 3 reflect typical crop- and water-management practices. When local water management and harvest timing deviate from those that are typical, adjustments are made to $K_{c_{mid}}$ and $K_{c_{end}}$.

When stand density, height, or leaf area of the crop are less than those attained under appropriate crop and irrigation management conditions, the value for K_c is reduced by 0.1–0.5, according to the amount of effective (green) leaf area relative to that of healthy vegetation having normal plant density:

$$K_c = K_{c_{table}} - A_{cm} \qquad [22]$$

where A_{cm} is the adjustment factor (0–0.5) that can be approximated through a green-cover ratio of the type:

$$A_{cm} = 1 - \mathrm{LAI_{actual}}/\mathrm{LAI_{normal}} \qquad [23]$$

where the LAI refers to the midseason period. Other procedures, including remote sensing, may be used to estimate the K_c values for nonpristine conditions.

The Dual-Crop Coefficient Approach

The basal crop coefficient, K_{cb} (eqn [8]), represents primarily the transpiration component of ET. Its use

provides for separate adjustment for evaporation from wet soil immediately following rain or irrigation events. This results in more accurate estimates of ET_c when computed on a daily basis. Recommended values for K_{cb} are listed in Table 3: these must be adjusted for climate using a similar equation to eqn [21].

The computation of the soil-water evaporation coefficient K_e is based on the fact that evaporation from the soil is governed by the amount of energy available at the soil surface, which depends, in turn, on the portion of total energy that has been consumed by plant transpiration. K_e decays after a wetting depending on the cumulative amount of water evaporated from the surface soil layer. Thus K_e can be calculated from:

$$K_e = K_r(K_{c_{max}} - K_{cb}) \quad [24]$$

where K_r is the evaporation-reduction coefficient (0–1) and $K_{c_{max}}$ is the maximum value for K_c after rain or irrigation. However, K_e is limited by the fraction of wetted soil exposed to sunlight, f_{ew} (0.01–1):

$$K_e \le f_{ew}K_{c_{max}} \quad [25]$$

$K_{c_{max}}$ represents an upper limit on ET from any cropped surface (1.05–1.35). It changes with climate similarly to K_c (eqn [21]), thus:

$$K_{c_{max}} = \max\left\{1.2 + [0.04(u_2 - 2) - 0.004(RH_{min} - 45)](h/3)^{0.3}\right\}, K_{cb} + 0.05\Big\langle \quad [26]$$

u_2, RH_{min}, and h may refer to the midseason period or, when more detailed computations are applied, be averaged for shorter periods (e.g., 5 days). h for the initial period can be considered the same as for the grass reference crop ($h = 0.12$ m).

The method used to estimate evaporation from soil is similar to the one used to compute $K_{c_{ini}}$, where the evaporation rate is at the maximum rate until the depth of water evaporated, D_e (millimeters), equals REW (Figure 4). When $D_e > REW$, the evaporation process is at stage 2, and its rate decreases in proportion to the remaining water. Therefore, K_r (eqn [24]) can be calculated as:

$$K_r = 1 \quad \text{for } D_e \le REW \quad [27a]$$

$$K_r = \frac{TEW - D_e}{TEW - REW} \quad \text{for } D_e > REW \quad [27b]$$

REW and TEW can be estimated with eqns [14]–[20]. D_e, the current depth of water depleted from the f_{ew} fraction of wetted soil exposed to sunlight, is computed from the daily water balance of the upper 100–200 mm of soil:

$$D_{e_i} = D_{e_{i-1}} - (P_i - RO_i) - \frac{I_i}{f_w} + \left(\frac{K_e ET_o}{f_{ew}}\right)_i + T_{s_i} \quad [28]$$

limited to ($0 \le D_{e_i} \le TEW$), where the subscript i refers to the day of estimation, P_i is the precipitation (millimeters), RO_i is runoff (millimeters; $0 \le RO_i \le P_i$), I_i is the net irrigation depth (millimeters) that infiltrates the soil (eqn [13]), $(K_e\ ET_o/f_{ew})_i$ is the evaporation from the f_{ew} fraction of the exposed soil surface (millimeters), and T_{s_i} is the transpiration from the f_w fraction of the evaporating soil layer (millimeters). To initiate the water balance, $D_{e_i} = 0$ immediately following heavy rain or irrigation, or $D_{e_i} = TEW$ if a long time has passed since the last wetting. When $P_i < 0.2\ ET_o$, it can be ignored. For most applications, $RO_i = 0$ and, for the majority of crops, except for very shallow-rooted crops, T_{s_i} can be neglected.

When the complete soil surface is fully wetted (e.g., by precipitation or sprinkler irrigation), $f_{ew} = (1 - f_c)$, where f_c is the average fraction of ground covered by vegetation (0–0.99). For irrigation systems where only a fraction of the soil surface is wetted, f_{ew} is calculated as:

$$f_{ew} = \min(1 - f_c, f_w) \quad [29]$$

When not observed, f_c can be estimated daily from:

$$f_c = \left(\frac{K_{cb} - K_{c_{min}}}{K_{c_{max}} - K_{c_{min}}}\right)^{1+0.5h} \quad [30]$$

where $K_{c_{min}}$ is the minimum K_c for dry, bare soil (0.15–0.20). The exponent $1 + 0.5h$ represents the effect of plant height on soil-shading and in increasing the K_{cb} given a specific value for f_c. $(K_{cb} - K_{c_{min}}) \ge 0.01$ for numerical stability.

K_{cb} values are reduced when the soil water content in the root zone is too low to sustain transpiration at potential levels. The reduction is made through the water stress coefficient, K_s (0–1):

$$K_s = \frac{\theta - \theta_{WP}}{\theta_p - \theta_{WP}} \quad (\theta < \theta_p) \quad [31]$$

where θ is average soil-water content in the root zone (millimeters per millimeter) and θ_p is the threshold θ below which transpiration is decreased due to water stress (millimeters per millimeter). By definition, $K_s = 1.0$ for $\theta > \theta_p$. The threshold θ_p is:

$$\theta_p = (1 - p)(\theta_{FC} - \theta_{WP}) \quad [32]$$

where p is the depletion fraction for no stress (0–1). Indicative values can be found in Table 3. The determination of K_s requires a daily balance of soil-water content.

K_c for Nonpristine and Unknown Conditions

For vegetation where the K_c is not known, but where estimates of the fraction of ground surface covered by vegetation can be made, $K_{cb_{mid}}$ can be approximated as:

$$K_{cb_{mid}} = K_{c_{min}} + (K_{cb_{full}} - K_{c_{min}}) f_{c_{eff}}^{1/1+h} \quad [33]$$

where $f_{c_{eff}}$ is the effective fraction of ground covered by vegetation (0.01–1), and $K_{cb_{full}}$ is the maximum value for K_{cb} for vegetation having complete ground cover:

$$K_{cb_{full}} = \min[(1.0 + 0.1h), 1.2] + 0.04(u_2 - 2) - 0.004(RH_{min} - 45)\left(\frac{h}{3}\right)^{0.3} \quad [34]$$

For small, isolated stand sizes, $K_{cb_{full}}$ may need to be increased beyond the value given by eqn [34]. $K_{cb_{full}}$ may be reduced for vegetation that has a high degree of stomatal control.

Irrigation Water Requirements

The soil-water balance is calculated for the effective rooting depth as:

$$\theta_i = \theta_{i-1} + \frac{(P_i - RO_i) + I_{w_i} - ET_{c_i} - DP_i + GW_i}{1000 z_{r_i}} \quad [35]$$

where, in addition to the symbols used previously, DP_i represents deep percolation (millimeters), GW_i is groundwater contribution (millimeters), and z_{r_i} is the rooting depth (meters), all referred to day i. DP is often estimated as $DP_i = 0$ when $\theta_i \leq \theta_{FC}$ and $DP_i = 1000(\theta_i - \theta_{FC})\, z_{r_i}$ otherwise. GW is estimated from soil hydraulic properties and the water table depth. z_{r_i} can be predicted assuming a linear variation from planting to maximum rooting. Maximum root depths for most common crops are presented in Table 3.

The latest date for scheduling irrigation to avoid water stress is when $\theta_i = \theta_p$ (eqn [32]). However, irrigation is often scheduled when the 'management-allowed depletion' (MAD) is attained. Generally, MAD $< p$ when there is risk aversion or uncertainty, and MAD $> p$ when plant water stress is intentional. Then:

$$\theta_i = \theta_{MAD} = (1 - MAD)(\theta_{FC} - \theta_{WP}) + \theta_{WP} \quad [36]$$

The net irrigation depth to be applied will be:

$$I_{w_i} = 1000\, z_{r_i}(\theta_{FC} - \theta_i) \quad [37]$$

which, summed for the entire season, leads to the IWR:

Table 6 Indicative values of irrigation efficiencies

System	Efficiency (%)
Irrigation methods	
Surface irrigation, precision leveling	
Furrow	65–85
Border	70–85
Basin	70–90
Surface, traditional	
Furrow	40–70
Border	45–70
Basin	45–70
Basin, rice fields	25–50
Sprinkler	
Solid set	65–85
Hand-move lateral	65–80
Side-roll wheel move	65–80
Traveler sprinkler	55–70
Lateral move systems, center pivot	65–85
Microirrigation	
Trickle, ≥3 emitters per plant	85–95
Trickle, <3 emitters per plant	80–90
Bubblers and sprayers	85–95
Line-source emitters	70–90
Distribution and transport systems	
Pipe	95–100
Lined canals	60–90
Nonlined canals	55–85

$$IWR = \frac{ET_c - P_e - GW - \Delta S}{1 - LR} \quad [38]$$

where P_e is the effective precipitation (gross precipitation less all runoff and deep percolation), GW is groundwater contribution, ΔS is the change in soil-water storage in the root zone between planting and harvesting, and LR is the leaching requirement (the percentage of irrigation water that must pass through the root zone to keep the salinity of the soil below a specified value). The soil water balance is currently computed through crop-water simulation models, which allow the selection of the best irrigation scheduling alternatives.

The gross irrigation water requirement (GIWR) is computed as:

$$GIWR = \frac{IWR}{Eff} \quad [39]$$

where Eff is the efficiency of the irrigation system. Indicative values of the efficiencies are presented in Table 6.

Summary

This article summarizes the essential definitions and methodologies for estimating crop water and irrigation requirements. The concept of reference

evapotranspiration is assumed relative to a crop canopy such as grass but with constant crop characteristics. The hypotheses on which this approach is based are discussed relative to crop surface and aerodynamic resistances to heat and vapor fluxes. The crop evapotranspiration is defined using crop coefficients applied to the reference evapotranspiration, which reflect the canopy differences between the crop and the reference crop. Both time-averaged and dual-crop coefficients are explained, the first when the coefficients relative to crop transpiration and evaporation from the soil are summed and averaged for the crop-stage periods, the latter when a daily calculation of transpiration and evaporation coefficients is adopted. Finally, essential information on the soil water balance to estimate crop irrigation requirements is provided.

List of Technical Nomenclature

β	Bowen ratio (dimensionless)
γ	Psychrometric constant (kPa°C^{-1})
Δ	Slope of the saturation vapor pressure–temperature function (kPa°C^{-1})
ΔS	Change in soil-water storage (mm)
θ	Soil-water content (mm mm^{-1})
θ_{FC}	Soil-water content at field capacity (mm mm^{-1})
θ_{WP}	Soil-water content at wilting point (mm mm^{-1})
λ	Latent heat of vaporization (kJ kg^{-1})
ρ	Air density (kg m^{-3})
c_p	Specific heat of dry air (kJ kg^{-1} °C^{-1})
D_a	Average total water available for evaporation (mm)
D_e	Evaporated water (mm)
DP	Deep percolation (mm)
d	Zero-plane displacement height (m)
E_s	Evaporation from bare soil (mm day^{-1})
E_{so}	Potential rate of soil evaporation (mm day^{-1})
ET	Evapotranspiration (mm s^{-1})
ET_c	Crop evapotranspiration (mm day^{-1})
ET_o	Reference evapotranspiration (mm day^{-1})
$e_s - e_a$	Vapor pressure deficit (VPD) (kPa)
f_c	Fraction of ground covered by vegetation (dimensionless)
f_{ew}	Fraction of exposed wetted soil (dimensionless)
f_w	Fraction of wetted soil fraction (dimensionless)
GIWR	Gross irrigation water requirement (mm)
GW	Groundwater contribution (mm)
h	Crop height (m)
I	Total irrigation depth (mm)
IWR	Irrigation water requirement (mm)
I_w	Infiltration depth from irrigation (mm)
K_c	Crop coefficient (dimensionless)
K_{cb}	Basal crop coefficient (dimensionless)
K_e	Soil-water evaporation coefficient (dimensionless)
K_r	Evaporation reduction coefficient (dimensionless)
K_s	Stress-reduction coefficient (dimensionless)
k	Von Karman constant (dimensionless)
LAI	Leaf area index (dimensionless)
LR	Leaching requirement (mm)
MAD	Management-allowed depletion (dimensionless)
n_w	Number of wetting events (dimensionless)
P	Precipitation (mm)
P_n	Net precipitation (mm)
p	Depletion fraction for no stress (dimensionless)
REW	Readily evaporable water (mm)
RH	Relative humidity (%)
RO	Runoff (mm)
$R_n - G$	Net balance of energy at the surface (kJ m^{-2} s^{-1})
r_a	Aerodynamic resistance (s m^{-1})
r_l	Bulk stomatal resistance (s m^{-1})
r_s	Bulk surface resistance (s m^{-1})
TEW	Total evaporable water (mm)
T	Mean daily temperature (°C)

u	Wind velocity ($m\,s^{-1}$)
W	Soil water (mm)
z_e	Depth of soil contributing to evaporation (m)
z_h	Height of air temperature measurement (m)
z_m	Height of wind velocity measurement (m)
z_{oh}	Roughness length for heat (m)
z_{om}	Roughness length for momentum (m)
z_r	Rooting depth (m)

See also: **Evapotranspiration**; **Irrigation:** Methods; **Plant–Soil–Water Relations**

Further Reading

Allen RG, Pereira LS, Raes D, and Smith M (1998) *Crop Evapotranspiration. Guidelines for Computing Crop Water Requirements*. FAO Irrigation and Drainage Paper 56. Rome, Italy: Food and Agriculture Organization of the United Nations.

Pereira LS, van den Broek B, Kabat P, and Allen RG (eds) (1995) *Crop-Water Simulation Models in Practice*. Wageningen, Germany: Wageningen Press.

Tiercelin JR (ed.) (1998) *Traité d'Irrigation*. Paris, France: Lavoisier, Technique & Documentation.

van Lier HN, Pereira LS, and Steiner FR (eds) (1999) *CIGR Handbook of Agricultural Engineering, vol. I, Land and Water Engineering*. St. Joseph, MI: ASAE.

Wooton TP, Cecilio CB, Fowler LC, Hui SL, and Heggen RJ (eds) (1996) *Hydrology Handbook*. ASCE Manuals and Reports on Engineering Practice 28. New York: ASCE.

CROP-RESIDUE MANAGEMENT

D C Reicosky and A R Wilts, USDA Agricultural Research Service, Morris, MN, USA

Published by Elsevier Ltd.

Introduction

Soils, along with water, air and sun, are the major resources that sustain our food supply and terrestrial ecosystems. Soil organic matter is one of the primary contributors to soil quality. Crop residues are precursors to soil organic matter (SOM). The stems, leaves, chaff and husks that remain in the fields after crops are harvested for grain, seed or fiber play a critical role in soil quality and environmental issues since they are primary inputs of elemental carbon (C) into soil systems (Figure 1). Crop residues have been referred to as 'wastes' but as a natural and valuable resource are also considered to be 'potential black gold.' The negative connotation of 'residues' may refer to the remains after a part is taken or something leftover or useless. On the contrary, crop residues offer a large, but finite potential mechanism for C sequestration and nutrient cycling.

Crop residue management (CRM) is a widely-used cropland conservation practice for wind and water erosion control. Crop residue provides significant quantities of nutrients for crop production. In addition to affecting soil physical, chemical and biological functions and properties, crop residues can also affect water movement, infiltration, runoff and quality. However, the decomposition of crop residues can have both positive and negative effects on crop production and the environment. The presence of crop residues on the surface generally results in wetter and cooler conditions, thus favoring disease and pests, and also provides pathogens with an additional source of energy to multiply. Agricultural managers aim to increase the positive environmental effects of CRM since each practice has drawbacks. Ideally, improved management practices should enhance crop yields and have minimal adverse effects on the environment.

Crop management recommendations for maximum residue production require basic scientific research information regarding site-specific soils, crops and climate. Soil conservation and CRM research covers many aspects including the factors affecting residue decomposition, effects on erosion control, nutrient cycling and plant availability, disease control problems, weed control problems, alternate uses of excess residue, selection of plant varieties for conservation tillage systems, machinery requirements and control of the soil–water–temperature regime.

Conservation, Carbon Cycle, Soil Organic Matter and Carbon Sinks

Conservation of soil resources requires proper management of crop residues. The primary limiting factor for microbial growth in most soils is the C energy source. An abundance of C and other nutrients are returned to the soil through decomposition of crop

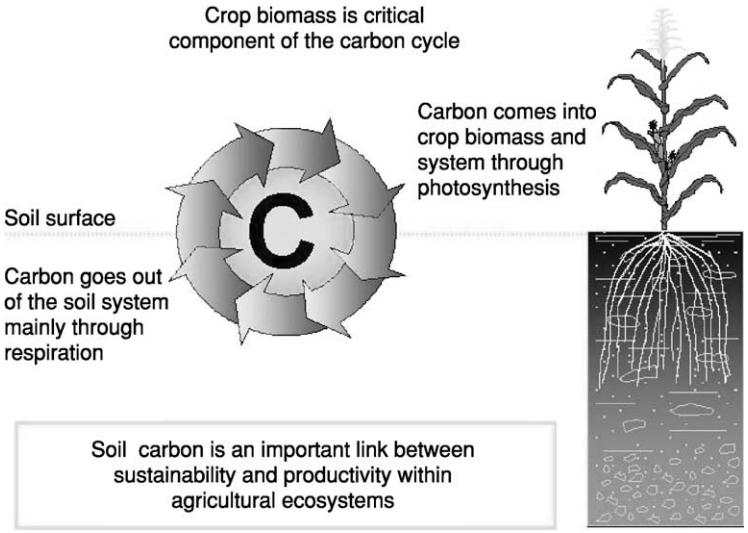


Figure 1 Schematic representing the role of crop biomass in the agricultural ecosystems carbon cycle.

residues and biological nutrient cycling. Since organic matter (OM) is known to maintain soil aggregate stability, the addition of crop residues often improves soil structure and aggregation. Crop residues and tillage management can also affect the leaching of the nutrients, which may pollute the groundwater or surface waters. Crop residue influences soil temperature primarily by insulating the soil surface from the sun's radiant energy. Studies have shown that soils high in SOM retain more moisture, especially when residues are retained on the soil surface as compared to when they are incorporated.

The recent interest in global climate change has prompted many to value C sources and sinks in agroecosystems. Soil C pools and fluxes in agroecosystems are influenced by a number of factors including amount and type of plant residue, temperature and precipitation and soil texture, pH and drainage. Carbon sequestration or storage in terrestrial ecosystems can be defined as the net removal of carbon dioxide (CO_2) from the atmosphere by crop photosynthesis into stable, long-lived pools of C. The soil organic carbon (SOC) pool is estimated to compose about two-thirds of the terrestrial biosphere C pool. Soil organic C storage in cropland soils is determined by tillage systems and the amount and placement of the crop residue. As grain and biomass yields increase and less intensive tillage systems are employed, farmers should gradually develop an enhanced C sink.

Crop Residues and Nutrient Cycling

The annual cycling of plant nutrients in the plant–soil ecosystem is essential to maintaining a productive agricultural system. The management of crop residues has important implications for the total

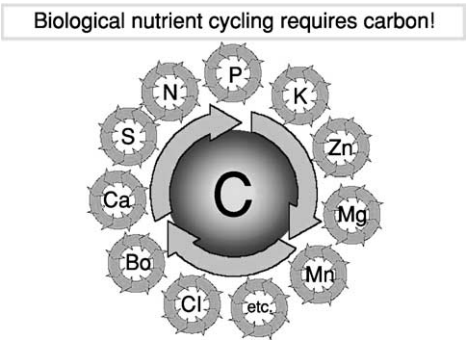


Figure 2 Soil carbon plays a critical role in biological nutrient cycling.

amounts of nutrients removed from and returned to the soil. Soil organic matter also improves the dynamics and bio-availability of main plant nutrient elements. The soil, water, and air, which are in contact with plants, contain various inorganic chemicals necessary for plant growth. Soil organic matter is the main determinant of biological activity because it is the primary energy source. The amount, diversity and activity of soil fauna and microorganisms are directly related to SOM content and quality.

Organic matter and the biological activity that it generates have a major influence on the physical and chemical properties of the soils. Each plant nutrient has its own C-dependent cycle (Figure 2) that controls its availability to the next generation of plants. Carbon compounds in the residue are the 'fuel' or energy sources for the soil microbes and fauna responsible for biological recycling of these inorganic plant nutrients. During microbial decomposition of crop residues, chemical elements are released into the immediate environment that may be utilized by living plants or other organisms. This constitutes the basic

framework of biological nutrient cycling in agricultural production systems. Carbon-enriched crop biomass becomes the primary food source for soil microorganisms and fauna and as a result 'nurtures' nutrient cycling.

Plant availability of nutrients (nitrogen (N), phosphorus (P), and potassium (K)) in crop residues is regulated largely by soil water, soil temperature, other soil physical and chemical properties, and soil and crop management practices. For N, activity of soil microorganisms is usually most important in determining the cycling and potential availability from crop residues; for P, both microbial activity and soil mineralogy are involved; and for K, mineralogy and soil water movement are important parameters. Management practices such as fertilization and the amounts of residue remaining after harvest determine the extent of cycling and plant availability of nutrients from crop residues. The shift from conventional to conservation tillage systems necessitates new research to determine the rate of cycling and plant nutrient availability.

Crop residues can be managed for increased OM levels, thereby sequestering C. The position and quantity of crop residue as well as N fertilization have variable influences on SOM. Several studies have indicated that when more crop residues were on or near the surface, as in no-till or reduced tillage systems, SOM content near the surface was increased, but when incorporated to depth by moldboard plow tillage, the quantity of crop residue had little or no influence on SOM. The physical incorporation and mixing of residue maximize soil-residue contact and result in rapid decomposition and C loss as CO_2.

The C:N ratio of the residue, an important key in soil management, also varies. Crop biomass is generally 40–50% C, but the nitrogen (N) content varies considerably within and among species. Thus, an adequate supply of N may be required to build SOM for crops with a high C:N ratio since C and N and their proportionality (i.e., C:N ratio) is relatively constant across a range of agricultural soils at about 10:1. For example, wheat straw has a C:N ratio around 90:1 and will require N addition or C loss to decompose to the soil equilibrium value of 10:1. Thus, soil C sequestration may be reduced when C and N inputs of the residue are out of balance.

Factors Controlling Residue Decomposition and Soil Quality

Soil microorganisms play a major role in the synthesis and degradation of crop residues into SOM. The decomposition rates of OM and crop residue depend primarily on chemical composition and on factors that affect the soil environment. Factors having the greatest effect on microbial growth and activity will have the greatest potential for altering the rate of residue decomposition in soil.

Soil factors that typically influence residue decomposition most include water, temperature, pH, aeration or oxygen supply and available nutrients. Residue factors include chemical composition of the residue, C:N ratio, age of plant material, particle size and the indigenous microflora. Additional factors that must be considered are the residue water content and the method of residue application to the soil (i.e., mixed with the soil, layered or banded in soil or left on the soil surface). Many of these factors are not independent and a change in one may affect a change in others.

Plant residue decomposition rates generally increase when residues are accessible to soil microbes. Many studies indicate that burying residues in soil increases the decomposition rate compared to placing residues on the soil surface. The effect of placement decreases with time. Thus, tillage maximizes residue–soil contact and enhances decomposition. Soil enzyme activities also respond to tillage practices. Crop residues and tillage have been reported to significantly and rapidly alter the composition, distribution and activity of the soil enzymes. Although straw amendments also contain some enzymes, the increase in activity in soils with organic residues most likely results from the stimulation of microbial and fungal activity rather than the direct addition of enzymes found in organic residues. Soil macroorganisms, such as earthworms, also play a role in crop residue decomposition. The nutritional quality and C:N ratios of plant material appear to be important factors that influence earthworm population.

Soil organic matter and humic substances exert physical, chemical and biological effects on soil quality by serving as soil conditioners, nutrient sources and substrates for microorganisms. Humic substances contribute to soil structure by acting as binding agents in the formation of soil aggregates. As a result, a stable soil structure ensures satisfactory drainage and aeration, provides protection against erosion and enhances soil properties. Humic substances are sources of nutrients that are essential for plant growth. These substances, which can be derived from crop biomass, impact soil quality and fertility and contribute to the vital role of SOM.

Crop Residues: Social and Environmental Benefits

Many of the social and environmental benefits of carbon cycled from crop residues are subtle and often difficult to detect. Integrating the numerous

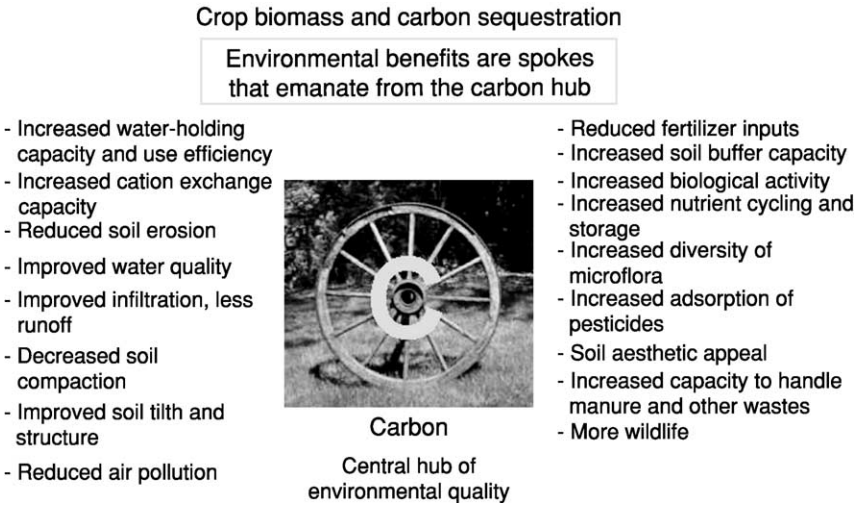


Figure 3 An environmental quality wheel illustrates results from numerous environmental benefits emanating from the carbon hub.

'small' benefits from C generated from crop residues yields important environmental benefits. Soil C from crop residues may be considered analogous to the central hub of a wheel (Figure 3). The spokes of the wheel represent incremental links to or benefits from soil C that lead to environmental improvement. Each of the secondary benefits that emanate from soil C contributes to environmental enhancement through improved soil C management. Crop biomass provides the C that becomes the supporting spokes of the environmental sustainability wheel.

Social benefits of agricultural residue management may include many off-site consequences of adopting new farm technologies and improving cultural practices. Reductions in runoff and soil erosion from cropland and rangeland enhance the functioning of streams, rivers and lakes through reduced flooding and sedimentation. The useful life of public facilities, a benefit to society, may be extended through improved water quality. Other important off-site benefits related to wind erosion have a direct impact on air quality. Air quality is primarily stressed in areas of industrial pollution and concentrated populations, but is often overlooked in rural areas as soil erosion by wind-driven particles. Additional benefits of CRM include providing both protective cover and a source of food for wild game.

Crop Residues, Research and Global Change

Agricultural crop residues and their proper management can also play an important role in helping society cope with increased greenhouse gas emissions from the burning of fossil fuels. The agricultural sector has a large capacity for removing CO_2 from the atmosphere and abating the emission of its own greenhouse gases. Croplands have the potential to offset a very significant portion of greenhouse gas emissions, but questions about climate change impacts on crop residue decomposition research need to be addressed. Specifically: (1) To what extent does climate change affect diversity of plant species and soil biota and residue decomposition processes? (2) What methods of decomposition management should be utilized to control C sequestration and CO_2 emissions? (3) What are the tillage methods and residue interactions important in carbon cycling for nutrient-use efficiency? Within a given ecosystem, the soils have a finite capacity to store carbon limited by natural soil formation factors. As a result, agriculture's contribution to these larger global climate change issues will likely be for the short term (25–50 years). Nevertheless, agriculture can help society buy time to develop new technologies and cleaner burning fuels.

Summary

Crop residue management through conservation agriculture can improve soil productivity and crop production by maintaining SOM levels. Two significant advantages of surface-residue management are increased OM near the soil surface and enhanced nutrient cycling and retention. Greater microbial biomass and activity near the soil surface acts as a reservoir for nutrients needed in crop production and increases structural stability for increased infiltration. In addition to the altered nutrient distribution within the soil profile, changes also occur in the chemical and physical properties of the soil. Improved soil C sequestration through enhanced CRM is a cost-effective option for minimizing agriculture's impact on the environment.

Ideally, CRM practices should be selected to optimize crop yields with minimal adverse effects on the

environment. Results from many experiments have indicated differing effects of CRM practices on harvested yield. Conflicting results occur due to the large number of complex interactions associated with residue quality, soil-related factors, health of the previous crop, potential susceptibility of the next crop and management options such as cultivar selection, crop rotation and planting date. Results suggest that no one CRM system is superior under all conditions. Thus, farmers have a responsibility in making management decisions that will enable them to optimize crop yields and minimize environmental impacts. Multidisciplinary and integrated efforts by a wide variety of scientists are required to design the best site-specific systems for CRM practices to enhance agricultural productivity and sustainability while minimizing environmental impacts.

Crop residues of common agricultural crops are important resources, not only as sources of nutrients for succeeding crops and hence agricultural productivity, but also for improved soil, water and air quality. The development of effective CRM systems depends on a thorough understanding of factors that control residue decomposition and their careful application within a specific crop production system. Maintaining and managing crop residues in agriculture can be economically beneficial to many producers and more importantly to society. Improved residue management and reduced tillage practices should be encouraged because of their beneficial role in reducing soil degradation and increasing soil productivity. Soil C sequestration contributes to these benefits and can play a significant role in mitigating global climate change. Food security and environmental improvement depend on soil C, a valuable resource, that can be sustainable in agroecosystems through improved, cost-effective CRM.

See also: **Carbon Cycle in Soils:** Dynamics and Management; **Carbon Emissions and Sequestration**; **Cover Crops**; **Cultivation and Tillage**; **Mulches**; **Nitrogen in Soils:** Cycle; **Nutrient Availability**; **Organic Matter:** Principles and Processes; Genesis and Formation; **Organic Residues, Decomposition**

Further reading

Allmaras RR, Schomberg HH, Douglas CL Jr, and Dao TH (2000) Soil organic carbon sequestration potential of adopting conservation tillage in U.S. croplands. *Journal of Soil and Water Conservation* 57(3): 365–373.

Cadisch G and Giller KE (1997) *Driven by Nature: Plant Litter Quality and Decomposition*. Cambridge: University Press, Wallingford: CAB International.

Crovetto Lamarca C (1996) *Stubble over the Soil: The Vital Role of Plant Residue in Soil Management to Improve Soil Quality*. Madison, WI: ASA.

CTIC (1995) *Survey Guide; National Crop Residue Management Guide*. West Lafayette, IN: Conservation Technology Information Center.

Guérif J, Richard G, Dürr B, Machet JM, Recous S, and Roger-Estrade J (2001) A review of tillage effects on crop residue management, seedbed conditions and seeding establishment. *Soil and Tillage Research* 61: 13–32.

Kumar K and Goh KM (2000) Crop residues and management practices: effects on soil quality, soil nitrogen dynamics, crop yield, and nitrogen recovery. *Advances in Agronomy* 68: 197–319.

Lal R (1997) Residue management, conservation tillage and soil restoration for mitigating greenhouse effect by CO_2-enrichment. *Soil and Tillage Research* 43: 81–107.

Lal R, Kimble JM, Follett RF, and Stewart BA (1997) *Management of Carbon Sequestration in Soil*. Boca Raton, FL: CRC Press, Lewis Publishers.

Lal R, Kimble JM, Follet RF, and Cole V (1998) *Potential of U.S. Cropland for Carbon Sequestration and Greenhouse Effect Mitigation*. USDA–NRCS Washington, DC. Chelsea, MI: Ann Arbor Press.

Oschwald WR (ed.) *Crop Residue Management Systems*. ASA Special Publication 31. Madison, WI: ASA.

Paul E, Paustian K, Elliott ET, and Cole CV (1997) *Soil Organic Matter in Temperate Agro-ecosystems: Long-term Experiments in North America*. Boca Raton, FL: CRC Press.

Reeder R (2000) Conservation tillage systems and management. In: *Crop Residue Management with No-till, Ridge-till, Mulch-till and Strip-till*. MWPS-45, 2nd edn. Ames, IA: MidWest Plan Service, Iowa State University.

Reicosky DC (1994) Crop residue management: soil, crop, climate interaction. In: Hatfield JL and Stewart BA (eds) *Advances in Soil Science, Crop Residue Management*, pp. 191–214. Special Series. Chelsea, MI: Lewis Publishers.

Unger P (1994) *Managing Agricultural Residues*. Boca Raton, FL: CRC Press.

Unger PW, Langdale GW, and Papendick RI (1988) Role of crop residues – improving water conservation and use. In: Hargrove WL (ed.) *Cropping Strategies for Efficient Use of Water and Nitrogen*, pp. 69–100. ASA Special Publication 51. Madison, WI: ASA.

CRUSTS

Contents
Biological
Structural

Biological

J Belnap, USGS Southwest Biological Science Center, Moab, UT, USA

Published by Elsevier Ltd.

Introduction

Biological soil crusts occur in a wide variety of climate regimes and vegetative communities around the globe (Figure 1). Wherever vascular plant cover is sparse enough for light to reach the soil surface, it will be colonized by some form of biological soil crust organisms. This is true even if the exposure of the soil surface is temporary, such as occurs when trees fall, or when volcanic activity or fire removes the vegetative cover. Soil crusts reach their greatest expression in arid and semiarid environments where sparse vegetation leaves large expanses of soils available for colonization (Figure 2). In the USA, this is mostly in western states (Figure 3). They are also found in other surprising places such as pine barrens, where infertile soils restrict vascular plant growth; subhumid grasslands where limited rainfall leaves open spaces between the plants; and tundra and alpine areas, where harsh conditions preclude high vascular plant cover.

Biological soil crusts (also referred to as cryptogamic, cryptobiotic, microbiotic, or microphytic soil crusts) consist of an interwoven community of cyanobacteria, green algae, microfungi, bacteria, lichens, and mosses (Figures 4 and 5). Cyanobacterial and microfungal filaments weave throughout the top few millimeters of soil, gluing loose soil particles together and forming a coherent crust that stabilizes and protects soil surfaces from erosive forces. These crusts have only recently been recognized as having a major influence on the functioning of terrestrial ecosystems.

Species Composition and Growth Forms

Globally, biological soil crusts have many similarities in species composition, despite occurring in

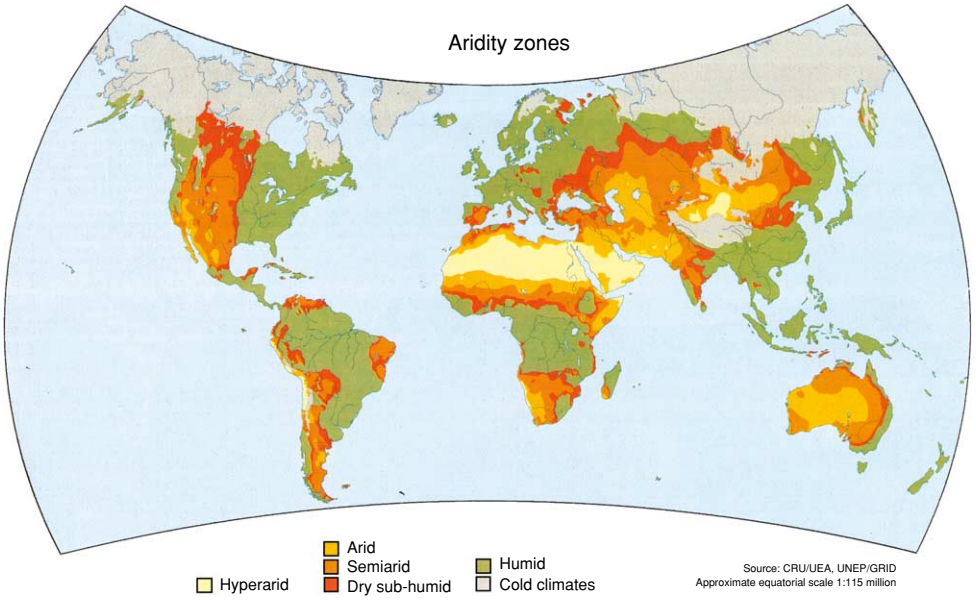


Figure 1 Aridity zones of the world. Shaded zones are regions where biological crusts have the potential to play pivotal roles in ecosystem functioning. Courtesy of United Nations Environment Program. Middleton N and Thomas D (eds) (1997) *World Atlas of Desertification*, 2nd edn. London: Arnold.

Figure 2 Biological soil crusts, completely covering the interspaces between plants, on the Colorado Plateau (northern Arizona, southern Utah, western Colorado).

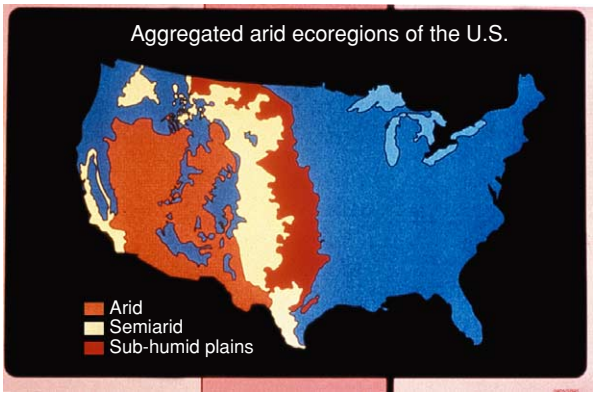


Figure 3 Arid, semiarid, and subhumid plains of the USA. When relatively undisturbed, soils across much of these regions are covered with biological soil crusts.

unconnected and seemingly dissimilar environments. Many of the dominant cyanobacteria, lichens, and moss species and genera found in soil crusts have a cosmopolitan distribution (Table 1). The proportional amount of the different species, however, varies with climate (Figure 6). In addition, climate influences what species are present. For instance, large filamentous species such as *Microcoleus* dominate the cyanobacterial flora in deserts where most rain falls during cool seasons. In contrast, the cyanobacterial flora in hot deserts with predominantly summer rainfall is often dominated by smaller genera such as *Scytonema*, *Nostoc,* and *Schizothrix*. Whereas hot desert soil crusts are dominated by cyanobacteria, cool desert soil crusts are dominated by lichens and mosses. Common lichens found in deserts throughout the world include *Fulgensia*, *Diploschistes*, *Psora*, *Placidium*, and *Collema*. Common mosses include *Tortula*, *Bryum*, and *Grimmia*.

Biological soil crusts have four general growth forms (Figure 7). Hot deserts that lack frost-heaving are generally characterized by smooth cyanobacterial crusts or rugose lichen-moss crusts. In cool deserts where frost-heaving is present, soil crusts with moderate (about 40%) lichen-moss cover are often pinnacled, due to frost-heaving upwards and differential erosion downwards. Crusts in cool deserts with a heavy lichen-moss cover are generally rolling, as frost-heaving and erosion are mitigated by the extensive lichen-moss cover.

Because the dominant components of biological soils crusts are photosynthetic organisms, they require sunlight. When soils are dry, the bulk of the crustal biomass is up to 0.5 mm below the soil surface, with some individuals found down to 4 mm. While mosses and lichens have ultraviolet (UV)-protective pigments or heavy coloration to protect them from UV radiation, only some cyanobacteria have such protection. The smaller, less motile cyanobacteria such as *Scytonema* and *Nostoc* manufacture UV-screening pigments, and are generally found on the soil surface. In contrast, the large, more motile filamentous species such as *Microcoleus*, *Lyngbya*, *Phoridium*, and *Oscillatoria* do not have UV-protective pigments, and are seldom found on the soil surface except on cloudy days when soils are moistened. Instead, these larger species are found tucked underneath the pigmented species, using them as a sunscreen. Despite sunscreening efforts by all species, mortality can be high in the summer, when UV radiation is high and moisture is limited.

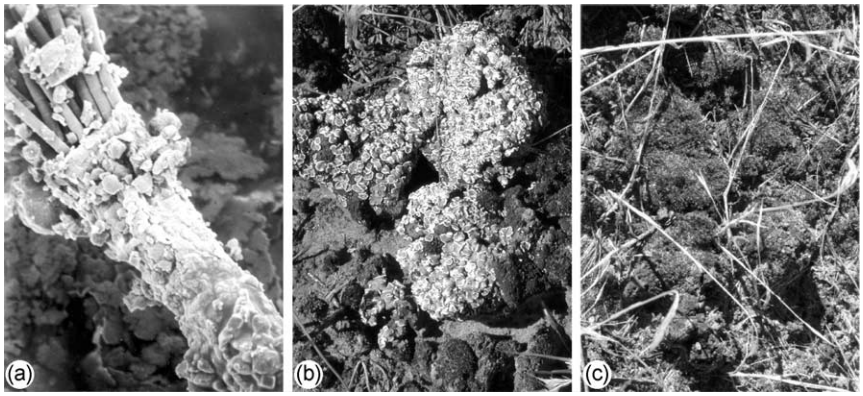


Figure 4 Close-up photos of (a) the cyanobacterium *Microcoleus vaginatus*, (b) the lichens *Psora decipiens* and *Collema coccophorum*, and (c) the moss *Syntrichia caninervis*.

Figure 5 Close-up of a biological crust community.

Table 1 Genera and species of lichens and cyanobacteria that have a cosmopolitan distribution

Lichen	*Cyanobacteria*
Fulgensia fulgens	*Microcoleus vaginatus*
Psora decipiens	*M. paludosus*
Squamarina spp.	*M. chtonoplastes*
Toninia sedifolia	*M. sociatus*
Catapyrenium spp.	*Nostoc commune*
Diploschistes spp.	*Calothrix* spp.
Endocarpon spp.	*Lyngbya* spp.
Collema spp.	*Oscillatoria* spp.
	Phormidium spp.
	Scytonema spp.
	Tolypothrix spp.

Ecological Roles

Carbon Fixation

Biological soil crusts are an important source of fixed carbon in the sparsely vegetated areas commonly found throughout the world. While vascular plants provide organic matter to the soils directly underneath them, large interspaces between plants have little opportunity to receive such input. Where biological soil crusts are present, they contribute carbon and help keep plant interspaces fertile. They do this both by leaking carbon-rich compounds into surrounding soils while alive, and contributing body carbon upon their death. These inputs help support carbon-limited microbial populations that are essential in the decomposition of plant materials.

Nitrogen Fixation

Nitrogen levels are low in desert ecosystems relative to other ecosystems, and many deserts have few nitrogen-fixing plants. Since nitrogen can limit plant productivity, the maintenance of normal nitrogen cycles is critical to maintaining the fertility of semiarid soils. Most soil crusts in deserts are dominated by complexes of organisms capable of fixing nitrogen, including *Scytonema*, *Nostoc* and the common soil lichens *Collema* and *Peltula*. In desert areas, rainfall events that are too small to promote plant growth can often stimulate crustal activity; thus, the overall time of soil crust activity can actually be fairly high. Soil crusts can be the dominant source of nitrogen for desert shrub and grassland communities. Input estimates range from 1 to 10 $\mathrm{kg\,ha^{-1}}$ annually. Nitrogen inputs are highly dependent on temperature, moisture, and crustal species composition; thus, climatic regimes and the timing, extent, and type of past disturbance which affects species composition are critical in determining fixation rates.

As with carbon, crusts contribute nitrogen to soils both underneath plants and in plant interspaces, counteracting the tendency of nutrients to concentrate around plants. Five to 88% of nitrogen fixed by crusts has been shown to leak into the surrounding soils. Nitrogen leaked from these organisms is

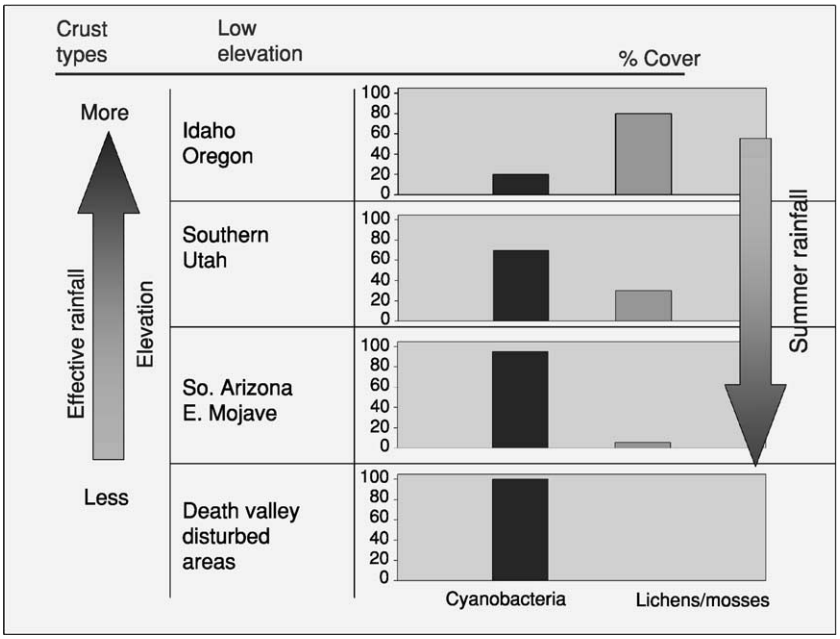


Figure 6 The cover of lichens and mosses, relative to that of cyanobacteria, increases with higher precipitation. Consequently, cyanobacteria dominate biological soil crusts in hotter deserts, whereas mosses and lichens dominate biological soil crusts in cooler deserts.

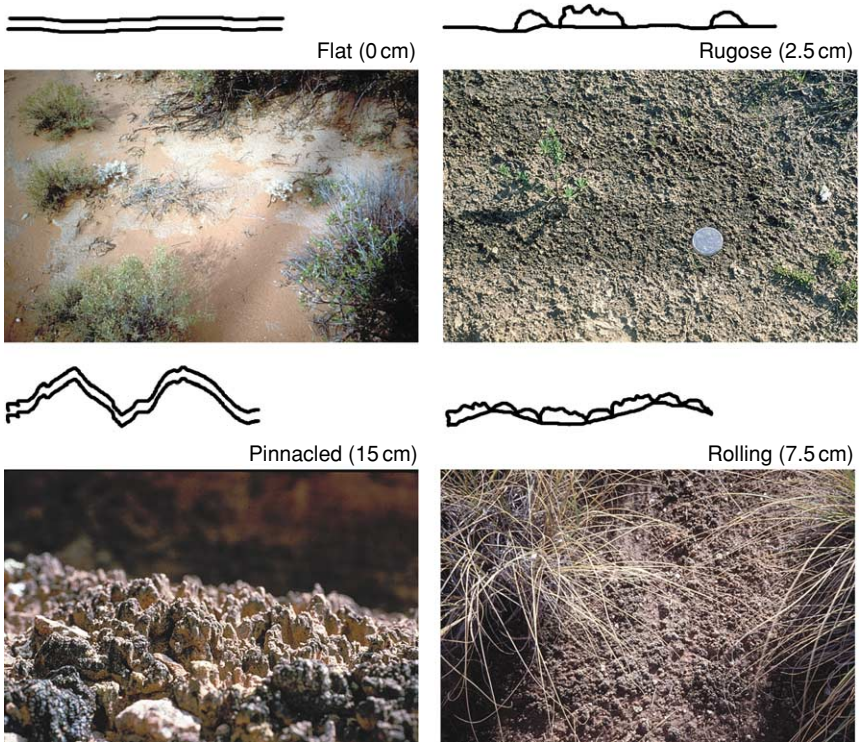


Figure 7 There are four distinct forms of soil crusts. Flat and rugose crusts are found in regions where soils do not freeze. Pinnacled and rolling crusts are found where soils do freeze. Flat crusts are also formed when any of the crust types are severely disturbed.

available to nearby vascular plants and microbial communities. Vascular plants growing in crusted areas show higher leaf concentrations of nitrogen when compared to plants growing in uncrusted soils. Leaked nitrogen has also been found in associated fungi, actinomycetes, and bacteria.

Dust Trapping

Dust can be an essential component of desert soil fertility, and soil crusts are effective in capturing eolian (wind-flown) dust deposits. Recent work in ecosystems throughout the world shows that dust inputs significantly increase levels of all major and minor soil nutrients in the tested soils. Nutrients are especially enriched in biologically crusted soils.

Effects on Vascular Plants

Soil crusts can influence the location of safe sites for seeds, as well as the germination and establishment of vascular plants. In hot deserts with flat cyanobacterial crusts, seeds generally skid off the smoothed surfaces and lodge under the nearest obstacle. The slight soil surface roughening of rugose crusts provides for some on-site seed retention. In contrast, the highly roughened soil surface created by soil crusts in cool and cold deserts results in very high retention of seeds and organic matter.

Crusts can influence seed germination, as they affect soil moisture, temperature, and stability. The increased soil moisture and temperature found in well-developed crusts generally favor germination. In deserts, low air humidity means most seeds need some form of cover in order to stay hydrated long enough for germination. Small seeds utilize small cracks for cover, while most large-seeded plants need a cover of soil or litter to germinate. Native seeds have self-burial mechanisms (such as hygroscopic awns) or are cached by rodents. However, there are nonnative seeds that lack such adaptations, and germination of these species can be inhibited by crusts.

Once seeds germinate, no studies have found that crusts affect root penetration or plant establishment. Survival of vascular plants is generally much higher, or unaffected, when crusted areas are compared with uncrusted ones. No studies have shown that biological crusts decrease vascular plant survival.

Many studies have attempted to correlate crust and vascular plant cover and results show negative, positive, and no relationship, depending on other site characteristics. At more arid sites, the correlation is generally positive, suggesting plants aid in the survival of crustal components, perhaps by providing shade. At higher elevations and/or areas with more plant cover, plants appear to inhibit crust cover by restricting the amount of light reaching the soil surface. No study has demonstrated a negative influence of crusts on overall plant cover.

Nutrient Levels in Vascular Plants

Plants growing on crusted soil generally show higher concentrations and/or greater total accumulation of various bioessential nutrients (including nitrogen, potassium, sodium, calcium, iron, and magnesium) when compared to plants growing in adjacent, uncrusted soils. Dry weight of plants in pots with cyanobacteria is up to four times greater than in pots without cyanobacteria, and dry weight of plants in untrampled areas can be twice that of trampled areas.

Several mechanisms have been postulated to explain the effect of soil crusts on vascular plants. Soil crusts contribute nitrogen and carbon, which directly increases soil fertility and also probably decomposition rates. Dust capture increases soil fertility as well. Cyanobacterial sheath material is negatively charged, binding positively charged macronutrients and thus preventing their leaching. Cyanobacteria secrete chelators that keep iron, copper, molybdenum, zinc, cobalt, and manganese more available in high-pH soils. In addition, dark-colored crusts increase soil temperature and therefore nutrient uptake rates.

Soil Hydrology and Stabilization

The effect of biological soil crusts on soil-water relations is heavily influenced by soil texture, soil structure, and the growth form of the crusts. In hot deserts, the presence of the mucilaginous cyanobacteria smooths the soil surface, decreasing soil permeability, and thus water infiltration. In contrast, the increased surface roughness that accompanies soil crust development in cool deserts increases water pooling and residence time, thus increasing the amount and depth of rainfall infiltration.

Soils in arid regions are often highly erodible, and soil formation is extremely slow, taking 5000 to 10 000 years or more. Consequently, reducing soil loss is very important in these regions. Crusts have been shown to reduce soil loss caused by wind and water erosion in all types of deserts. Polysaccharides extruded by the cyanobacteria and green algae, in combination with lichen and moss rhizines, entrap and bind soil particles together (**Figure 8**). These larger soil aggregates are heavier, have a greater surface area, and are more difficult to move by wind or water, thus reducing soil loss (**Figure 9**). When wetted, cyanobacterial sheath material swells and covers the soil surface even more extensively

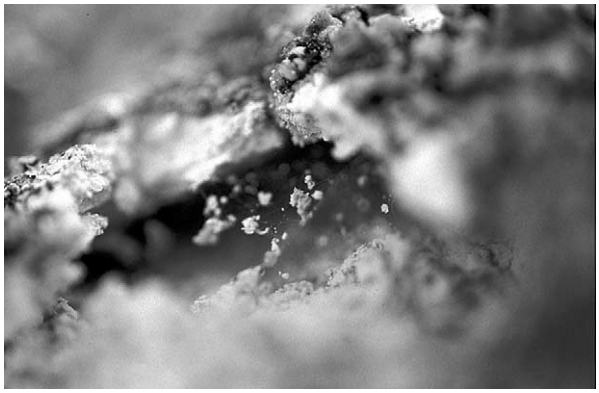


Figure 8 Clumps of soil held in place by cyanobacterial filaments. Reproduced from Belnap J and Gardner JS (1993) Soil microstructure in soils of the Colorado Plateau: the role of the cyanobacterium *Microcoleus vaginatus*. *Great Basin Naturalist* 53: 40–47.

Figure 10 When wetted, *Microcoleus vaginatus* swells, casting a net over soil surface particles (×70).

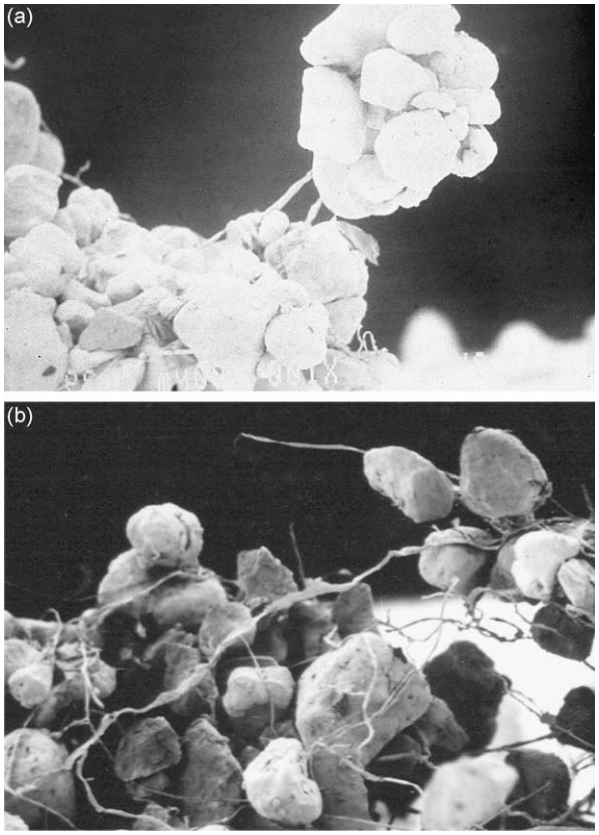


Figure 9 Scanning electron micrographs of biological soil crusts. Filaments of *Microcoleus vaginatus* can be seen connecting sand grains together and forming large soil aggregates. Top panel ×90; bottom panel ×100. Reproduced with permission from Belnap J and Gardner JS (1993) Soil microstructure in soils of the Colorado Plateau: the role of the cyanobacterium *Microcoleus vaginatus*. *Great Basin Naturalist* 53: 40–47.

than when dry, protecting soils from both raindrop erosion and overland water flow during rainstorms (Figure 10). Resistance to wind erosion increases with biological crust development, with well-developed soil crusts able to withstand all recorded ground wind speeds.

Effects of Disturbance

Species Composition

Trampling of crusted surfaces generally reduces the cover, biomass, and species richness of soil crusts. Whereas a well-developed crust can have up to 70 species of cyanobacteria and green algae and 10 or more species of soil lichens and mosses, severely trampled areas generally contain only a few species of cyanobacteria.

Water Erosion

As crustal components are brittle and easily crushed when dry, the soil aggregates formed by crust organisms are disrupted when trampled. When the roughened microtopography of undisturbed cool desert crusts is flattened, the velocity of surface water flows is increased and surfaces are subjected to sheet erosion. Seed and dust retention is reduced. Surface disturbance also crushes the buried cyanobacterial sheath material which binds soil particles, adsorbs nutrients, and increases soil moisture retention, despite lacking living filaments. Damage to this abandoned material is irreparable, since living cyanobacteria are no longer present at these depths to regenerate sheath materials. Consequently, trampling can greatly accelerate desertification processes through increased soil loss and water runoff.

Wind Erosion

Wind is a major erosive force in deserts, where there is little soil surface protection by organic matter or vegetative cover, and where soil formation is slow.

Experiments have demonstrated that, while well-developed, undisturbed crusts protect soil surfaces from wind erosion, any compressional disturbances to these crusts leave surfaces vulnerable. A decrease in soil wind resistance is directly associated with increased sediment movement. In addition, nearby biological soil crusts can be buried by blowing sediment, resulting in the death of the photosynthetic organisms. Because most photosynthetic productivity occurs in organisms in the top 1 mm of soil, very small losses can dramatically reduce site fertility and further reduce soil surface stability.

Nutrient Cycles

Nitrogen fixation in crusts shows long-term reductions in response to all types of experimentally applied disturbance, including human feet, mountainbikes, four-wheel-drive trucks, tracked vehicles (tanks), and shallow and deep raking. Consequently, disturbance can result in large decreases in soil nitrogen through a combination of reduced biological nitrogen input and elevated gaseous loss of nitrogen and soil loss. Short-term reductions in nitrogen fixation range up to 100%. Long-term studies show a 42% decrease in soil nitrogen 25 years following disturbance.

Albedo

Trampled surfaces, when compared to untrampled crusted surfaces, show up to a 50% increase in reflectance of wavelengths ranging from 0.25 to 2.5 μm. This represents a change in the surface energy flux of approximately 40 $W m^{-2}$. Large acreages of trampled areas, combined with lack of urban areas to offset this energy loss, may lead to changes in regional climate patterns in many semiarid regions. Increased albedo can radically decrease soil surface temperatures.

Surface temperatures can regulate many ecosystem functions. Lower temperature decreases carbon and nitrogen fixation rates, microbial activity, plant nutrient uptake rates, and plant growth rates. It can also delay seed germination. Because timing of these events is often critical in deserts, relatively small delays can reduce species fitness and seedling establishment, which may eventually affect community structure. Foraging times are often partitioned among ants, arthropods and small mammals based on surface temperature. Many small desert animals are weak burrowers, and soil surface microclimates are of great importance to their survival. Consequently, altering surface temperatures can affect nutrient availability and community structure for many desert organisms, thus increasing susceptibility to desertification.

Recovery from Disturbance

Natural Recovery Rates

Recovery rates are related to the type of soil crust present and the evolutionary history of the site; the type, timing, and intensity of disturbances; climate; and soil characteristics (Figure 11). In addition, 'recovery' must be defined. Visual recovery, which is not species-dependent, will occur more quickly than nitrogen flxation, which requires specific species. Nitrogen fixation in turn, recovers more rapidly than soil stability, as it depends on the recovery of all species. Faster recovery is found where: (1) the disturbance type has been present in evolutionary time and the flora is pre-adapted to disturbance; (2) disturbance is light enough to crush crust material in place, as opposed to removing it, and is infrequent; (3) surface-to-volume ratios are low, so colonizing organisms can reach the site more quickly; (4) sites are relatively stable (fine-textured soils, low or no slope as opposed to coarse soils and steep slopes); and (5) effective rainfall is high. Recovery times are also very species-dependent, as cyanobacteria are both more resistant to disturbance, and faster to recover, than lichens or mosses (Figure 12). Estimates of recovery times are shown in Table 2.

Assisted Recovery

The use of inoculants can substantially hasten recovery of soil crusts. However, development of inoculant in the laboratory has not been widely successful.

Conclusion

Unfortunately, the increasing activities of humans in desert areas are often incompatible with the well-being of biological soil crusts. The cyanobacterial fibers that confer such tensile strength to these crusts are no match for the compressional stresses placed on them by vehicles or trampling. Crushed crusts contribute less nitrogen and organic matter to the ecosystem. Impacted soils are left highly susceptible to both wind and water erosion. Raindrop erosion is increased, and overland water flows carry detached material away. Relatively undisturbed biological soil crusts can contribute a great deal of stability to otherwise highly erodible soils. Unlike vascular plant cover, crustal cover is not reduced in drought, and unlike physical crusts, these organic crusts are present year-round. Consequently, biological crusts offer stability over time and under adverse conditions, features often lacking in other soil surface protectors. Unfortunately, ever-increasing recreational and commercial uses of semiarid and arid areas are resulting in a rapid loss of well-developed biological crusts around the world.

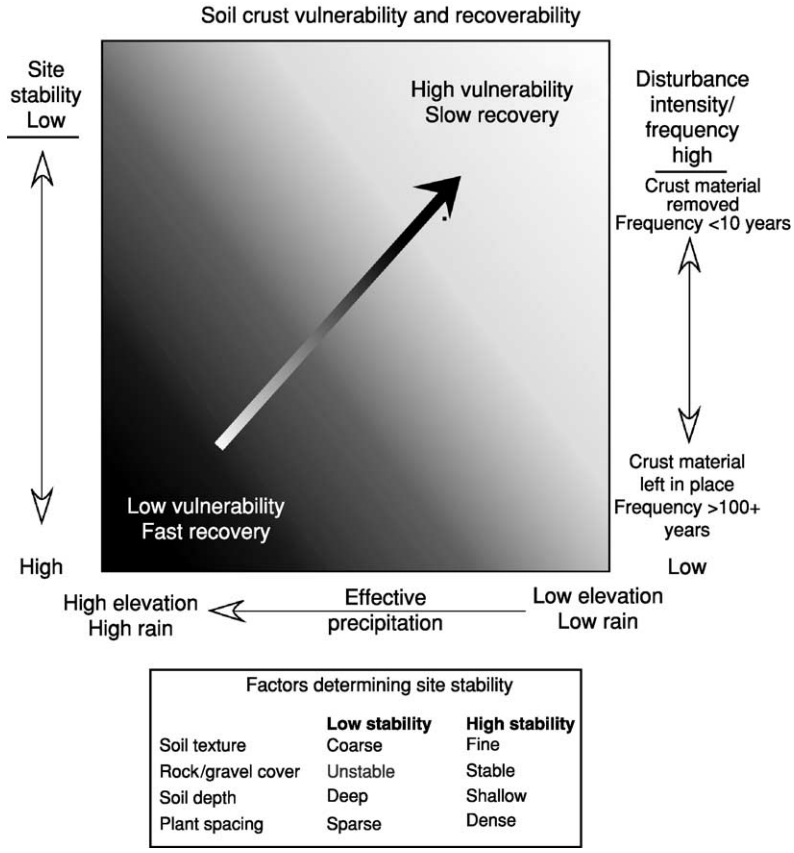


Figure 11 Summary chart of factors that influence susceptibility and recoverability of biological soil crusts to disturbance. Reprinted from Belnap J and Lange O (eds) *Biological Soil Crusts: Structure, Function, and Management*, p. 375. Berlin: Springer Verlag.

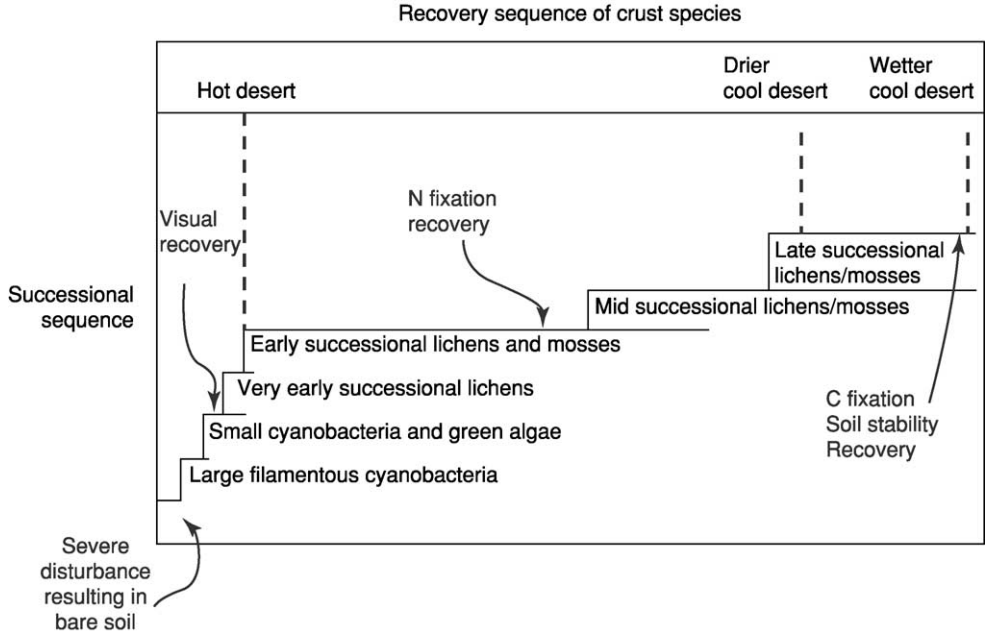


Figure 12 Colonization sequence of individual components of biological soil crusts after severe disturbance. This chart also represents the vulnerability of these components to disturbance: large filamentous cyanobacteria are the most resistant to disturbance, whereas late successional mosses and lichens are the most vulnerable to disturbance. Reprinted with permission from Belnap J and Lange O (eds) *Biological Soil Crusts: Structure, Function, and Management*. Berlin: Springer Verlag.

Table 2 Estimates of recovery time for biological soil crusts on sandy soils after severe disturbance that resulted in the removal of all crust material. Estimates are based on linear extrapolations of data collected from plots disturbed 20–80 years ago

	Years to recovery						
Desert type	*Very early successional*		*Early successional*		*Mid successional*		*Late successional*
Lower mojave (<1500 m, 100 mm)	1200	→	3800				
High mojave (1500 m, 200 mm)	200	→	800		NK		NK
Colorado plateau (1500 m, 200 mm)	50	→	500		NK		NK
Northern great basin (1000 m, 350 mm)	20	→	25	→	60	→	125

Reproduced from Belnap J and Lange O (eds) (2001) *Biological Soil Crusts: Structure, Function, and Management.* Berlin: Springer Verlag.

Further Reading

Belnap J (1995) Surface disturbances: their role in accelerating desertification. *Environmental Monitoring and Assessment* 37: 39–57.

Belnap J (2001) Factors influencing nitrogen fixation and nitrogen release in biological soil crusts. In: Belnap J and Lange OL (eds) *Biological Soil Crusts: Structure, Function and Management*, pp. 241–261. Berlin: Springer-Verlag.

Belnap J and Eldridge D (2001) Disturbance and recovery of biological soil crusts. In: Belnap J and Lange OL (eds) *Biological Soil Crusts: Structure, Function and Management*, pp. 363–383. Berlin: Springer-Verlag.

Belnap J and Gillette DA (1998) Vulnerability of desert biological soil crusts to wind erosion: the influences of crust development, soil texture, and disturbance. *Journal of Arid Environments* 39: 133–142.

Belnap J, Büdel B, and Lange OL (2001) Biological soil crusts: characteristics and distribution. In: Belnap J and Lange OL (eds) *Biological Soil Crusts: Structure, Function and Management*, pp. 3–30. Berlin: Springer-Verlag.

Crawford CS (1991) The community ecology of macroarthropod detritivores. In: Polis G (ed.) *Ecology of Desert Communities*. Tucson: University of Arizona Press.

Dregne HE (1983) *Desertification of Arid Lands*. New York: Harwood Academic Publishers.

Harper KT and Marble JR (1988) A role for nonvascular plants in management of arid and semiarid rangeland. In: Tueller PT (ed.) *Vegetation Science Applications for Rangeland Analysis and Management,* pp. 135–169. Dordrecht: Kluwer Academic Publishers.

McKenna-Neuman C, Maxwell CD, and Boulton JW (1996) Wind transport of sand surfaces crusted with photoautotrophic microorganisms. *Catena* 27: 229–247.

Sagan C, Toon OB, and Pollack JB (1979) Anthropogenic albedo changes and the earth's climate. *Science* 206: 1363–1368.

Wallwork JA (1982) *Desert Soil Fauna*. London: Praeger Scientific Publishers.

Williams JD, Dobrowolski JP, West NE, and Gillette DA (1995) Microphytic crust influences on wind erosion. *Transactions of the American Society of Agricultural Engineers* 38: 131–137.

Structural

R L Baumhardt and R C Schwartz, USDA Agricultural Research Service, Bushland, TX, USA

Published by Elsevier Ltd.

Introduction

Many of the processes that define soil productivity are governed at the soil–atmosphere interface. That is, soil otherwise capable of supplying nutrient and water needs of a crop may become deficient because of physical or transport limitations within a thin surface layer overlying the bulk soil. Frequent and often extensive changes in the physical conditions of the soil surface affect various fundamental hydrologic and biological processes, further emphasizing the importance of this very dynamic interfacial layer. This modified surface layer is called a 'crust' when dry or a 'seal' when wet.

Crusts occur as biological or physical surface layers that modify plant growth and various exchange processes in virtually all soils. Biotic, i.e., biological soil crusts are 10–100-mm-thick microbial plant communities that usually develop on semiarid and arid soils, which remain relatively undisturbed for decades. In contrast, physical soil crusts develop in response to a combination of raindrop impact and aggregate dispersion to form rapidly, in minutes, a 1–10-mm surface layer that is less porous and less conductive than the underlying bulk soil. Soil crusts typically reduce infiltration, thus causing increased runoff and, hence, water erosion. Additionally, physical soil crusts can impede emerging crop seedlings and reduce plant stand density and seedling vigor. The impact of this thin surface layer on hydrologic processes and agronomic practices has stimulated investigation of physical soil crusts. These investigations link soil mineralogy, physical properties, and the interacting soil and water chemical conditions to

crust formation and the development of practical management practices.

Crust Types and Morphology

Biological crusts found in arid or semiarid regions are formed from microbial plant communities of cyanobacteria, green algae, lichens, mosses, and microfungi that colonize the near-surface soil during an establishment period of many years. (*See* **Crusts:** Biological.) This network of bacterial filaments, fungal hyphae, and moss rhizines binds an unconsolidated, usually sandy, soil matrix into cryptobiotic, cryptogamic, microbiotic, or microphytic soil crusts that range in thickness from 10 to 100 mm. Biotic crusts can constitute up to 70% of the soil cover in nutrient-poor areas between localized growths of vascular plants, thus modifying albedo and, subsequently, carbon and nitrogen fixation. Other functions performed by biotic crusts include stabilizing the soil surface against wind or water erosion, and promoting seedling establishment and plant growth by concentrating water and nutrients in microdepressions near crust fractures. The worldwide distribution of biotic crusts in desert-like environments highlights the importance of their complex ecologic role in promoting the growth of vascular plants, which in turn benefits various animal communities, including ungulates such as deer and elk. The primary emphasis of this article, however, will be about physical crusts.

Physical soil crusts, in contrast to biotic crusts, are thin, 1–10-mm-deep, layers that, compared with the underlying bulk soil, are less conductive, denser, and often cemented between soil aggregates and primary soil particles. Physical soil crusts may be divided into depositional or structural categories depending on the primary process or agent of formation. Depositional soil crusts are formed primarily in association with ponded water, e.g., flood and furrow irrigation conditions. Water slakes aggregates and disperses the soil, which is subsequently deposited in sedimentation layers. Surge irrigation makes use of depositional crusts formed by periodic water application to depress in-furrow infiltration and modify irrigation advance rates to distribute water applications more evenly (*see* **Irrigation:** Methods).

In contrast, structural crusts form during rain or irrigation wetting processes that initially weaken and then slake or disperse soil aggregates. Impacting raindrops simultaneously fracture aggregates and detach soil particles. Soil aggregates, primary particles, and colloids are carried, subsequently, into and through the bulk soil matrix by infiltrating water. As the water infiltrates, soil particles and aggregates precipitate and occlude the near-surface conducting pores of the bulk soil, thus forming a 'washed-in layer,' as shown in Figure 1. This gradual

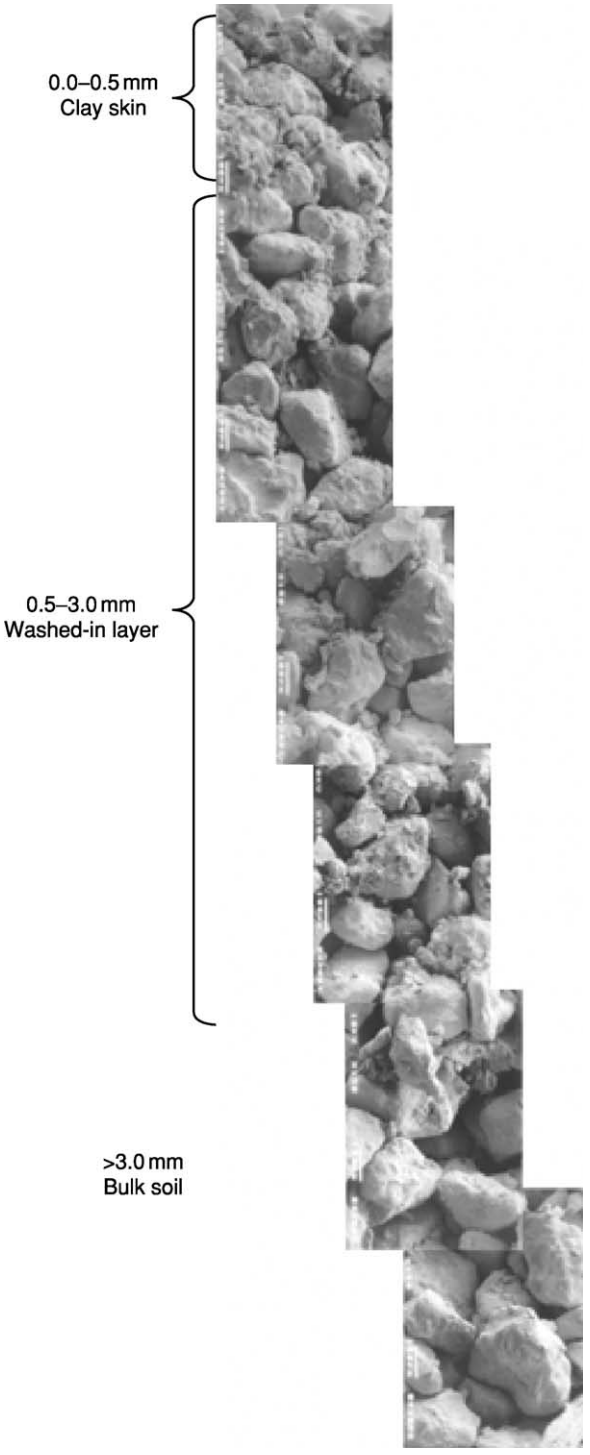


Figure 1 Scanning electron micrographs of a rain-formed structural crust, showing the typical morphology of skin and washed-in layers overlying the bulk soil. Adapted from Wakindiki IIC and Ben-Hur M (2002) Soil mineralogy and texture effects on crust micromorphology, infiltration, and erosion. *Soil Science Society of America Journal* 66: 897–905.

occlusion of conducting pores further reduces infiltration sufficiently to cause surface ponding that further accelerates aggregate slaking and dispersion. The resulting deposition of fine soil material at the soil surface forms a thin 'skin' layer (approximately 0.1 mm) over the remaining conducting pores and completes soil crust formation. The soil pore micromorphology of depositional crusts often reflects sand or silt bedding planes and occasional vesicular structure in contrast with the more complex changes found in structural crusts that include pore-size reduction or partial blocking of conducting pores.

Factors Governing Crust Formation

Rainstorm Characteristics

Virtually all soil surfaces are subject to mechanical changes attributed to either dispersion or the cumulative beating action of raindrop impact that rearranges aggregates and particles into a less-conductive, compacted crust layer. Intercepting drop impact with an energy-absorbing material such as crop residues appreciably reduces crust formation. For example, Figure 2 illustrates the ability of raindrop impact on a bare soil to form a smooth, crusted surface in contrast with the case where raindrop energy is intercepted before reaching the soil (no drop impact). Cumulative rainstorm impact energy calculated as the product of rain intensity and the kinetic energy density (drop size) is frequently used, as a first approximation, to describe the potential for physical changes in the surface soil matrix during crust formation. One such relationship between increasing cumulative rainstorm kinetic energy and the decreasing hydraulic conductance through the crust or seal is shown in Figure 3. Crust formation is delayed during initial wetting and softening of the surface soil aggregates that absorb drop impact without any corresponding change in the surface conductance. Continued drop impact on the wetted soil surface, however, fractures aggregates and releases primary soil particles into the infiltrating water, resulting in pore occlusion of the developing crust or seal and a rapid decrease in the crust conductance to a final value. Increasing storm energy density independent of rain intensity accelerates crust or seal formation and, consequently, depresses the infiltration at an accelerated rate (Figure 4). In this example, the final infiltration rate converged to approximately the same value for all nonzero drop impact conditions.

Alternatively, when storm energy density remains constant, increasing rain intensity increases kinetic energy accumulation and similarly accelerates crust or seal formation, which results in a steep decline in infiltration rates. Higher rainfall intensities also shorten the time to ponding independent of energy density. Infiltration rate plotted with time in Figure 5 decreased more rapidly with increasing rain intensity; however, the final infiltration rate was greater with the higher rain intensity. Comparison of rainstorm intensity and kinetic energy density effects on both final infiltration rate and the corresponding hydraulic conductance through the crust or seal are shown in

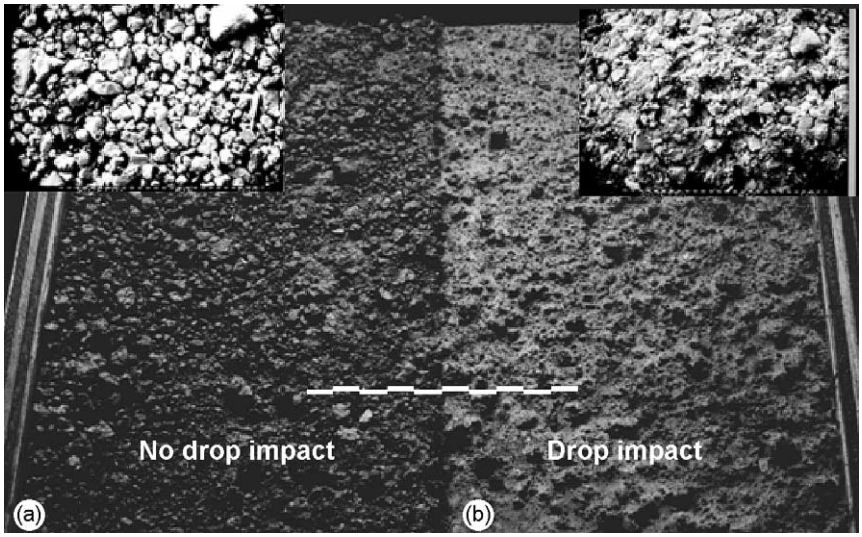


Figure 2 The surface of a silt loam soil exposed to simulated rain with no drop impact (a) has negligible crust formation in contrast to normal drop impact conditions (b). Scale 10 cm. Fractured soil aggregates provide primary particles and microaggregates needed to develop a crust and render a smooth surface (inset, b) compared with the practically undisturbed soil aggregates not exposed to raindrop impact (inset, a). Reproduced with permission from Gantzer CJ (1980) *Physical and Morphological Investigations of Surface Seal Formation on Selected Soils*. PhD Thesis. Saint Paul, MN: University of Minnesota Press.

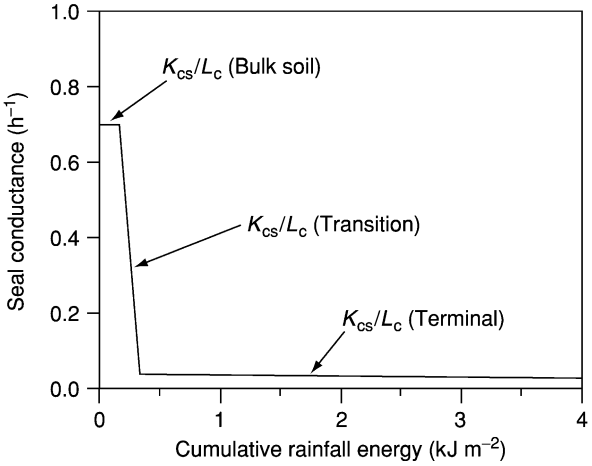


Figure 3 Seal or crust conductance through a silty clay loam crust calculated as the ratio of the crust's saturated hydraulic conductivity, K_{cs}, divided by crust thickness, L_c, and plotted as a function of cumulative rainfall energy. Initially, crust conductance is equivalent to the bulk soil; however, surface wetting and the raindrop impact modifies the soil surface matrix, resulting in a rapidly decreasing crust conductance that changes to a terminal value. Adapted from Baumhardt RL, Römkens MJM, Whisler FD, and Parlange J-Y (1990) Modeling infiltration into a sealing soil. *Water Resources Research* 26: 2497–2505.

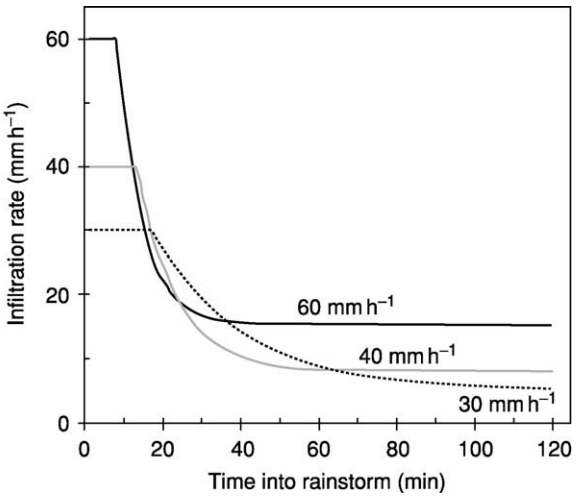


Figure 5 Modeled infiltration rate (millimeters per hour) during a rainstorm with a kinetic energy density of 0.0275 kJ m^{-2} is plotted against time for application intensities of 30, 40, and 60 mm h^{-1}. As rain intensity increases, observed infiltration decreases more rapidly; however, infiltration rate after 120 min is greater with higher rain intensity that erodes the seal or crust and increases its conductance. Adapted from Baumhardt RL, Römkens MJM, Whisler FD, and Parlange J-Y (1990) Modeling infiltration into a sealing soil. *Water Resources Research* 26: 2497–2505.

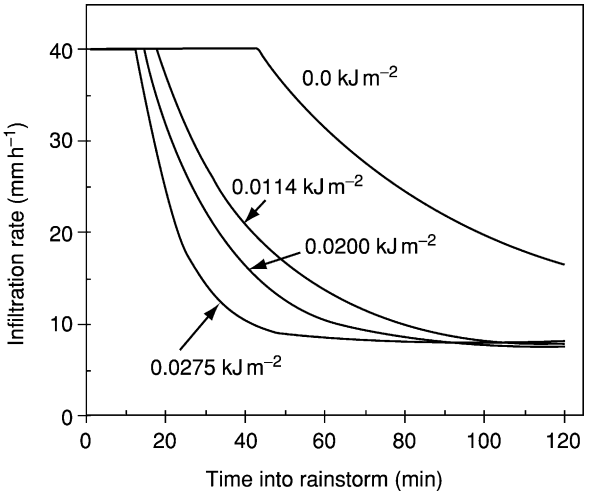


Figure 4 Modeled infiltration rate (millimeters per hour) is plotted with time into rainstorm for intercepted raindrop impact 0.0 kJ m^{-2} compared with storms having progressively higher kinetic energy densities, 0.0114–0.0275 kJ m^{-2} (larger drop sizes). Increasing rainstorm energy density progressively decreases infiltration more rapidly because of accelerated crust formation and decreased crust conductance. Intercepting drop impact delays time of ponding and increases infiltration rate. Adapted from Baumhardt RL, Römkens MJM, Whisler FD, and Parlange J-Y (1990) Modeling infiltration into a sealing soil. *Water Resources Research* 26: 2497–2505.

Table 1. These data show that both final infiltration rate and the corresponding crust hydraulic conductance increased with increasing rainstorm intensity for different energy densities. Possible explanations for these and similar observations of rain-intensity effects include formation of a thinner crust, erosion of the developing crust, and the formation of a water film that absorbs drop impact energy. Higher rain intensities cause more rapid crust formation, thus preserving a higher gradient to drive infiltration, and may not independently modify the surface hydraulic conductivity. Because rainstorm energy density does not vary widely within common rain intensities, crust properties may be estimated as a function of the cumulative rainfall or infiltration rather than from more cumbersome rainstorm intensity and cumulative energy relationships.

The mechanical crusting process is accompanied by a chemical interaction between rainwater and the soil surface that contributes to soil flocculation or dispersion and subsequent aggregate formation or slaking. Water and soil chemical properties interacting to flocculate soil delay and diminish crust formation, while monovalent cations such as sodium that disperse clays on wetting collapse soil aggregation and increase crusting. For example, applying distilled water as simulated rain to a sandy loam soil with an exchangeable sodium percentage (ESP) of 1.0 produces a 7.5-mm h^{-1} final infiltration rate. Increasing the soil ESP from 1.0 to 2.2 and 4.6, a more dispersive condition, results in the formation of crusts that depress the final infiltration rate to 2.3 and 0.7 mm h^{-1}, respectively. Water with a high polyvalent electrolyte

Table 1 Measured final infiltration rate and crust conductance for three kinetic energy densities (drop sizes) and increasing simulated rainstorm application intensities

Application intensity ($mm\,h^{-1}$)	Kinetic energy ($kJ\,m^{-2}$ per mm rain)					
	Final infiltration rate ($mm\,h^{-1}$)			Final conductance (h^{-1})		
	0.0275	0.0200	0.114	0.0275	0.0200	0.114
30	5.4	6.2	6.9	0.025	0.030	0.035
40	8.1	7.7	7.4	0.035	0.035	0.040
60	15.3	14.0	9.7	0.070	0.050	0.060
90	19.3	16.6		0.110	0.080	

Adapted from Baumhardt RL, Römkens MJM, Whisler FD, and Parlange J-Y (1990) Modeling infiltration into a sealing soil. *Water Resources Research* 26: 2497–2505.

concentration (EC) flocculates the soil and consequently stabilizes soil aggregates, reducing crust or seal formation. Infiltration rates increase from 1.2 to 7.5 $mm\,h^{-1}$ as the EC of water applied as simulated rain to the sandy loam soil (ESP = 1.0) increases from 0.0 to 5.6 $dS\,m^{-1}$. Understanding the dispersive effects of low EC water in rainfall allows researchers to develop soil-amendment applications of powdered phosphogypsum, which enters solution rapidly enough to increase the rainwater EC sufficiently to reduce soil dispersion and crust formation.

Soil Properties

Crust or seal formation is governed by interacting chemical, biological, and physical properties that affect soil susceptibility to mechanical changes in surface aggregation, pore structure, and consequently depressed final infiltration rate. Physical soil properties affecting crust or seal formation include dynamic conditions and essentially fixed features, e.g., antecedent water content (dynamic) compared with soil texture (fixed). The final infiltration rate of two silt loam soils wetted for 24 h under 0.5-kPa suction is depressed approximately 60% compared with dry soil exposed to identical crusting conditions. Although soil crust formation is observed over various textures, including coarse-textured, sandy soils, those soil textures with greater silt concentrations are the most subject to crusting. For example, as soil silt content increases from 51 to 84%, infiltration decreases as much as 300%, with a corresponding 700% increase in crust strength. This may be due, in part, to similarities in the size of the conducting pores and soil particles forming the washed-in layer. Sandy soil readily forms weak structural crusts in contrast to soil having clays that provide bonding surfaces, which stabilize aggregates and reduce soil susceptibility to crust formation.

Soil chemical properties interact with rainwater to disperse and slake aggregates during structural crust formation, but aggregate stabilization with natural and synthetic flocculating or bonding agents delays formation and reduces the strength of dry crusts. For example, the soil amendment polyacrylamide (PAM) is a potent flocculating agent that stabilizes aggregates by offsetting the dispersive effects of water on primary particles. This amendment is frequently applied in irrigation water to reduce soil crust formation while increasing infiltration and irrigation efficiency. Animal manure and other composted organic materials increase nutrient content and bioavailability, which stimulates microbial activity and promotes the production of complex organic polymers and chemical metabolites that serve as bridges and bonding agents for cementing flocculated clay particles into aggregates (*See* **Aggregation:** Microbial Aspects). Soil aggregates are consequently stabilized naturally by the accumulation of organic matter produced by microorganisms such as fungi, whose hyphae hold soil particles together and generate a glycoprotein (glomalin) cementing agent that helps bond primary soil particles. Other microbial by-products and secretions of complex organic polymers or chemical metabolites similarly strengthen soil aggregates. Polymer soil amendments such as polyvinyl alcohol are applied to cement soil particles synthetically and strengthen aggregates, thus increasing soil aggregate stability and reducing crust formation.

Hydrologic Impact of Crusts

Soil crusts determine field infiltration, runoff, and erosion processes on practically all landscapes, but the dynamic processes involved in surface crusting vary spatially and often result in crusts with different characteristics on the same landscape. The degree to which crusting modifies field hydrology is influenced largely by tillage and cropping practices. For example, the interception of raindrop impact by surface residues and plant canopies can prevent or at least delay the formation of surface crusts. Desirable tillage practices retain sufficient residue to protect

approximately 40–60% of the soil surface area by intercepting raindrop impact, thus, maintaining an aggregated surface soil with high infiltration rates. With sufficient biomass production, residue-conserving, no-tillage practices frequently increase saturated conductivities through improved aggregate stability imparted by greater soil organic carbon content.

In semiarid regions, crop residue production is often limited and, generally, provides insufficient cover (less than 40%) to intercept raindrop impact and protect the soil surface from crusting regardless of conservation or no-tillage practices. Soil crusts commonly develop in fields under no-tillage residue management during infrequent but intense rainstorms. Thus, crust development progresses rapidly in soils with dispersible clays and low residue cover and consequently infiltration rates of no-tillage soils are frequently similar to tilled soils after 1 or 2 h of rainfall. A spatially clumped residue distribution may promote higher, effective one-dimensional infiltration rates than fields with evenly distributed residues.

Infiltration – Developed Crusts

Improved understanding of the relationship between crust formation and infiltration processes has permitted the projection of crust impact on field hydrology through the use of progressively more sophisticated models. Empirical models that predicted infiltration as an exponential decay function, such as the Kostiakov- or Horton-type equations, have been replaced largely by mechanistic approaches to calculating infiltration fluxes (*See* **Infiltration**). Infiltration of water through soil with a developed surface crust is calculated as a special, simplest, case of flow through a two-layered system. As steady-state flow is approached, the flux through the crust governs the flux into the more rapidly draining, bulk-soil transmission zone, with a water potential gradient approaching unity. Darcy flux through the crust layer can be described in the expression:

$$q = K_c \frac{h_0 - h_c + L_c}{L_c} \qquad [1]$$

where q is infiltration flux (meters per second), K_c is the flow-averaged conductivity across the crust (meters per second), h_0 is the positive hydraulic pressure head (meters) imposed at the surface and usually taken as zero, h_c is the water potential (meters) at the crust interface with bulk-soil, and L_c is crust thickness (meters). Using measurements of the water potential at or just below the crust and the infiltration flux during steady flow, we can rewrite Eqn [1] to estimate crust conductance C (per second) or hydraulic resistance r_c (seconds) as:

$$C = \frac{1}{r_c} = \frac{q}{h_0 - h_c + L_c} \qquad [2]$$

Water flow through crusted soils, therefore, becomes a self-regulating system whereby the physical properties of the crust and bulk soil control the infiltration rate and developing water potentials within the transmission zone. Drainage through the underlying bulk-soil transmission zone under a unit gradient causes the water potential beneath the crust to decrease as the hydraulic conductance through the crust decreases.

Infiltration – Developing Crusts

In contrast to steady-state infiltration through developed soil crusts with fixed conductive properties, forming soil crusts present complex challenges in quantifying infiltration and consequently field hydrology. Approximations of infiltration into a crusting soil use simplified, quasianalytical methods based on variants of the Green–Ampt equation derived to account for the presence of a hydrologically developed surface crust. Assuming a constant water potential at the wetting front (h_f), a uniformly wetted transmission zone to a depth, L_f, with a constant water content $\theta(h_f)$ and conductivity $K(h_f)$, and a soil profile with an initially uniform water content (θ_i), one can approximate the rate of infiltration into a crusted or sealed soil as:

$$\frac{dI}{dt} = \frac{K(h_f)[h_0 - h_f + L_f]}{L_f + K(h_f)r_c} \qquad [3]$$

where I is cumulative infiltration. Setting I equal to $L_f \cdot [\theta(h_f) - \theta_i]$ permits a time-dependent solution for infiltration from known crust resistance and the $K(\theta)$ and $h(\theta)$ functions. However, initial infiltration is overestimated with Eqn [3] because it implicitly assumes a constant water potential h_f at the crust–soil interface, when, in fact, it increases asymptotically toward a steady-state water potential. Piecewise implementations of Green–Ampt-type equations (and other derivations) have been developed to address the dynamics of potential and conductivity changes in the transmission zone. One such formulation is:

$$\frac{dI}{dt} = K(h_f)\frac{\int_{h_f}^{h_i} K(h)/K(h_f)dh + SL_f}{SL_f} \qquad [4]$$

where S is a shape factor that ranges from 1.0 to 1.2. To solve for I and h_f with time, Eqn [4] is combined with the surface boundary condition:

$$q = -K(h_f)\left[\frac{\partial h}{\partial z} - 1\right] = \frac{-h_f + L_c}{r_c} \qquad [5]$$

Although infiltration fluxes calculated using Eqn [4] with an r_c measured after 2 h achieve satisfactory agreement with two-layer solutions of the Richards equation (Figure 6), the applied r_c was estimated before steady-state flow conditions. Because the r_c calculated with Eqn [2] decreases asymptotically over time for a stable crust, crust resistances measured during near steady-state flow conditions (e.g., after 20 h) are underestimated. Consequently, use of the steady-state r_c leads to serious overestimation of earlier infiltration fluxes with the Green–Ampt approach (Figure 6). This estimation error may be attributed to the effects of rapidly increasing crust resistances as the soil surface matrix changes and because of the concomitant changes in the unsaturated conductivity with increasing water content of the crusting or sealing surface.

Field-scale models of infiltration into crusted landscapes often apply a Green–Ampt approach. The decrease in the soil surface conductance through a developing crust is typically described with an exponential decay function that incorporates the cumulative kinetic rainfall energy since the last tillage event and a soil stability factor. Not surprisingly, raindrop intercepting residue cover greatly influences the Green–Ampt parameters describing crust effects on infiltration. This approach has been found sufficiently sensitive to simulate the effects of tillage, residue cover, and crusting when projecting infiltration for field-scale applications. Added precision in calculating field infiltration requires more complex techniques to account for the changing hydraulic properties of a dynamic crust. This is further complicated by the spatial uncertainty in describing the crust-formation process.

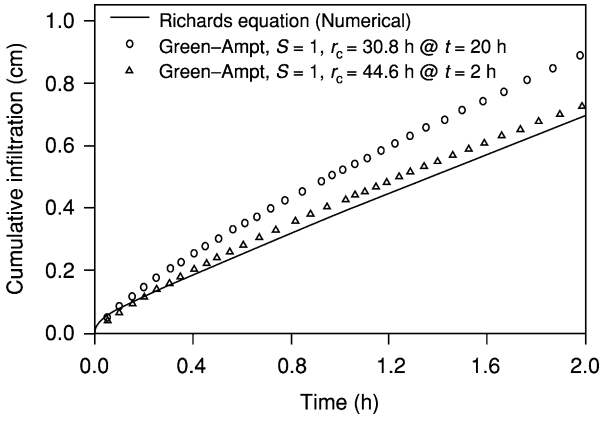


Figure 6 Cumulative infiltration for a fine-textured soil calculated using a two-layer numerical solution to the Richards equation with a 5-mm seal and using the Green–Ampt formulation (Eqns [4] and [5]). Crust resistances for the Green–Ampt solutions were calculated from the soil–crust interface water potentials and surface fluxes obtained from the numerical solution at $t = 20$ h (near steady state) and $t = 2$ h.

Numerical Infiltration Models for Developing Crusts

Crust formation begins when the energy imparted by the first falling raindrops transforms the bulk soil matrix into a more dense and less-conductive surface layer and continues beyond incipient ponding. Simultaneously, raindrops wet the soil and increase the unsaturated hydraulic conductivity of the developing crust. The earlier relationship between the decreasing crust hydraulic conductance and cumulative rainstorm kinetic energy is revisited in Figure 7; however, the unsaturated hydraulic conductivity of the crust (dashed line) initially increases as the soil wets before being reversed as the crusted soil matrix saturates. These dynamic surface boundaries are suited to numerical infiltration models.

Finite-difference solutions of the Richards equation are typically used to obtain numerical approximations of infiltration into crusted soils. In these solutions, the crust layer is treated either as a separate layer with distinct hydraulic properties or as a hypothetical membrane with a hydraulic resistance. The latter method is identical to the boundary condition, Eqn [5], and implicitly assumes that the crust layer becomes rapidly saturated. Because the crust or seal–soil interface may never become saturated, r_c does not represent an independent crust or seal hydraulic property, and imposing it under differing flow conditions can result in errors of undetermined magnitude.

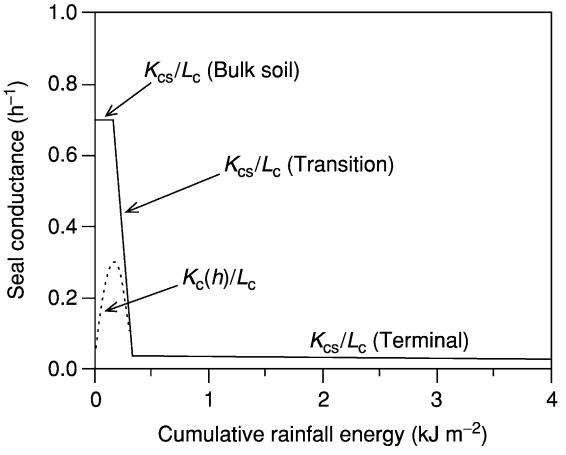


Figure 7 Model-calculated unsaturated hydraulic conductivity, $K_c(h)$, during crust formation increases as the soil is wetted (dashed line); however, as the surface soil matrix crusts, K_{cs} rapidly decreases with increasing cumulative rainfall energy and causes a corresponding reduction in $K_c(h)$ that eventually follows the previously plotted conductance through a silty clay loam crust. Adapted from Baumhardt RL, Römkens MJM, Whisler FD, and Parlange J-Y (1990) Modeling infiltration into a sealing soil. *Water Resources Research* 26: 2497–2505.

Treating the crust as a thin layer with dynamic bulk density and $K(h)$ and $\theta(h)$ functions that are independent of the underlying soil has stimulated the use of numerical models to estimate infiltration into crusting soils. In these models, the hydraulic properties of the crust change in response to the cumulative kinetic energy, rainfall intensity, and soil-rainwater chemical properties. For example, the changes in crust hydraulic conductance are functionally related to rainstorm cumulative kinetic energy, thus providing an estimate of the saturated crust or surface soil matrix hydraulic conductivity K_{sc} (Figure 7). Corresponding adjustments to crust porosity can be calculated from K_{sc} using the Kozeny–Carman equation. Assuming that decreasing porosity results from isometric reduction in pore radius, concomitant adjustments to the air-entry value may be calculated using the Poiseuille equation in conjunction with the capillary equation. An alternative approach describes the increased bulk density of the surface soil matrix changes resulting from raindrop impact and pore occlusion with sediments carried by infiltrating water to varying depths. The $K(h)$ and $\theta(h)$ functions are related to the crust bulk density and therefore are time-dependent.

A fundamental difficulty with describing changes in the crust matrix is that the physical properties of crusts developing over time generally cannot be directly measured. Crust properties are typically estimated using inverse methods, whereby parameters describing the matrix are adjusted so that calculated infiltration rates match observations. Calculating infiltration using numerical methods that adjust crust or seal conductivity in response to wetting and soil matrix changes must rely on researcher insight to quantify the effects of the crust-formation process.

Agronomic Importance of Crusts

Reduced infiltration through a less-conductive surface crust is accompanied by a corresponding reduction in gas and vapor transport; however, crusts have a complex effect on the soil surface energy balance. For example, the denser-crusted soil surface can increase its thermal conductivity and potential to increase the energy driving evaporation. Conversely, the lower exposed area and higher albedo of the crusted surfaces absorb less irradiant energy, thus reducing the potential evaporation. While tillage to fracture soil surface crusts may reduce albedo and increase infiltration, no-tillage residue management practices retain surface residue to intercept both raindrop impact and irradiant energy. No-tillage residues minimize crust formation and increase storage of precipitation as plant-available soil water.

Soil crusting reduces infiltration and increases the potential for water-driven soil erosion. As the amount of runoff increases, the potential entrainment and transport of sediment increases, which is manifested as increased soil erosion. Alternatively, crusted soil is 40–70 times less erodible to wind than the corresponding unconsolidated material. This benefit is, generally, offset by the increased surface wind speeds and the loss of large-aggregate shelter angle effects with a smooth crust surface. In the absence of standing residue to reduce surface wind speed and provide a nonerodible barrier against saltating sand grains, the distribution of fine soil particles within a crust exposes sand particles at the surface that may move with the wind and further contribute to saltation or surface-creep wind erosion. Soil erosion increases as the abrasive action of the moving sand grains on the crust provides more material from the soil surface, thus escalating wind erosion. Tillage to fracture the soil crust is required to disrupt sand saltation and abrasion of the eroding soil. The 'blowing sand' observed from eroding crusted soils produces severe collateral crop damage to tender seedling tissues.

Reduced and delayed seedling emergence resulting in poor crop establishment is a common agronomic production hazard caused by soil crusts (*See* **Germination and Seedling Establishment**). The cemented soil crust also acts as a barrier to an emerging crop (Figure 8) which must either grow through a fracture or exert sufficient force to lift or penetrate the crust. Soil crust strength varies with formation conditions. For example, the strength of crusts formed by overhead sprinkler irrigation with lower

Figure 8 The reduction of seedling emergence is, possibly, the most visible and the most significant agronomic impact by a soil crust. Emergency tillage operations are performed to fracture the crust and, when ineffective, are frequently followed by replanting decisions.

drop impact energy and less-dispersive well water will have a lower penetration resistance (approximately 0.5 MPa using a fine-probe penetrometer) compared with crusts formed during intense natural rainstorms (approaching 1.2 MPa). Likewise, the crust impact on seedling emergence varies with the crop. The force exerted by a seedling varies from 0.15 N for alfalfa (crust susceptible) to more than 4.0 N for corn (crust tolerant) depending on water imbibition and growth-limiting factors such as temperature and seed mass. For cotton seedlings exerting a maximum emergence force of 3.02–4.63 N, the corresponding measured axial pressure (or maximum penetrable crust strength) ranges from 1.25 to 1.90 MPa. As the crust penetrometer resistance increases from 0 to 1.0 MPa, cotton seedling emergence decreases from 78 to 21% 2 days after planting, thus illustrating the impact of soil crusts to injure sensitive crops.

Agronomic practices to improve seedling emergence through a crust rely on tillage to fracture the crust or small irrigation applications (when available) to wet and soften the crust. Directed tillage to fracture soil crusts improves seedling emergence, but also imposes injury risks to emerged seedlings. Irrigation to soften the crust may not be available or practical because of the resulting depression in seed-zone soil temperature. Some cotton producers in California (USA) and elsewhere control severe crusting effects on seedling emergence by removing previously formed soil caps above the seed (Figure 9). The cap may also reduce the amount of force required to penetrate and/or fracture the crust by providing a fracture zone along the cap peak or reducing the force required to move the crust.

Agronomic management paradigms achieving the greatest success in minimizing the impact of soil crusts attempt to reduce crust formation. Residue-retaining tillage practices reduce soil crust formation by intercepting raindrop impact and ameliorate crust impact on seedling emergence by delaying surface drying. Likewise, agronomic cultural practices that retain residues on the soil surface can decrease runoff and surface wind speeds attributed to soil crust formation and, consequently, decrease water and wind erosion while increasing plant-available soil water. The continued study of soil crusts will continue to result in agronomic management practices to reduce crust formation and minimize crust impact.

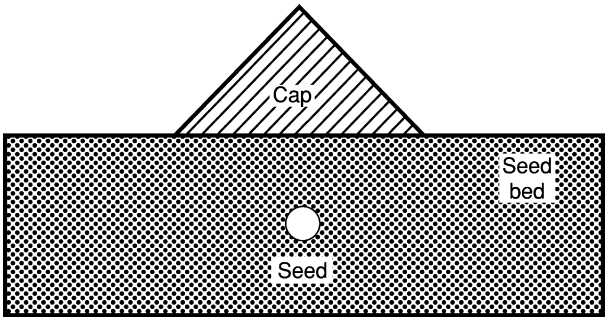


Figure 9 Diagram of seed and seedbed with a soilcap (small hill). Soil caps can be removed by tillage to encourage seedling emergence under crusting conditions.

See also: **Aggregation:** Microbial Aspects; **Compaction**; **Crusts:** Biological; **Flocculation and Dispersion**; **Germination and Seedling Establishment**; **Infiltration**; **Irrigation:** Methods

Further Reading

Ahuja LR (1983) Modeling infiltration into crusted soils by the Green–Ampt approach. *Soil Science Society of America Journal* 47: 412–418.

Assouline S and Mualem Y (2000) Modeling the dynamics of soil seal formation: analysis of the effect of soil and rainfall properties. *Water Resources Research* 36: 2341–2349.

Baumhardt RL, Römkens MJM, Whisler FD, and Parlange J-Y (1990) Modeling infiltration into a sealing soil. *Water Resources Research* 26: 2497–2505.

Belnap J, Kaltenecker JH, Rosentreter R, Williams J, Leonard S, and Eldridge D (2001) *Biological Soil Crusts: Ecology and Management*. Technical Reference 1730. Denver, CO: US. Department of the Interior.

Cary JW and Evans DD (1974) *Soil Crusts*. Technical Bulletin 214. Tucson, AZ: University of Arizona Agricultural Experiment Station.

Duley FL (1939) Surface factors affecting the rate of intake of water by soils. *Soil Science Society of America Proceedings* 4: 60–64.

Gantzer CJ (1980) *Physical and Morphological Investigations of Surface Seal Formation on Selected Soils*. PhD Thesis. Saint Paul, MN: University of Minnesota Press.

Hillel D and Gardner WR (1969) Steady infiltration into crust-topped profiles. *Soil Science* 108: 137–142.

Hillel D and Gardner WR (1970) Transient infiltration into crust-topped profiles. *Soil Science* 109: 69–76.

Mannering JV (1967) *The Relationship of Some Physical and Chemical Properties of Soils to Surface Sealing*. PhD Thesis. West Lafayette, In: Purdue University Press.

McIntyre DS (1958) Permeability measurements of soil crusts formed by raindrop impact. *Soil Science* 85: 185–189.

Poesen JWA and Nearing MA (1993) *Soil Surface Sealing and Crusting*. Cremlingen, Germany: Catena-Verlag.

Römkens MJM, Baumhardt RL, Parlange MB, Whisler FD, Parlange J-Y, and Prasad SN (1985) Rain-induced surface seals: their effect on ponding and infiltration. *Annales Geophysicae* 4: 417–424.

Shainberg I (1985) The effect of exchangeable sodium and electrolyte concentration on crust formation. In: Stewart BA (ed.) *Advances in Soil Science*, pp. 101–122. New York: Springer-Verlag.

Sumner ME and Stewart BA (1992) *Soil Crusting: Chemical and Physical Processes*. Boca Raton, FL: Lewis Publishers.

Tackett JL and Pearson RW (1965) Some characteristics of soil surface seals formed by simulated rainfall. *Soil Science* 99: 407–412.

Wakindiki IIC and Ben-Hur M (2002) Soil mineralogy and texture effects on crust micromorphology, infiltration, and erosion. *Soil Science Society of America Journal* 66: 897–905.

CULTIVATION AND TILLAGE

M R Carter, Agriculture and Agri-Food Canada, Charlottetown, PE, Canada
E McKyes, McGill University, Ste-Anne de Bellevue, PQ, Canada

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Cultivation and tillage are activities related to soil manipulation and movement. Soil movement can occur due to natural (e.g., freeze–thaw processes) and biotic (e.g., earthworm movement) processes; however, both 'cultivation' and 'tillage' generally refer to the direct intervention of a 'soil manager,' who uses tillage tools to modify, manipulate, or ameliorate a soil condition. Although both terms are used interchangeably, 'cultivation' is usually restricted to soil manipulation to create a seedbed suitable for crop seeding and establishment. In comparison, the term 'tillage' serves a wider function than cultivation. As a subsystem of a crop-production system, tillage is employed to facilitate crop establishment, modify soil structure, incorporate fertilizer and soil amendments (e.g., lime and manure), control plant residues and weeds, and alleviate both climatic and soil constraints. Over time, basic tillage tools have been developed to perform specific soil-moving operations. A combination of tillage operations and their timing results in the development of a tillage system to provide specific functions in given situations.

Goals of Soil Tillage

The traditional aims of tillage are to improve soil structure for crop growth, incorporate organic amendments into the soil, and to control weeds. The last goal can often be met by use of herbicides, and this has led to the development of no-tillage systems, where tillage is confined to soil disturbance associated with crop seeding or planting. Soil tillage is also involved in soil-water conservation and regulation, and in soil-erosion control. Table 1 outlines some of the various goals of tillage in agricultural systems.

The role of tillage in soil structure improvement has various facets related to crop growth and productivity. Tillage is conducted to improve soil functions such as water and air regulation and flow, to enhance the water-storage capacity of the soil, and to create a desirable aggregation size distribution conducive for crop seed–soil contact. Tillage is often needed to increase soil-water infiltration and thus to improve soil drainage. Soil loosening to ameliorate soil

Table 1 Specific goals of soil tillage in agricultural systems

Reasons for use of soil tillage	*Critical processes and characteristics*
Preparation of a seedbed	Creation of a soil aggregate size distribution to suit specific seed size
Weed control	Removal of unwanted plant species that compete with the crop for soil resources. Use of herbicides can remove the requirement for tillage
Soil erosion or degradation	Excess tillage can create a potential for soil erosion. Reduced tillage systems can retain crop residues that can help decrease the potential for both water and wind erosion
Conserve or regulate soil water	Tillage can conserve soil water content by manipulation of both soil structure and crop-residue retention. Excess soil water can also be removed by deep tillage
Improve plant root penetration and growth	Shallow tillage can remove soil-surface crusts that impede plant germination, while deep tillage can remove compact, impervious soil layers. Tillage can improve soil aeration
Incorporate crop residue, fertilizer and lime, and organic amendments	Tillage can bury or mix crop residues and incorporate fertilizer additions

compaction is another role of tillage. Tillage is also used to manipulate crop residues, either by burying and mixing them into the soil, or by retaining them at the soil surface to serve as mulch for soil protection or as a barrier to reduce soil and water movement.

The amount and degree of tillage needed are not the same for all soil types and cropping situations. Some soils, due to climate (e.g., cool, wet soils) or type (e.g., high clay or sand content, or poor permeability), have a relatively high 'tillage requirement.' A combination of easily compactable soils and excessive vehicular traffic can lead to compacted subsoils or soil layers that present the need for deep tillage. In contrast, some soils have a relatively low tillage requirement, as they are able to regenerate soil structure under natural processes associated with soil wetting and drying, and freeze–thaw processes. To meet the goals of soil tillage, agricultural engineers have developed various tools and implements which have certain advantages and limitations.

Main Types of Tillage Tools

Moldboard Plow

The term 'moldboard plow' describes an implement that cuts soil, lifts it, and turns it at least partly upside down by means of a curved plate, or moldboard (Figure 1). The concept of the moldboard plow is quite ancient. Wooden plows have been in existence in Asia and Africa for more than 5000 years and adapted versions were in use in Europe 500 years ago, featuring a drawbar for animals, wheels, a leading cutting-coulter, the soil-cutting blade, and a moldboard.

The precursor to the modern moldboard plow was invented by Thomas Jefferson, of Virginia, USA, around 1790, after he had observed European plow designs. He first designed and tested a wooden moldboard that could be duplicated easily. By 1814 he had them cast in iron, and they soon became known throughout North America and Western Europe.

In 1837 John Deere, of Vermont, USA, invented the modern moldboard plow, in Grand Detour, Illinois, using smooth, self-cleaning steel for the moldboard rather than cast iron. By 1847 his company was manufacturing more than 1000 plows per year, and his Moline Plow Works factory was producing 75 000 per year by 1875. This basic design is still being manufactured today, although with numerous improvements.

Applications The moldboard plow performs the following tillage operations and soil-conditioning functions:

- It cuts, lifts, breaks up, and loosens soil that has been compacted through machinery traffic or natural causes to a depth of usually 100–200 mm below the soil surface. Plowed soil is easier to form into a loose seedbed, drains water better, and warms up more quickly after winter;
- It overturns vegetation, buries it, and exposes topsoil. This is useful when developing new arable fields in grasslands or burying crop residues, weeds, and insect pests;
- It has a flat bottom that covers the entire width of the tillage zone, and thus all soil down to the depth of tillage is cut, lifted, and loosened.

Disadvantages and limitations The action of burying most surface vegetation leaves the soil surface much more vulnerable to erosion by wind and water, and leads to an increased loss of soil and fertilizer resources. Furthermore the eroded soil material

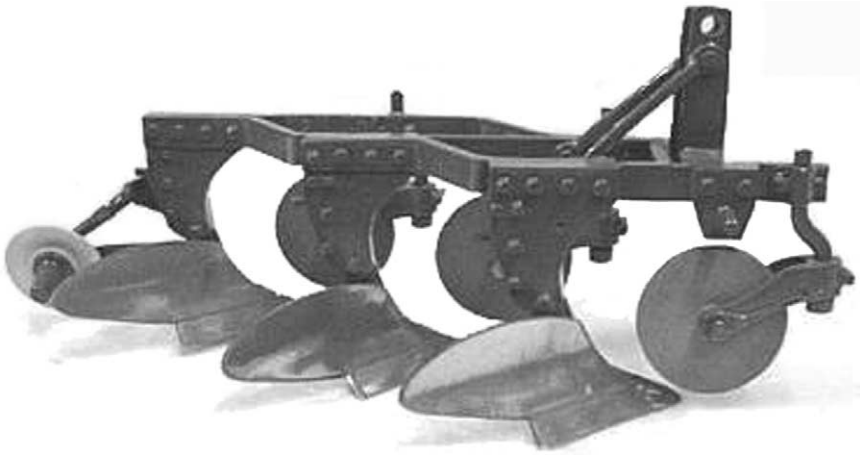


Figure 1 A three-bottom moldboard plow.

and nutrients cause pollution in their recipient surface watercourses. The sealing of the soil surface by impacting rainfall and the resulting reduced water infiltration can be a problem. In addition, continuous plowing can destroy the soil aggregate structure considerably, leading to reduced water and air movement, root growth, and crop yields. The moldboard plow requires the most energy to operate per unit field area of all tillage implements, with the exception of the deep subsoiler.

As a result of these disadvantages, the regular use of the moldboard has been reduced considerably in most of the world over the past 20 years. For example, the US Department of Agriculture (USDA) Agricultural Research Service has estimated that the use of the moldboard plow by American farmers had decreased from 75–85% in 1980 to less than 10% in 1993. Problems of weed and insect pests are offset in the absence of this plow by the use of pesticides or cover crops. Abandoning the moldboard plow has also been reported to result in an increase in the accumulation of organic matter in soils. This implement remains in use, however, in northern cool and humid climates, where the enhanced soil drainage and spring warming outweigh the disadvantages.

Chisel Plow

A basic chisel plow (Figure 2) was used for agricultural tillage at least 5000 years ago in Northern Africa and the Middle East. Wooden chisel ploughs were replaced by iron ones in the 1700s, and by smooth steel in the 1800s. Most chisel ploughs have a shank or leg width of 63.5 mm and an angle at the bottom tip of 40–60° to the horizontal.

Applications The chisel plow also cuts, lifts, and loosens soil prior to the preparation of a suitable seedbed, but differs from the moldboard plow in the following ways:

- Individual chisels are usually spaced at approximately 300 mm laterally, and their width is approximately 63.5 mm, so not all the soil to the tillage depth is loosened. The cross-section of soil loosened is V-shaped, with greater width at the soil surface; therefore there is a resulting lack of homogeneity in the tilled soil profile;
- The chisel plow leaves much more crop residue or stubble on the soil surface, usually four or five times more depending on the crop. This has the advantage of reducing soil erosion from wind and water by 40–80%, decreasing the sealing of the soil surface by rainfall, decreasing the amount of soil moisture that can be lost in a dry spring season, and increasing the organic matter content of the soil surface layer.

Disk Harrow

The disk harrow (Figure 3), consisting of a set of curved, circular rotating disks attached to a frame, has been in use for over 100 years. This implement can be used for secondary tillage, following moldboard or chisel plowing, to smooth the soil surface and reduce the size of aggregates in preparation for planting. It can also be used as the sole tillage implement when the goal is to chop crop residues to facilitate their decomposition and the subsequent planting of crops.

Because of its reversed angle of attack on the soil, each disk relies on its weight to penetrate vegetation and soil. As a result, disks do not lift and loosen soil

Figure 2 A chisel plow.

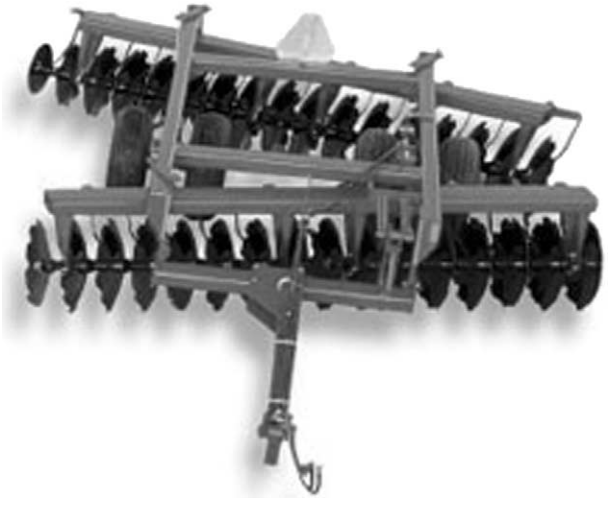


Figure 3 An offset disk harrow. (Photo courtesy of Case IH Corporation.)

much, and, in soils that are subject to compaction, the disk when used alone can produce an excessively dense topsoil layer, with resulting problems in infiltration, drainage, air movement, and root growth.

Toothed Cultivator

The toothed cultivator shank or leg, usually made from flexible spring steel, has the same shape as the chisel plow, except that it is smaller and operates at less depth, typically 20–50 mm (Figure 4). In most cases a shovel-shaped point is fixed to the bottom of each shank, with a width of 50–100 mm and a low attack angle of 20–30° to the horizontal. The cultivator removes weeds from between rows of crops and loosens the top layer of soil.

Subsoiler

The subsoiler is so named because it cuts and loosens soil below the normal tillage depth of 100–200 mm. Its shape is similar to the chisel plow except that it is made with a stronger shank or leg in order to resist the higher force required to till soil at greater depth. The bottom cutting tip can be narrow, 25–50 mm, or a low-angle sweep or foot can be attached as shown in Figure 5. The sweep has the advantage of loosening a greater volume of soil at the expense of only a small increase in the required pulling force or draft.

Use of a subsoiler is indicated when there is a dense and/or hard layer of soil below the normal tillage depth. Such a layer can result from physical causes such as machinery traffic or from physicochemical sources that produce cementation among soil aggregates. Such a hard layer, sometimes called a plow pan, can lead to poor water infiltration and drainage, reduced soil aeration, and diminished growth of crop roots.

No-tillage

No-till or no-tillage agriculture is conducted without the use of any of the tillage implements described above in most years, although occasional tillage may be required to control weeds that are tolerant to

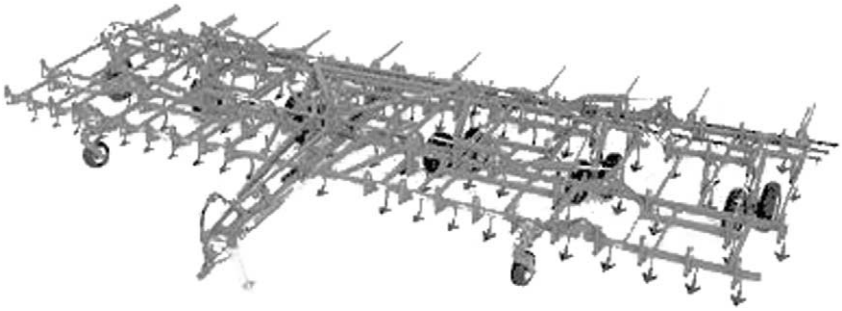


Figure 4 A spring-toothed cultivator.

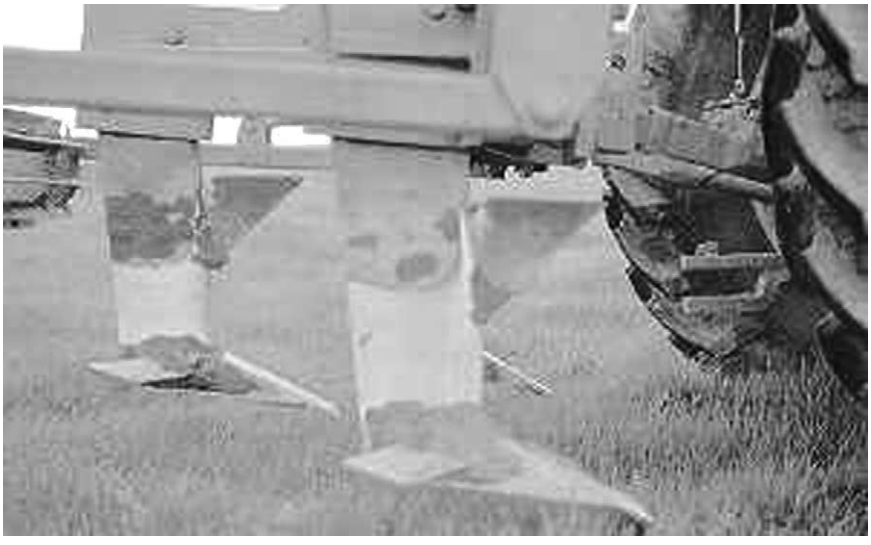


Figure 5 A three-shanked subsoiler. (Photo by K C Cameron, Lincoln University, Canterbury, New Zealand.)

herbicides. Crop seeds are planted directly into the stubble and residue of the previous year's crop. A special planter is generally required to penetrate residual vegetation and a stronger soil surface, and many manufacturers have developed effective no-till planters in the past 30 years. This production system has gained favor on many thousands of farms, especially in the plains of North America, where temperatures and rainfall are moderate.

No-till agriculture can have the following advantageous results:

- Less compression and breaking down of soil aggregates;
- Greater amounts of organic residue on and near the soil surface, resulting in less surface-sealing by rainfall, reduced soil erosion, enhanced moisture retention, and more organic matter accumulation;
- Similar crop yields to those of conventional-till and reduced-till systems, and even higher yields in relatively dry regions such as midwestern North America.

Basic Tillage Operations

Traditionally, most tillage operations consist of primary tillage followed by secondary tillage. This traditional combination is referred to as 'conventional tillage,' although this term is also applied to any tillage system that has been adopted for a period of time. Primary tillage usually consists of relatively deep (15–25 cm) tillage, using a moldboard or chisel plow. In contrast, secondary tillage is relatively shallow (10–15 cm) and used to pulverize and consolidate the soil to form a suitable seedbed. Both primary and secondary tillage are usually conducted uniformly across a field. The last few decades of the twentieth century have seen concerted efforts to reduce the amount and degree of tillage by adoption of reduced, minimum, or shallow tillage. Reduced tillage allows a reduction in depth, degree, and frequency of tillage, while 'minimum tillage' refers to the minimum soil manipulation necessary for crop production or other soil tillage requirement. Shallow tillage, sometimes called 'noninversion tillage,' involves restricting tillage to a shallow (less than 15 cm) soil depth. Usually, the soil is not turned over or inverted, thus the tillage operation is often termed 'ploughless tillage.' Overall, the main impetus behind such developments is to reduce the degree of soil degradation often associated with excess tillage, and also to reduce tillage costs in commercial agriculture.

In many cases, uniform tillage across a field may not be required or desirable. For row crops such as maize (*Zea mays* L.) and potato (*Solanum tuberosum* L.), tillage can be restricted to zones across a field. Partial-width tillage is based on the concept that seedbed conditions can be confined to the planting row, while interrow areas receive less or a different type of tillage treatment. For example, in semiarid climates, for maize production, tillage can be used to create a 'planting zone,' characterized by optimum seedbed conditions conducive to seed germination and establishment, while in the interrow zone a coarse soil structure is maintained to allow optimum water intake. Further to this, crop residue cover can be maintained in the interrow zone to facilitate water and soil conservation.

Basic soil tillage operations consist of inverting, loosening, mixing, and pulverizing, or clod-breaking (Table 2). Specific tillage tools operating at a range of soil depths are used to achieve these operations. Soil inversion is best accomplished using a moldboard plow, while soil-loosening can be achieved using several tillage tools. Rotary cultivators are well suited for mixing soil and thus are often used to mix soil fertilizer and organic amendments into the soil. Plant residues can often pose a problem for tillage tools in regard to interference with the tillage operation. Under such conditions, coulters or disks are often used in combination with the tillage tool to cut or divert crop residue away from the immediate tillage zone.

Types of Tillage Systems

In many cropping situations and climates, innovations in tillage practices lead to the development of conservation tillage and other closely related systems. 'Conservation tillage' is a generic term used to describe tillage systems that have the potential to

Table 2 Main role and function of tillage implements and their efficiency

Tillage tool	*Tillage depth (cm)*	*Inverting*	*Loosening*	*Mixing*	*Clod breaking*
Moldboard plow	15–25	Good	Good	Partial	None
Chisel plow	10–20	Partial	Good	Partial	Partial
Disk plow	10	Partial	Good	Partial	Partial
Rotary cultivator	10	None	Partial	Good	Good
No tillage	5	None	None	None	None

conserve soil and water by reducing their loss relative to some form of conventional tillage. Generally, there are four main types of conservation tillage: mulch tillage, ridge tillage, strip tillage, and no-tillage. Precise definitions of conservation tillage are only possible within the context of known crop species, soil types and conditions, and climates. A well-accepted operational definition of conservation tillage is tillage or a tillage and planting combination which retains 30% or greater cover of crop residue on the soil surface. Combinations of tillage tools can be developed to meet the various requirements posed by both soil and crop needs, and the constraints related to farming systems and climate.

List of Technical Nomenclature

Conventional tillage Combined primary and secondary tillage operations normally performed in preparing a seedbed for a given crop and area

Deep tillage A primary tillage operation that manipulates soil to a greater depth than normal plowing (i.e., more than 100–200 mm; e.g., heavy-duty deep moldboard plow, deep disk plow, heavy chisel, or subsoiler)

Inversion tillage Primary tillage that 'inverts' or 'turns over' soil, causing much soil mixing. The main tillage implement used for inversion tillage is the moldboard plow

Land-forming tillage Tillage operations that move soil in order to create the desired configurations, including level ground, slopes for irrigation, contouring and terracing, small-scale ridging, or pitting

Minimum tillage The minimum use of primary and secondary tillage necessary to meet crop-production requirements under existing soil and climatic conditions, usually resulting in fewer tillage operations than for conventional tillage

Primary tillage The tillage that constitutes the first major soil-working operations, normally designed to reduce soil strength, cover plant and insect materials, and rearrange soil aggregates (e.g., by moldboard plow, chisel plow, and/or disk plow)

Reduced tillage Reduction in total number of primary and secondary tillage operations usually used in conventional tillage, to prepare a soil for crop establishment; combining the primary tillage operation with special planting operations in order to reduce or eliminate secondary tillage operations

Secondary tillage Any of a group of different tillage operations, following primary tillage, that are designed to create refined soil conditions before seed planting. Examples include disk harrow, cultivator chisels or sweeps, and roller harrow

Shallow or non-inversion tillage Primary tillage confined to shallow soil depth (less than 15 cm); absence of soil inversion tillage

Soil cultivation Shallow tillage operations to create improved soil aeration, water infiltration, moisture conservation, a level surface, or to control weeds to promote the growth of crop plants

See also: **Carbon Emissions and Sequestration**; **Compaction**; **Crop-Residue Management**; **Drainage, Surface and Subsurface**; **Germination and Seedling Establishment**; **Shifting Cultivation**; **Structure**; **Tilth**; **Weed Management**; **Zone Tillage**

Further Reading

Cannell RQ and Hawes JD (1994) Trends in tillage practices in relation to sustainable crop production with special reference to temperate climates. *Soil Tillage Research* 30: 245–282.

Carter MR (ed.) (1994) *Conservation Tillage in Temperate Agroecosystems*. Boca Raton, FL: CRC Press.

Gill WR and Vanden Berg GE (1968) *Soil Dynamics in Tillage and Traction*. Washington, DC: US Government Printing Office.

Gregorich EG, Turchenek LW, Carter MR, and Angers DA (2001) *Soil and Environmental Science Dictionary*. Boca Raton, FL: CRC Press.

Hillel D (1980) *Applications of Soil Physics*. New York: Academic Press.

Koolen AJ and Kuipers H (1983) *Agricultural Soil Mechanics*. Berlin, Germany: Springer-Verlag.

Kuipers H (1963) The objectives of soil tillage. *Netherlands Journal Agricultural Science* 11: 91–96.

Lal R (1991) Tillage and agricultural sustainability. *Soil Tillage Research* 20: 133–146.

Larson WE (1964) Soil parameters for evaluating tillage needs and operations. *Soil Science Society of American Proceedings* 28: 119–122.

McKyes E (1985) *Soil Cutting and Tillage*. Amsterdam, The Netherlands: Elsevier.

Spoor G (1975) Fundamental aspects of cultivations. In: *Soil Physical Conditions and Crop Production*, pp. 128–144. London, UK: HMSO.

Unger PW and Van Doren DM Jr (eds) (1982) *Predicting Tillage Effects on Soil Physical Properties and Processes*. ASA Special Publication 44. Madison, WI: American Society of Agronomy.

D

DARCY'S LAW

D Swartzendruber, University of Nebraska-Lincoln, Lincoln, NE, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

In a porous medium consisting of interconnected pores amid the solid particles, one of the foremost physical properties is the liquid content within the network of pores. If the porous medium is soil and the liquid is water, the implications are crucial not only for the growth of plants, but also for the land phase of the hydrologic cycle (e.g., water infiltration, drainage, and water movement to plant roots) and many other aspects of soil physics. The soil water is also a liquid harbor for dissolved or suspended substances, thus enabling pollutants to move through or to accumulate in the soil. In still another context, the water content can have a major influence on the strength of the soil mass in its ability to resist compaction in agricultural soils and to provide foundational support for buildings, bridges, and earthworks such as highways and dams, these being of prime concern in the field of soil mechanics.

For the quantitative study of the soil water content as it changes within the soil mass as a function of position and time, to have basic flow relationships in mathematical form of reasonable simplicity is of great utility for analysis and measurement. The simple example is provided by a water-saturated porous medium, and the more complicated example by an unsaturated porous medium – in which the soil pores are not completely filled with water. This is the more common condition for soils generally, except for wetlands and some lowland soils subject to high water tables or frequent flooding.

Water-Saturated Soil Conditions

The mathematical–physical foundation for analytical descriptions of liquid flow through a porous medium was laid by Henry P.G. Darcy in 1856, in an appendix to a major and famous treatise setting forth a comprehensive design for the municipal water system of the city of Dijon, France. The porous medium was sand, the liquid was water from the system of a local hospital, and the experiments were conducted by two engineers, Ritter and Baumgarten. Darcy was a highly regarded engineer and an esteemed public-spirited citizen. He died in 1858, not long after the publication of his book.

In simplest terms, Darcy's classic law, or equation, states that the water flux q in one-dimensional flow is directly proportional to the driving hydraulic gradient i, or mathematically in scalar form:

$$q = Ki \qquad [1]$$

where the proportionality constant K is called the hydraulic conductivity of the porous medium or soil and is recognized as a composite property of both the porous medium (soil) and the flowing liquid (water). Referring to Figure 1, where A is the constant bulk cross-sectional area of the soil column and Q is the volumetric time rate of water flow perpendicular to A, then $q = Q/A$. Let h_1 and h_2, both being measured from the same but arbitrary datum plane, be the respective hydraulic heads (in equivalent height of

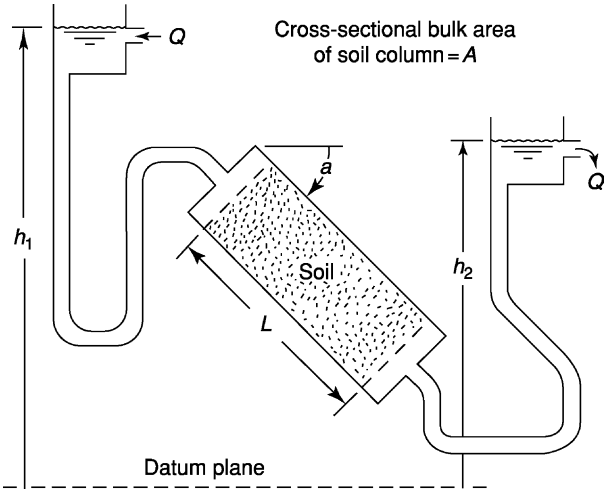


Figure 1 A water-saturated soil column to illustrate Darcy's equation for one-dimensional saturated flow.

water column) at the inlet and outlet ends of the soil column of constant bulk length L, assuming no loss of head in the tubings connecting the inlet and outlet ends to the reservoirs. Then, $i = (h_1 - h_2)/L = -\Delta h/L$. Substituting these expressions for q and i back into eqn [1] and rearranging yields:

$$Q = -KA\Delta h/L \qquad [2]$$

which is essentially the form employed by Darcy. The definition of $\Delta h = h_2 - h_1$ is the standard one of mathematics.

Some textbooks have presented separate diagrams of each of three special angular orientations of the flow column ($a = \pi/2$, 0, and $-\pi/2$ in **Figure 1**), with eqn [2] written for each orientation, but with h_1 and h_2 expressed as sums of components (i.e., pressure head and elevation head). Unfortunately, these diagrams may foster an illusory impression of Q being affected by the direction of flow. We therefore examine these three angular orientations as follows. Assume in **Figure 1** that there is sufficient length and flexibility of the connecting tubing from the inlet and outlet water-level reservoirs, so that Δh can remain fixed (in sign and magnitude) for any angle a between $-\pi/2$ and $\pi/2$. Then, for the first extreme of $a = \pi/2$, the direction of water flow through the now-vertical soil column is vertically upward, maximally counterdirectional to gravity. For the intermediate $a = 0$, the water-flow direction through the now-horizontal soil column is horizontal in the positive direction (left to right), independent of gravity. For the second extreme of $a = -\pi/2$, the water-flow direction through the once-again-vertical soil column is vertically downward, maximally codirectional with gravity, and was the column orientation used in Darcy's experiments. Noting Q for each of these three a values with Δh unchanged, observe from eqn [2] that Q will be the same for all three orientations, because K, A, Δh, and L remain the same. This same Q would also hold for all values of a such that $-\pi/2 < a < \pi/2$. Thus, for a given soil column (fixed K, A, and L), Q only changes with Δh, so that Δh is the sole determiner of Q regardless of flow direction due to soil-column orientation. Hence, there is no need to present the three separate orientation diagrams and their corresponding equations for Q, because **Figure 1** and eqn [2] encompass them all.

On the basis of viscous fluid flow through small-bore cylindrical tubes or between flat parallel plates, the flux of fluid is ideally proportional to the density ρ of the fluid and inversely proportional to its absolute viscosity η. This suggests the possibility of partitioning the K of eqns [1] and [2] by:

$$K = k\rho g/\eta \qquad [3]$$

where g is the acceleration of gravity, and k is a constant independent of the flowing fluid and hence dependent only on the pore system of the soil. The name for k is either 'intrinsic permeability' or merely 'permeability' if there is no likelihood of confusion with K. In principle, the permeability k should be independent of the fluid used to determine it, provided that no interaction occurs between the fluid and the soil particles to change the internal geometry of the soil-pore system. In practice, this seems to hold reasonably well for sand and polar (e.g., water) or nonpolar fluids and air. But for soil containing swelling colloids or clays, with the fluid being water alone or water with dissolved electrolytes, the strong fluid–soil interaction can alter the internal geometry of the soil-pore system, so that k will not be independent of the flowing fluid.

Validity of Darcy's Equation

Darcy himself was well aware that eqn [1] would not hold for values of i and q large enough to create and be affected by inertia and turbulence effects in the flowing fluid. Such effects would cause the flux–gradient curve to depart from proportionality by bending toward the gradient axis, meaning that the flux would increase less than proportionally with gradient. But unless the soil particles were unusually large (e.g., gravel or very coarse sand), it was generally assumed that such less-than-proportional effects would seldom be encountered.

As steadily increasing attention was being focused on Darcy's equation in soil physics and hydrology up through the first half of the twentieth century, there seemed to have been but little inhibition against accepting an extension of Darcian validity from sand to soil. During the decades of 1952–1972, however, several extensive sets of experimental data were published that did not adhere to Darcian proportionality. Moreover, these flux–gradient departures from proportionality were the very opposite of the less-than-proportional behavior at large gradients as just discussed, and occurred in clays and clay-bearing porous sandstones. As shown in **Figure 2** for such a sandstone, the experimental points of q versus i curved upward in a more-than-proportional manner, as described rather precisely by:

$$q = B\{i - J[1 - exp\,(-Ci)]\} \qquad [4]$$

where B, J, and C are the characterizing constants. As i becomes large, the term exp $(-Ci)$ vanishes and eqn [4] reduces to the linear form:

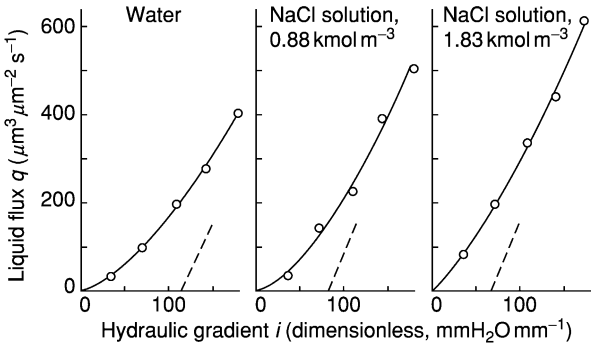


Figure 2 Flux–gradient values and relationships for clay-bearing sandstone sample No. 15L of Von Engelhardt and Tunn, published in 1954/1955. Each solid-line curve is eqn [4] fitted to its data set, with each short, broken straight line being the initial portion of the linear asymptote, eqn [5], for each solid-line curve.

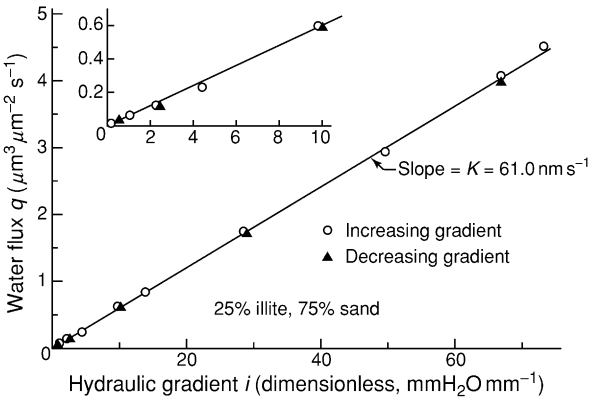


Figure 3 Experimental values of water flux versus hydraulic gradient for a mixture of one part illite in three parts fine quartz sand. Reproduced from Russell DA and Swartzendruber D (1971) *Soil Science Society of America Proceedings* 35: 25 with permission.

$$q = B\{i - J\} \qquad [5]$$

which is the linear asymptote approached at large i by the generally curvilinear eqn [4], where J is the i-intercept when the linear asymptote is extended back to the i-axis. The broken straight lines in Figure 2 are the linear asymptotes in the vicinity of the i-intercepts (or J-values). Note also that the flow behavior at large gradient, although linear, is still not Darcian, because eqn [5] does not pass through the origin as does eqn [1].

If $J = 0$ or $C = 0$, then eqn [4] reduces to $q = Bi$, which is simply Darcy's equation with B replacing K in eqn [1]. Hence, B has the character of hydraulic conductivity and could be partitioned in the manner of eqn [3] if desired, using the constant b to replace k. Again, b would have the character of permeability, and would equal k for Darcian flow ($J = 0$ or $C = 0$). A measure of non-Darcian behavior, N_d, for J and C generally, has been defined by:

$$N_d = J^2 C \qquad [6]$$

in which $N_d = 0$ for Darcian behavior ($J = 0$ or $C = 0$). The values of N_d for the three subgraphs in Figure 2 are 104.6, 73.3, and 35.8 mmH$_2$O mm^{-1}, in progressing from water (zero electrolyte) through the two successively higher concentrations of salt (NaCl). Clearly, N_d is maximal for water alone and decreases steadily with increasing salt concentration, although Darcian behavior is not completely restored ($N_d = 0$) even at the largest concentration of salt.

The most extreme expression of non-Darcian behavior would occur if there were an initial gradient range over which $q = 0$, starting at the origin and extending to some finite, nonzero value i_0 called the threshold gradient, after which $q > 0$ for $i > i_0$. This would imply the soil water initially to behave like a solid until the threshold gradient were exceeded. This drastic idea has not been well received by soil physicists and soil hydrologists. An i_0 value as large as 65.2 mmH$_2$O mm^{-1} has been reported for water in a sodium-saturated montmorillonite clay. Note that i_0 must not be confused with the gradient intercept J of eqn [5].

To account for the curvature of the graphs of Figure 2, the original postulate was that the pore water in the vicinity of the clay surfaces was changed from a Newtonian to a non-Newtonian viscosity by forces emanating from these clay surfaces. This would also have explained the effect of salt, inasmuch as increasing salt concentration is known to break down the water structure.

To obtain further and refined experimental data, a special technique, with highly sensitive flowmeters and differential manometers, yielded very rapid, simultaneous measurements of q and Δh (and hence i) for mixtures of fine quartz sand, quartz silt, kaolinite, illite, and montmorillonite. For the various mixtures of sand, silt, and kaolinite or illite, the results are typified in Figure 3, which is unequivocally Darcian for both the main graph and the small, expanded-scale inset graph. Also, the decreasing-gradient points, obtained immediately after those for increasing gradient, replicated very well. Neither is there any hint of a threshold gradient, nor of curvature in the flux–gradient relationship.

In contrast, however, if the mixture of sand, silt, and clay is 5% montmorillonite, the results are rather different, as displayed in Figure 4. With generally good agreement between increasing and decreasing gradients, the non-Darcian behavior of a sigmoidal character is obviously more complex than the monotonic curves of Figure 2; the modified-water postulate

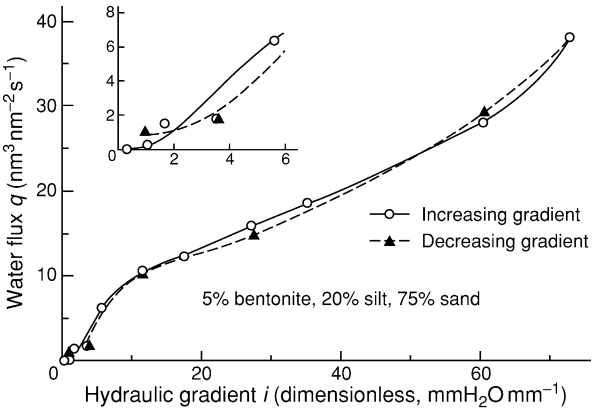


Figure 4 Experimental values of water flux versus hydraulic gradient for a mixture of 5% montmorillonite (bentonite) with 20% quartz silt and 75% fine quartz sand. Reproduced from Russell DA and Swartzendruber D (1971) *Soil Science Society of America Proceedings* 35: 25 with permission.

could therefore not apply. Lastly, in the inset graph (Figure 4), the first point is suggestive of a threshold gradient.

The refined experimentation indicates Darcy's equation to be applicable with confidence to sands, silts, nonswelling clays, and mixtures of these, but that Darcian behavior in the presence of swelling clays is much more problematic. Admittedly, a straight line through the origin could be drawn through the experimental points of Figure 4 as a first approximation, but doing this would seem less applicable to the data in Figure 2. Ultimately, however, in a pragmatic or field setting, the determining factor would become the error attendant to measuring K. If such error is much greater than the errors associated with approximating the data (Figures 2 and 4) with proportional lines, then it would be unreasonable to deny the use of Darcy's equation in such a case. Conversely, nonetheless, the soil physicist or hydrologist should not lose sight of the non-Darcian manifestations observed in swelling clays.

Solving Saturated-Flow Problems

To describe flow in the three cartesian dimensions, the components of water flux in the mutually perpendicular x, y, and z directions are employed. For the x direction:

$$q_x = -K\partial h/\partial x \qquad [7]$$

with K considered constant. The equations for q_y and q_z are just like eqn [7], but with x replaced by y and z, respectively. The use of the same K for all three dimensions (x, y, and z) means that the soil is assumed to be isotropic, in distinction from the less-simple anisotropic case, wherein K is not the same for the three dimensions.

The conservation of mass of an incompressible fluid, also called the equation of continuity, is:

$$\frac{\partial q_x}{\partial x} + \frac{\partial q_y}{\partial y} + \frac{\partial q_z}{\partial z} = -\frac{\partial f}{\partial t} \qquad [8]$$

where f is the total porosity of the soil. Taking the soil to be incompressible yields $\partial f/\partial t = 0$, and introducing the q_x from eqn [7] along with the similar equations for q_y and q_z into eqn [8] yields finally:

$$\frac{\partial^2 h}{\partial x^2} + \frac{\partial^2 h}{\partial y^2} + \frac{\partial^2 h}{\partial z^2} = 0 \qquad [9]$$

which is the classic Laplace equation, for which the shortened 'del squared' operator notation, $\nabla^2(\bullet)$, is used to write eqn [9] as:

$$\nabla^2 h = 0 \qquad [10]$$

There are an infinite number of solutions to Laplace's equation, but only when a given solution also satisfies the boundary conditions can it be claimed as the unique solution to the problem at hand. Also, the particular form of $\nabla^2(\bullet)$ depends on the coordinate system.

The simple, proportional form of Darcy's equation takes its place alongside Ohm's law of electrical flow, the Newton–Fourier heat-flux equation, and Ficks' first law of diffusion. Solutions from these other fields are thus available for scrutiny and potential use or modification in saturated water-flow problems. Historically, this has been an advantage and motivation for preferring Darcy's equation. Clearly, if one attempted to use eqn [4] in eqn [8], the result would be far more complicated and difficult to use than eqn [9] and would probably still remain difficult even with modern computer capabilities.

Water in Unsaturated Soils

When the total porosity f of the soil is no longer completely occupied with water, the soil is said to be unsaturated. This means that part of the pore space is then occupied by air and, consequently, that liquid water flow can only take place through the remaining porosity that contains liquid water. We therefore expect the K of the unsaturated soil to be less than the saturated K. Defining θ to be the volumetric water content of the soil, we formulate this first aspect of unsaturation by taking K to be a function of θ alone, $K = K(\theta)$, rather than a constant as in eqn [7]. Note that if $\theta = f$, $K = K(f)$ is simply the saturated K once again.

The second aspect of unsaturation is that the reduction of θ from f creates an attraction of the unsaturated soil for water, akin to a blotter that 'sucks up' ink. This means that the hydraulic head h is also affected by θ being less than f. Therefore, we take $h = z - \tau(\theta)$, where the position coordinate z is called the elevation head as measured from the datum plane placed at the xy plane, and $\tau = \tau(\theta)$ is called the soil-water suction-head function. Because $\tau(\theta)$ is different for wetting of soil than for drying of soil – a phenomenon known as hysteresis – the use of a given $\tau(\theta)$ must be restricted to monotonic changes in water content. In principle, $K(\theta)$ may also be envisaged as subject to hysteresis, but its effect is usually negligible in practice. Finally, the functions $K(\theta)$ and $\tau(\theta)$ have been called the Buckingham functions in honor of their originator. Buckingham's discussion and explanations of almost a century ago, which we have followed here, are so instructive and apt that they can scarcely be improved upon today, and thus stand as a remarkable and striking example of prescience.

Incorporation of $K = K(\theta)$ and $h = z - \tau(\theta)$ into eqn [7] yields:

$$q_x = K(\theta)\frac{\partial \tau}{\partial x} \qquad [11]$$

where $\partial\tau/\partial x$ is the suction-head gradient in the x direction. Expanding $\partial\tau/\partial x$ by making use of $\tau(\theta)$, eqn [11] can be written for homogeneous soil as:

$$q_x = -K(\theta)\left[-\frac{d\tau}{d\theta}\right]\frac{\partial \theta}{\partial x} \qquad [12]$$

Now $d\tau/d\theta$ is also a function of θ, since τ is a function of θ. Therefore, since $K(\theta)$ is obviously a function of θ, the composite multiplier of $\partial\theta/\partial x$ in eqn [12] is also a function only of θ, which we label $D(\theta)$, setting $D(\theta) = K(\theta)[-d\tau/d\theta]$, so that eqn [12] becomes:

$$q_x = -D(\theta)\frac{\partial \theta}{\partial x} \qquad [13]$$

$D(\theta)$ is called the soil-water diffusivity function, and eqn [13] expresses the horizontal water flux in terms of a response to a water-content gradient, $-\partial\theta/\partial x$. The physical process, however, is not one of diffusion, since nothing has been done to alter the hydrodynamic basis of eqn [7], from which the derivation began. With the same analysis conducted for q_y, the results will be the same at every stage corresponding to eqns [11–13], except that y replaces x. For q_z, however, the results are different, because $\partial z/\partial z = 1$ and not zero. The counterpart of eqn [11] is:

$$q_z = K(\theta)\frac{\partial \tau}{\partial z} - K(\theta) \qquad [14]$$

and the counterpart of eqn [12] also has $K(\theta)$ subtracted from the right-hand side. The counterpart of eqn [13] is:

$$q_z = -D(\theta)\frac{\partial \theta}{\partial z} - K(\theta) \qquad [15]$$

Eqns [11–15] are all forms of the Buckingham–Darcy flux-gradient equation, which fills the same role for unsaturated soil as does Darcy's equation for saturated soil. For horizontal flux, only one soil-characterizing function is needed, either $K(\theta)$ of eqn [11] or $D(\theta)$ of eqn [13]. For vertical flux, the single soil-characterizing function $K(\theta)$ employed in eqn [14] is sufficient, whereas if $D(\theta)$ is employed in eqn [15] it is still necessary to use $K(\theta)$ as well. Experimentally, however, it may be easier to measure θ and its gradient than τ and its gradient, especially at low values of θ, when τ values are very large.

Early experimental tests that successfully validated the Buckingham–Darcy equation were carried out, ironically enough, on sandy soils or materials, reminiscent of Darcy-equation verification on sands rather than soils. In the early 1960s, however, some very precise gamma-ray measurements of horizontally transient θ values were reported for a nonswelling silty clay loam. These data were amenable for a test of eqn [13]. To visualize this test, note in eqn [13] that, at a fixed value of θ, D is also fixed, so that a plot of q_x versus $(-\partial\theta/\partial x)$ is asserted to be a straight line through the origin with a slope equal to the fixed D. But, instead of just a single proportional line as in the saturated case illustrated in **Figure 3**, there will now be a family of proportional lines that constitute eqn [13], one proportional line for each value of fixed $D(\theta)$. The steepest slope will be for the largest (wettest) θ, with slopes becoming progressively less steep as the fixed θ is closer to air dryness. For the first three largest water contents in the silty clay loam, $\theta = 0.45$, 0.40, and 0.35 $mm^3\,mm^{-3}$, the experimental points fell on straight lines through the origin and with progressively decreasing slope, entirely in accord with eqn [13]. At $\theta = 0.30\,mm^3\,mm^{-3}$, however, the experimental points curved upward in more-than-proportional fashion similar to the curves of **Figure 2**, and this behavior accentuated progressively as θ decreased to 0.15 $mm^3\,mm^{-3}$. Hence, eqn [13] was obeyed for the wetter but not the drier regions of the soil.

An alternative explanation for these deviations was developed from the time-dependent premise that $K = K(\theta, t)$, $\tau = \tau(\theta, t)$, and $D = D(\theta, t)$, rather than $K = K(\theta)$, $\tau = \tau(\theta)$, and $D(\theta)$, as used here in deriving eqns [11–15]. This makes it less compelling to

invoke failure of the Buckingham–Darcy equation. Admittedly also, precise experiments in unsaturated flow are even more difficult than in saturated flow. Once again, then, general errors of measurement may sufficiently exceed those due to departures from the Buckingham–Darcy equation, so that its use in this pragmatic or practical sense is still permissible.

Solving Unsaturated-Flow Problems

The cartesian equation of continuity for unsaturated soil conditions is the same as eqn [8] except that the right-hand side is changed to $-\partial\theta/\partial t$. Using q_x from eqn [13], $q_y = -D(\theta)\partial\theta/\partial y$, and q_z from eqn [15], eqn [8] becomes:

$$\frac{\partial}{\partial x}\left[D(\theta)\frac{\partial\theta}{\partial x}\right] + \frac{\partial}{\partial y}\left[D(\theta)\frac{\partial\theta}{\partial y}\right] + \frac{\partial}{\partial z}\left[D(\theta)\frac{\partial\theta}{\partial z}\right] + \frac{\partial K(\theta)}{\partial z} = \frac{\partial\theta}{\partial t} \qquad [16]$$

An alternative form emerges if we note that, since $\tau = \tau(\theta)$, then conversely $\theta = \theta(\tau)$, whereupon $K(\theta) = K[\theta(\tau)]$ so that we can replace $K(\theta)$ with $K(\tau)$. Also, differentiation of $\theta = \theta(\tau)$ yields $\partial\theta/\partial t = (\mathrm{d}\theta/\mathrm{d}\tau)(\partial\tau/\partial t)$, and $(\mathrm{d}\theta/\mathrm{d}\tau)$ is a function of τ because of $\theta = \theta(\tau)$. Making use of these relationships along with q_x from eqn [11], $q_y = K(\tau)\partial\tau/\partial y$, and q_z from eqn [14], eqn [8] becomes:

$$\frac{\partial}{\partial x}\left[K(\tau)\frac{\partial\tau}{\partial x}\right] + \frac{\partial}{\partial y}\left[K(\tau)\frac{\partial\tau}{\partial y}\right] + \frac{\partial}{\partial z}\left[K(\tau)\frac{\partial\tau}{\partial z}\right] - \frac{\partial K(\tau)}{\partial z} = \frac{\mathrm{d}\theta}{\mathrm{d}\tau}\frac{\partial\tau}{\partial t} \qquad [17]$$

As before, the use of a single $K(\tau)$ in eqn [17] or of a single $D(\theta)$ in eqn [16] means that the unsaturated soil is taken to be isotropic in its water-transport properties.

Eqns [16] and [17] are both forms of the Richards equation, so named in honor of the originator of eqn [17]. Including eqn [16] is appropriate, however, because Richards explicitly mentioned that the choice of either $\tau = \tau(\theta)$ or $\theta = \theta(\tau)$ was a matter of mathematical expediency. The Richards equation, along with the necessary boundary and initial conditions, fills the same role as a problem-solving framework for unsaturated flow as does the Laplace equation (eqn [9]) for saturated flow. But the solutions to Richards' equation are much more difficult – recourse to numerical and computer solutions has been needed almost from the beginning.

List of Technical Nomenclature

η	Absolute viscosity of flowing water or fluid (eqn [3]) ($\mathrm{mN\,s\,m^{-2}} = \mathrm{g\,m^{-1}\,s^{-1}}$)
θ	Volume of water in a soil sample divided by the sample bulk volume, 'volumetric water content' ($\mathrm{mm^3\,mm^{-3}} = 1$)
π	3.1416 (radian)
ρ	Mass density of flowing water or fluid ($\mathrm{kg\,m^{-3}}$)
$\tau(\theta)$	The suction head created by the attraction of unsaturated soil for water and which is a function of water content θ ($\mathrm{mmH_2O}$)
$\nabla^2(\bullet)$	Cartesian Laplacian operator (left-hand side of eqn [9]) ($\mathrm{mmH_2O\,mm^{-2}}$)
A	Bulk cross-sectional area of soil column ($\mathrm{mm^2}$)
a	Angle of soil column orientation (**Figure 1**) (radian)
B	Non-Darcian hydraulic conductivity (eqn [4]) ($\mu\mathrm{m\,s^{-1}}$)
b	Non-Darcian permeability of a soil ($\mu\mathrm{m^2}$)
C	Non-Darcian parameter (eqn [4]) ($\mathrm{mm\,mmH_2O^{-1}}$)
$D(\theta)$	Soil-water diffusivity function (eqns [13] and [15]) ($\mathrm{mm^2\,s^{-1}}$)
f	Soil porosity: volume of pores in a sample of soil divided by the bulk volume of the sample ($\mathrm{mm^3\,mm^{-3}} = 1$)
g	Acceleration due to gravity ($\mathrm{m\,s^{-2}}$)
h	Height of water column sustained by the soil water at any point in the soil, measured above an arbitrary datum plane and called hydraulic head ($\mathrm{mmH_2O}$)
h_1, h_2	Inlet and outlet hydraulic heads, respectively, at the ends of the soil column (**Figure 1**) ($\mathrm{mmH_2O}$)
i	Reduction in hydraulic head per unit distance along the path of water flow, e.g., $-(h_2 - h_1)/L$, called 'hydraulic gradient' ($\mathrm{mmH_2O\,mm^{-1}}$)
i_0	Nonzero hydraulic gradient at and below which no detectable water flux (flow) occurs ($\mathrm{mmH_2O\,mm^{-1}}$)
J	Non-Darcian parameter (eqn [4]) ($\mathrm{mmH_2O\,mm^{-1}}$)

K	Darcian hydraulic conductivity (eqn [1]) ($\mu m\,s^{-1}$)
k	Darcian permeability of a soil (eqn [3]) (μm^2)
$K(\theta)$	Darcian hydraulic conductivity of an unsaturated soil as a function of water content ($\mu m\,s^{-1}$)
L	Bulk length of the soil column (**Figure 1**) (mm)
N_d	A measure of non-Darcian behavior (eqn [6]) ($mmH_2O\,mm^{-1}$)
Q	Volume of water (or fluid) flowing per unit time through the soil column (**Figure 1**) ($mm^3\,s^{-1}$)
q	Volume of water (or fluid) flowing per unit time through unit bulk cross-sectional area, Q/A, called water (or fluid) flux ($mm^3\,mm^{-2}\,s^{-1} = mm\,s^{-1}$)
q_x, q_y, q_z	Components of water (or fluid) flux in the three cartesian directions x, y, and z, respectively ($mm\,s^{-1}$)
t	Time after the beginning of water or fluid flow (s)
x, y, z	Cartesian position coordinates, respectively (mm)

See also: **Capillarity**; **Evaporation of Water from Bare Soil**; **Infiltration**; **Isotropy and Anisotropy**; **Porosity and Pore-Size Distribution**; **Thermodynamics of Soil Water**; **Vadose Zone:** Hydrologic Processes; **Water Potential**; **Water Retention and Characteristic Curve**; **Water, Properties**

Further Reading

Buckingham E (1907) *Studies on the Movement of Soil Moisture. US Department of Agriculture Bureau of Soils Bulletin 38*. Washington, DC: Government Publishing Office.

Darcy H (1856) *Les Fontaines Publique de la Ville de Dijon*. Paris, France: Victor Dalmont.

Guerrini IA and Swartzendruber D (1992) Soil water diffusivity as explicitly dependent on both time and water content. *Soil Science Society of America Journal* 56: 335–340.

Kutilek M (1965) Influence de l'interface sur la filtration de l'eau dans les sols. *Science du Sol* 1: 3–14.

Lutz JF and Kemper WD (1959) Intrinsic permeability of clay as affected by clay–water interaction. *Soil Science* 88: 83–90.

Miller RJ and Low PF (1963) Threshold gradient for water flow in clay systems. *Soil Science Society of America Proceedings* 27: 605–609.

Olson TC and Swartzendruber D (1968) Velocity–gradient relationships for steady-state unsaturated flow of water in nonswelling artificial soils. *Soil Science Society of America Proceedings* 32: 457–462.

Philip JR (1995) Desperately seeking Darcy in Dijon. *Soil Science Society of America Journal* 59: 319–324.

Richards LA (1931) Capillary conduction of liquids through porous mediums. *Physics (NY)* 1: 318–333.

Russell DA and Swartzendruber D (1971) Flux–gradient relationships for saturated flow of water through mixtures of sand, silt, and clay. *Soil Science Society of America Proceedings* 35: 21–26.

Swartzendruber D (1962a) Modification of Darcy's law for the flow of water in soils. *Soil Science* 93: 22–29.

Swartzendruber D (1962b) Non-Darcy flow behavior in liquid-saturated porous media. *Journal of Geophysical Research* 67: 5205–5213.

Swartzendruber D (1963) Non-Darcy behavior and the flow of water in unsaturated soils. *Soil Science Society of America Proceedings* 27: 491–495.

Swartzendruber D (1968) The applicability of Darcy's law. *Soil Science Society of America Proceedings* 32: 11–18.

Swartzendruber D (1969) The flow of water in unsaturated soils. In: de Wiest RJM (ed.) *Flow Through Porous Media*, pp. 215–292. New York: Academic Press.

Von Engelhardt W and Tunn WLM (1955) *The Flow of Fluids Through Sandstones* (translated by Witherspoon PA from *Heidelberger Beträge zur Mineralogie und Petrographie* 2: 2–25, 1954). *Illinois State Geological Survey* Circular 194.

DEGRADATION

C J Ritsema, Alterra, Wageningen, The Netherlands
G W J van Lynden, ISRIC, Wageningen, The Netherlands
V G Jetten and S M de Jong, Utrecht University, Utrecht, The Netherlands

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Soil is under increasing threat from a wide range of human activities that are undermining its long-term availability and viability. One third of the world's agricultural soils, or approximately 2 billion hectares of land are affected by soil degradation. Water and wind erosion account for most of the observed damage, while other forms such as physical and chemical degradation are responsible for the rest. Appropriate soil and water conservation strategies are needed to prevent and combat the effects of soil degradation in the field and at the planning level.

Soil degradation is "a process that describes human-induced phenomena which lower the current and/or future capacity of the soil to support human life." In a general sense, soil degradation could be described as the deterioration of soil quality, or in other words: the partial or entire loss of one or more functions of the soil. Quality should be assessed in terms of the different potential functions of the soil.

Land degradation is the reduction in the capability of the land to produce benefits from a particular land use under a specified form of land management. Seven main groups of land-degradation processes are normally distinguished: (1) mass movement (such as debris flows and avalanches), (2) water erosion (sheet, rill, gully erosion), (3) wind erosion, (4) excess of salts (salinization, sodification), (5) chemical degradation (acidification, contamination, toxicity), (6) physical degradation (crusting, compaction, oxidation), and (7) biological degradation (loss of soil biodiversity).

An important aspect of many soil and land degradation processes are the so called off-site effects; for example, dust storms or eroded sediment cause problems such as damage by mudflows, siltation of dams, or pollution of drinking water in downwind or downstream areas.

Factors and Processes Affecting Degradation of Soils

Various types of human activities may lead to soil degradation. Although some degradation processes may also occur naturally, many degradation types are the result of human disturbance of either a natural or anthropogenic state of equilibrium. Some of these are:

Agricultural causes: Defined as the improper management of cultivated arable land. It includes a wide variety of practices, such as insufficient or excessive use of fertilizers, shortening of the fallow period in shifting cultivation, use of poor quality irrigation water, absence or bad maintenance of erosion-control measures, improper use of heavy machinery, etc. Degradation types commonly linked to this causative factor are erosion (water or wind), compaction, loss of nutrients, salinization, and pollution (by pesticides or fertilizers).

Deforestation or removal of natural vegetation: Defined as the near complete removal of natural vegetation (usually primary or secondary forest) from large stretches of land, for example by converting forest into agricultural land (hence sometimes followed by agricultural mismanagement), large-scale commercial forestry, road construction, urban development, etc. Deforestation often causes erosion and loss of nutrients.

Overexploitation of vegetation for domestic use: Contrary to 'deforestation or removal of natural vegetation,' this causative factor does not necessarily involve the (near) complete removal of the 'natural' vegetation, but rather a degeneration of the remaining vegetation, thus offering insufficient protection against erosion. It includes activities such as excessive gathering of fuel wood, fodder, (local) timber, etc.

Overexploitation of natural water resources: This leads to water shortages for the natural ecosystem and in the long term to the removal of the natural vegetation cover. The result is an increased vulnerability of the land for surface runoff, soil erosion, and soil surface crusting. As soon as the process of vegetation deterioration starts, it normally has a self-enhancing effect which is difficult to stop or to reverse.

Overgrazing: Besides actual overgrazing of the vegetation by livestock, other phenomena of excessive livestock amounts are also considered here, such as trampling. The effect of overgrazing usually is soil compaction and/or a decrease in plant cover, both of which may in turn give rise to water or wind erosion.

Industrial activities: All human activities of an industrial or bioindustrial nature are included: industries, power generation, infrastructure and urbanization, waste handling, traffic, etc. It is most often linked to pollution of different kinds (either point source or diffuse) and loss of productive function.

Types of Soil Degradation

The type of soil degradation refers to the nature of the degradation process. Soil particles may be displaced by the action of water or wind (erosion and sedimentation), which may cause damage to crops, infrastructure, buildings, and the environment in general. Erosion can be linear, i.e., concentrated along certain channels (rill or gully erosion and mass wasting such as landslides), sometimes creating very deep scars in the landscape (Figure 1). Less conspicuous, but often even more detrimental to crops is the gradual removal of the topsoil layer (sheet erosion). Off-site effects of erosion may consist of siltation of reservoirs and river beds and/or flooding, or dune formation and 'overblowing' in the case of wind erosion. Degradation *in situ*, i.e., without movement of soil particles, can be chemical (soil pollution by chemical wastes or excessive fertilization; fertility decline due to nutrients being removed by harvesting, erosion and leaching; salinization due to irrigation with saline groundwater and/or without proper drainage in semiarid and arid areas, acidification due to pH-lowering additions to the soil from fertilizers or from the atmosphere), or physical (compaction due to the use of heavy machinery; deteriorating soil structure such as crusting of the soil surface; waterlogging due to increased water table or its opposite, aridification).

Assessment of Degradation

Approaches

The status of soil degradation can be assessed in a qualitatively broad manner or in a more detailed quantitative manner. The former generic approach is better suited for small-scale assessments, such as for entire countries, continents, or global overviews. A quantitative approach is required for more specific and detailed assessments, e.g., to determine the erosion status for a watershed or the pollution status for a province. Qualitative assessments are based on expert judgement and hence more liable to subjectivity than quantitative methods. A method does not have to be fully qualitative or quantitative, mixtures may occur. Some frequently used methods or tools are:

Figure 1 Severely degraded soils on the Loess Plateau of China.

1. Expert opinion: Qualitative assessment on a controlled mapping base and semiquantitative definitions, as employed for instance in the Global Assessment of Human-induced Soil Degradation (GLASOD) survey. GLASOD and related methods are based on an assessment of land suitability by national experts that use defined, semiquantitative class limits on a given mapping base. Its major disadvantage is the inevitable degree of subjectivity. Its major advantage is its capacity to produce results, such as achieving complete world coverage (Figure 2), in a short time and on a small budget. Costs per unit area are relatively low. In Figure 3 an integrated global soil degradation severity map is shown, indicating areas with different degradation rates;

2. Remote sensing: Analysis of low- and high-resolution satellite data and airborne imagery (e.g., analysis of composite indices such as the Normalized Difference Vegetation Index (NDVI)). Remote sensing always includes linkages with ground observations. The basis of this method is comparison of remotely sensed imagery of different dates, for regional coverage, mainly low-resolution imagery; and, specifically, comparison of the NDVI, derived from imagery collected by the sensor aboard the National Oceanographic and Atmospheric Administration (NOAA) satellite, and more detailed imagery. This method was tested amongst others in Saudi Arabia and shows areas where vegetation response to rainfall is decreasing (degradation of resources) or increasing (rehabilitation of resources). It has been applied particularly to early warning systems. For longer-term comparisons, some form of calibration for preceding rainfall is needed. Costs are relatively low. It is recognized that remote sensing cannot be used alone.

Spectral mixture analysis (SMA): Since 1985 hyperspectral remote sensing has been developed, opening new methods to survey and assess degradational state of the soil surface. Hyperspectral remote sensing refers to the collection of images in the solar spectrum, with many narrow spectral bands allowing the collection of very accurate spectra of objects and the earth surface and identification of absorption features of plants and of soil minerals in these spectra. SMA is a technique to unravel the spectral information in the remote-sensing images by assuming that the spectral variation is caused by a limited number of surface material (green vegetation, senescent vegetation, a

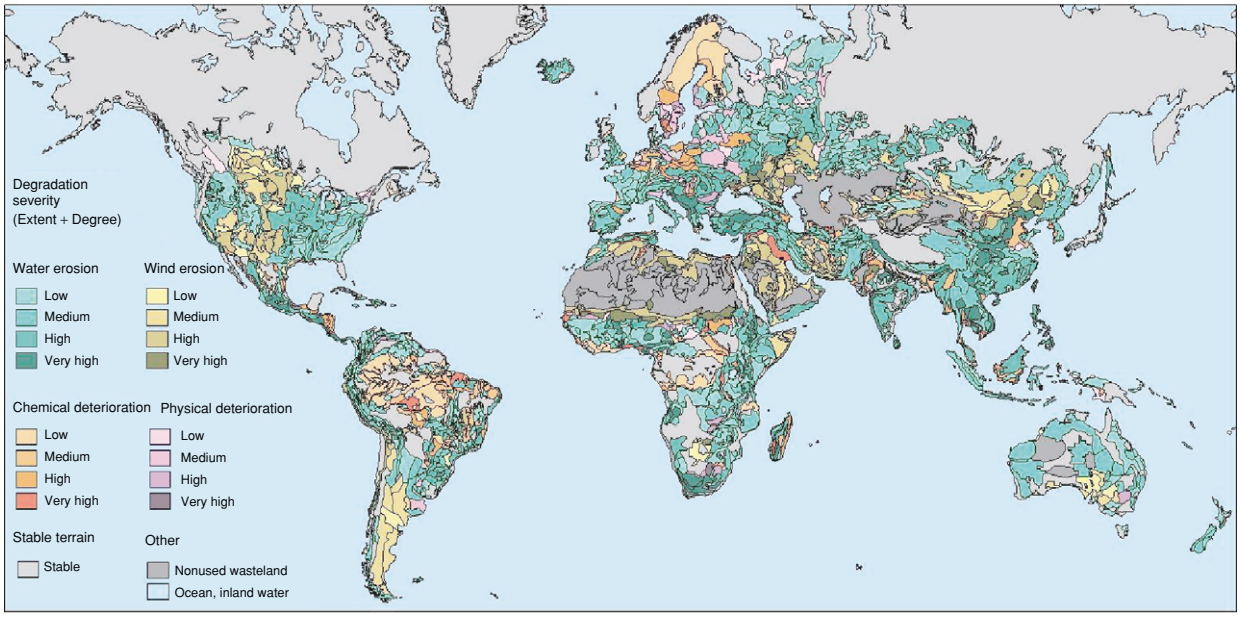


Figure 2 Global assessment (in 1990) of the status of human-induced soil degradation. (Reproduced with permission from Oldeman LR, Sombroek WG, and Hakkeling R (1991) *World Map on the Current Status of Human-Induced Soil Degradation. An Explanatory Note*, 2nd edn. Wageningen, the Netherlands: ISRIC/Nairobi, Kenya: UNEP.)

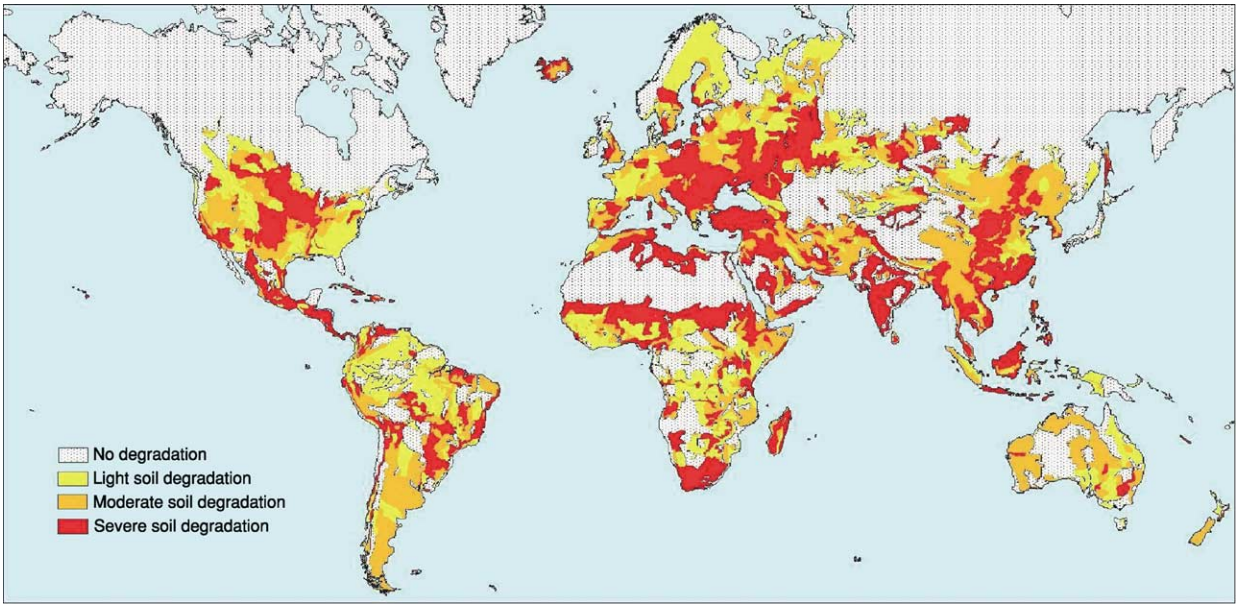


Figure 3 Global soil degradation severity map as produced by the GLASOD initiative.

number of soil types, and water). A reference library of these surface materials collected in the field or in the laboratory yields the basis for SMA of the remote-sensing images. This approach has been applied successfully in a number of case studies to survey soil conditions and to identify classes of degradation. The SMA approach normally improves on results using the NDVI but requires more spectral bands: SMA is successfully applied to separate, in images, bare soil surfaces from senescent vegetation and yellow vegetation from green vegetation. These three factors are important inputs in soil-erosion models because they act differently with respect to raindrop interception.

3. Field monitoring: Stratified soil sampling and analysis, and field observation of vegetation and biodiversity under certain land-use or management practices and climate variability. To date, soil monitoring has been applied mainly in developed countries, and tests are needed of its cost-effectiveness in developing countries. In areas where baseline studies have

been established, monitoring of changes will be undertaken; in other areas, establishment of a baseline will be a priority. Stratified soil-sampling with analysis, and/or benchmark sites, repeated over 5- to 10-year intervals, has been advocated as a basic activity for national soil survey organizations. Examples of application to date (2003) are few, but successful: the method has been applied to 20 000 sites over a 25-year period in Japan; is currently being used for a national 16-km-grid in France; and has been started in Denmark and Switzerland. The same approach has been applied to field observations of vegetation, along transects or in sampling plots, and to biodiversity. Costs per unit area are relatively high, but could be reduced by application to priority areas only, on a stratified sampling basis.

4. Productivity changes: Observation of changes in crop yields, biomass production, and livestock output, which directly apply to the definition of land degradation in terms of lowered productivity, although they are influenced by many other factors. There is a range of possibilities: At national level, use might be made of national yield statistics (of which the reliability is still under debate), adjusted for fertilizer use and climate. At local level, yield monitoring is possible by comparisons with a standard crop, either without fertilizer or with standard fertilizer and management. Substantial problems arise in that productivity decline could be due to factors other than land degradation, e.g., removal of fertilizer subsidy or civil strife. The same cost constraints apply as for soil monitoring.

5. Sample studies at farm level, based on field criteria and the expert opinion of land users. Even at national level, such detailed studies are essential on a sample basis, to obtain grass roots views both of the severity of degradation and its causes, together with practicable remedies (Stocking and Murnaghan, 2001). Field indicators of soil degradation were developed about 20 years ago, and could be extended to condition of vegetation. Talking with farmers means getting the views of farmers, and other land users, on whether things have got worse – which are of course, subjective and perhaps systematically biased, but still essential to get grass roots view at local level. The method is clearly applicable only at a local scale, and thus on a selective sampling basis. Observations of the state of the land can be combined with assessment of driving factors and impacts.

6. Modeling: Based on data obtained by other methods, modeling can be used in many ways, such as: (1) prediction of degradation hazard; (2) operational definition of degradation in terms of unfavorable changes in plant productivity, soil properties, and hydrology; (3) design of conservation measures using climatic data with a specific return period (worst-case scenario modeling); (4) extending the range of applicability of results; (5) integrating biophysical with socioeconomic factors. Much research has been put into devising models for the prediction of soil-erosion hazard. There are established methods for the modeling of both water and wind erosion, which have been widely applied, in part because it is vastly cheaper than any form of field observation. The modeling approach is mainly relevant to degradation hazard, but can be applied to actual degradation first, as a means of calibration of the model to the specific requirements of an area, optimizing sampling design, or to extrapolate the applicability of results obtained on a sampling basis. Risk reflects a potential development in the future, while status reflects the development to date. Models vary widely in complexity and data requirements, depending on the type of degradation they are addressing and the size of the area under investigation. Models are useful to learn and understand degradation processes, but both the model and input data are a simplification of reality, hence extrapolation of models should be done with care. Very often models are developed for experimental plots or pilot zones of a restricted size and under more or less controlled conditions, which should be taken into account when applying the model elsewhere. The data requirements and structure of a model, and the type of processes included in it, depend on many things: (1) the temporal scale of the research objectives: Is an annual, daily, or event-based result required? (2) the spatial scale: Are predictions needed for a single plot, a field, complex spatial catchment, or an entire region? (3) Is the emphasis on the on-site effects of land degradation (e.g., soil erosion or crop yield changes) or on the off-site effects such as water sediment levels and pollution? Spatial and temporal scales are often linked, as, for example, is the case for physically based spatial erosion models (Figure 4) that simulate single events for first-order catchment with a high level of detail. They can be used to answer subtle questions about the effects of specific land-use changes or soil and water conservation measures in the catchment upon reducing runoff and erosion (Figure 5). On the other hand there are less-complex empirical models that can simulate continuous periods mostly for fields or hillslopes, but they can only show the change in annual erosion or soil loss.

Potentials and Limitations

It is useful to emphasize some potentials and limitations of land degradation assessments. It is obvious that an assessment at a small scale (e.g., 1:1 M) does not have a direct value for activities at the field level,

but can be highly useful (if well done) to planners, government agencies, legislative bodies, educational institutions, nongovernment organizations (NGOs), and the general public in highlighting (potential) problem areas and decision-making for further action.

Besides geographic coverage and scale, another factor that determines the usefulness of an assessment methodology is the range of degradation issues the assessment tries to cover. Land degradation is a very broad issue, covering a wide range of degradation issues, which makes its assessment a rather generic or alternatively highly unwieldy exercise. It is already quite complicated to assess one specific type of soil pollution, not to mention the various other types of soil degradation, which in itself is just one aspect of land degradation. This also means that the frequently observed desire to have 'simple' assessment methods is not entirely realistic, if this is supposed to be anything more than just a general awareness-raising tool.

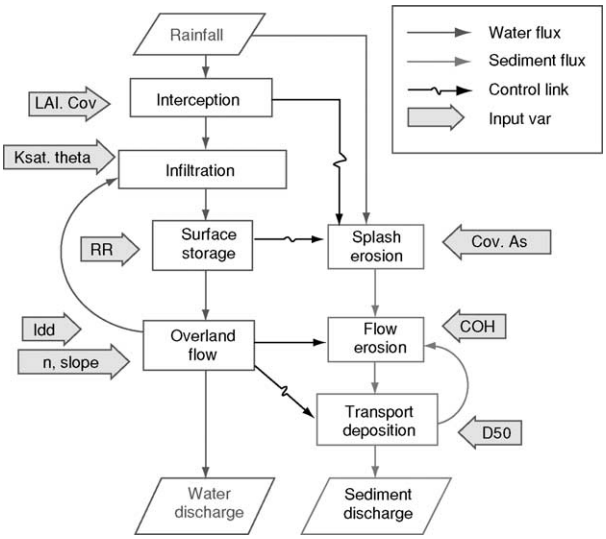


Figure 4 Model structure of a physically based spatial hydrologic and soil-erosion model, in which water and sediment are routed to the outlet of a catchment and produced as discharge. Input var, input variable; LAI, leaf area index; Cov, soil cover; Ksat, hydraulic conductivity; theta, moisture content; RR, surface roughness; Idd, runoff network; n, flow resistance; slope, terrain slope; As, aggregate stability; COH, Cohesion; D50, median grain size of suspended sediment.

Two types of assessments have been identified one is 'backward-looking' and determines the result of degradation over a recent past period. The other approach is forward-looking in the sense that it makes predictions for the future based on models and scenarios. Although the backward-looking approach considers the current status, it does not necessarily reflect the result of the degradation process over that period, but the net result of a number of acting and counteracting factors. Degradation is one of these, but remedial activities compensating the degradation effect to some extent is another one.

Though the wish to have 'simple' degradation assessment methods is often expressed, it should be realised that soil degradation is a complex process, determined by a range of factors of a natural and socioeconomic character. Hence a simple method will tend to correspond less with reality than a more complex and comprehensive one.

Degree and Impact of Degradation

Degree of Degradation

Degree is defined as the intensity of the soil degradation process, e.g., in the case of erosion, the amount

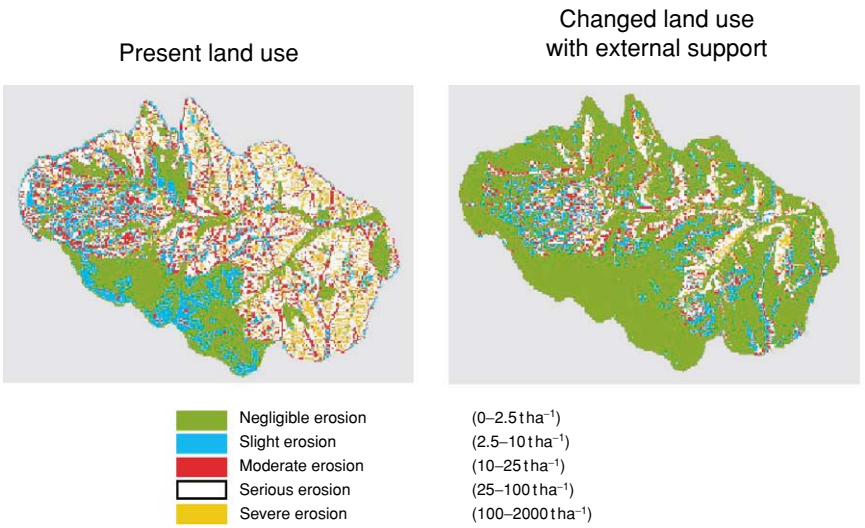


Figure 5 Computed soil losses in a first-order watershed for the current land use and management conditions and for an alternatively defined land-use distribution.

of soil washed or blown away. The FAO has proposed values for maximum acceptable limits of soil loss by erosion with respect to decreased agricultural productivity. Four classes are distinguished, ranging from no loss of productivity to severe loss of productivity. The classes are <12, 12–25, 25–50, and $>50\,t\,ha^{-1}\,year^{-1}$. Relative changes of the soil properties are other good indicators of soil degradation: the percentage of the total topsoil lost, the percentage of total nutrients and organic matter lost, the relative decrease in soil moisture-holding capacity, changes in buffering capacity, etc. However, although such data may exist for experimental plots and pilot areas, precise and actual information is often lacking at a regional scale.

Rate of Soil Degradation

The recent rate of degradation relates to the rapidity of degradation over the past 5–10 years or, in other words, the trend of degradation. A severely degraded area may be quite stable at present (i.e., low rate, hence no trend toward further degradation), while other areas that are now only slightly degraded may show a high rate, hence a trend toward rapid further deterioration. From a purely physical point of view, the latter area would have a higher conservation priority than the former. Areas where the situation is improving (through soil conservation measures, for example), can also be identified. A comparison of the actual situation with that of the preceding decade may suffice, but often it is preferable to examine the average development over the last 5–10 years to level out irregularities. Whereas the degree of degradation only indicates the current, static situation (measured by decreased or increased productivity compared with some 10–15 years ago) the rate indicates the dynamic situation of soil degradation, namely the change in degree over time.

Impact of Degradation

Impact refers to the effects of soil degradation on the various soil functions. Changes in soil and terrain properties (e.g., loss of topsoil, development of rills and gullies, exposure of hardpans in the case of erosion) may reflect the occurrence and intensity of soil degradation but not necessarily the seriousness of its impact. Removal of a 5-cm layer of soil may have a greater impact on a poor shallow soil than on a deep fertile soil. The impact depends on the function and/or use of the soil: a heavily compacted soil is unsuitable for agriculture, but may be an appropriate basis for road construction.

Whereas the degree of degradation mainly refers to the degradation process, the impact of degradation can be manifold, depending on the current function (or use) of the soil. In many cases, the impact of degradation types will be on its biotic functions, or more specifically on its productivity. A significant complication in indicating productivity losses caused by soil degradation is the variety of reasons that may contribute to yield decline. Falling productivity may be caused by a wide range of factors such as erosion, fertility decline, improper management, drought, or waterlogging, quality of inputs (seeds, fertilizer), pests, and plagues, often in combination with each other. However, if one considers a medium- to long-term period (e.g., 25 years), large aberrations resulting from fluctuations in the weather pattern or pests should be leveled out.

The effects of soil degradation can be partially hidden by various management measures such as soil conservation, use of improved varieties, fertilizers, and pesticides. Some of these inputs are used to compensate for the productivity loss caused by soil degradation, for example application of fertilizers to compensate for lost nutrients. In other words, yields could have been much higher in the absence of soil degradation (and/or costs could have been reduced). Therefore, productivity changes should be seen in relation to the degree of input or level of management. The latter may include use of fertilizers, biocides, improved varieties, mechanization, various soil conservation measures, and other important changes in the farming system.

Changes in productivity should be expressed in relative terms, i.e., the current average productivity compared with the average productivity in the non-degraded situation and in relation to inputs. For instance, if previously an average yield of 2 t of wheat ha^{-1} was attained while at present only 1.5 t is realized in spite of high(er) inputs – and all other factors being equal – this would be an indication of strong soil degradation. Sometimes the impact may be ranked as negligible, even when degradation occurs, because of the capacity of the soil to resist a certain amount of degradation. Although for most degradation types the dominant impact is on productivity, some types (pollution in particular) may have additional or different impacts, e.g., on human or animal health or on entire ecosystems.

Preventing and Combating Degradation

There are a wide variety of measures to prevent or combat land degradation. These measures are generally known as soil conservation or soil and water conservation (SWC), especially when related to aspects like erosion, soil-moisture problems and soil fertility. More broadly applicable are names such as land husbandry or sustainable land management.

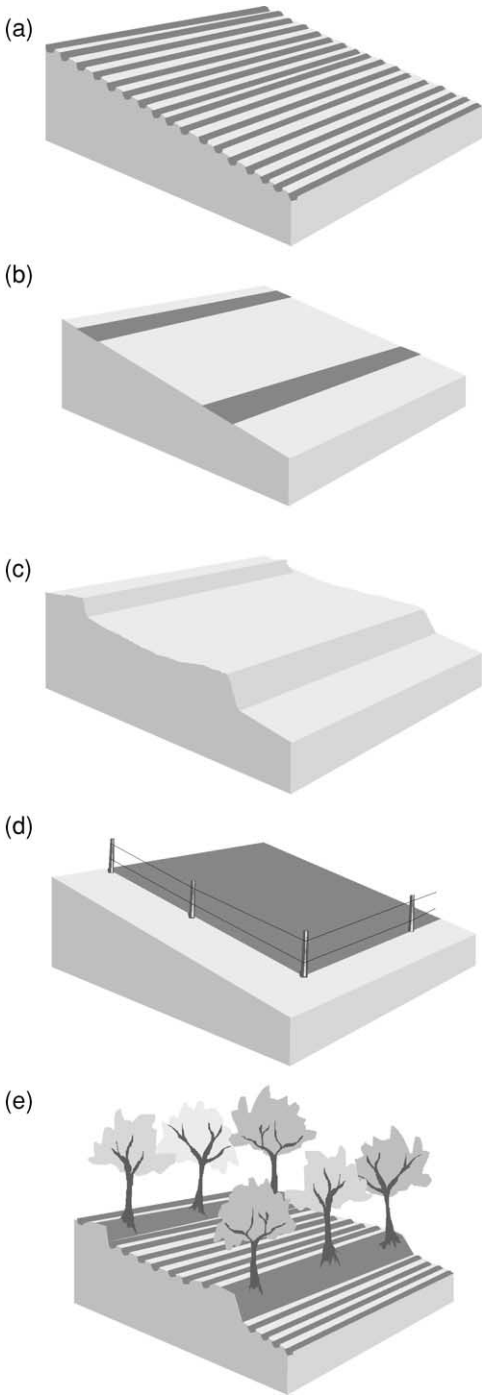


Figure 6 Categorization of soil and water conservation measures according to the World Overview of Conservation Approaches and Technologies (WOCAT) initiative: (a) agronomic measures such as mixed cropping, contour cultivation, mulching, etc. which: (1) are usually associated with annual crops, (2) are repeated routinely each season or in a rotational sequence, (3) are of short duration and not permanent, (4) do not lead to changes in slope profile, (5) are normally not zoned, (6) are normally independent of slope; (b) vegetative measures such as grass strips, hedge barriers, windbreaks, etc. which: (1) involve the use of perennial grasses, shrubs, or trees, (2) are of long duration, (3) often lead to a change in slope profile, (4) are often zoned on the contour or at right angles to wind, (5) direction, (6) are often spaced according to slope; (c) structural measures such as terraces, banks, bunds, constructions, palisades, etc. which: (1) often lead to a change in slope profile, (2) are of long duration or permanent, (3) are carried out primarily to control runoff, wind velocity, and erosion, (4) require substantial inputs of labour or money when first installed, (5) are often zoned on the contour/against wind direction, (6) are often spaced according to slope, (7) involve major earth movements and/or construction with wood, stone, concrete, etc.; (d) management measures such as land use change, area closure, rotational grazing, etc. which: (1) involve a fundamental change in land use, (2) involve no agronomic and structural measures, (3) often result in improved vegetative cover, (4) often reduce the intensity of use; (e) combinations in conditions where they are complementary and thus enhancing each other. Any combinations of the above measures are possible, e.g.: structural: terrace, with vegetative: grass and trees, with agronomic: ridges.

The WOCAT (World Overview of Conservation Approaches and Technologies) network, which constitutes an international consortium of institutions and individuals from all over the world, provides an evaluation tool for SWC activities, an information-management system designed to collect, analyze, present, and disseminate SWC knowledge and a decision-support system designed to assist in the search for SWC options appropriate to the prevailing biophysical and socioeconomic settings. WOCAT was initated in 1992 and has developed a common framework and methodology, consisting of three comprehensive questionnaires (in English, French, and Spanish) for the documentation and evaluation of SWC.

The WOCAT methodology consists of three major modules:

1. Questionnaire and database on SWC technologies;
2. Questionnaire and database on SWC approaches;
3. Questionnaire and database on the geographic distribution of SWC (mapping).

The first two modules aim at a comprehensive and detailed description of specific technologies, i.e., agronomic, vegetative, structural, and/or management measures used in the field (Figure 6), and the ways and means used to implement an SWC technology on the ground. The mapping module is more or less similar to the qualitative methodology for degradation assessment described earlier. In this approach, information is collected for individual units of a (physiographic or other) base map on the following items:

- Land use: type, extent, trend in area, trend in intensity;
- (Per land use type) degradation, as above, but only for water and wind erosion and fertility decline in erosion-prone areas;

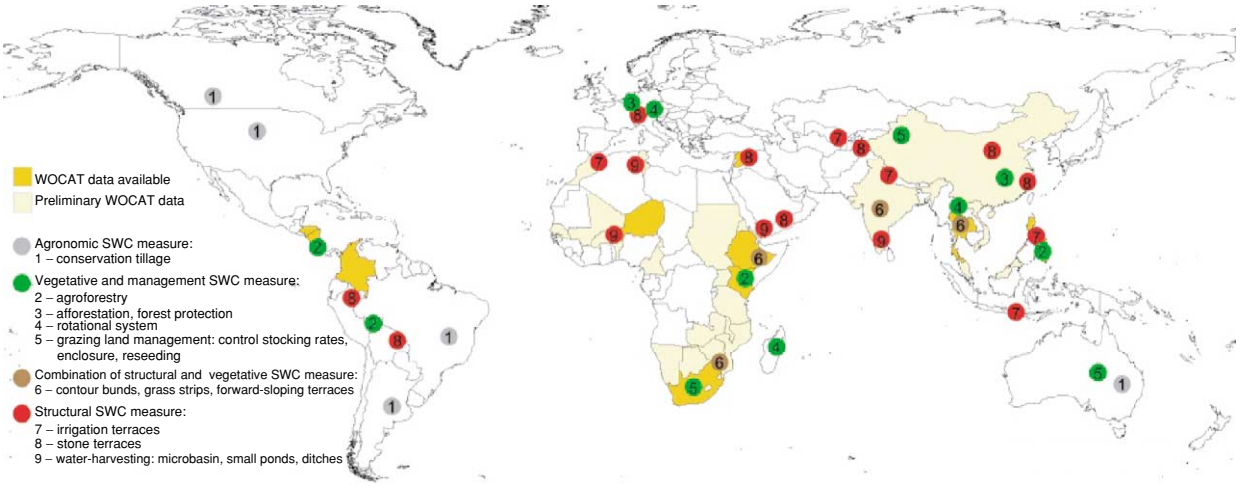


Figure 7 Reported soil conservation measures as recently compiled by the WOCAT initiative.

- (Per land use type) conservation: type, extent, period of implementation, effectiveness, trend in effectiveness, and reference to a corresponding questionnaire in the technology database for more detailed information;
- (Per land use type) productivity: trend, contribution of SWC or degradation to this trend, average production value, average input value.

An overview of soil and water conservation measures applied in different geographic regions of the world is shown in Figure 7.

See also: **Desertification**; **Erosion:** Water-Induced; Wind-Induced; **Salination Processes**

Further Reading

Blaikie PM and Brookfield HC (1987) *Land Degradation and Society*. London, UK: Methuen.

Blum WEH (1988) *Problems of Soil Conservation*. Nature and Environment Series 39. Strasbourg, France: Council of Europe.

Crosson P (1997) The on-farm economic costs of erosion. In: Lal R, Blum WEH, Valentin C, and Stewart BA (eds) *Methods for Assessment of Soil Degradation*. Boca Raton, FL: CRC Press.

Dregne HE (1997) Desertification assessment. In: Lal R, Blum WH, Valentin C, and Stewart BA (eds) *Methods for Assessment of Soil Degradation*. Boca Raton, FL: CRC Press.

Lal R, Blum WH, Valentine C, and Stewart BA (eds) (1997) *Methods for Assessment of Soil Degradation*. Boca Raton, FL: CRC Press.

Liniger HP, van Lynden GWJ, and Schwilch G (2002) Documenting field knowledge for better land management decisions – Experiences with WOCAT tools in local, national and global programs. *Proceedings of ISCO Conference 2002*, vol. I, pp. 259–267, Beijing, China: Tsinghua University Press.

Oldeman LR, Sombroek WG, and Hakkeling R (1991) *World Map on the Current Status of Human Induced Soil Degradation. An Explanatory Note*, 2nd edn. Wageningen, the Netherlands: ISRIC/Nairobi, Kenya: UNEP.

Ritsema CJ (ed.) (2003) Soil erosion and participatory land use planning on the Loess Plateau in China. (Special issue.) *Catena* 754.

Stocking M and Murnaghan N (2001) *Handbook for the Field Assessment of Land Degradation*. Sterling, VA: Earthscan Publications.

Van der Meer FD and de Jong SM (eds) (2001) *Imaging Spectrometry: Basic Principles and Prospective Applications*. Bookseries Remote Sensing and Digital Image Processing, vol. 4. Dordrecht, the Netherlands: Kluwer Academic Publishers.

Van Lynden GWJ (1995) *European soil resources. Current status of soil degradation, causes, impacts and need for action*. Nature and Environment Series 71. Strasbourg, France: Council of Europe.

DENITRIFICATION

D A Martens, USDA Agricultural Research Service, Tucson, AZ, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Denitrification is the major biological process through which the soil N available to plants is returned to the nonavailable atmospheric N pool as various N oxides (NO, N_2O) and dinitrogen (N_2). Despite the central role of microbial denitrification for N-cycle regulation of plant-available N, this process remains one of the least quantified mechanisms of soil N transformation. Rates of N loss from fertilizer applied to agricultural soils through denitrification can vary tremendously, ranging from 0 to 70% of applied N. In the USA, N_2O from the use of N fertilizers in 1998 accounted for 45% of the total N_2O emission budget.

Definitions and Pathways

Denitrification is defined as the "microbial reduction of nitrate or nitrite coupled to electron transport phosphorylation resulting in gaseous N either as molecular N_2 or as an oxide of N." The key to denitrification as defined is the availability of the N oxides, nitrite (NO_2^-) or nitrate (NO_3^-), which are formed from the autotrophic nitrification pathway substrate, ammonia (NH_3), which is derived from ammonium (NH_4^+). Chemical fertilizer inputs and soil organic-matter mineralization are the main sources of NH_4^+ in the environment. Nitrous oxide production from soils has been linked to two biological processes. The first is during the process of nitrification of NH_4^+ under aerobic conditions and the second is the coupled nitrification/denitrification (denitrification) pathway that occurs under anaerobic conditions. The pathway proposed for nitrifying organisms to release N_2O during the nitrification process has been defined as nitrifier denitrification.

The denitrification pathway is prevalent once NO_3^- is formed and the correct environmental conditions (low- or no-O_2 concentrations, and high soluble C content) are imposed with microbial reduction of the N oxides. The term 'denitrification' or 'respiratory denitrification' has been defined as a bacterial respiratory process and as such, a clear distinction between the denitrification pathway and the nitrifier denitrification pathway is necessary, because the relative proportion of N_2O production from the pathways is affected by different environmental conditions. A third pathway involving the chemical decomposition of NO_2^- has also been found in soils and can be prevalent in low-pH environments. The nonbiological pathways, or chemodenitrification, are so closely linked with nitrification that is often difficult to determine whether the nitric oxide (NO) and N_2O produced are formed through nitrification or chemodenitrification. Early research on denitrification was the result of the inability to mass balance total inputs of N and outputs of N in agricultural systems. The large portion of N unaccounted for was determined to be lost as gaseous N, which agronomically, resulted in decreased N fertilizer efficiency.

Research into global-change processes has found another impact of the N gas loss during the nitrification–denitrification process. Nitrous oxide is now known to be a potent greenhouse gas with important impacts on our environment. Nitrous oxide has a global-warming potential about 320 times as strong as carbon dioxide mainly due to an atmospheric lifetime of about 120 years. As the amounts of industrial or biologically fixed N_2 used for crop production increase, the production of N_2O due to nitrification–denitrification processes will also increase and potentially cause significant depletion of the Earth's stratospheric ozone layer and contribute to a warming of the Earth's surface by influencing the radiative budget of the troposphere.

Organisms and Substrates

Denitrification

Denitrification is the stepwise reduction of N oxides with gaseous products such as N_2O or N_2 under conditions of limited O_2 (Figure 1). The process results from the use of N oxide as a terminal electron acceptor instead of molecular O_2 by bacteria and is irreversible once NO is formed. Thus, a source of organic C is required for bacteria metabolism and sufficient NO_3^- to act as an electron acceptor must be available in an environment that has limited O_2. The requirement of all three factors, C source, low O_2, and sufficient NO_3^- must be present for the occurrence of denitrification. The large coefficient of

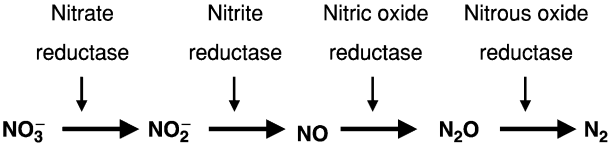


Figure 1 Pathway and enzymes involved in denitrification.

variation for measurement of gaseous N loss from field soils has been accredited to the formation of soil 'hotspots' that exhibit increased soil respiration. Spatial variability in N_2O fluxes typically causes coefficients of variation from 30 to 100% to be measured at a spatial scale of several meters with field plots. Hotspots of organic debris can stimulate denitrification due to rapid microbial mineralization and release of NH_4^+ and NO_3^- formation with concomitant depletion of O_2 concentrations. The best field indicator for denitrification has been found to be drainage class, which determines the potential for the soil becoming waterlogged at times of high precipitation. Seasonal differences generally noted for denitrification (lower in spring compared with higher in autumn) have also been attributed to low spring NO_3^- concentrations in temperate soils. Microbial activity has also been found to result in increased denitrification at the soil surface compared with deeper subsoils due to the greater organic inputs to the soil surface.

The genera of denitrifying bacteria are diverse. The capacity to denitrify is present in about 23 genera of bacteria. The list of denitrifying genera in Table 1 includes 13 genera (considered facultative aerobes) for which there is confirmed or multiple documentation of denitrifying activity. Although the distribution of denitrifiers is ubiquitous, the activity or measurement of denitrifier biomass and synthesis of denitrifying enzyme activity from soil field cores is poorly correlated with measured N gas loss. Recent laboratory research has also found that yeast and filamentous fungi are also capable of gaseous N production, but field confirmation of importance is lacking. Fungal N_2O production may be of great significance, because fungi often dominate the microbial biomass of temperate soils, yet fungi are not tolerant of waterlogged conditions. Since denitrifiers are not fermentative (facultative anaerobes), growth under O_2-limited conditions is solely dependent on the presence of N oxides. Enzymes involved in the reactions are nitrate reductase, nitrite reductase, nitric oxide reductase, and nitrous oxide reductase (Figure 1). During this process, N_2O is an intermediate and so the ratio of N_2O to N_2 fluxes can be affected by various environmental factors. The proportion of N_2O compared with N_2 is increased if the soil pH is decreased, if the amount of NO_3^- is increased, and if lower O_2 concentrations are present. In each case, the environmental factors have an influence on the enzyme nitrous oxide reductase. To summarize, the release of N_2O as an intermediate of denitrification can be an important N loss mechanism in low-O_2 environments with sufficient NO_3^- coupled with rapid mineralization of C.

Table 1 Genera of documented denitrifying bacteria

Genus	*Characteristics*
Alcaligenes	Common soil isolate
Agrobacterium	Some species are plant pathogens
Azospirillum	Associated with grass species N_2 fixation
Bacillus	Thermophilic denitrifiers reported
Flavobacterium	Recent denitrifier identification
Halobacterium	Requires high salt concentrations for growth
Hyphomicrobium	Grows on one-carbon substrates
Paracoccus	Capable of lithotrophic and heterotrophic growth
Propionibacterium	Fermenters capable of denitrification
Pseudomonas	Common soil isolates
Rhizobium	Capable of N_2 fixation in symbiosis with legumes
Rhodopseudomonas	Photosynthetic bacteria
Thiobacillus	Generally grow as chemoautotrophs

Nitrifier Denitrification

Nitrification is the bacterial oxidation of NH_4^+ or NH_3 via NO_2^- to NO_3^-. This process is carried out by two groups of autotrophic bacteria. The first step, from NH_3 to NO_2^-, is catalyzed by NH_3 oxidizers, and the second step, from NO_2^- to NO_3^-, is carried out by NO_2^- oxidizers. The best-studied member of the first group is *Nitrosomonas europaea*, and *Nitrobacter winogradskyi* is a representative of the NO_2^- oxidizers. Use of molecular probes specific for genes encoding for heme-type dissimilatory nitrite and nitrous oxide reductase gene homologies derived from *Nitrosomonas europaea* gene sequences has found that prominent nitrifiers of the genera *Nitrosospiria* and *Nitrosolobus* may differ from *Nitrosomonas europaea* in the mechanism and capacity for denitrification. Nitrous oxide can be released during the synthesis of the nitrification intermediates (Figure 2). The first possible step can form hydroxylamine (NH_2OH) as a potentially unstable intermediate, and N_2O can be released at this step. The formation of NO_2^- can serve as a second point where N_2O can

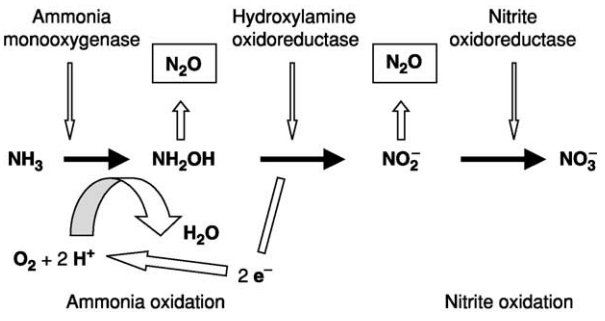


Figure 2 Pathway and enzymes involved in nitrification.

be released. The NO_2^- pathway to N_2O via nitrifier denitrification has been observed to be consistent with the kinetics exhibited by the denitrification pathway, but can occur in an environment with higher O_2 concentrations. In contrast to the traditional pathway of denitrification previously outlined, autotrophic and heterotrophic microorganisms are often able to release N_2O under aerobic conditions, because the intermediate NH_2OH can act as an electron donor. The relative proportion of nitrifier denitrification to the N_2O budget ranges from 0 to 30% of the total N_2O released.

Nitrification has also been found to occur in heterotrophic fungi and bacteria that utilize the same substrate, intermediates, and products as autotrophic nitrification, and again in contrast to denitrifiers, heterotrophic nitrifying bacteria are often able to form N_2O under aerobic conditions. Release of N_2O by heterotrophic nitrification is considered a minor contribution to N_2O budget.

Environmental Factors

Denitrification supports microbial life through respiration, because mineralization of a reduced substrate (mainly organic C, but reduced iron or sulfur may also donate electrons) donates electrons via carriers to an oxidized N oxide. The decline in free energy is coupled with the generation of adenosine triphosphate (ATP). In the presence of O_2, the ATP yield per number of electrons transported is about 1.7 times greater than if NO_3^- is the terminal electron acceptor, with the final efficiency apparently being species-dependent. Thus the rate of denitrification processes depends on the interactions and availability of reduced substrates, limited O_2 potential and presence of N oxides to function as terminal electron acceptors.

Organic C Availability

The availability of a suitable reduced C source is important because of the strong relationship between denitrification and soil organic matter. More important, the water-soluble portion of the soil organic matter has been found to be a good indicator for potential denitrification. Recent work has also found that organic litter with low C to N ratios has a greater potential to promote higher denitrification activity than litter with higher C to N ratios. Animal manures with a readily decomposable C source and a high N content can greatly enhance soil denitrification rates.

The presence or absence of plants has also been found to affect denitrification rates. Higher denitrification rates have been noted in the plant rhizosphere as compared with bulk soil several millimeters away from the roots if NO_3^- was sufficient; but if NO_3^- availability was limited, the presence of plants decreased denitrification activity. Plant roots have the potential to impact denitrification activity by providing C to support microbial populations and supply electrons for NO_3^- reduction. Roots can also create anaerobic zones through respiratory O_2 consumption when O_2 supply is limited by water conditions. The rooting environment can create a drier soil through evapotranspiration processes and result in increasing O_2 diffusion in the drier soil. Roots can also assimilate NO_4^- and NO_3^- from the soil solution reducing the availability of substrate to the denitrification pathways.

Due to a typical decrease in available soil C with depth, denitrification activity has also been found to decrease with soil depth and is potentially limited by a lack of available C below 60–75 cm. Other soil characteristics can also change with depth concomitant with organic C content such as porosity/permeability, pH, temperature, and depth to water table that may also have an impact on denitrification rates.

Controls on O_2 Availability

Reduced availability or absence of O_2 is required for both the synthesis and activity of enzymes involved in denitrification. Soil parameters can control overall O_2 availability and influence denitrification rates. The universally recognized inverse relationship between water content and denitrification is due to the slower diffusion of O_2 to metabolically active sites with the greater amounts of water present in the soil matrix. Bulk soil parameters that have been used to describe O_2 control of denitrification include soil moisture content, soil water potentials, partial pressure of O_2, O_2 concentration in solution, percentage air-filled porosities, and redox potentials. Bulk measurements have shown promise for predicting denitrification activity in laboratory studies, but, in field measurements, spatial variability due to hot spots can mask the predictive nature of O_2 diffusion and/or C mineralization rates.

Soil texture is an important factor for determining critical air-filled porosity, as a finer-textured soil (less sand content) will have a higher critical air-filled porosity value than found for coarser-textured soils. O_2 concentrations in soil atmospheres at which denitrification has been observed for different soils types range from 4 to 17%. Although the bulk soil O_2 concentrations are important, it has been noted that the O_2 concentration at the microsite level is more important for the determination of denitrification activity.

The occurrence of anaerobic microsites can explain how an O_2-sensitive process is possible in a

well-structured soil whose bulk O_2 measurements show aerobic conditions. The microsite concept proposes that anaerobic sites occur within centers of water-saturated soil aggregates because of limited diffusion of O_2 and consumption of O_2 by microbial metabolism. Research has shown that the occurrence of anaerobic microsites is largely a function of localized C availability and is possible within aggregates exceeding 9 mm. The microsites are more likely to occur in the rhizosphere and with particles of decomposing plant litter in zones where water can limit O_2 diffusion. Again, denitrification activity is a function of available C, limited O_2 potentials, and a source of N oxides.

Nitrate Supply

Research has indicated that the rate of denitrification activity in pure-culture studies was independent of NO_3^- concentration. Recent research has reported first-order kinetics with NO_3^- concentrations of less than 40 mg l^{-1}, and the kinetics changed from first-order to zero-order for NO_3^- concentrations of more than 100 mg. Microbial NO_3^- reduction is an enzyme-catalyzed reaction, and use of Michaelis–Menten kinetics has found K_m values for NO_3^- reduction to be less than 15 μmol l^{-1} NO_3^- or approx. 93 μg NO_3^- l^{-1} for several bacterial isolates. The use of K_m values determines that concentrations required for NO_3^- reduction are higher for soils than those found for pure culture or enzyme studies. The rate kinetics also change if the soil has recently received chemical fertilizers. However, Michaelis–Menten kinetics has not been found to describe NO_3^- reduction dependence accurately on available C and NO_3^- concentrations in soil. Several studies have found that kinetic analysis may not reflect denitrification rates, but the rate of NO_3^- diffusion from solution into the actively denitrifying soil aggregate. The evidence indicates that even very low NO_3^- concentrations present in soil solutions are in excess of enzyme and microbial needs for terminal electron acceptors during O_2 stress, but denitrifying activity may actually be limited by the diffusion of NO_3^- into the soil microsites.

Temperature

Denitrification, because of enzyme dependence on temperature for the activity rate, would be expected to increase exponentially with increased temperature within the range of enzyme activity. The temperature effect is impacted by the interaction of temperature on O_2 solubility and O_2 consumption as well as O_2 diffusion. Decreasing temperature will increase O_2 solubility and decrease O_2 consumption. Research has found a minimum soil temperature range of 2.7–10°C for denitrifying activity and a maximum temperature of approx. 50°C. Above 50°C, chemical decomposition reactions of NO_2^- can complicate the denitrification–temperature relationship. Pure culture experiments with soil isolates found 30°C to be the optimum temperature for growth, but evidence suggests that denitrifiers are adapted to soils and growth conditions of the specific soil. Bacterial isolates from tropical soils (33 in total) show no growth when incubated at 10°C, whereas 68% of temperate soil isolates (95 in total) grow at 10°C.

pH

The pH requirement for denitrifying organisms appears to be similar to heterotrophic organisms. In the neutral pH range from 6 to 8, there is a limited effect of pH on the activity of denitrification, but denitrification rates can be limited in acid pH ranges. Adjusting a naturally acid soil (pH 3.5) with reduced microbial activity to pH 6.5 rapidly stimulates a strong increase in denitrification rates and suggests that a general population with neutral growth optimum is present in a low-pH environment. Research has suggested that the decrease in denitrification activity in a moderately acid environment favors NO_3^- reduction to NO_2^- and that increased soil acidity may favor NO_3^- reduction rather than denitrification. Isotopic evidence has suggested that, although a moderate change in pH (pH 6.7 to 5.2) has little effect on the total rate of N_2 and N_2O production, the pH adjustment results in a rapid change from predominantly N_2 production at pH 6.7 to predominantly N_2O at pH 5.2. Due to the ratio change noted with change in soil pH value, the enzyme N_2O reductase may be extremely sensitive to pH.

Conclusion

Three possible soil-based mechanisms can give rise to gaseous loss of N. Denitrification and nitrifier denitrification are biological pathways that are very closely linked by the production of NO_3^- in soils and sediments. The chemodenitrification pathway is also linked to the biological processes, but occurs due to NO_2^- instability in an acid environment. The requirements for denitrification activity include a reduced C source, a NO_3^- pool, and an O_2-limited environment. The interactions with the three factors can result in different proportions of N_2O and N_2 being produced. The N_2O to N_2 ratio produced by the denitrification pathway can be increased by increased NO_2^- and NO_3^- concentration. The ratio can also be increased by decreased O_2 concentration, pH values, and increasing the C availability. The spatial variation of soil C and N and low O_2 tensions required

for denitrification during field quantification and the importance of nitrifier denitrification processes are crucial areas of research that need to be understood before management of gaseous N loss can be achieved.

See also: **Fertilizers and Fertilization**; **Greenhouse Gas Emissions**; **Microbial Processes:** Kinetics; **Nitrogen in Soils:** Cycle; Nitrification; **Organic Matter:** Principles and Processes; **Organic Residues, Decomposition**

Further Reading

Arah JRM (1997) Apportioning nitrous oxide fluxes between nitrification and denitrification using gas phase mass spectrometry. *Soil Biology and Biochemistry* 29: 1295–1299.

Firestone MK (1982) Biological denitrification. In: Stevenson FJ (ed.) *Nitrogen in Agricultural Soils*, pp. 289–326. Agronomy Monograph no. 22, Madison, WI: American Society of Agronomy.

Focht DD (1982) Denitrification. In: Burns RG and Slater JH (eds) *Experimental Microbial Ecology*, pp. 194–211. Oxford: Blackwell Scientific Publications.

Hochstein LI and Tomlinson GA (1988) Enzymes associated with denitrification. *Annual Review of Microbiology* 42: 231–261.

Laughlin RJ and Stevens RJ (2002) Evidence for fungal dominance of denitrification and codenitrification in a grassland soil. *Soil Science Society of America Journal* 66: 1540–1548.

Parkin TB (1990) Characterizing the variability of soil denitrification. *FEMS Symposium of the Federal European Microbiological Society* 56: 213–228.

Parkin TB and Meisinger JJ (1989) Denitrification below the crop rooting zone as influenced by surface tillage. *Journal of Environmental Quality* 18: 12–16.

Poth M and Focht DD (1986) ^{15}N kinetic analysis of N_2O production by *Nitrosomonas europaea*: an examination of nitrifier denitrification. *Applied and Environmental Microbiology* 49: 1134–1141.

Tiedje JM (1982) Denitrification. In: Page AL, Miller RH, and Keeney DR (eds) *Methods of Soil Analysis, part 2, Chemical and Microbiological Properties*, 2nd edn, pp. 1011–1026. Madison, WI: American Society of Agronomy.

Wrage N, Velthof GL, van Beusichem ML, and Oenema O (2001) Role of nitrifier denitrification in the production of nitrous oxide. *Soil Biology and Biochemistry* 33: 1723–1732.

DESERTIFICATION

D Hillel, Columbia University, New York, NY, USA
C Rosenzweig, NASA Goddard Institute for Space Studies, New York, NY, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Ecosystems in semiarid and arid regions around the world appear to be undergoing various processes of degradation commonly described as 'desertification.' According to the United Nations Environmental Program (UNEP), all regions in which the ratio of total annual precipitation to potential evapotranspiration (P/ET) ranges from 0.05 to 0.65 should be considered vulnerable to desertification. Such regions constitute some 40% of the global terrestrial area. They include northern Africa, southwestern Africa, southwestern Asia, central Asia, northwestern India and Pakistan, southwestern USA and Mexico, western South America, and much of Australia, and are home to an estimated sixth of the world's population.

'Desertification' is a single word used to cover a wide variety of interactive phenomena – both natural and anthropogenic – affecting the actual and potential biological and agricultural productivity of ecosystems in semiarid and arid regions. It is an emotive term, conjuring up the specter of a tide of sand swallowing fertile farmland and pastures. Apparently with this somewhat simplistic image in mind, UNEP sponsored projects in the early 1980s to plant trees along the edge of the Sahara, with the aim of warding off the invading sands. While there are places where the edge of the desert can be seen encroaching on fertile land, the more pressing problem is the deterioration of the land due to human abuse in regions well outside the desert. The latter problem emanates not from the expansion of the desert *per se* but from the centers of population outside the desert, owing to human mismanagement of the land. A vicious cycle has begun in many areas: as the land degrades through misuse, it is worked or grazed ever more intensively, so its degradation is exacerbated; and as the returns from 'old' land diminish, 'new' land is brought under cultivation or under grazing in marginal or even submarginal areas.

As defined in recent dictionaries, desertification is the process by which an area becomes (or is made to become) desert-like. The word 'desert' itself is derived

from the Latin *desertus*, being the past participle of *deserere*, meaning 'to desert,' 'to abandon.' The clear implication is that a desert is an area too barren and desolate to support human life. An area that was not originally desert (e.g., a steppe or savanna) may come to resemble a desert if it loses so much of its usable resources that it can no longer provide adequate subsistence to a given number of humans. This is a very qualitative definition, since not all deserts are the same. An area's resemblance to a desert does not make it a permanent desert if it can recover from its damaged state, and, in any case, the modes of human subsistence and levels of consumption differ greatly from place to place.

In recent decades, the very term 'desertification' has been called into question as being too vague, and the processes it purports to describe too ill-defined. Some critics have even suggested abandoning the term, in favor of what they consider to be a more precisely definable term, namely 'land degradation.' However, 'desertification' has already entered into such common usage that it can no longer be revoked or ignored. It must therefore be clarified and qualified so that its usage is less ambiguous.

'Land degradation' itself is a vague term, since the land may be degraded with respect to one function and not necessarily with respect to another. For example, a tract of land may continue to function hydrologically – to regulate infiltration, runoff generation, and groundwater recharge – even if its vegetative cover is changed artificially from a natural species-diverse community to a monoculture and its other ecologic functions are interrupted. Perhaps better than 'land degradation' is the term 'semiarid ecosystem degradation.' A semiarid ecosystem encompasses the diverse biotic community sharing the domain. Included in this community is the host of plants, animals, and microorganisms that interact with one another through such modes as competition or symbiosis, predation, and parasitism. It also includes the complex physical and chemical factors that condition the lives of those organisms and are in turn influenced by them. Each ecosystem performs a multiplicity of ecologic functions. Included among these are photosynthesis, absorption of atmospheric carbon and its incorporation into biomass and the soil, emission of oxygen, and regulation of temperature and the water cycle, as well as the decomposition of waste products and their transmutation into nutrients for the perpetuation of diverse, interdependent forms of life.

A semiarid ecosystem may be more or less natural, relatively undisturbed by humans, or it may be artificially managed, such as an agroecosystem. An agroecosystem is a part of the landscape that is managed for the economic purpose of agricultural production. The transformation of a natural ecosystem into an agroecosystem is not necessarily destructive if the latter is managed sustainably and productively, and if it coexists harmoniously alongside natural ecosystems that continue to maintain biodiversity and to perform vital ecologic functions. In too many cases, however, the requirements of sustainability fail, especially where agricultural systems expand progressively at the expense of the remaining, more-or-less natural ecosystems. The appropriation of ever-greater sections of the remaining native habitats, impelled by the increase in population as well as by the deterioration of farmland or rangeland due to overcultivation or overgrazing, decimates those habitats and imperils their ecologic functions. In the initial stages of degradation, the deteriorating productivity of an agroecosystem can be masked by increasing the inputs of fertilizers, pesticides, water, and tillage. Sooner or later, however, if such destructive effects as organic matter loss, erosion, leaching of nutrients, and salination continue, the degradation is likely to reach a point at which its effects are difficult to overcome either ecologically or economically.

Key processes related to desertification include drought, primary production and carrying capacity, soil degradation, and use of water resources. The role of social factors is also important.

The Role of Drought

A typical feature of arid regions is that the mode (the most probable) amount of annual rainfall is generally less than the mean; i.e., there tend to be more years with below-average rainfall than years in which the rainfall is above average, simply because a few unusually rainy years can skew the statistical mean well above realistic expectations for rainfall in most years. The variability in biologically effective rainfall is yet more pronounced, as years with less rain are usually characterized by greater evaporative demand, so the moisture deficit is greater than that indicated by the reduction of rainfall alone. Timing and distribution of rainfall also play crucial roles. Below-average rainfall, if well distributed, may produce adequate crop yields, whereas average or even above-average rainfall may fail to produce adequate yields if the rain occurs in just a few ill-timed storms with long dry periods between them.

In semiarid agricultural regions, 'drought,' like desertification, is a broad, somewhat subjective term that designates years in which cultivation becomes an unproductive activity, crops fail, and the productivity of pastures is significantly diminished. Drought is a constant menace, a fact of life with which rural dwellers in arid regions must cope if they are to survive. The occurrence of drought is a certainty, sooner or later; only its timing, duration, and severity are ever

in doubt. And it is during a drought that ecosystem degradation in the form of denudation and soil erosion occur at an accelerated pace, as people try to survive in a parched habitat by cutting the trees for fuel and browse, and by animals overgrazing the wilted grass. The topsoil, laid bare and pulverized by tillage or the trampling of animals, is then exposed to a greatly increased risk of wind erosion. When the coveted rains recur, they tend to scour the erodible soil.

Any management system that ignores the eventuality of drought and fails to provide for it ahead of time is doomed to fail in the long run. That provision may take the form of grain or feed storage (as in the biblical story of Joseph in Egypt), or of pasture tracts kept in reserve for grazing when the regular pasture is played out, or of the controlled migration of pastoralists to other regions able to accommodate them for the period of the drought.

A much-debated question is whether the frequency, duration, and severity of droughts have been increasing in recent decades. One possibility is that the process of desertification, once begun, produces a feedback effect that is self-exacerbating. Some have hypothesized that the increase in atmospheric dust from denuded and wind-eroded drylands (the so-called dust-bowl effect), as well as from air pollution (as denudation of an area's vegetation is associated with biomass burning, which releases smoke into the air), may have changed the patterns of air mass movement and hence of precipitation. Another hypothesis is that droughts can be worsened by the increased reflectivity of the bared surface to incoming sunlight. That reflectivity, called 'albedo,' may rise from approximately 20% for a well-vegetated area to perhaps 35% or more for an exposed, bright, sandy soil. As a larger proportion of the incoming sunlight is reflected skyward rather than absorbed, the surface becomes cooler than it would be otherwise, and so the air in contact with the surface has less of a tendency to rise and condense its moisture so as to yield rainfall.

An additional effect of denudation is to decrease the interception of rainfall by vegetation and the infiltration rate, while increasing surface runoff, thereby reducing the amount of soil moisture available for evapotranspiration. Crops and grasses, which have shallower roots than trees and in any case transpire less than the natural mixed vegetation of the savanna, transpire even less when deprived of moisture during a drought. The 'biophysical feedback' hypothesis is that such changes may reduce regional precipitation. Lower rainfall leads in turn to more overgrazing and less regrowth of biomass, and to further reduction in reevaporated rain owing to the decline in soil moisture. Thus, the feedback hypothesis offers its own explanation as to why the drought in the African Sahel, for example, has tended to persist.

Primary Production and Carrying Capacity

The biological productivity of any ecosystem is due to its primary producers (known as autotrophs), which are the green plants growing in it. They alone are able to create living matter from inorganic materials. They do so by combining atmospheric carbon dioxide with soil-derived water, thus converting radiant energy from the sun into chemical energy in the process of photosynthesis. Green plants also respire, which is the reverse of photosynthesis, and in so doing they utilize part of the energy to power their own growth. The net primary production then becomes available for the myriad of heterotrophs, which subsist by consuming (directly or indirectly) the products of photosynthesis. A stable ecosystem is one in which production and consumption, synthesis and decomposition, are in balance over an extended period of time.

When humans enter into an ecosystem and appropriate some of its products for themselves, they do so in competition with other potential consumers. As populations increase, the tendency is to intensify the use of resources by promoting the production of desired goods while suppressing the species that do not serve that end. In the process, the ecosystem's biodiversity and natural productivity are profoundly affected. Especially affected are areas within the semiarid and arid regions, which, because of the paucity of water and the fragility of the soil (typically deficient in organic matter, structurally unstable, and highly erodible) are most vulnerable and least resilient.

The term 'carrying capacity' has been used to characterize an area's productivity in terms of the number of people or grazing animals it can support per unit area on a sustainable basis. However, the productive yield obtainable from an area depends on how the area is being used. Under the hunter-gatherer mode of subsistence, an area may be able to carry only, for example, 1 person per square kilometer, whereas under shifting cultivation it may carry 10, and under intensive agriculture perhaps 100. The intensive forms of utilization also involve inputs of capital, energy, and materials, such as fertilizers and pesticides, which are brought in from the outside to enhance an area's productivity. As the usable productivity is strongly affected by the supply of water (i.e., by seasonal rainfall), it varies from year to year and from decade to decade, so a long-term average is difficult to determine, especially given the prospect of climate change. It is therefore doubtful if any given area can be assigned an intrinsic and objectively quantifiable carrying

capacity. By whatever measure, however, the capacity of an area to support a given population is clearly diminished by human mismanagement.

Wherever human pressure on the land ceases or is diminished, even a severely eroded soil may recover gradually. However, on the time-scale of years to a few decades, especially if overgrazing and overcultivation continue, soil erosion may become, in effect, irreversible.

Soil Degradation and Rehabilitation

An important criterion of soil degradation is the loss of soil organic matter. Compared with soils in more humid regions, those in warm arid regions tend to be inherently poor in organic matter from the outset, owing to the relatively sparse natural vegetative cover and to the rapid rate of decomposition. The organic matter present is, however, vitally important to soil productivity. Plant residues over the surface protect the soil from the direct impact of raindrops and from deflation by wind, and help to conserve soil moisture by minimizing evaporation. Plant and animal residues that are partially decomposed and that are naturally incorporated into the topsoil help to stabilize its structural aggregates, which in turn enhance infiltrability, reduce water loss by runoff, and enable seed germination and root growth. The organic matter present also contributes to soil fertility by the gradual release of nutrients.

When the natural vegetative cover is removed, and especially when the soil is tilled and/or trampled repeatedly, there follows a rapid process of organic matter decomposition and depletion. Accelerated erosion also removes the layer of topsoil that is richest in organic matter. Consequently, the destabilized soil tends to form a surface crust that further inhibits infiltration. Water losses by both runoff and evaporation increase, and the soil loses an important source of nutrients. These destructive processes can be countered or ameliorated by methods of conservation management, including minimum or zero tillage, maintenance of crop residues, the periodic inclusion of green manures in the crop rotation, and agroforestry.

The destructive processes induced by soil mismanagement, and, in contrast, the constructive processes induced by conservation management, though seemingly local, may have – when practiced on a regional scale – an impact on climate charge. Soils subject to accelerated decomposition of organic matter tend to release carbon dioxide and thus contribute to the enhanced greenhouse effect. Conversely, soils that are being revegetated and enriched with organic matter can absorb and sequester quantities of carbon that are extracted from the atmosphere in photosynthesis.

Potentialities and Problems of Irrigation

Where fresh water resources (from riverine or underground sources) are available and can be utilized economically, irrigation can be an effective way to intensify and stabilize production in semiarid or arid regions, and thus to relieve the pressure on extensive areas of rain-fed land that are most vulnerable to degradation. Irrigation is the deliberate supply of water to agricultural crops, designed to permit farming in arid regions and to offset drought in semiarid regions. Even in areas where total seasonal rainfall is adequate on average, it may be poorly distributed during the year and variable from year to year.

Wherever traditional rain-fed farming is a high-risk enterprise owing to scarce or uncertain precipitation, irrigation can help to ensure stable production. It not only raises the yields of specific crops but also prolongs the effective crop-growing period in areas with dry seasons, thus permitting multiple cropping (two, three, or even four crops per year) where only a single crop could be grown otherwise. With the security provided by irrigation, additional inputs needed to intensify production further (pest control, fertilizers, improved varieties, etc.) become economically feasible. Irrigation reduces the risk of these expensive inputs being wasted by crop failure resulting from lack of water. Although irrigated land amounts to only some 17% of the world's total cropland, it contributes well over 30% of the total agricultural production. That vital contribution is most important in arid regions, where the supply of water by rainfall is least, even as the demand for water imposed by the bright sun and the dry air is greatest.

Irrigation, however, is not without its own problems. From its inception in the Fertile Crescent some six or more millennia ago, irrigated agriculture, especially in poorly drained river valleys, has brought about processes of degradation that have threatened its sustainability. The application of water to the land tends to raise the water table, which in turn induces the self-destructive twin scourges of waterlogging and salination.

Some investigators include the degradation of irrigated lands in the category of desertification. Though the processes taking place differ fundamentally from those in rain-fed lands, the damage done to injudiciously irrigated lands is indeed in the category of ecosystem degradation. Processes occurring off-site (upstream as well as downstream of the irrigated area) strongly affect the sustainability of irrigation. For example, denudation of upland watersheds by forest clearing, cultivation, and overgrazing induces erosion and the subsequent silting of reservoirs and canals, thereby reducing the water supply. The

construction of reservoirs often causes the submergence of natural habitats as well as of valuable scenic and cultural sites. Concurrently, the downstream disposal of drainage from irrigated land tends to pollute aquifers, streams, estuaries, and lakes with salts, nutrients, and pesticide residues. Finally, the irrigation system itself (its reservoirs, canals, and fields) may harbor and spread water-borne diseases, thus endangering public health. Thus the very future of irrigated agriculture has been called into question.

Experience shows that irrigation can be sustained, but at a cost. The primary cost is the necessary investment in systems of efficient irrigation (avoiding excessive application of water such as causes water-table rise, waterlogging, and salination), as well as in the timely provision of effective land drainage and the safe disposal of its salt-laden effluent.

Social Factors

Social factors are necessarily involved in both semiarid ecosystem conservation and its inverse, which is ecosystem degradation. Farmers who do not have tenure to the land are not likely to invest in its conservation or improvement. Neither are communities that lack stable institutional structure likely to establish and maintain essential infrastructure and services that enable, encourage, and coordinate farmers' efforts to implement land improvement and conservation measures. And no effective action at all may be possible in the absence of a proactive governmental policy, including the provision of credit or subsidies, professional guidance and training, as well as the preparation and implementation of national and regional drought contingency plans for both farmers and herders. The conservation of soil, water, and biotic resources is a collective societal concern, and an intergenerational one, not merely a private concern of the people utilizing the land directly at any particular time.

Finally, there looms the most difficult, yet inescapable, problem of population numbers. No system of management, however efficient it may be, can be sustained if the population continues to grow without limit. A crucial aspect of population control is the empowerment of women, through education and equal rights, as full participants in the management of their societies' physical, biological, and human resources. The issue is extremely sensitive inasmuch as it carries cultural and traditional, as well as social and economic, implications.

Monitoring Desertification

The techniques of remote sensing have made possible the monitoring of changes to ecosystems on a regional scale. Studies based on remote sensing of the African Sahel have shown that, contrary to many alarmist reportings, there has been no progressive change of the Saharan desert boundary. Rather, there has been a back-and-forth shifting of vegetative density during alternating spells (sometimes lasting several years) of below-average and of above-average rainfall. In principle, however, statistical criteria designed to test the probability levels of differences (between sites or between successive measurements on the same site) should not be used to 'prove' the opposite, namely that there are no differences. Measurements made at various times on large areas may obscure subtle local changes that could have taken place on a smaller scale.

One of the main indexes in use at present is the so-called normalized difference vegetation index (NDVI). It is obtained from the ratio between the red and near-red infrared reflectance bands, obtained from high-resolution radiometer data generated by the polar-orbiting satellite of the US National Oceanic and Atmospheric Administration (NOAA). In arid and semiarid regions, NDVI evidently correlates with the density of the vegetative cover and its biomass, as well as with its 'leaf area index' (LAI) and its photosynthetic activity. Care is needed, however, in applying NDVI to the assessment of net primary production, since the measurement of NDVI is oblivious to the amount of vegetation harvested by humans and by their animals prior to the time of the measurement.

Taken to be a general indicator of the 'greenness' of an area, NDVI has also been conjectured to correlate with biological productivity, but that correlation may not necessarily hold. In principle, the amount of vegetation present per unit area should depend on the amount produced *in situ* minus the amount removed from it. Therefore, the relation between an area's productivity and its vegetative biomass at any time must depend on whether the vegetation has been or is being 'harvested' (grazed by livestock or cut and carried away by humans). An area could be quite productive, yet relatively bare, if it had been harvested just prior to the NDVI measurement. Even if there is no discernible change in the density of an area's overall vegetative cover, there might well be a considerable change in the composition of the vegetation (i.e., its biodiversity, ecologic function, and feed value). For example, an overgrazed area may exhibit a proliferation of less nutritious plants at the same time that it loses the most palatable species of grasses and legumes that had contributed to the area's original carrying capacity.

Clearly, the most decisive factor affecting the overall density of vegetative cover in an arid region is the fluctuation of rainfall amounts. Taking the African Sahel as an example once again, we see that the annual precipitation has fluctuated widely over the decades. The amounts (as seen in Figure 1) for the last

three decades of the twentieth century appear to be generally lower than those in the preceding decades. Although an analysis based on any particular short period may be misleading, the question does arise as to the possible effects of global climate change.

The Role of Climate Change

Ecosystems in arid and semiarid regions are likely to be increasingly influenced by global climate change. Emissions of radiatively active gases and aerosols due to human activity are altering the Earth's radiation balance and hence the temperature of the lower atmosphere. One of the manifestations of the change may be an increase in climate instability. In a warmer world, climatic phenomena are likely to intensify. Thus, episodes or seasons of anomalously wet conditions (violent rainstorms of great erosive power) may alternate with severe droughts, in an irregular and unpredictable pattern.

A more unstable climatic regime will make it harder to devise and more expensive to implement optimal land use and agricultural production practices, including drought-contingency provisions. Failure to prepare for such contingencies may exacerbate the consequences of extreme events such as floods and droughts, to the effect of worsening land degradation and periods of severe food shortages. Especially vulnerable is the continent of Africa, where the issues of greatest concern pertain to human health and food security, water resources, natural ecosystems and biodiversity, and, not least, land degradation or desertification (Table 1).

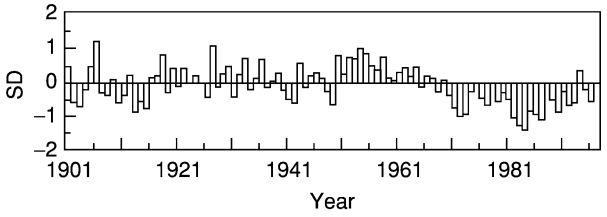


Figure 1 Rainfall fluctuations in the African Sahel during the period 1901–1998, expressed as a regionally averaged standard deviation, SD (departure from the long-term mean divided by the standard deviation). Reproduced from Densanker P and Magadza C (2001). Africa in climate change 2001: impacts, adaptation and vulnerability. In: McCarthy JJ, Canzania OF, Leary NA, Dokker DD, and White KS (eds) *International Panel of Climate Change*.

Climate change appears likely to cause further semiarid ecosystem degradation through alteration of spatial and temporal patterns in temperature, rainfall, solar insolation, winds, and humidity. Analyses based on global and regional climate models suggest that droughts may become more frequent, severe, and prolonged. It is still impossible, however, to ascribe exact probabilities to any of the various climate change scenarios, owing to uncertainties regarding future emissions of the radiatively active trace gases and tropospheric aerosols, and the potential responses of the climate system to those changes.

The question is sometimes posed: is it human exploitation of the land or is it overall climate change that constitutes the predominant cause of desertification? The answer is that the two sets of factors or processes interact and may well have become mutually reinforcing. Ultimately, both are impelled by human intervention and therefore can only be redressed by coordinated actions at the local, regional, and global levels.

Overview

The pressures generated by growing populations, intensified land use, and overall environmental change are evidently causing a progressive degradation of natural and managed ecosystems, especially in arid and semiarid regions. To define and quantify the nature, degree, and extent of the degradation, national and international agencies are working to implement consistent monitoring programs. These consist not only of remote sensing from above but also of ground-based observations.

Table 1 Sectors vulnerable to climate change in Africa

Sector	*Projected impacts*
Water resources	Dominant impact is predicted to be a reduction in soil moisture in the subhumid zones and a reduction in runoff
Food security	There is wide consensus that climate change, through increased extremes, will worsen food security in Africa
Natural resources and biodiversity	Climate change is projected to exacerbate risks to already threatened plant and animal species, and fuelwood
Human health	Vector-borne and water-borne diseases are likely to increase, especially in areas with inadequate health infrastructure
Desertification	Changes in rainfall, increased evaporation, and intensified land use may put additional stresses on arid, semiarid, and dry subhumid ecosystems

Reproduced from Densanker P and Magadza C (2001) Africa in climate change 2001: impacts, adaptation and vulnerability. In: McCarthy JJ, Canzania OF, Leary NA, Dokker DD, and White KS (eds) *International Panel of Climate Change*.

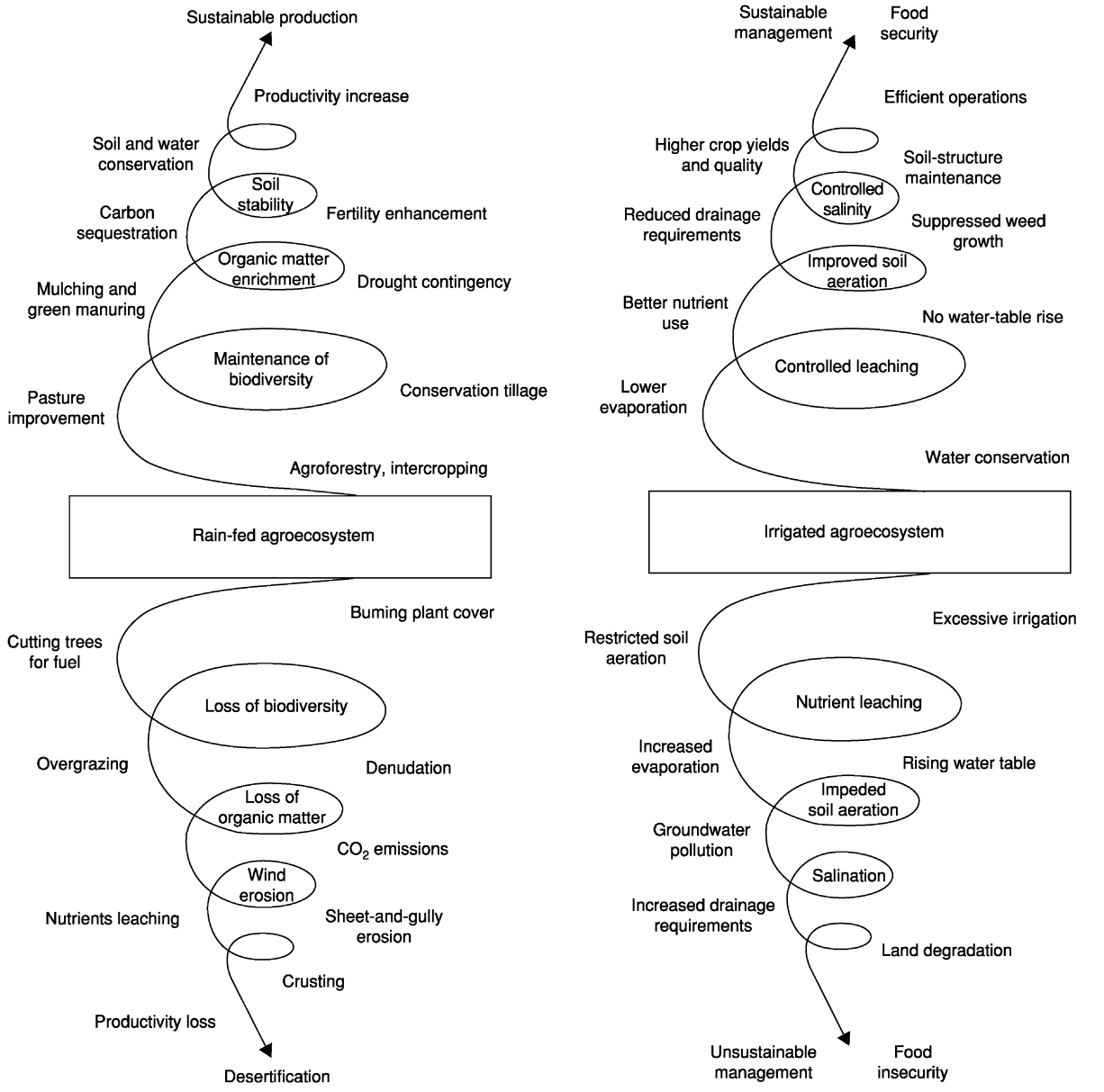


Figure 2 Upward and downward spirals of sustainable versus unsustainable patterns of management in rain-fed and irrigated agriculture in arid regions. Reproduced with permission from Hillel D and Rosenzweig C (2002) Desertification in relation to climate variability and change. *Advances in Agronomy* 77: 1–38.

To redress or rehabilitate degraded ecosystems, vulnerable countries are beginning to institute appropriate policies and programs. These include keeping reserve areas to protect biodiversity, avoiding overgrazing on managed lands, reseeding of pastures, implementing soil and water conservation measures, and – in the social arena – land tenure, family planning, and contingencies for droughts. Determining the availability of fresh water resources (surface water, renewable groundwater, and, in some cases, nonrenewable groundwater as well) and planning their careful utilization are important components of such programs. Inappropriate patterns of management may lead to a downward spiral of ecosystem degradation, whereas appropriate measures of conservation and sustainable use hold the promise of sustainable development in the context of both rain-fed and irrigated agriculture (Figure 2).

See also: **Degradation**; **Erosion:** Water-Induced; Wind-Induced

Further Reading

Bouwman AF (ed.) (1992) *Soil and the Greenhouse Effect*. Chichester, UK: John Wiley.

Charney JG (1975) Dynamics of deserts and drought in the Sahel. *Quarterly Journal of the Royal Society* 101: 193–202.

Dregne HE (1994) Land degradation in the world's arid zones. In: Baker RS, Gee GW, and Rosenzweig C (eds) *Soil and Water: Keys to Understanding Our Global Environment*. SSSA Special Publication No. 41. Madison, WI: Soil Science Society of America.

Glantz MH and Orlovsky N (1983) Desertification: a review of the concept. *Desert Control Bulletin* 9: 15–22.

Hillel D and Rosenzweig C (2000) Desertification in relation to climate variability and change. *Advances in Agronomy* 27: 1–38.

Justice CD (ed.) (1986) *Monitoring the Grassland of Semi-arid Africa Using NOAA–AVHRR Data*. London, UK: Taylor & Francis.

Lal R (2001) Potential of desertification control to sequester carabon and mitigate the greenhouse effect. *Climatic Change* 51: 35–72.

Le Houreou HN (1977) Biological recovery versus desertization. *Economic Geography* 63: 413–420.

Mainguet M (1994) *Desertification: Natural Background and Human Mismanagement*. Berlin, Germany: Springer-Verlag.

McCarthy JJ, Canziani OF, Leary NA, Dokken DJ, and White KS (eds) (2001) *Intergovernmental Panel on Climate Change*. Cambridge, UK: Cambridge University Press.

Nicholson S (2000) Land surface processes and Sahel climate. *Reviews of Geophysics* 38: 117–139.

Rind D, Rosenzweig C, and Peteet D (1989) African drought: history, possible causes, prognosis. In: Lemma A and Malaska P (eds) *Africa Beyond Famine*. London, UK: Tycooly.

Tucker CJ, Dregne HE, and Newcomb WW (1991) Expansion and contraction of the Sahara desert from 1980 to 1990. *Science* 253: 299–301.

UNEP (1992) *World Atlas of Desertification*. Seven Oaks, UK: Edward Arnold.

DIFFUSION

T M Addiscott, Rothamsted Research, Harpenden, UK
P Leeds-Harrison, Cranfield University, Silsoe, Bedford, UK

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

In soil science the diffusion of gases and solutes needs to be understood. A gas in a closed vessel distributes itself such that its pressure is the same at all points. If there is a mixture of gases, each exerts its own 'partial pressure' and distributes itself such that its partial pressure is the same throughout the vessel. If you change the partial pressure of one gas in any part of the vessel there will be a gradient in its partial pressure, and the resulting flow of that gas within the other gases will tend to equalize its partial pressure throughout the volume. A solute in a solution behaves in several ways like a gas in a closed vessel. It has a chemical potential, which depends on its concentration and is conceptually related to the partial pressure of a gas through the thermodynamic concept of free energy. The solute distributes itself so that its concentration and chemical potential are the same throughout the solution, as a gas does with its partial pressure in a closed vessel. Changing the concentration in any part of the solution leads to a gradient in chemical potential which causes a flow of solute that tends to equalize the concentration. You can readily watch this happening by placing a crystal of potassium permanganate at the bottom of a beaker of water. There is a gradient of chemical potential between the solution adjacent to the crystal and the rest of the water, and the rich purple color of the permanganate gradually spreads out into the water. This process, which is essentially the same for gases and solutes, is diffusion, which involves gases that diffuse in the soil atmosphere, that is, within other gases, and solutes that diffuse in the soil solution, often interacting with soil solids. Although diffusion in solution is caused in principle by gradients in chemical potential, the topic is usually treated in terms of the concentration.

How Diffusion Occurs

Diffusion occurs because of Brownian motion, the random movement of ions or molecules in a state of thermal motion. Although the molecules move randomly, the probability is that gas molecules will move from high-pressure to low-pressure zones, and solute molecules from zones of high concentration to zones of lower concentration in a solution. Because gradients in partial pressure and concentration are gradually lessened by diffusion but never eliminated, partial pressures and concentrations of solute become uniform only at infinite time (although in practice many diffusive flows become too small to measure after a relatively short time). This is why many diffusion

equations include the concentration at infinite time C_∞ or the mass diffusing in infinite time M_∞.

Fick's Laws

The first mathematical description of diffusion was given by Fick in 1885. He based his ideas on the mathematical treatment of the conduction of heat given by Fourier in 1822. Fick's equations are similar to other flux equations in which the flux is proportional to the gradient in some property of the system. The first law of Fick is stated in mathematical form as:

$$F = -D\frac{\partial C}{\partial x} \qquad [1]$$

where F is the amount of diffusing substance passing perpendicularly through a reference surface in unit time per unit cross-sectional area, C is the concentration, x is the space coordinate, and D is the diffusion coefficient, which has dimensions of L^2T^{-1} (a surface area per unit time). The value of D depends on the units of distance and time, but is independent of the units in which C is measured. In eqn [1], D is often assumed constant. However, in some circumstances it depends on the concentration of diffusing substance in the medium.

A quantitative prediction of the rate at which a diffusion process occurs can be obtained from eqn [1] when the diffusion coefficient is known. The equation expressing the change in concentration with time is obtained by combining the concept of diffusion with that of continuity:

$$\frac{\partial C}{\partial t} = -\frac{\partial F}{\partial x} \qquad [2]$$

to give the transient state equation in one dimension:

$$\frac{\partial C}{\partial t} = \frac{\partial}{\partial x}\left[D\frac{\partial C}{\partial x}\right] \qquad [3]$$

When the diffusion coefficient D is constant, eqn [3] becomes the second law of Ficks:

$$\frac{\partial C}{\partial t} = D\frac{\partial^2 C}{\partial x^2} \qquad [4]$$

Initial and Boundary Conditions

To solve eqns [1–4] for real systems, initial and boundary conditions must be known for the system under study. Very often we set our initial conditions to those of uniform concentration throughout the system or within a subsystem. A perturbation at a boundary at time zero then sets up concentration gradients and fluxes of diffusing material, and concentrations within the system can be found at subsequent times.

The boundary conditions of the region of interest are of critical importance. The boundary conditions most usually encountered are:

1. The Dirichlet or concentration condition. Here the concentration is maintained constant at the boundary of the system. As an example, when considering oxygen diffusion into soil, such a boundary occurs at the soil surface, where the concentration of oxygen in the air is constant and equal to that of the atmosphere;
2. The Neuman or flux condition. For the case where a flux density, $F(t)$, is imposed for $t \geq 0$ on a boundary, the Neuman condition is written (after eqn [1]):

$$-D\frac{\partial C}{\partial x} = F(t)$$

When the flux at a surface is zero, as occurs at impermeable barriers, the Neuman boundary condition is:

$$\frac{\partial C}{\partial x} = 0$$

3. The Robin or radiation condition. For the case where the flux across the surface is proportional to the difference between the surface concentration, C_s, and the surrounding medium concentration, C_0, the boundary condition is:

$$-D\frac{\partial C}{\partial x} = \beta(C_s - C_0)$$

where β is a proportionality constant.

Often the complexity of real soil systems means that some simplifying assumptions are made about the nature of the boundaries so that analytical solutions approximating to the real situation can be found. These may include the simplification of the surface geometry, for instance, assuming aggregates of soil to be spherical, or assuming the solute concentration in macropores surrounding an aggregate to be constant. Interpretation of the results from experiments on diffusion in soils within the framework of theoretical analysis always needs to take account of the assumptions that have been made about initial and boundary conditions.

Components of the Diffusion Coefficient in Soil

The diffusion coefficient can in principle be measured in aqueous solution, although the practice is not always simple. Soil introduces further complications,

particularly if the diffusing substance is a solute which is sorbed by the soil. Nye and his colleagues developed a theory for assessing the components of the diffusion coefficient of such a solute. They started with the assumption that both the solid and liquid phases contributed to the diffusion coefficient, but experiments with chloride, phosphate, sodium, and strontium ions suggested that the contribution by the solid phase was not important. There was a suggestion that exchangeable (in contrast to solution) sodium made a contribution to the diffusion constant for sodium but only at the lowest water contents used. Exchangeable strontium seemed to make no contribution, and neither did sorbed phosphate. This suggested that for most situations likely to occur in practice all diffusive flow occurred in the solution phase.

The relation between the diffusion constant in soil D and that in free solution D_l has to take account of three soil characteristics:

1. The fraction of the area perpendicular to the direction of flow in which diffusive flow occurs. This is the same as the fraction of the soil volume occupied by the solution v_l and it is dimensionless;
2. The impedance factor f_l which allows for two factors, the tortuous path that diffusive flow has to follow round soil particles and the probability that some pathways are 'dead ends.' For anions, both factors are likely to be accentuated by anion exclusion (the repulsion of anions from the negatively charged surfaces that predominate in nonacid soils). The impedance factor is also dimensionless;
3. The slope of the isotherm relating the solute concentration C_l in solution to the concentration in the whole soil C. The isotherm is usually taken to be linear.

Combining these factors gives the following:

$$D = D_l v_l f_l (C_l/C) \qquad [5]$$

Of these components of the diffusion coefficient, D_l will usually be known and v_l and (C_l/C) can usually be measured. The impedance factor f_l cannot usually be measured independently and often has to be obtained by measuring D_l when the other components are known. It might be possible to estimate it from acoustic or conductivity measurements.

Diffusion and Geometry

The nature and mathematical treatment of diffusion depend on the shape of the volume into or from which the diffusion occurs. Analytical solutions of eqn [4] have been developed for various geometric shapes, which are represented by the appropriate boundary conditions. Although these solutions can be applied only to simple geometries and constant diffusion coefficients, they have proved widely useful. The mathematical treatment for a sphere is presented below. That for a cylinder is somewhat similar.

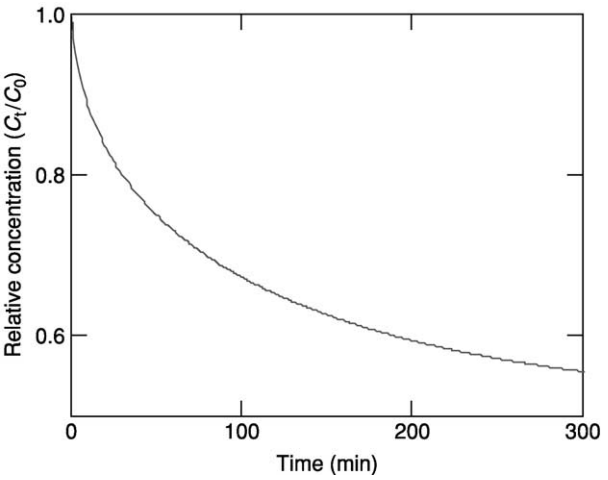


Figure 1 Change in concentration of solute within an aggregate as a function of time.

Diffusion in a sphere needs to be described in spherical coordinates r, θ, ϕ:

$$\frac{\partial C}{\partial t} = \frac{D}{r^2}\left\{\frac{\partial}{\partial r}\left(r^2\frac{\partial C}{\partial r}\right) + \frac{1}{\sin\theta}\frac{\partial}{\partial\theta}\left(\sin\theta\frac{\partial C}{\partial\theta}\right) + \frac{1}{\sin^2\theta}\frac{\partial^2 C}{\partial\phi^2}\right\} \qquad [6]$$

Here we assume that the concentration of solute within the sphere is C_0 and the solution is free from solute. If the solution is well stirred, the concentration of the solution depends only on time, while the total amount of solute in the solution and sphere remains constant.

The time rate of change of solute concentration, C, within the sphere (Figure 1) is described by eqn [5], and, since there is only variation of concentration radially, we have:

$$\frac{\partial C}{\partial t} = D_b\left(\frac{\partial^2 C}{\partial r^2} + \frac{2}{r}\frac{\partial C}{\partial r}\right) \qquad [7]$$

where r is the radial coordinate in the sphere, since the concentration depends on the radius only and is independent of both angles, θ and ϕ.

Some Useful Analytical Solutions for Various Geometries

Diffusion from a Semiinfinite Medium

A semiinfinite medium is notionally one which is large enough for diffusion to occur from it for infinite

time without becoming limited by the length dimension. It is semiinfinite because if it was completely infinite it would have no face across which diffusion could occur. Neither C_∞ nor M_∞ appears, and the solution for the mass M_t diffusing from the medium in time t is given by Crank as:

$$M_t = 2C\sqrt{\frac{Dt}{\pi}} \qquad [8]$$

where C is concentration of the diffusing substance within the medium, and D is again the diffusion coefficient. Eqn [8] provides an example of the general rule that all diffusion processes are characterized by a relation to the square root of time.

Diffusion in a Cylinder

Only radial diffusion is considered. The cylinder is assumed to be long enough for diffusion along it to be ignored. In the solution supplied by Crank, the mass M_t diffusing in time t from a cylinder of diameter a is given by the series:

$$\frac{M_t}{M_\infty} = \left(\frac{4}{\pi^{1/2}}\right)\left(\frac{Dt}{a^2}\right)^{1/2} - \frac{Dt}{a^2} - \left(\frac{1}{3}\pi^{1/2}\right)\left(\frac{Dt}{a^2}\right)^{3/2} + \ldots \qquad [9]$$

where D is the diffusion coefficient. If the series converges sufficiently rapidly for the third and higher terms to be ignored:

$$\left(\frac{1}{t}\right)\frac{M_t}{M_\infty} = \left(\frac{4}{\pi^{1/2}}\right)\left(\frac{D}{a^2 t}\right)^{1/2} - \frac{D}{a^2} \qquad [9a]$$

Plotting $(1/t)(M_t/M_\infty)$ against $t^{-1/2}$ should enable D and a to be estimated from the slope and intercept of the resulting relation.

Diffusion in a Sphere

Crank has supplied an analytical solution to eqn [6] in which the mass of solute diffusing from a sphere into a well-stirred solution of limited volume at a constant concentration is:

$$\frac{M(t)}{M_\infty} = 1 - \sum_{n=1}^{\infty}\frac{6}{\pi^2 n^2}\exp\left(\frac{-9n^2\pi^2 Dt}{a^2}\right) \qquad [10]$$

If we apply the following initial and boundary conditions to eqn [6]:

$$C = C_0 \qquad 0 \le r \le a, t = 0$$
$$C_M(t) = 0 \qquad r = a, t \ge 0$$
$$C = C_M(t) \qquad r = a, t \ge 0$$

where $C_M(t)$ is the concentration in the solution external to the sphere at time t and a is its radius. An analytical solution for the above diffusion equation for uniform size spheres is given by Crank:

$$\bar{C}(t) = C_\infty + (C_0 - C_\infty)\sum_{n=1}^{\infty}\frac{6\alpha(1+\alpha)\exp(-D_b q_n^2 t/a^2)}{9 + 9\alpha + q_n^2\alpha^2} \qquad [11]$$

where C_∞ is the equilibrium concentration at infinite time and q_n is the positive root of:

$$\tan q_n = \frac{3q_n}{3 + \alpha q_n^2} \qquad [11a]$$

and $\alpha = V_M/V_A$, the ratio of the volume of external solution, V_M, to that of the sphere, V_A.

Simple Model for Various Geometries

The analytical approaches outlined above are limited in their applicability by their inability to deal with externally imposed changes in concentration. For example, diffusion from spherical aggregates into the interaggregate solution is likely to be interrupted in the real world by rainfall, which changes the concentration of the solution. The theory cannot allow for this. Furthermore, few soils comprise uniformly sized aggregates. Most aggregate sizes are log-normally distributed. Analytical approaches typically deal with this problem by using a single aggregate size computed on a volume-weighted basis. This, however, is not really adequate because of the interaction between aggregates of different sizes. Small aggregates lose, for example, 50% of their solute very much more rapidly than larger aggregates. This is taken into account in the volume-weighting process, but what is not accommodated is that the solute lost rapidly from the smaller aggregates decreases the concentration gradient determining diffusion from the larger ones.

A simple nonanalytical model was developed for cubes and other regularly shaped aggregates. This is based on a cube which is divided into concentric equal volumes. The model uses the first law of Ficks to compute the flux between adjacent volumes and between the outermost volume and the external solution. The model can be used without amendment for diffusion in a sphere and with minor amendments for a regular tetrahedron or a rhombic octahedron. Shapes with larger numbers of faces are unlikely to be more relevant than a sphere for simulating soil aggregates. This model can accommodate changes imposed from outside the system in the concentration of the external solution – or changes in its volume. It can

also cope with normal or log-normal distributions of aggregate sizes. Its main disadvantage is that it is less exact than the analytical solutions.

Applications in Soil Science

Plant Nutrition

Many studies have been made on the mechanisms by which plant nutrients reach roots. Three mechanisms are postulated: diffusion, convection (mass flow), and root extension. The dominant mechanism depends on the nutrient studied and the experimental conditions. The nutrients with the largest concentrations in the soil solution, often calcium and nitrate, tend to be carried to the roots in sufficient quantities by convection. Where convective flow fully supplies the needs of the plant, no appreciable concentration gradient builds up and no diffusion occurs. Nutrients such as potassium and phosphate strongly held by the soil and therefore at low concentration in the soil solution are less likely to be supplied adequately by convection. This can result in a low concentration at the root surface and the development of a concentration gradient which makes diffusion the main transport mechanism. Diffusion to a root is often studied using the theory for diffusion in a cylinder and considers a cylinder of soil around the root. This is satisfactory for isolated roots, but roots rarely exist in isolation from other roots, and dealing with interacting cylinders is far more difficult. This is part of a general problem: diffusion to roots is difficult to measure other than in specially designed systems that are somewhat artificial. The studies have generally suggested that root extension is not important by comparison with the other mechanisms, but this result could reflect the fact that they were made at a fairly small scale, usually in pots in a glasshouse. Root extension could have been more important at the field scale.

Kinetics of Potassium Release From Clays

The release of nonexchangeable potassium from illites and related clays has been treated as very slow diffusion in a cylinder, with a diffusion coefficient of the order of 10^{-18}–$10^{-20}\,\mathrm{cm^2\,s^{-1}}$. Eqns [9] and [9a] are used for this purpose.

Gaseous Diffusion

Air occupies approximately one-third of the volume of the soil but is not very free to move. There is therefore little convective movement of gases in the soil. Because of the consumption of oxygen by aerobic microbes and the generation of carbon dioxide by most microbes, there can be substantial gradients in the partial pressures of these gases. Diffusion therefore plays an important role in the transport of these gases and also of dinitrogen and nitrous oxide when denitrification occurs. We are often concerned with diffusion through the whole soil or, more correctly, through the larger pores of the soil. But diffusion of oxygen within aggregates is also important, because it determines whether or not anoxic zones develop at the centers of the aggregates, with the consequent risk of denitrification and other anoxic processes. This oxygen diffusion is usually treated from a theoretical standpoint as diffusion within spheres.

Leaching

The nature of solute leaching depends greatly on the texture and structure of the soil. In soils in which there is appreciable aggregation, water moves to a large extent around the aggregates and solutes within the aggregates enjoy temporary protection from leaching. This only lasts until they diffuse out of the aggregates into the water flowing round them. Leaching and intraaggregate diffusion therefore need to be simulated together. This has been done using the theory for diffusion in spherical aggregates and also the simple model for various geometries. As noted above, the latter is better able to cope with a distribution of aggregate sizes but is less exact.

Of particular interest in leaching is the behavior of pesticides. Many pesticides partition between the solid (soil particles, particularly organic matter and clays) and the liquid in the pores. To describe the movement of reactive solute, a retardation factor, R, is introduced which accounts for the partitioning of the solute. It is defined as:

$$R = 1 + \frac{\rho K_\mathrm{d}}{\theta}$$

where K_d is the partition coefficient describing the ratio of solute sorbed to solid particles to the solute concentration in solution at equilibrium, ρ is the bulk density of the soil, and θ is the water content. The effective diffusion coefficient, D_eff, is then defined as $D_\mathrm{eff} = (D/R)$.

List of Technical Nomenclature

Diffusion coefficient	Proportionality coefficient required by Fick's law. Dimensions L^2T^{-1}
Impedance	Factor allowing for tortuous path of diffusion, including dead-end pores. Dimensionless

Reference area Area perpendicular to direction of flow in which diffusion occurs. For solutes this is multiplied by the fraction of the soil volume occupied by the soil solution to give the area across which diffusion occurs, and for gases by the fraction occupied by air. Dimensions L^2

Further Reading

Addiscott TM (1982) Simulating diffusion within soil aggregates: a simple model for cubic and other regularly shaped aggregates. *Journal of Soil Science* 33: 37–45.

Addiscott TM, Armstrong AC, and Leeds-Harrison PB (1998) Modeling the interaction between leaching and intraped diffusion. In: Selim HM and Ma L (eds) *Physical Nonequilibrium in Soils. Modeling and Application*, pp. 223–241. Chelsea, MI: Ann Arbor Press.

Crank J (1975) *The Mathematics of Diffusion*, 2nd edn. Oxford, UK: Oxford Scientific Publications.

Currie JA (1961) Gaseous diffusion in the aeration of aggregated soils. *Soil Science* 92: 40–45.

Hartley GS and Graham-Bryce IJ (1980) *Physical Principles of Pesticide Behaviour*, vol. 1, *The Dynamics of Applied Pesticides in the Local Environment in Relation to the Biological Response*, ch. 3. London, UK: Academic Press.

Nye PH and Tinker PB (1977) *Solute Movement in the Soil–Root System*. Oxford, UK: Blackwell.

Olsen SR and Watanabe FS (1963) Diffusion of phosphorus as related to soil texture and plant uptake. *Soil Science Society of America Proceedings* 27: 648–653.

Rao PSC, Jessup RE, Rolston DE, Davidson JM, and Kilcrease DP (1980) Experimental and mathematical description of nonadsorbed transport by diffusion in spherical aggregates. *Soil Science Society of America Journal* 44: 684–688.

Rowell DL, Martin MW, and Nye PH (1967) The measurement and mechanism of ion diffusion in soils. III. The effect of moisture content and soil solution concentration on the self-diffusion of ions in soils. *Journal of Soil Science* 18: 204–222.

Smith KA (1980) A model of the extent of anaerobic zones in aggregated soils, and its potential application to estimates of denitrification. *Journal of Soil Science* 31: 263–277.

DISINFESTATION

A Gamliel, The Volcani Center, Bet Dagan, Israel
J Katan, The Hebrew University of Jerusalem, Rehovot, Israel

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Soil disinfestation is one of the most effective means of controlling soilborne pests and improving plant health. Soil disinfestation is a drastic means applied to soil before planting in order to reduce or eliminate soilborne pests. Effective soil disinfestation aims to promote healthy and productive crops. All crops are sensitive to one or more harmful biotic or abiotic soilborne agents that affect plant health and productivity. The incidence of these agents increases with frequent cropping, especially under monoculture. The biotic agents consist of major and minor pathogens, including fungi, bacteria, nematodes, soil viruses, arthropods, parasitic plants, and weeds. These are referred to here as 'pests.' The abiotic agents include, among others, the accumulation of harmful chemicals from various sources, deficiencies in essential mineral nutrients, and deterioration in physical status of the soil. In addition, the population of beneficial microorganisms such as mycorrhizae and plant growth-promoting rhizobacteria may also decrease with continuous cropping. These harmful effects are reflected in either typical disease symptoms and eventually plant death or in poor plant health and growth retardation. The latter phenomena (which are especially connected with continuous cropping of the same crop in the same plot) are also referred to as soil 'fatigue,' soil 'sickness,' and 'replant disease' (the latter typical in tree plantations). Soil disinfestation, crop rotation, and specific treatments are potential tools for controlling these phenomena, improving plant health, and resuming soil productivity.

The increase in the incidence of harmful biotic and abiotic agents with a long history of cropping stems from the fact that frequent planting of the same crop causes the enrichment and consequently rapid buildup of detrimental biological, chemical, and physical agents in the soil. Soil disinfestation aims to control these agents and thereby improve plant health, and reestablish high and sustainable levels of yield. These goals are especially important in regions where both crop options and agricultural land are limited. Therefore soil disinfestation can be regarded as soil reclamation.

Disinfestation in its present form was established at the end of the nineteenth century, in the early stages of the establishment of crop-protection sciences. The spread of the soil pest *Phylloxera* in vineyard soils in France during that period led to the introduction of carbon disulfide (CS_2), the first soil fumigant, for soil disinfestation. A physical means of soil disinfestation, steaming, was also developed. Approximately 100 years later, a third approach for soil disinfestation, 'soil solarization,' was developed.

Principles of Soil Disinfestation

The main goal of soil disinfestation is to eradicate, usually before planting, harmful, soilborne biotic agents, uniformly to the desired depth, with minimal disturbance of the biological equilibrium and with minimal effect on chemical or physical soil properties. This has to be achieved using an effective, feasible, economic and environmentally acceptable technology. With respect to the spectrum of pest control, we face a dilemma: wide-spectrum pest control is economically and practically desirable, but usually involves the use of means that can be biologically destructive. This is the case when soil disinfestation is carried out using drastic physical or chemical means to ensure the control of the harmful biotic agents in the deeper soil layers, as well as in all other soil niches. Therefore, soil disinfestation has to be carried out before planting to avoid harm to the crop. Also, the control agent, or its decomposition product, has to dissipate before planting to avoid residual harmful effects. Soil disinfestation was initially developed for the control of harmful biotic agents, although possible effects on abiotic soil components, e.g., improvement of mineral nutrition status of the plant, should not be excluded.

The basic principles of soil disinfestation are:

1. Treating the soil with a wide-spectrum control agent in order to eliminate (or strongly reduce) populations of a variety of soilborne pests;
2. Causing minimal disturbance to the biotic (especially beneficial microorganisms) and abiotic components of the soil;
3. Effectively reaching and controlling the pests with the control agent to the desired soil depth, usually to 30–50 cm, or even deeper in some cases. In soil-less culture, this layer is shallow;
4. Using most of the existing technologies, treating the soil before planting;
5. Avoiding contamination from outside sources. soil disinfestation can only control the existing populations of the soil pests.

Methods of Soil Disinfestation

Physical

Steaming, aerated steam, and hot-water treatments are used in greenhouses, especially with container (growth) media. Steam has been utilized for soil disinfestation for more than a century. Resting structures (i.e., cells or structures of microorganisms that can survive for a long period in soil) of plant pathogens and other pests, e.g., chlamydospores, sclerotia, oospores, and bacterial spores, are eliminated by steaming if heated to lethal temperatures, even in cases of heavy soil contamination. Moreover, steaming frequently has a growth-stimulating effect on the subsequent crop.

Careful soil preparation is essential for good steam penetration. The soil should be tilled as deeply as possible and then left to dry completely before steaming. It is important to reduce the amount of plant debris. Good preparation permits good steam penetration and enables pest control even in heavy soils. The steaming of aerated growth substrates is usually effective, but peat poses difficulties owing to its high water content.

The soil is steamed by either 'passive' or 'active' techniques. In passive steaming, the steam is blown to the surface, under a covering sheet, and left to heat the upper layer. Lower layers are then heated by heat transmission. This process continues until 100°C is reached at a depth of 10 cm. Disinfestation of deeper layers, especially in sandy soil, may only be partial.

Active steaming can be performed using either positive or negative pressure. Both techniques employ drainage systems, based on pipes laid at 50–70 cm depth, approx. 80 cm apart. With the positive-pressure technique, the steam is blown through holes located along the pipes. The negative pressure involves an improved technique, utilizing the advantages of the other two application methods. As in passive steaming, the steam is released over the treated area under plastic sheeting, ensuring rapid and even distribution over the entire surface of the plot, followed by active suction into the deeper layers of the soil, achieved by negative pressure applied through the drainage system. This technique, widely used in the Netherlands, is cheaper than the other two, due to the energy savings incurred by the faster heat transfer. Nevertheless, steaming treatments are expensive and are feasible mainly in places where there are heating systems (for heating the greenhouse during the cold season) or if applied by contractors. Steaming, however, can be useful and economic for the disinfestation of shallow layers of growth media, such as are usually found in nurseries.

Chemical

Soil fumigation is achieved by applying toxic pesticides to the soil by various means, and these fumigants move down and across the soil profile to reach the target organisms, either directly or by very efficient secondary distribution due to their relatively high vapor pressure.

Methyl bromide The most powerful soil fumigant available is methyl bromide (MBr), with a very broad spectrum of activity. Many soilborne fungi, nematodes, and weeds are sensitive to MBr; in contrast, some soilborne bacteria such as *Clavibacter michiganensis* sp. *michiganensis*, are not satisfactorily controlled at regular (commercial) rates of application. The duration of the application depends on soil temperature (1–2 days at 15°C, in the 0- to 20-cm soil layer, 3 days at 10–15°C, and more than 4 days at 8–10°C at the same depth). Possible problems due to the toxicological hazard of MBr are related mainly to health hazards for the applicators and to the increase in inorganic bromine residues in edible plant products. In a few cases, MBr has been found in water near greenhouses in the Netherlands, where PVC water pipes were improperly placed only 10 cm deep in the ground. Concerns regarding the possible role of MBr in ozone depletion and the forthcoming phase-out of MBr have triggered research efforts to develop alternative chemical and nonchemical methods for soil disinfestation.

Metham sodium Effective against several soilborne pathogens in both covered and open outdoor cultivation, metham sodium (MS, sodium methyldithiocarbamate) is a fumigant that generates methylisothiocyanate (MITC). In water solutions, MS rapidly changes to MITC. The spectrum of control includes fungi, free-living nematodes, and some weeds. Since most pest resting structures are present in the upper 40 cm of the soil profile, and since MS is 100% water-soluble, it is most effective when applied via the irrigation system (chemigation). However, special care should be taken to avoid contamination of the main irrigation system with MS. The chemical is used at various doses according to the target pathogen and/or the soil type to be treated. Soil temperature is also a critical factor in the effective application of the chemical: a range of 18–30°C at a soil depth of 10 cm is best.

Dazomet (3,5-dimethyltetrahydro-1,3,5,(2H)-thiodiazinothione) A second fumigant that generates MITC is dazomet, a product formulated as either a powder or granules. The chemical is gradually hydrolyzed to at least four subproducts, MITC being the main one. Dazomet is effective against various fungal pathogens at a rate of 400–600 kg a.i. ha^{-1}. This fumigant can be used for the control of several diseases in seed beds and greenhouses, or in field-grown vegetables, cotton, tobacco, and ornamentals. It is applied to the soil by spreading or irrigating followed by mechanical mixing (such as rotovator cultivation or shovel plough) into the soil. The chemical, which is not applicable at temperatures lower than 8°C, is also partially effective against insects, various nematodes, and weed seeds. One of the disadvantages of dazomet is the long period (3 weeks) needed after application of the chemical before planting or sowing is permissible.

Other fumigants and mixtures Fumigants other than those listed, having a narrower range, are registered and used in various cropping systems. These include nematicides such as 1,3-dichloropropene (1,3-D), fungicides such as chloropicrin (CP), and bactericides such as formlin. These are still in use on relatively small scales. Recently, the potential of iodomethane as soil fumigant has also been examined. It is clear that, with the currently available fumigants, there is no satisfactory replacement for MBr. The use of other fumigants involves identification of the causal agent and, in many cases, the use of a mixture of two or more chemicals, to control a wider range of disease agents, pests, and weeds in the treated plots. Methylisothiocyanate 20 + dichloropropane-dichloropropene 80 may serve as an example of this type of product, as this pesticide was formulated to control both pests controlled by MS and the root-knot nematode. Furthermore, data regarding residual effects of these alternative fumigants before planting are needed and their environmental impact is not yet fully clear.

Application Soil fumigants should be applied to well-prepared soil before planting. Most chemicals are injected into the soil in a liquid form to the desired depth with special application machinery. The chemical vaporizes in the soil and diffuses to reach every niche in which pests exist. Application of a fumigant through an irrigation system with delivery via the water to deep soil layers is also common. Such application, however, requires special formulation of the fumigants.

A fumigated field is usually covered with plastic mulch following fumigation to minimize gas escape. Standard polyethylene films are permeable to fumigants and the fumigants dissipate quickly by escaping through the film shortly after application. The permeability of fumigants such as MBr, MITC, CP, and 1,3-D through impermeable film is only

0.001–0.0001 g m^{-2} per hour, depending on the barrier formula, compared with an emission of 5 g m^{-2} per hour for regular, low-density polyethylene. Pest control is determined by the factors of pesticide concentration (C) and exposure time (T). Thus, extending fumigant retention in the soil under impermeable films for a longer period allows the use of reduced fumigant dosages with the same CT values, without reducing control efficacy. Further reduction is possible by burying the film edges more deeply in the soil and by continuous mulching. Machinery for the continuous mulching of plastic film over a large area is in commercial use.

Soil Solarization

The basic idea of soil solarization is to heat the soil by means of solar energy, e.g., by mulching it with transparent polyethylene under the appropriate climatic conditions, thereby killing soilborne pests, improving plant health, and consequently increasing yields. Under the appropriate conditions, the results obtained by soil solarization can be comparable with those obtained by the widely used chemical disinfestation. The first publication on the method of soil solarization was in 1976, although ideas on using solar heating in crop protection have been known for centuries.

The term 'solarization' has several meanings. "To solarize" is defined in *Webster's Third New International Dictionary* as "to expose to sunlight; to affect or alter in some way by the action of the sun's rays." Although many terms have been used to describe this process since 1976 – solar heating of the soil, polyethylene or plastic mulching (tarping), solar pasteurization, and solar disinfestation – here the term 'soil solarization,' is preferred (which was introduced by American plant pathologists), because it is both widely accepted and concise. In addition, 'soil solarization' implies an active process of solar heating of the soil, rather than the usual passive heating of a soil exposed to sunlight.

Principles of soil solarization The principles of soil solarization are summarized as follows:

1. Solarization heats the soil through repeated daily cycles. At increasing soil depths, maximal temperatures decrease, are reached later in the day, and are maintained for longer periods (Figure 1);

2. The best time for soil mulching, i.e., when climatic conditions are most favorable, can be determined experimentally by tarping the soil and measuring the temperatures. Meteorological data from previous years and predictive models further aid in this task;

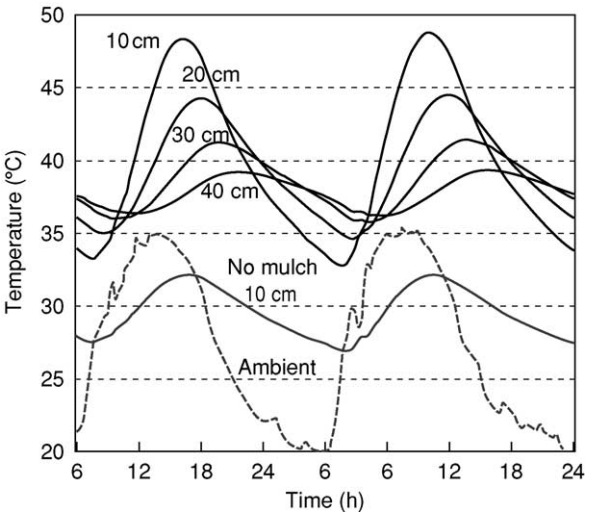


Figure 1 The daily course of soil heating by polyethylene at three soil depths, as compared to nonsolarized (no mulch) soil at a depth of 10 cm. Typical results obtained during July–August in Rehovot, Israel.

3. Adequate soil moisture during solarization is crucial to increase the thermal sensitivity of the target organisms, improve heat conduction in the soil, and enable biological activity during solarization. The soil can be moistened by a single irrigation shortly before tarping. Additional irrigation during solarization via drip system or furrow irrigation is usually not necessary, except for very light soils; in addition to which it may reduce soil temperatures unless carried out during the night;

4. Proper preparation of a soil ready for planting is essential. This is the case because, after plastic removal, the soil should be disturbed as little as possible to avoid recontamination;

5. The soil is mulched with thin, transparent polyethylene sheets or other plastic material. Another method of solarization involves a closed glasshouse (or plastic house), provided climatic conditions are suitable and the soil is kept wet. Novel technologies such as the use of sprayable plastics can replace plastic mulching of the soil;

6. Successful pathogen control in various regions of the world is usually obtained within 20–60 days of solarization. Extending the solarization period enables control in deeper soil layers, as well as of pathogens that are less sensitive to heat;

7. Solarization causes chemical, physical, and biological changes in the soil that affect pest control, plant growth, and yield.

Although both solarization and artificial soil heating involve soil heating, there are important biological and technical differences between these two methods of soil disinfestation. With soil solarization,

there is no need to transfer the heat from its source to the field. It can therefore be carried out directly in the open field or in the greenhouse. Solar heating is carried out at relatively mild temperatures (Figure 1), as compared to artificial heating, which is usually carried out at 70–100°C; thus, the former's effects on living and nonliving soil components are likely to be less drastic. Indeed, negative side-effects observed with soil steaming in certain cases, e.g., phytotoxicity due to the release of manganese or other toxic products and rapid soil reinfestation due to the creation of a 'biological vacuum,' have rarely been reported with solarization. This term refers to a situation where microbial populations are much reduced, resulting in an unbalanced population of soil microflora. Nevertheless, the possibility of the occurrence of such negative side-effects should not be excluded *a priori*. Under appropriate conditions, many soilborne pathogens such as fungi (e.g., *Verticillium, Fusarium, Phytophthora*, *Pythium*, *Pyrenochaeta*), nematodes (e.g., *Pratylenchus* and *Ditylenchus*), and bacteria (e.g., *C. michiganensis*), as well as a variety of weeds, especially annuals, are controlled by soil disinfestation and consequently yields are increased.

As with any soil disinfestation method, soil solarization has advantages and limitations. It is a nonchemical method with less drastic effects on the biotic and abiotic components of the soil; it is simple (and is therefore suitable for both developing and developed countries); and it is frequently less expensive than chemical soil disinfestation. The limitations of this method stem from its dependence on climate and it can therefore be used only in certain climatic regions and during limited periods of the year. In addition, during solarization, the soil remains without a crop for several weeks. Nevertheless, this method has attracted many researchers in more than 60 countries and it is used by farmers, especially in combination with other methods.

Soil heating: simulation models for the prediction of soil temperatures The principles of soil heating and the energy balance of bare and mulched soils have been described in various publications. The main factors involved in soil heating are climatic (e.g., solar radiation, air temperature, air humidity, and wind speed), soil properties, and photometric and physical characteristics of the mulch. The models described below have been developed by Mahrer (The Hebrew University of Jerusalem) and his coinvestigators. The fluxes of energy which have to be considered are:

- R_g, global radiation (waveband of approximately 0.3–4 μm);
- R_L, atmospheric (long-wave) radiation (waveband of approximately 4–80 μm);
- S, conduction of heat in the soil (soil heat flux);
- H, vertical heat exchanges with the air enclosed between mulch and soil by conduction and with the surrounding air by convection (sensible heat flux);
- E, condensation and evaporation of water (latent heat flux).

The two basic equations describing the energy balance of bare and mulched soils, respectively, are:

$$R_{sn} + R_{Ln} - H - E - S = 0 \qquad [1]$$

$$R_{snm} + R_{Lnm} - H_m - E_m - S_m = 0 \qquad [2]$$

where subscript m stands for mulched soil, and R_{sn} and R_{Ln} are the net fluxes of short- and long-wave radiation at the bare soil surface, respectively; S, H, E are soil heat flux, vertical heat exchanges, and condensation and evaporation of water, respectively.

Using these equations, a one-dimensional numerical model to predict the diurnal cycle of soil temperature in mulched and bare soils was developed by Mahrer, with good agreement between observed and predicted soil temperatures. Calculations show that increased soil temperature in wet mulched soil is mainly due to the reduction of heat loss through sensible and latent heat fluxes during the day, and partially due to the greenhouse effect of the wet cover (owing to the formation of small water droplets on its inner surface). A two-dimensional model has been used to study the spatial temperature regime in the soil. Results show that soil heating at the edge of the mulch is lower than at its center. It has also been found that a narrow strip of mulch is less effective in heating the soil than a wide one. Predicted results agree well with observations.

Combining Disinfestation Methods

Soil fumigation with chemicals may have negative effects on the environment, could be extremely dangerous to humans, and may leave toxic residues in plant products. Thus, innovative approaches are urgently needed by farmers and consumers. This can be achieved by combining fumigants with pesticides, at reduced dosages, or with nonchemical methods, e.g., solarization or alternative methods.

Combining solarization with chemicals at reduced dosages or other measures, e.g., biocontrol agents, can reduce the limitations of solarization. The control efficacy may be increased owing to additive effects or to a synergistic effect caused by the hotter environment, which increases vapor pressure and

chemical activity of the added pesticide. Another reason for the improved activity of the pesticide is the weakening of the pathogenic resting structure by the heat. Solarization combined with fumigants could reduce the required duration of solarization, thus making the method more acceptable by the farmers. Furthermore, sublethal fumigation in combination with solarization is especially useful for areas that are marginal for the application of solarization.

Benefits and Limitations

Plant growth in disinfested soil is, as expected, enhanced as compared to that in untreated, infested soil, as a result of pest control. Less expected is a phenomenon of plant growth enhancement in disinfested soils in the absence of known pests, discovered at the end of the nineteenth century, which has since been repeatedly reported with all disinfestation methods, including solarization. It is possible that some cases of increased plant growth response (IGR) might be found to have resulted from the control of hitherto unknown pests, which could not be identified by the procedures available at the time. Nevertheless, different mechanisms, not related to pathogen control, have been suggested to explain IGR in disinfested soils: increased micro- and macroelements in the soil solution; elimination of minor pathogens or parasites; destruction of phytotoxic substances in the soil; and release of growth regulator-like substances, including soluble organic matter and humic substances. Stimulation of mycorrhizae, fluorescent pseudomonads, or other beneficial microorganisms has been frequently observed in solarized soils under greenhouse conditions, and to a lesser extent under field conditions. Soil solarization has been found to result in increased electrical conductivity (EC) in many treated soils, while having little or no effect on others. In saline soils, however, where the level of saline groundwater is relatively close to the soil surface, soil solarization has resulted in a significant decrease in the EC of the surface-soil extract. This is attributed to the prevention of evaporation in mulched (solarized) soil, thereby eliminating salt transport toward the soil surface and its eventual deposition.

Microbial changes take place in the soil during and after solarization. These have been studied specifically in relation to the biological control of pathogens stimulated by solarization, beyond the killing by heat. In drastically disinfested soils, a biological vacuum usually occurs. This can lead to reinfestation by pathogens, which can occur at a faster rate than in nondisinfested soils.

Thus, IGR has very important economic implications that should be taken into account when considering the use of soil disinfestation. The major difficulty is that we cannot predict whether or to what extent a soil will respond with an IGR.

The MBr Crisis and Its Implications

Soil fumigation with MBr has become the major method used for controlling soilborne pests in intensive agriculture worldwide. However, since MBr was listed under the Montreal Protocol in 1992 as an ozone-depleting substance, regulations on its use and consumption have been imposed. In developed countries, the reduction of MBr consumption was started in 1999 and, except for certain exemptions for critical uses, MBr is to be phased out by the year 2005 (or earlier in certain countries). This is being done to protect the vital ozone layer in the stratosphere from depletion. This is a crucial issue, since the ozone layer has already been depleted by a variety of substances such as chlorofluorocarbons. The impending phaseout of MBr poses new and unprecedented challenges for the agricultural research community and the authorities, since many major crops, especially in intensive agriculture, have become totally dependent on MBr use. To avoid an economic and social crisis, alternatives have to be developed relatively quickly.

During a period of more than 100 years of accelerated development in crop-protection sciences, involving the development of hundreds of highly effective pesticides, only three approaches for soil disinfestation have been developed. In practice, only a small number of fumigants have been used in any given period. We need to understand the reasons for this situation in order to avoid additional crises in the future.

There are lessons to be learnt from the MBr crisis. First, we have to avoid dependence on a single soil disinfestation method. Second, we have to combine and alternate methods of control in order to improve them, reduce pesticide dosage, and minimize-side effects. Last, but not least, disinfested soils should be continuously monitored to detect, at the earliest possible stages, negative effects. Soil disinfestation is an expensive, but highly effective, method of control and should be practiced in the best way possible from both plant protection and environmental points of view.

Further Reading

Bell CH, Price N, and Chakrabarti (1996) *The Methyl Bromide Issue*. Chichester, UK: John Wiley.

Gamliel A, Grinstein A, and Katan J (1997) Improved technologies to reduce emissions of methyl bromide

from soil fumigation. In: Grinstein A, Ascher KRS, Mathews G, Katan J, and Gamliel A (eds) Improved application technology for reduction of pesticide dosage and environmental pollution. *Phytoparasitica* 25: 21–30.

Gullino A, Katan J, Garibaldi A, and Matta (eds) (2002) Chemical and non-chemical soil disinfestation. *Acta Horticulturae* 532: 9–256.

Katan J (1981) Solar heating (solarization) of soil for control of soilborne pests. *Annual Review of Phytopathology* 19: 211–236.

Katan J (1984) The role of soil disinfestation in achieving high production in horticultural crops. *British Crop Protection Conference*, pp. 1189–1196. Croydon, UK: British Crop Protection Council.

Katan J (1999) The methyl bromide issue: problems and potential solutions. *Journal of Plant Pathology* 81: 153–159.

Katan J and DeVay JE (1991) *Soil Solarization*. Boca Raton, FL: CRC Press.

Ristaino JB and Thomas W (1997) Agriculture, methyl bromide and the environment. *Plant Disease* 81: 964–978.

Runia WT (2000) Steaming methods for soil and substrates. *Acta Horticulturae* 532: 115–123.

Stapleton JJ (2000) Soil solarization in various agricultural production systems. *Crop Protection* 19: 837–841.

Wilhelm S (1966) Chemical treatments and inoculum potential of soil. *Annual Review of Phytopathology* 4: 53–78.

Dispersion *See* **Flocculation and Dispersion**

DISSOLUTION PROCESSES, KINETICS

K G Scheckel and C A Impellitteri, USEPA, Cincinnati, OH, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Chemistry by its very nature is concerned with change. There are simple but significant interactions between air, water, and minerals that impact our natural environment. Minerals with well-defined structure are converted by various environmental chemical reactions into their elemental building blocks with, perhaps, differing chemical properties relative to the original crystal configuration. The influence of natural phenomena may cause minerals to dissolve in aqueous solutions, thus a solution of atoms is eventually formed, succumbing to the fundamental natural law of element-cycling. Further, these reactions are rarely at equilibrium. They proceed at varying kinetic rates as any dissolved material in solution may be removed (e.g., mineral recrystallization) or added by further mineral dissolution. Recent studies have shown that mineral-like, divalent metal surface precipitates exhibit similar dissolution behaviors to clay and oxide minerals. The examination of mineral dissolution can become even more complex when interactions of surfaces with microorganisms and charged compounds are considered that may induce reductive dissolution of redox-sensitive minerals.

The study of kinetics can be defined broadly as the rate of change of concentrations of reactants in a chemical reaction. The rates are affected by both physical (e.g., diffusion of reacting species) and chemical processes. The kinetics involved in dissolution are often ignored in environmental studies by eliminating time as a variable and assuming a state of equilibrium or 'pseudo' equilibrium. For example, the myriad chemical extractions utilized to assess the potential mobility of metals do so in a set time period. Generally, the rigorousness of a particular extraction acts as a surrogate for time. However, the assumptions employed may be more suitable for some soils than others because of the high variation in dissolution kinetics as a function of the minerals and compounds present. In the short term, (e.g., time of a typical soil chemical extraction) a soil containing relatively small concentrations of weakly sorbed Pb may be assessed as a higher risk than a soil containing a high concentration of galena (PbS). In the long term, as the kinetics of dissolution approach equilibrium, the actual risk may be much higher for the galena soil. Thus, understanding the fundamental rates of dissolution is necessary when pondering the comprehensive health and sustainability of the natural environment.

Rates and Limits of Element-Cycling

Mineral formation and dissolution are two major processes influencing the overall cycle of elements within the natural environment. Dissolution of minerals involves several key steps which are either rate-limiting or not, depending on whether the dissolution is transport (diffusion)- or surface-controlled (Figure 1). A transport-controlled, rate-determining step involves the movement of a reactant or a weathering product through a diffusion layer on the surface of the mineral that often results in a buildup of concentration at the surface interface greater than concentrations found in the bulk solution at distances from the surface; this is often best described by the parabolic kinetic rate law. Likewise, if surface-controlled processes are in command of the dissolution reactions, the concentrations adjacent to the surface build up to values essentially identical to those in the surrounding bulk solution and may approach steady-state conditions conforming to zero-order kinetics. Typically, surface-controlled dissolution mechanisms dictate the kinetics of mineral dissolution. The possible rate-limiting steps for dissolution are: (1) mass transfer of the reactants in the bulk solution to the surface, (2) adsorption of the reactants, (3) interlattice transfer of reacting species, (4) chemical reactions at the surface, (5) movement of the reaction products away from the surface, and (6) mass transfer of products and excess reactants into the bulk solution. Under normal system conditions, steps 1 and 6 (transport controlled) are not usually rate-limiting, while steps 3, 4, and 5 (surface controlled) are often rate-limiting. Figure 1 shows a comparison of concentration in solution as a function of distance from the surface for transport- and surface-controlled dissolution mechanisms.

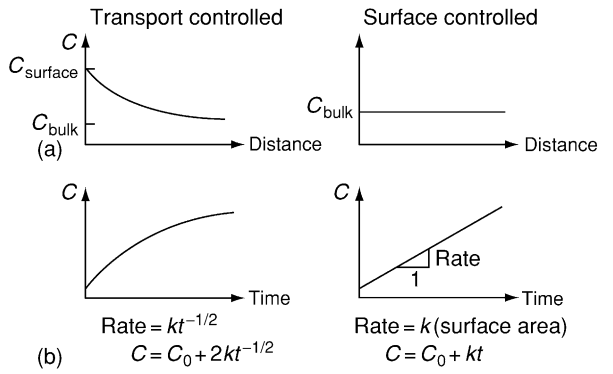


Figure 1 Transport- versus surface-controlled dissolution. Schematic representation of concentration in solution, C, with respect to surface concentration ($C_{surface}$) and bulk concentration (C_{bulk}), as a function of (a) distance from the surface of the dissolving mineral; and (b) of time, t. Kinetic rate of dissolution is a function of the rate coefficient (k) and time for transport controlled or surface area for surface controlled dissolution. (Reproduced with permission from Stumm W (1992) *Chemistry of the Solid–Water Interface*. New York: John Wiley & Sons, Inc.)

Dissolution is known to occur by interactions of surfaces with ligands, protons, water, and metals. The surface protonation of O and OH lattice sites governs the dissolution of silicates. The dissolution rate increases with decreasing pH. For example, the points of attack of the protons in layered phyllosilicates are the O atoms that interlink the Al-oxide groups with the Si-oxide structures. The protonation slowly releases Al from the phyllosilicate surface followed by the subsequent detachment of $Si(OH)_4$ species. Ligands such as organic acids from biological decomposition and/or root exudates are known to promote the dissolution of phyllosilicates. Dissolution of phyllosilicates and other minerals has been widely studied on a macroscopic scale: kaolinite, muscovite, δ-Al_2O_3, and α-FeOOH as influenced by pH, orthosilicates by divalent metal–oxygen bonds, Fe(III) (hydr)oxides via reductive dissolution, albite promoted by temperature and pressure, and biotite with acids to name a few.

Kinetics of Proton- and Ligand-Promoted Mineral Dissolution

All environmental dissolution processes are time-dependent to varying degrees, thus, in order to comprehend the functioning interactions of solid minerals with respect to their fate with time, an understanding of the kinetics of mineral dissolution is important. Key reasons for examining the kinetics of environmental processes include determination of reaction rates, assessment of time needed to attain equilibrium, and determination of possible reaction mechanisms. A basic approach for describing dissolution kinetics is the development of rate laws based on differential equations that establish the premise that the rate of the reaction is proportional to some power of the concentrations of reactants or intermediate species in the system. For example, the proton-promoted dissolution reaction of gibbsite with a known surface area in an acidic solution can be represented by the following:

$$\gamma\text{-Al(OH)}_3 + 3\text{H}^+ \underset{k'_{-\text{H}}}{\overset{k'_{\text{H}}}{}} \text{Al}^{3+} + 3\text{H}_2\text{O} \qquad [1]$$

for which the kinetic rate expression for eqn [1] can be written as:

$$\text{d}[\text{Al}^{3+}]_{\text{aq}}/\text{d}t = -\text{d}[\gamma\text{-Al(OH)}_3]_{\text{ss}}/\text{d}t \qquad [2]$$

meaning that the change in solution concentration of Al^{3+} with time is directly related to the change in concentration of available reactive surface sites (ss; in this case for protons) on the gibbsite surface.

The forward reaction rate law can be written as:

$$\mathrm{d}[\mathrm{Al}^{3+}]_{aq}/\mathrm{d}t = k_H[\gamma\text{-Al(OH)}_3]_{ss}[\mathrm{H}^+]^3 = k_H(\equiv\mathrm{Al(OH)}_2^+)^3 = k_H(C_H^s)^3 \quad [3]$$

where k_H is the proton-promoted forward rate constant, $[Al^{3+}]_{aq}$ is the dissolved Al concentration in solution as a result of gibbsite dissolution by H^+, $[\gamma\text{-Al(OH)}_3]_{ss}$ is the amount of gibbsite in the system, often expressed in terms of reactive surface area, and $[H^+]$ is the proton concentration in solution directly related to pH. This relationship is further defined as $\equiv Al(OH)_2^+$, which conforms to surface protonation leading to polarization of the lattice sites in the proximity of the surface metal (Al) center and is proportional to the power of the surface protonation rate, in this case $3([H^+]^3)$, and equates to C_H^s, the concentration of the surface proton complex (in moles per square meter).

The reverse reaction rate law for eqn [1] is:

$$\mathrm{d}[\mathrm{Al}^{3+}]_{aq}/\mathrm{d}t = -k_{-H}[\mathrm{Al}^{3+}]_{aq} \quad [4]$$

where k_{-H} is the proton-promoted reverse rate constant.

Individually, eqns [3] and [4] are only true far from equilibrium where the influences of back reactions are inconsequential; however, if both reactions occur near equilibrium then eqns [3] and [4] must be combined in order to describe the reaction as the difference between the sums of all forward and reverse reaction rates (eqn [5]):

$$\mathrm{d}[\mathrm{Al}^{3+}]_{aq}/\mathrm{d}t = k_H[\gamma\text{-Al(OH)}_3]_{ss}[\mathrm{H}^+]^{3+} + -k_{-H}[\mathrm{Al}^{3+}]_{aq} \quad [5]$$

Designing experiments that measure only initial rates of reactions can ensure that back reactions are not significant. By employing an initial rate method for eqn [5], one may plot the concentration of Al^{3+} against time over a short reaction period, during which the forward reaction (described by eqn [3]) prevails and the concentration of gibbsite changes very little, resulting in no alteration of the initial rate.

The proton-promoted dissolution rate, R_H, can be expressed as:

$$R_H = k'_H(\equiv\mathrm{MOH}_2^+)^j = k'_H(C_H^s)^j \quad [6]$$

where k'_H is the rate constant for proton-promoted dissolution, $\equiv MOH_2^+$ represents the metal–proton complex, C_H^s is the concentration of the surface proton complex, and j is the oxidation state of the metal ion in the surface structure. Drawing upon the previous example of proton-promoted dissolution of gibbsite, the relationship:

$$R_H = \mathrm{d}[\mathrm{Al}^{3+}]_{aq}/\mathrm{d}t = k_H(\equiv\mathrm{AlOH}_2^+)^3 = k_H(C_H^s)^3 \quad [7]$$

can be linearized to a $y = b + mx$ equation where:

$$\log R_H = \log k_H + 3\log(C_H^s) \quad [8]$$

for which a plot of $\log(C_H^s)$ versus $\log R_H$ should yield a straight line with a slope of 3 (oxidation state of Al^{3+}) and a y-intercept of k_H (as $C_H^s) \rightarrow 1, \log(C_H^s \rightarrow 0)$.

One can express the rate of ligand-promoted dissolution, R_L, as:

$$R_L = k'_L(\equiv\mathrm{ML}) = k'_L C_L^s \quad [9]$$

where k'_L is the rate constant for ligand-promoted dissolution (per time), $\equiv$ML represents the metal–ligand complex, and C_L^s is the surface concentration of the ligand complex (moles per square meter).

The overall rate of dissolution is the sum of ligand-promoted, proton-promoted, deprotonation-promoted, and pH-independent dissolution, which is expressed as:

$$R = R_L + R_H + R_{OH} + R_{H_2O} \quad [10]$$

where R_{OH} is the deprotonation-promoted dissolution rate and R_{H_2O} is defined as the pH-independent dissolution rate.

The overall rate of dissolution can be expressed as:

$$R = k'_L C_L^s + k'_H(C_H^s)^j + k'_{OH}(C_{OH}^s)^i + k'_{H_2O} \quad [11]$$

where k'_{OH} is the rate constant for deprotonation-promoted dissolution, C_{OH}^s is the concentration of the surface deprotonation sites, and i is the oxidation state of the metal ion in the surface structure. The pH-independent dissolution rate is represented by k'_{H_2O}. It should be noted that temperature plays a crucial role in the kinetic rate of reactions; maintaining a constant temperature during kinetic experiments is vital. Furthermore, the employment of environmentally reasonable temperatures helps to ensure that kinetic rate data are relevant to natural environmental conditions.

The rate of ligand-promoted and proton-promoted dissolution of $\delta\text{-Al}_2O_3$ has been investigated. The ligand-promoted dissolution of $\delta\text{-Al}_2O_3$ by the aliphatic ligands oxalate, malonate, and succinate (**Figure 2**) follows a linear relationship between the ligand-promoted dissolution rate, R_L, and the surface concentration of the ligand complexes, C_L^s. While all the ligands examined in **Figure 2** are bidentate (possessing two donor atoms, oxygen, and able to occupy two sites in a coordination sphere) chelating chemicals, the overall size of the molecules plays an important role in the kinetic rate of $\delta\text{-Al}_2O_3$ dissolution. As molecular-ring size increases (oxalate < malonate < succinate), the rate of $\delta\text{-Al}_2O_3$ dissolution decreases accordingly. Likewise, as the number

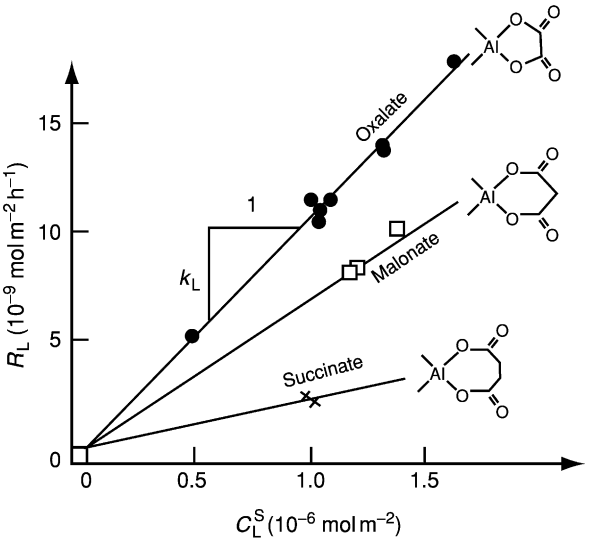


Figure 2 The rate of ligand-catalyzed dissolution of δ-Al_2O_3 by the aliphatic ligands oxalate, malonate, and succinate, R_L (nanomoles per square meter per hour), can be interpreted as a linear dependence on the surface concentration of chelate complexes, C_L^s (moles per square meter). In each case the individual values for C_L^s were determined experimentally. (Reproduced with permission from Furrer G and Stumm W (1986) The coordination chemistry of weathering. I. Dissolution kinetics of δ-Al_2O_3 and BeO. *Geochimica et Cosmochimica Acta* 50: 1847–1860.)

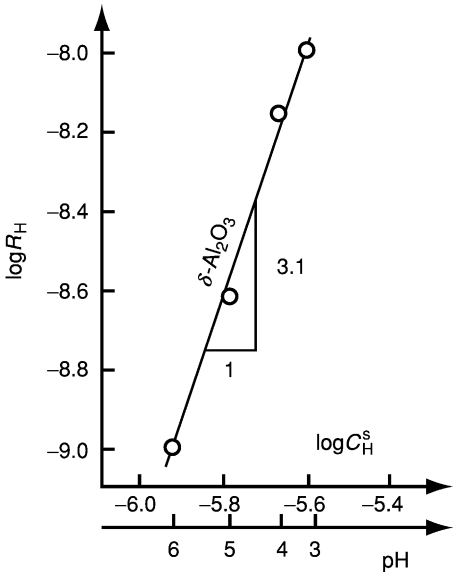


Figure 3 Dependence of the rate of proton-promoted dissolution of δ-Al_2O_3, R_H (nanomoles per square meter per hour), on the surface concentration of the proton complexes, C_H^s (moles per square meter). (Reproduced with permission from Furrer G and Stumm W (1986) The coordination chemistry of weathering. I. Dissolution kinetics of δ-Al_2O_3 and BeO. *Geochimica et Cosmochimica Acta* 50: 1847–1860.)

of donor atoms within a chelating agent increases, the rate of dissolution typically increases (i.e., ethylenediaminetetraacetic acid (EDTA) > oxalate). This phenomenon is known as the 'chelate effect.' Thermochemical studies of complex formation in aqueous solution show that in nearly all cases the chelate effect is due to a more favorable entropy change for complex formation involving multidenate ligands. Proton-promoted dissolution of δ-Al_2O_3 (Figure 3) exhibits a linear dependence with the slope (j) equal to the oxidation state of Al(III) when plotted as logR_H (rate of proton-promoted dissolution) versus logC_H^s, the surface concentration of the proton complexes (eqn [8]).

Figure 4 shows the enhanced and inhibited influence of protons, ligands, cations, and oxyanions on dissolution. Enhancement of dissolution is evident from surface protonation and deprotonation reactions as well as surface complexation with multidentate ligands that are mononuclear in structure. Surface-associated reactions with bi- or multinuclear complexes and some metals result in the inhibition of dissolution due to blockage of surface site reactivity. Hydrophobic compounds can also inhibit dissolution by obstructing surface groups. In studies on the reactivity of Fe(III) (hydr)oxides, dissolution is severely inhibited by phosphate, arsenate, and selenite at near-neutral pH values; however, at pH <5, dissolution is accelerated by the presence of phosphate, arsenate, and selenite.

Metal-Promoted Mineral Dissolution

While some studies have shown that the steric effects of metals can inhibit dissolution, metal-promoted dissolution mechanisms have also been proposed. Sorbing metal cations may induce the dissolution of cations held within the sorbent lattice structure and the released cations then become incorporated in multinuclear surface precipitates with the sorbing metal cations in the bulk solution. Figure 5 shows the kinetics of Ni sorption on pyrophyllite and congruent Si release during Ni sorption as a function of time, as well as Si release from pyrophyllite under identical sorption conditions without Ni present. During Ni sorption in Figure 5, the curve of Ni uptake is mimicked closely by Si release, and that Si release is distinctly different from Si dissolution from pyrophyllite when Ni is not present in solution. It is speculated that the Ni sorption-promoted dissolution of Si and Al from the pyrophyllite structure follows a similar mechanism of proton-promoted dissolution where Ni binding to surface oxide ions causes the lattice bonds to weaken, enhancing the detachment of lattice metal species into solution. Also, X-ray absorption fine-structure (XAFS) spectroscopy has determined

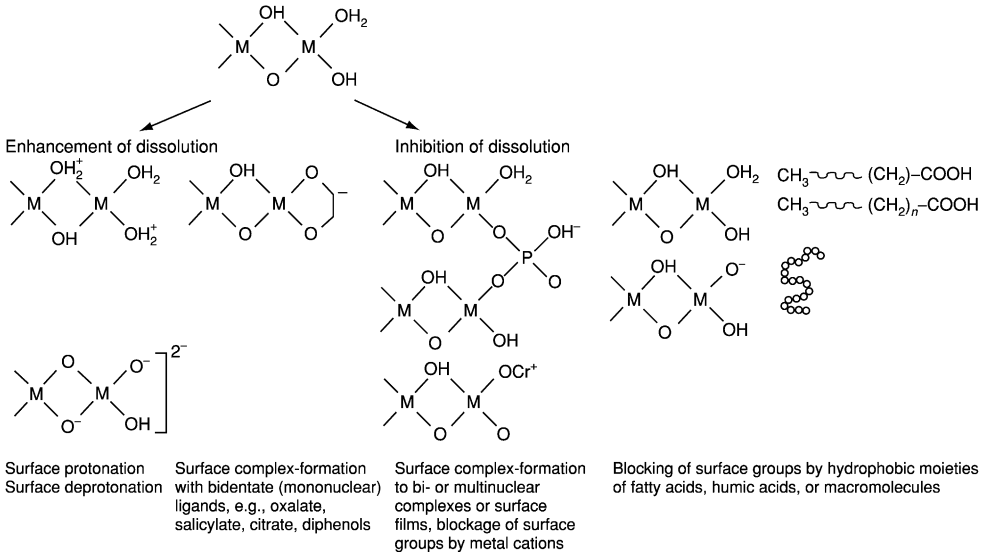


Figure 4 The dependence of surface reactivity and kinetic mechanisms on the coordinative environment of the surface groups. (Reproduced with permission from Stumm W and Wollast R (1990) Coordination chemistry of weathering: kinetics of the surface-controlled dissolution of oxide minerals. *Reviews of Geophysics* 28: 53–69.)

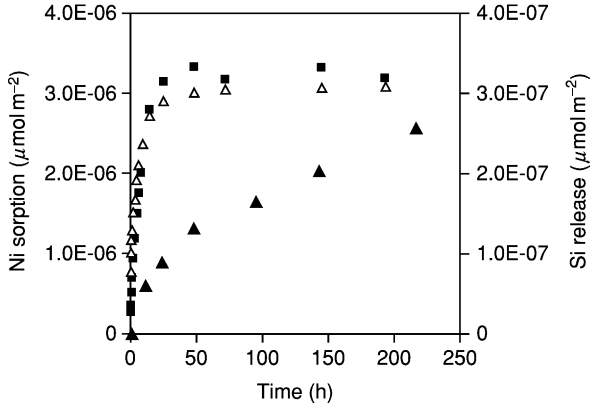


Figure 5 The kinetics of Ni sorption on pyrophyllite from a 3×10^{-3} M Ni solution at pH 7.5. Squares, the amount of sorbed Ni (micromoles per square meter), and empty triangles, the amount of simultaneously dissolved Si (micromoles per square meter). The dissolution of untreated pyrophyllite at pH 7.5 is shown for comparison (filled triangles). (Reproduced with permission from Scheidegger AM, Lamble GM, and Sparks DL (1997) Spectroscopic evidence for the formation of mixed-cation hydroxide phases upon metal sorption on clays and aluminum oxides. *Journal of Colloid Interface Science* 186: 118–128.)

that the released Al and Si are incorporated by the newly formed Ni surface precipitates into the octahedral oxide layer and interlayer, respectively, forming a mixed Ni–Al layered double hydroxide (LDH) precipitate.

Reductive Dissolution of Minerals

Changes in the oxidation state of metals have a dramatic impact on the solubility of metal oxide minerals. The oxide minerals of iron and manganese are more soluble and dissolve more quickly when the surface metal centers are reduced by naturally occurring or anthropogenic reductants such as organic chemicals from microorganisms or industrial pollution. The mode of action for the reductive dissolution of metal hydroxide minerals involves a suite of surface reactions that contribute to the overall dissolution rate, including precursor complex formation with the reductant, electron transfer, release of oxidized organic product, and release of the reduced metal ion (Figure 6). This dissolution mechanism assumes that, prior to the detachment step, the reduction of the metal ion and the protonation of the nearest-neighbor oxide or hydroxide must take place.

Dissolution of Metal Surface Precipitates

In recent years, studies have shown that sorption of metals onto natural materials results in the formation of new, mineral-like precipitate phases. Formation of these precipitate phases can reduce metal concentration in soil and sediment solutions. These three-dimensional structures can co-occur with adsorption processes, and may, in some instances, be the product or extension of sorption reactions. However, the stability of the precipitates and the potential long-term release of the metal back into the soil solution have not been extensively examined. While it is evident that surface precipitates often form on mineral surfaces, investigations on the dissolution of surface precipitates are not common. In light of the knowledge that formation of polynuclear metal complexes on

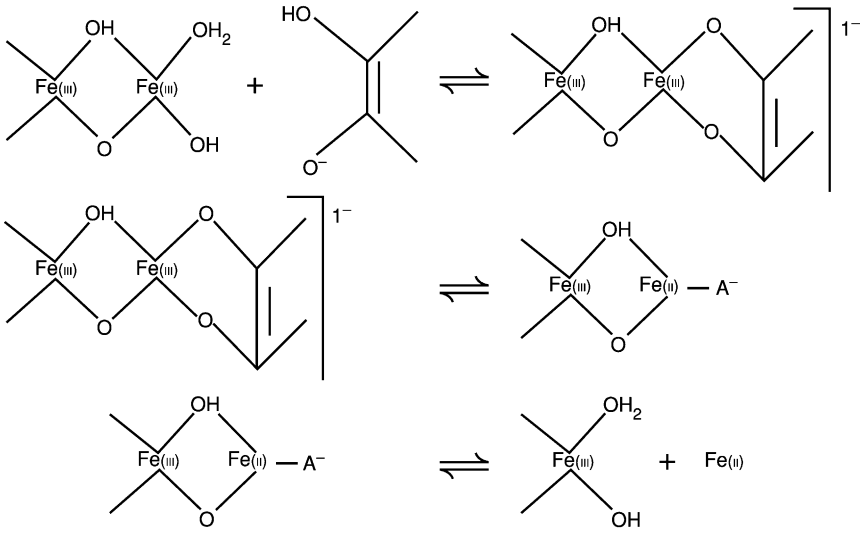


Figure 6 Schematic representation of the dependence of surface reactivity and the kinetic mechanisms on the reductive dissolution surface metal sites.

natural materials is common, there is a need to better understand the degree and mechanism(s) of metal dissolution from surface precipitates. Such information is vital to the better assimilation of the fate of metals in the subsurface environment.

Sorption reactions at the mineral–water interface largely determine the mobility and bioavailability of metals in soils and sediments. Spectroscopic and microscopic studies since the 1990s have consistently shown the importance of metal hydroxide precipitate formation on a variety of clay mineral and oxide surfaces reacted with Ni, Co, and Zn. In spite of the difficulties studying trace amounts of low-crystalline surface precipitates, substantial advances could be achieved in characterizing their chemical composition and structure. Where the sorbent phase releases Al from the lattice structure during reaction with the metal solutions, the precipitates are predominantly Al-containing, layered double hydroxide phases. In the case of Al-free or inert sorbents, however, metal sorption results in α-type metal hydroxide precipitates. Both types of precipitate consist of brucite-like metal hydroxide layers. In contrast to the highly crystalline hydroxide minerals, however, the layers are separated from each other by water and anions, leading to a turbostratic structure, and metal–metal distances within the layer are significantly reduced with respect to the well crystalline metal hydroxides.

The formation of Ni–Al LDH precipitates, as well as the dissolution of these polynuclear Ni(II) surface complexes, from pyrophyllite via proton-promoted dissolution with HNO_3, has been investigated. Nickel detachment from surface complexes is rapid initially at both pH values (with less than 10% of total Ni released) and is attributable to desorption of specifically adsorbed, mononuclear-bound Ni. Dissolution then slows tremendously, primarily owing to the gradual dissolution of the multinuclear surface precipitates. A reference compound, crystalline $Ni(OH)_2$, has also been examined for its dissolution potential. The replenishment method has been employed to simulate steady-state flow and a conventional batch method is also used to compare the influence of reaction products present in solution after dissolution. The replenishment method is more effective in removing Ni from the surface precipitates (approximately 12% at pH 6 and approximately 48% at pH 4) due to removal to reaction products to keep the system from equilibrium. Compared with the dissolution of crystalline $Ni(OH)_2$ (approximately 96% dissolved), Ni release from pyrophyllite is extremely slow.

Studies on the dissolution of Ni–Al LDH surface precipitates on pyrophyllite as a function of residence (aging) time have shown that detachable Ni drastically decreases when the age of the precipitate increases from 1 h to 1 year (Figure 7). By employing high-resolution thermogravimetric analysis, which is sensitive to changes in the interlayer composition of LDH, and by paralleling the results of the surface precipitates with those of reference compounds, a substantial part of the aging effect is shown to be due to replacement of interlayer nitrate by silicate, which transforms the initial Ni–Al LDH into a Ni–Al phyllosilicate precursor (Figure 8). The source of the silicate is the dissolving surface of the pyrophyllite. Studies have also investigated the dissolution of Ni–Al LDH phases on pyrophyllite and gibbsite, and α-$Ni(OH)_2$ precipitates on talc and a mixture of

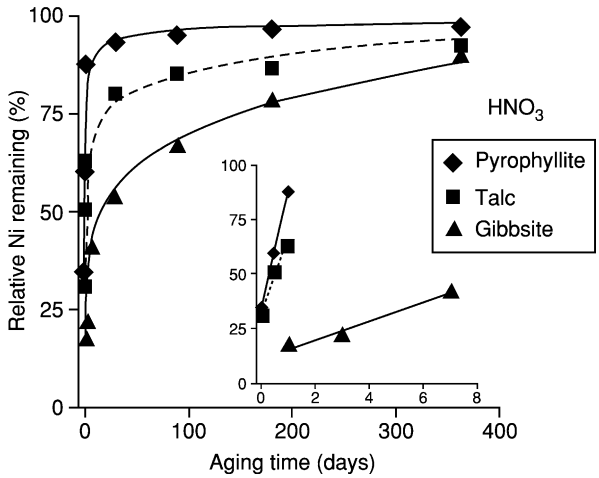


Figure 7 Ni remaining on solids after 10 replenishment steps as a function of aging of the Ni-reacted clay minerals. The inset shows the early aging times. (Reproduced with permission from Scheckel KG, Scheinost AC, Ford RG, and Sparks DL (2000) Stability of Ni hydroxide surface precipitates – a dissolution kinetics study. *Geochimica et Cosmochimica Acta* 64: 2727–2735.)

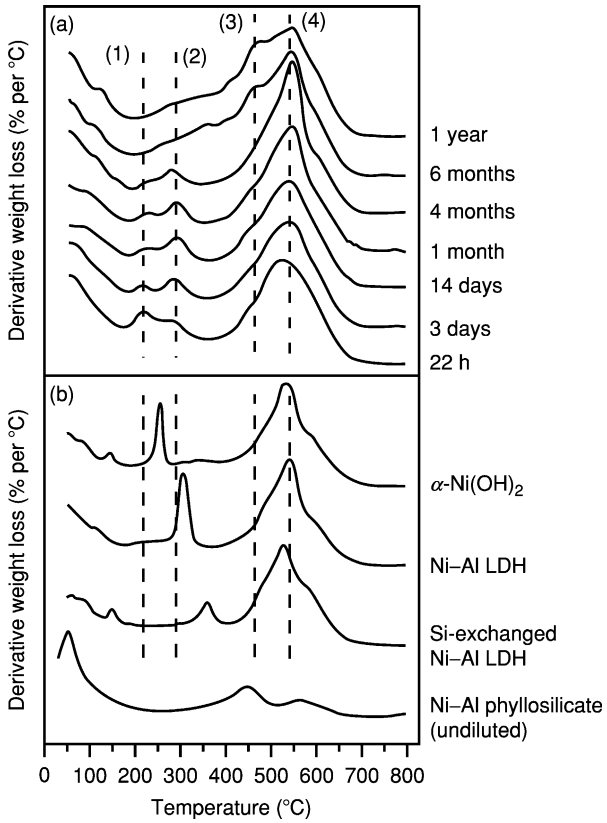


Figure 8 Changes in the thermal stability of the Ni surface precipitate with aging. The derivative of the weight loss curve is shown for (a) aged Ni-pyrophyllite samples and (b) reference precipitates physically diluted with pyrophyllite to match surface loading in sorption samples (2% w/w). Weight-loss events: (1) expulsion of H_2O and nitrate from the layered double hydroxide (LDH) interlayer, (2) dehydroxylation of nitrate-bearing LDH, (3) decomposition of the precursor Ni–Al phyllosilicate, and (4) dehydroxylation of pyrophyllite. (Reproduced with permission from Ford RG, Scheinost AC, Scheckel KG, and Sparks DL (1999) The link between clay mineral weathering and the stabilization of Ni surface precipitates. *Environmental Science and Technology* 33: 3140–3144.)

gibbsite–amorphous silica (gibbsite–silica), employing EDTA (pH 7.5) and nitric acid (pH 4.0) for sorption aging times that range from 1 h to 1 year. In these studies, an array of analytic techniques, including diffuse reflectance spectroscopy (DRS) (Figure 9) and XAFS (Figure 10), have been applied to examine the dissolution of Ni surface precipitates formed on pyrophyllite, talc, gibbsite, and the gibbsite–silica mixture. The differences in stability due to aging of the surface precipitates are speculated to be a combination of Al-for-Ni substitution in the hydroxide layers (for pyrophyllite and gibbsite) and silicate-for-nitrate substitution in the interlayer (for pyrophyllite, talc, and gibbsite–silica mixture).

Macroscopic dissolution studies have demonstrated increased stability in Ni surface precipitates with aging and, to understand this phenomenon better, dissolution of synthetic Ni precipitates of varying composition have been examined (Figure 11). Ni–Al LDH phases with nitrate interlayers are more stable than α-$Ni(OH)_2$ precipitates with nitrate interlayers. Upon changing the interlayer ion from nitrate to silica, there is a dramatic increase in stability for both Ni–Al LDH and α-$Ni(OH)_2$ precipitates, so much so that the Si-containing interlayer α-$Ni(OH)_2$ is more stable than the nitrate interlayer Ni–Al LDH precipitates. Macroscopic and spectroscopic data also show that Al-containing sorbents yield Ni–Al LDH phases and Si-containing sorbents lead to silica-for-nitrate exchanged interlayers as aging time increases. Therefore, the increase in stability with residence time is attributed to three mechanisms: (1) Al-for-Ni substitution in the octahedral sheets of the brucite-like hydroxide layers, (2) Si-for-nitrate exchange in the interlayers of the precipitates, and (3) Ostwald-ripening (size increase) of the precipitate phases. It appears that the second factor, Si-for-nitrate exchange in the interlayers, contributes largely to the increase in stability (Figure 11). The Ni–Al LDH precipitates on pyrophyllite, which possess all three aging factors, result in the most stable complexes. However, Ni–Al LDH on gibbsite, which cannot undergo Si-for-nitrate exchange in the interlayers, are the least stable precipitates, but do increase slightly in stability with aging time due to Ostwald-ripening. The α-$Ni(OH)_2$ precipitates on a gibbsite–silica mixture, amorphous silica, and talc fall between pyrophyllite and gibbsite in regard to stability and are probably a result of the degree of interlayer silication. These results show that increased aging (or residence) times

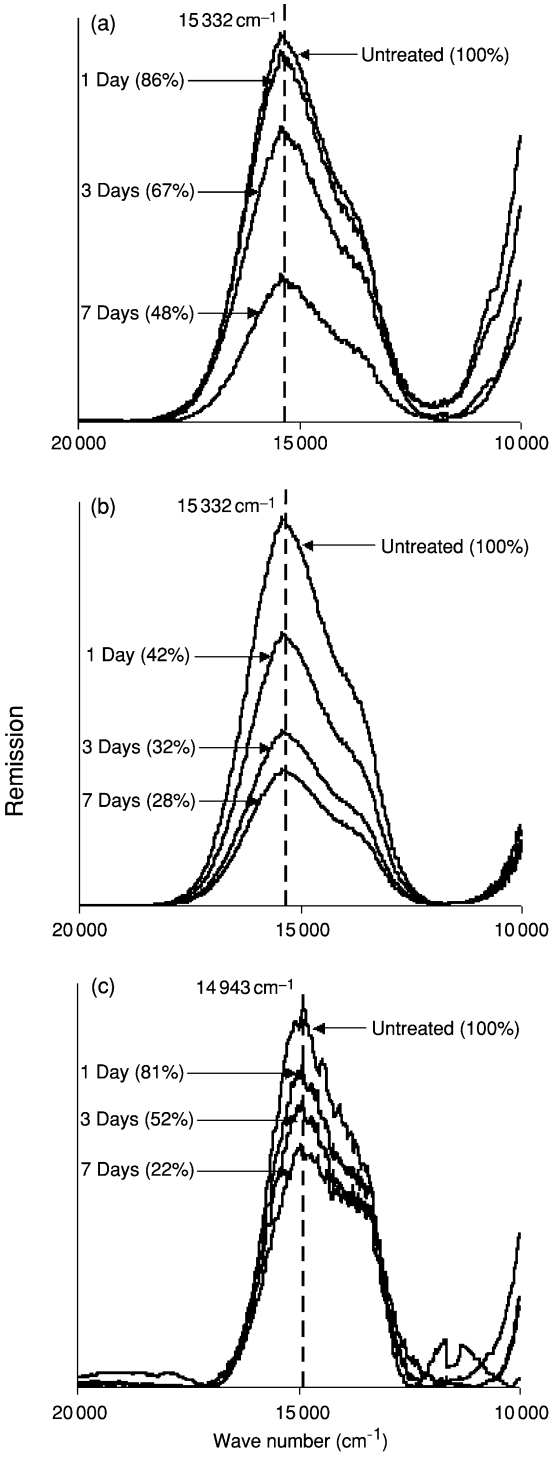


Figure 9 The diffuse reflectance spectroscopy (DRS) $\nu 2$ absorption band for the Ni surface precipitates on (a) pyrophyllite, (b) gibbsite, and (c) talc aged 1 month ('untreated') and subsequently extracted with EDTA (pH 7.5) for 1, 3, and 7 days. The relative amount of Ni remaining on the clay mineral is also given. (Reproduced with permission from Scheckel KG, Scheinost AC, Ford RG, and Sparks DL (2000) Stability of Ni hydroxide surface precipitates – a dissolution kinetics study. *Geochimica et Cosmochimica Acta* 64: 2727–2735.)

under optimal conditions lead to progressively more stable surface precipitates. When environmental conditions change toward those more favorable for dissolution, the rate of dissolution is generally much slower for the precipitates that age for longer periods of time.

The Good and Bad of Dissolution Kinetics

The rate of dissolution of solids in the environment influences a great number of processes. For soil fertility, the rate at which some minerals break down in the presence of plant root exudates, such as oxalate, has a significant impact on plant survival. Some soils are naturally sustainable because of this phenomenon. The mineral components for soil fertility in these soils are slowly extracted over time, requiring little need for anthropogenic inputs of fertilizer; usually nitrogen is the exception. Other soils containing very stable mineral components in climates that are not conducive to dissolution may need regular supplements of required trace elements and nutrients to support crops.

Another issue of recent concern is the dissolution of minerals associated with high amounts of arsenic (As), particularly in parts of Bangladesh and Vietnam. This problem seems to be coupled with the stability of Fe-containing minerals (e.g., pyrite) and Fe oxides. The rate of As recharge into water wells suggests that the dissolution rate is rapid.

Acid mine-drainage directly results from the oxidation of sulfide-containing minerals. While the reaction rate is generally mediated by microorganisms, the protons released in the process can attack other minerals, including nonsulfur-bearing minerals, by proton-promoted dissolution. Thus the production of acid mine-drainage increases pollutants directly by the oxidation of sulfides and indirectly by providing high concentrations of protons that can accelerate the dissolution of minerals.

Recent remediation attempts to immobilize heavy metals in precipitated forms as LDHs or pyromorphite precipitates would not benefit from rapid dissolution kinetics. The objectives of such work are to extensively sequester the metals in solid phases that possess little chance of releasing the metal back into the natural environment where fate, transport, and bioavailability entities may be affected. If the metal can be transformed into a precipitate form with an extremely slow kinetic dissolution rate, then serious issues pertaining to risk assessment and biological availability considerably diminish.

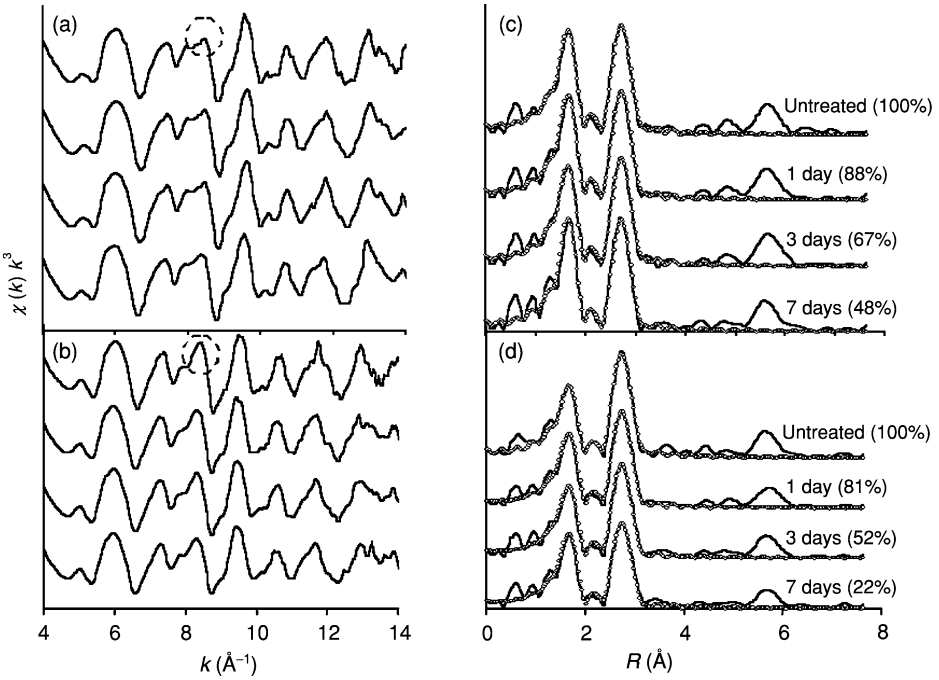


Figure 10 Ni-K_α X-ray absorption fine-structure spectroscopy (XAFS) spectra of pyrophyllite (a and c) and talc (b and d) aged with Ni for 1 month and subsequently treated with EDTA (pH 7.5) for 1, 3, and 7 days. The k^3-weighted χ functions are shown on the left side, and the measured (solid lines) and fitted (dotted lines) radial structure functions are shown on the right side (uncorrected for phase shifts). The circle shows a key identification for Ni–Al LDH versus α-Ni hydroxide. (Reproduced with permission from Scheckel KG, Scheinost AC, Ford RG, and Sparks DL (2000) Stability of Ni hydroxide surface precipitates – a dissolution kinetics study. *Geochimica et Cosmochimica Acta* 64: 2727–2735.)

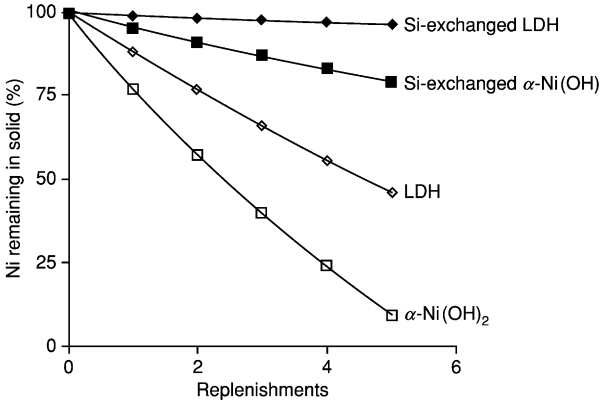


Figure 11 Ni release by HNO_3 at pH 4.0 from homogeneous synthetic Ni-Al layered double hydroxide (LDH) and α-Ni hydroxide, both with either predominantly nitrate or silicate in the interlayer.

List of Technical Nomenclature

$[Al^{3+}]_{aq}$	Dissolved Al concentration in solution
$[\gamma\text{-}Al(OH)_3]$	Amount of gibbsite in the system
$[H^+]$	Proton concentration in solution
$=ML$	Metal–ligand complex
$=MOH_2^+$	Metal–proton complex
Al	Aluminum
Co	Cobalt
C_H^S	Concentration of the surface proton complex (mol m^{-2})
C_L^s	Concentration of the surface ligand complex (mol m^{-2})
C_{OH}^s	Concentration of the surface deprotonation complex (mol m^{-2})
DRS	Diffuse reflectance spectroscopy
EDTA	Ethylenediaminetetraacetic acid
Fe	Iron
HNO_3	Nitric acid
HRTGA	High-resolution thermogravimetric analysis
i	Oxidation state of the metal ion in the surface structure during deprotonation-promoted dissolution
j	Oxidation state of the metal ion in the surface structure during proton-promoted dissolution
k_H	Proton-promoted forward rate constant
k_{-H}	Proton-promoted reverse rate constant

k_{H_2O}	pH independent-promoted forward rate constant
k_{-H_2O}	pH independent-promoted reverse rate constant
k_L	Ligand-promoted forward rate constant
k_{-L}	Ligand-promoted reverse rate constant
k_{OH}	Deprotonation-promoted forward rate constant
k_{-OH}	Deprotonation-promoted reverse rate constant
LDH	Layered double hydroxide
mol m^{-2}	Moles per meter squared
Ni	Nickel
Ni(OH)$_2$	Nickel hydroxide
pH	Negative log of the proton concentration ($-\log[H^+]$)
R_H	Proton-promoted dissolution rate
R_{H_2O}	pH independent-promoted dissolution rate
R_L	Ligand-promoted dissolution rate
R_{OH}	Deprotonation-promoted dissolution rate
Si	Silicon
ss	Surface sites
XAFS	X-ray absorption fine-structure spectroscopy
Zn	Zinc

See also: **Chemical Equilibria**; **Heavy Metals**; **Kinetic Models**; **Metal Oxides**; **Minerals, Primary**; **Precipitation–Dissolution Processes**; **Redox Reactions, Kinetics**; **Sorption:** Metals; Oxyanions; **Sorption-Desorption, Kinetics**

Further Reading

Ford RG, Scheinost AC, Scheckel KG, and Sparks DL (1999) The link between clay mineral weathering and the stabilization of Ni surface precipitates. *Environmental Science and Technology* 33: 3140–3144.

Furrer G and Stumm W (1986) The coordination chemistry of weathering. I. Dissolution kinetics of δ-Al_2O_3 and BeO. *Geochimica et Cosmochimica Acta* 50: 1847–1860.

Laidler KJ (1965) *Chemical Kinetics*, 2nd edn. New York: McGraw-Hill.

Lasaga AC (1998) *Kinetic Theory in the Earth Sciences*. Princeton, NJ: Princeton University Press.

Scheckel KG, Scheinost AC, Ford RG, and Sparks DL (2000) Stability of Ni hydroxide surface precipitates – a dissolution kinetics study. *Geochimica et Cosmochimica Acta* 64: 2727–2735.

Scheidegger AM, Lamble GM, and Sparks DL (1997) Spectroscopic evidence for the formation of mixed-cation hydroxide phases upon metal sorption on clays and aluminum oxides. *Journal of Colloid Interface Science* 186: 118–128.

Sparks DL (1989) *Kinetics of Soil Chemical Processes*. San Diego, CA: Academic Press.

Sparks DL and Suarez DL (eds) (1991) *Rates of Soil Chemical Processes*. SSSA Special Publication 27. Madison, WI: Soil Science Society of America, Inc.

Sposito G (1994) *Chemical Equilibria and Kinetics in Soils*. New York: John Wiley & Sons, Inc.

Stumm W (ed.) (1990) *Aquatic Chemical Kinetics*. New York: John Wiley & Sons, Inc.

Stumm W (1992) *Chemistry of the Solid-Water Interface*. New York: John Wiley & Sons, Inc.

Stumm W and Morgan JJ (1996) *Aquatic Chemistry*. New York: John Wiley & Sons, Inc.

Stumm W and Wollast R (1990) Coordination chemistry of weathering: kinetics of the surface-controlled dissolution of oxide minerals. *Reviews of Geophysics* 28: 53–69.

DRAINAGE, SURFACE AND SUBSURFACE

N R Fausey, USDA Agricultural Research Service, Columbus, OH, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Soil drainage is a natural process by which water moves across, through, and out of the soil as a result of the force of gravity. Drainage is a component of the global hydrologic cycle; and streams and rivers are the naturally developed drainage conduits through which some of the water arrives at the land surface as precipitation is transported across the landscape and eventually to the oceans. This natural process also provides the water that supports seeps, springs, stream baseflow, and aquifer recharge. As water leaves the soil, air moves into the space previously occupied by the water; this process is called aeration.

Adequate soil aeration is vital for maintaining healthy plant roots and the many beneficial organisms that live in the soil and require oxygen for respiration. As the proportion of water and air in the soil changes as a result of drainage, the ability of the soil to provide support and traction for animals and vehicles (trafficability) is altered as the strength of the soil changes with water content.

The natural drainage of the soil may limit human use of the resource. Poor drainage has social and economic impacts. The drainage of the soil can be accelerated by the use of surface and subsurface drainage practices. Surface drainage diverts excess water from the soil surface directly to streams, thereby reducing the amount of water that will move into and possibly through the soil. Subsurface drainage, provided by ditches and drainpipes, collects and diverts water from within the soil directly to streams.

Hydrology and Drainage

Precipitation, as well as irrigation, delivers water to the land surface. The type, rate, and amount of precipitation and the properties of the soil determine the relative amounts of the water that will infiltrate, runoff, or remain on the surface temporarily as surface storage. The water that runs off does so by the natural process of surface drainage, sometimes called runoff or overland flow. The water that remains temporarily stored on the surface will eventually either evaporate or infiltrate. Water that infiltrates into the soil increases the soil water content and may ultimately be evaporated back to the atmosphere directly, be taken up by plants, or move deeper into the soil.

The forces of adhesion, cohesion, and gravity control redistribution of the water that infiltrates into the soil. Water infiltrates into the pores between soil particles and adheres to these particles. Additional soil water accumulates as water coheres to water already in contact with the soil particles. Within small pores, enough water can be held against the force of gravity to fill the pores completely. Pores that hold water against the force of gravity are called capillary-size pores. In larger pores, all of the water that enters cannot be held against the force of gravity and some of the water moves downward through these pores. This process is called percolation; percolation is a natural drainage process. Water that percolates deeper into the soil may: enter capillary-size pores at deeper depths and increase the soil-water content; encounter restrictive features that block its descent and cause accumulation of water in the form of a perched water table; reach and add to an existing water table; and reach and replenish an aquifer.

Accelerated Drainage

The history of human adoption of accelerated drainage is not well documented. However, there is evidence from archeological investigations that indicates early adoption and use of manmade drainage facilities. Some examples are raised bed and furrow systems used by the Mayan culture and hillside terraces used by ancient Asian cultures. Although the latter example causes decelerated drainage, it is clearly an effort to alter the natural drainage process for the benefit of humans. There is written record from the Roman era indicating early use of subsurface drainage practices, including instructions for drain installation for agricultural areas.

Surface Drainage Principles and Practices

Surface drainage systems typically consist of an outlet channel (existing natural stream or constructed channel emptying to a natural stream), lateral ditches, and field ditches. Such systems are used primarily in flat areas having poor natural drainage to remove water that collects on the land surface when the rainfall rate exceeds the infiltration capacity. Additional improvement for surface drainage may include land smoothing or land grading to fill in shallow depressions and to assure a continuous slope in the field toward the field ditches. A primary goal in the design and construction of surface drainage systems is to remove the water from the surface as quickly as possible while avoiding soil erosion that can occur when the water moves too rapidly. To avoid soil erosion during surface runoff (drainage), designers look for ways to 'walk the water' off the surface.

Because steep-sided ditches impede the movement of the modern, large machines used in farming, many of these ditches have been replaced with grassed waterways to convey the concentrated runoff flows. This type of channel can be easily crossed by machines and can follow the natural contour of the land surface, thereby minimizing the amount of land taken out of production and the expense involved in construction of the surface runoff collectors. Drop structures are then often needed to dissipate the energy in the runoff water as it drops from the shallow grassed waterways into deeper, open-ditch field drains.

In flat, poorly drained landscapes, surface drains are typically spaced approximately 100 m apart, as shown schematically in Figure 1. The field drains are perpendicular to the prevailing land slope and empty into lateral drains that carry the runoff to larger collectors and main ditches or outlet ditches. Smoothing or grading is used between the field drains to minimize surface storage and ponding. Typical

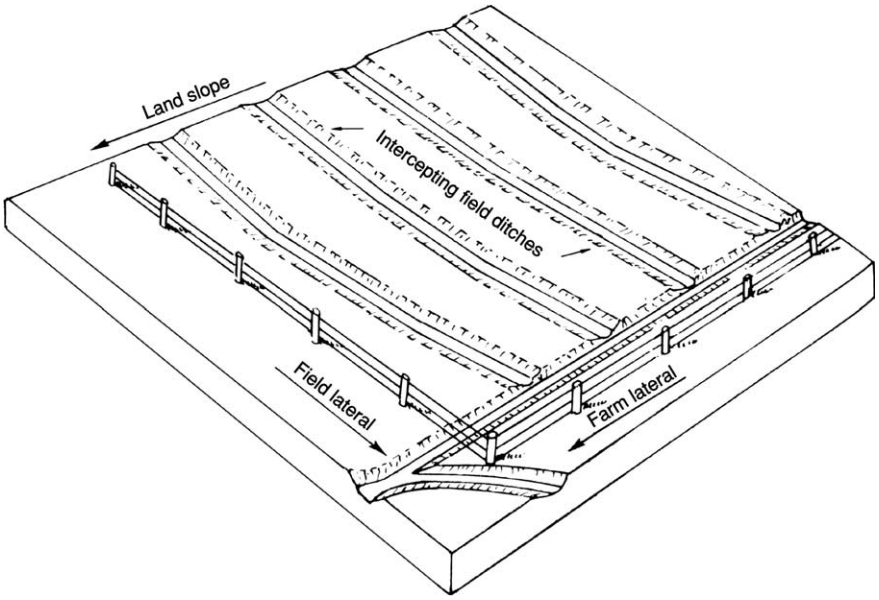


Figure 1 Schematic view of a field surface drainage system with field drains and lateral drains. Reprinted from Pavelis GA (ed.) (1987) *Farm Drainage in the United States: History, Status, and Prospects*. Miscellaneous Publication No. 1455. Economic Research Service. US Department of Agriculture. Washington, DC: US Government Printing Office.

Figure 2 Landplane used for smoothing the soil between surface drains. Reprinted from Pavelis GA (ed.) (1987) *Farm Drainage in the United States: History, Status, and Prospects*. Miscellaneous Publication No. 1455. Economic Research Service. US Department of Agriculture. Washington, DC: US Government Printing Office.

equipment used for this type of work is shown in Figure 2.

Minimizing surface storage is an important aspect of surface drainage. Filling and/or providing outlets for depressions located above and between the grassed waterways and lateral ditches can accomplish this. Land smoothing is typically accomplished by using a land plane to drag small amounts of soil from around the edges of depressions into the depression to fill it in. Land grading involves greater amounts of soil disturbance to construct connections between deeper depressions and to fill up some of the depression storage. This is done with earthmoving equipment.

Subsurface Drainage Principles and Practices

Subsurface drainage describes the process of removal of that water which has infiltrated into the soil in excess of the amount that can be held by capillary forces against the force of gravity. Soils that require accelerated subsurface drainage typically have some impermeable or slowly permeable feature below the surface that prevents water that has entered the soil from moving deeper into the soil and underlying

materials at a rate that allows agricultural production to be economically viable. Other criteria may involve the stability of roads and building sites. The obstruction to rapid percolation of water through the soil may be shallow bedrock, highly dense glacial till, depositional clay layers, and other similar causes. In other words, there is no natural outlet for the water, and the soil becomes saturated by the accumulated infiltration of water.

A primary goal in the design and construction of subsurface drainage systems is to remove noncapillary water from the upper layers of the soil profile as quickly as possible to ensure an adequately aerated root zone and trafficability for critical field operations such as planting and harvesting. An illustration of how subsurface drains lower the water table in the soil is given in **Figure 3**. The depth and spacing of subsurface drains are dependent upon many factors, including especially the availability of an outlet, the soil texture, and the crops to be grown.

The basic equation describing movement of water in the soil was derived by Darcy in 1856. His equation describing the flow rate of water through a soil column is:

$$Q = KA(H_1 - H_2)/L$$

where Q is the flow rate, A is the cross-sectional area of the column, L is the column length in the direction of flow, H_1 and H_2 are the hydraulic heads at the ends of the column, and K is the hydraulic conductivity of the soil. Following this approach, others contributed to the development of an equation to calculate the design spacing between subsurface drains. Subsurface drain spacing varies by soil type and ranges from approximately 10 m for heavy clay soils to as much as 50 m for highly permeable soils.

The path of water flow through the soil is tortuous, around and between soil particles and aggregates, especially in the horizontal direction, where there are no residual channels from roots or insects or animals. Thus, although at the bank of a stream water may drain out of the soil and into the stream due to the force of gravity, this will not continue for very much distance horizontally from the streambank. The stream only provides an outlet for water from a narrow strip of soil near the streambank. There is no outlet for the water in the soil further from the stream. A ditch dug away from the stream into the soil can provide an outlet for the water in a narrow strip of soil along the ditch. The ditch lies at an elevation below the surface of the soil and provides an outlet that allows drainage of the soil above the bottom of the ditch and a short distance away from the side of the ditch. Thus, the ditch is also a subsurface drain. Of course a conduit that admits water through its walls can be placed into the ditch, the ditch can be filled in above the conduit, and the conduit will continue to function as a subsurface drain. And so, by connecting ditches and buried conduits, subsurface drainage systems can be created to remove noncapillary water from broad areas of the soil.

Early subsurface drains were probably always open ditches, but eventually various types of conduits were placed in the bottoms of the ditches and they were covered over. Tree branches tied in bundles, stones, fired clay products in various shapes, extruded clay and concrete pipes, and eventually smooth- and corrugated-wall plastic pipes were used to create subsurface drainage conduits. Drainage ditches were at first dug by hand, but this process has been mechanized greatly over the years. Large machines, like those shown in **Figure 4**, with automated control of the depth and slope of the ditch bottom or conduit, are now in routine use around the world for installing subsurface drainage systems.

Recent Innovations in Drainage Practice

Traditionally, drainage systems have been installed and operated with open unmanaged outlets that allow free and unrestricted discharge of any water that reaches the drainage conduit or channel. Recent concerns about the delivery of nonpoint source pollutants to streams have encouraged the practice of closing drainage outlets during times of the year when trafficability

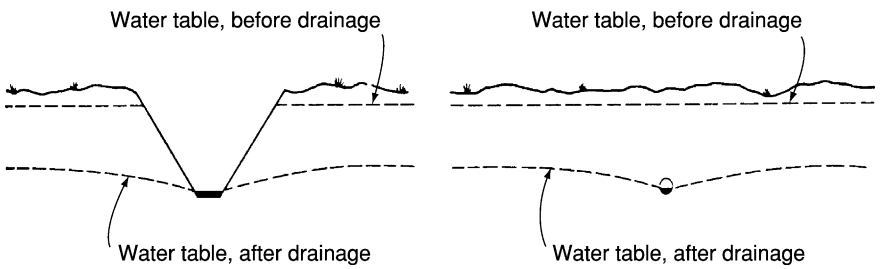


Figure 3 Water table drawdown by ditch and pipe subsurface drains. Reprinted from Pavelis GA (ed.) (1987) *Farm Drainage in the United States: History, Status, and Prospects*. Miscellaneous Publication No. 1455. Economic Research Service. US Department of Agriculture. Washington, DC: US Government Printing Office.

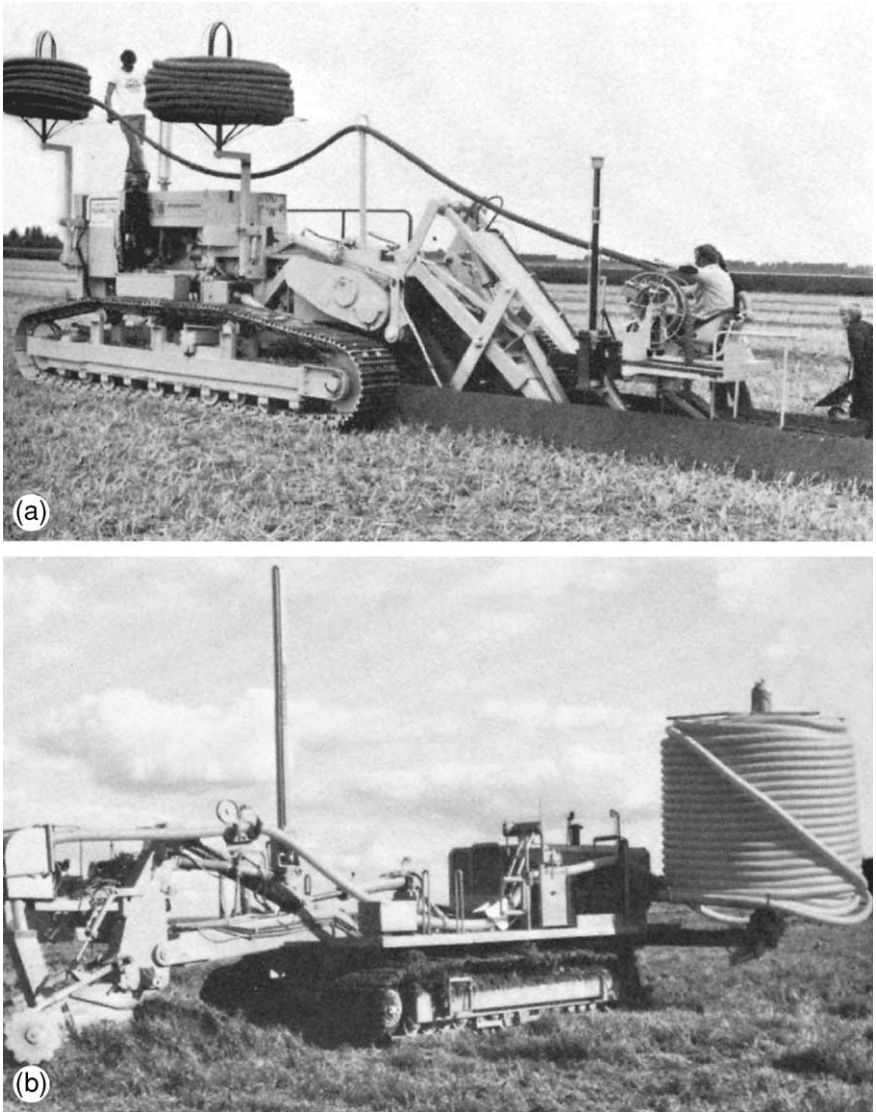


Figure 4 (a) Trench- and (b) plow-type machines for installing subsurface drainage. Reprinted from Pavelis GA (ed.) (1987) *Farm Drainage in the United States: History, Status, and Prospects*. Miscellaneous Publication No. 1455. Economic Research Service. US Department of Agriculture. Washington, DC: US Government Printing Office.

and aeration of crop roots are not required. The practice is known as controlled drainage and involves using a control mechanism at the outlet of surface and subsurface drainage systems to close and open the drains as needed. In some cases when control structures are in place, subirrigation water can be added back into the soil during the growing season through the same conduits that are used to drain the soil. Such systems have been shown to reduce the delivery of pollutants to streams and to increase crop yields. However, a much greater level of management input is required to operate controlled drainage systems.

See also: **Childs, Ernest Carr**; **Hooghoudt, Symen Barend**; **Irrigation:** Environmental Effects; Methods

Further Reading

Pavelis GA (ed.) (1987) *Farm Drainage in the United States: History, Status, and Prospects*. Miscellaneous Publication No. 1455. Economic Research Service. US Department of Agriculture. Washington, DC: US Government Printing Office.

Skaggs RW and van Schilfgaarde J (eds) (1999) *Agricultural Drainage*. Monograph No. 38. Madison, WI: American Society of Agronomy.

van Schilfgaarde J (1974) Non-steady flow to drains. In: *Drainage for Agriculture*, pp. 245–270. Monograph No. 17. Madison, WI: American Society of Agronomy.

Weaver MM (1964) *History of Tile Drainage (In America Prior to 1900)*. Waterloo, NY: M.M. Weaver.

DRYLAND FARMING

G A Peterson, Colorado State University, Fort Collins, CO, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Dryland farming can be described as crop production in semiarid regions of the world with no irrigation. Historically these regions have provided much of the world's grain supply. The world population today still depends heavily on dryland regions for much of their wheat, barley, millet, sorghum, and pulse supply. Judicious management is critical to the sustainability of dryland farming.

Definition

Dryland farming is frequently defined as crop production in areas with less than 500 mm of annual precipitation, but this definition omits a critical component of the equation, evaporation potential. Operatively, dryland farming is practiced where annual potential water evaporation exceeds annual precipitation. The example for the Central Great Plains of the USA in Figure 1 illustrates this. Note that the deficit between precipitation and potential evaporation is large and is at its peak in the middle of the summer crop growing season. As the water deficit increases (i.e., the difference between annual precipitation and potential evaporation becomes more negative), the difficulty of producing crops increases proportionally. Worldwide dryland farming areas are characterized by deficits between precipitation and potential evaporation, but differ in the size of the deficit and the time of the year it occurs. For example, in Morocco in northwestern Africa the deficits are so large in the summer that no dryland crop production can occur. Overall dryland farming productivity is inversely related to the size of the deficit between annual precipitation and annual potential evaporation. A large deficit indicates more plant stress and lower yields.

World Scope of Dryland Farming

Dryland farming is an ancient practice with a fascinating historical record. Today, one can find dryland farming practices that are basically unchanged since ancient times. Practices used since 500 BC stand in sharp contrast to dryland farming practices involving the latest discoveries of chemistry and plant genetics. In many instances, the productivity of the soils under the older management techniques has decreased, especially in situations where population pressures on food supply have multiplied.

All continents, with the exception of Antarctica, have substantial amounts of land where dryland farming is practiced. The climates and soils where dryland farming is practiced vary widely from continent to continent and even within continents. A primary factor for dryland farmers is when they receive their annual precipitation and how that fits with the growing season of the particular crops they produce. There are many combinations of precipitation pattern and crop production. Figure 2 illustrates two very different precipitation environments, both of which are used for dryland production. The US Central Great Plains has a summer rainfall pattern, while

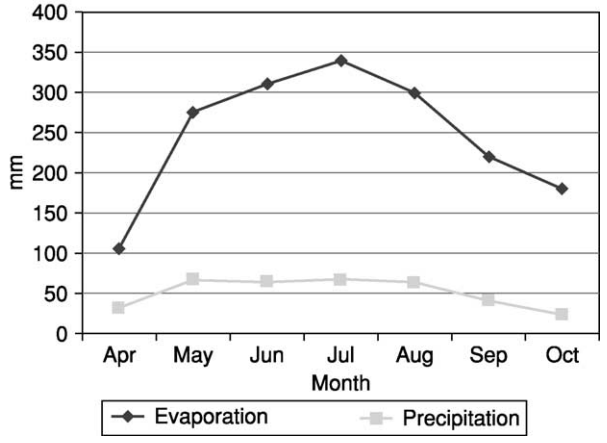


Figure 1 Evaporation and evaporation potential in the US Central Great Plains.

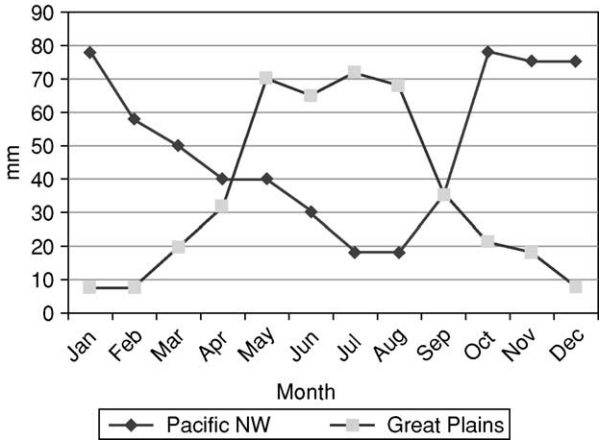


Figure 2 Rainfall patterns in the US Pacific Northwest and the Great Plains.

the US Pacific Northwest has a winter precipitation pattern. Winter precipitation patterns are found in both cold and warm climates, greatly altering management. For example, in the Pacific Northwest the winter precipitation is primarily snow, whereas the winter precipitation pattern in Northwest Africa or Southern Spain is all rainfall with no snow.

In areas with a winter precipitation pattern, all precipitation is received in the winter, and the water is conserved in the soil for the crop that follows. In summer precipitation pattern areas, most of the precipitation is received during the crop growing season, but is accompanied by high temperatures that encourage evaporation. These examples illustrate the complexity of the management issues where dryland farming is practiced.

A wide diversity in soils exists in dryland farming areas but, in general, soils are neutral to basic in pH, low in organic matter, and low in both nitrogen and phosphorus supply capacity. They are relatively unweathered because of the climatic conditions under which they have been formed, and thus have adequate amounts of basic cations such as calcium, magnesium, and potassium. Physically, these soils have weak structure, which is easily destroyed by tillage or natural events such as raindrop impact. The low organic matter contents contribute to the weak structure.

Major Dryland Farming Issues

Despite the wide climate and soil variations where dryland farming is practiced, there are similarities that create identifiable management issues. The two major commonalities that exist for dryland farming areas are: (1) annual crop production is always limited by water supply; and (2) soil erosion hazards are high, both from water and wind.

Lack of water results in relatively limited yields and general lack of vegetative cover, which in turn creates soil conditions that are highly vulnerable to erosion. Thus, dryland farming has often been associated with 'dust bowls' and ecosystem degradation. Susceptibility to erosion by water also remains high because many dryland farming areas receive short bursts of high-intensity rainfall, which can cause immediate soil erosion.

The primary means of improving the stability of dryland farming is to maintain cover on the soil surface for as much of the year as possible. Maintaining cover on the land, either as vegetative canopy or crop residue from previous production, improves water conservation both between crop seasons and during the crop growing season. Improved water conservation in turn improves plant productivity and increases the opportunity to keep the soil from eroding, which is a positive feedback to the agroecosystem.

Dryland Farming Management Techniques

Fallowing

Fallowing has been the primary yield stabilizing tool used by farmers in the majority of the dryland areas worldwide. Simply stated, this means leaving the land idle as frequently as every other year to store water and to allow nitrate-nitrogen to accumulate in the soil. The stored water plus the nitrate-nitrogen accumulation increase the probability of producing an adequate crop in the year after fallow. Fallows are generally managed as weed-free as possible if water conservation is the primary objective, but in some areas of the world weeds are allowed to grow during the fallow year and are grazed by livestock. In the latter situation, little soil water is stored because the weeds transpire the water. However, some organic residues are returned to the soil, and the farmer has a source of forage for animals. The practice of 'weedy fallow' is most commonly found in subsistence farming areas.

Unfortunately, fallowing in any form is highly inefficient in terms of water storage. Depending on atmospheric evaporative demand and cover management, the proportion of precipitation stored in the soil during fallow ranges from near 0 to 50%. Most often, however, less than 20% of the precipitation is stored in the soil even if weeds are controlled during the fallow period. Despite this inefficient precipitation management, fallowing has remained popular because it ensures some measure of crop yield, which can be critical to farmer survival. Introduction of no-till management of fallow has improved water storage potential, but in most cases precipitation storage is still less than 50% of the precipitation, meaning that more than 50% of the precipitation is lost to evaporation, weed growth, or water runoff. No-till is a farming system that maximizes water conservation and minimizes soil erosion. Herbicidal weed control is used, no tillage events of any kind are allowed, and thus residues from the previous crop remain as protection on the soil surface.

Sustainability of dryland farming in the long term requires improved precipitation-use efficiency. Present research efforts focus on three aspects of precipitation management: (1) precipitation capture in the soil, whether from rainfall or from snow melt; (2) water retention in the soil after capture, which involves weed control and minimization of evaporation; and (3) efficient water use by the crop plants,

which involves agronomic practices associated with soil fertility, variety choice, etc.

Water Capture Management

Capturing the incident precipitation, whether it is rainfall or snow melt, is the first water conservation step in dryland farming. It is critical to system sustainability that capture be maximized within the economic constraints of the particular situation. Water capture is governed by soil properties like soil texture, aggregation, pore size, and pore-size distribution. Large amounts of surface soil macroporosity are needed to maximize water infiltration. This soil characteristic is highly governed by soil texture; fine-textured soils have less macropore space than coarse-textured soils, and subsequently lower infiltration rates than coarse-textured soils. Degree of aggregation within a given soil texture also governs water capture. Well-aggregated soils have larger amounts of macropore space, and thus improved water capture relative to poorly aggregated soils of the same texture. Degree of soil aggregation is possible to manage, while differences due to texture are not feasibly altered.

Structural stability of the aggregates also affects water capture. Soils with weak soil structure quickly lose their ability to absorb water as the surface aggregates disintegrate upon wetting and the surface pore spaces become smaller (e.g., soil crusting). Soils in dryland regions tend to have low organic matter content, which contributes to weak aggregation.

Soil water content at the beginning of a precipitation event also modifies the effects of other properties. As the water content of a soil increases, the water infiltration rate decreases because the surface soil pores are full of water and there is little space for water entry. The maximum water-intake rate is obtained at the beginning of a rainfall event and then decreases rapidly as water fills the surface pore space. If the soil is dry, there is a large storage capacity, relative to the same soil in a moist condition; hence water will flow rapidly into the soil storage reservoir, thus accounting for a higher infiltration rate in dry soil.

Maintaining cover on the soil surface is a key element for maximization of water capture in dryland systems. Cover, whether plant canopy during the growing season or crop residue after harvest, protects the soil aggregates from raindrop forces and thus soil pores remain open for water capture. Cover also slows the running water after a storm and increases opportunity time for water absorption by the soil. Historically, dryland farmers could not maintain cover on the soil between cropping seasons because tillage for either weed control or seedbed preparation incorporated the remaining crop residues into the soil. Cultivation for weed control during the cropping season further depleted any residue cover. New scenarios for cover management have evolved with the advent of no-till management. Herbicidal weed control permits residue retention between seasons and during the growing season, and modern no-till planting equipment incorporates only small amounts of residue ($<5\%$), leaving most of the residue on the soil surface. Thus the soil has maximum protection year-round.

Tillage has a secondary negative effect on water capture because it physically grinds aggregates and reduces their overall size distribution, which slows water absorption. It also increases soil organic matter oxidation rates because of aggregate destruction and subsequent increased exposure of organic compounds to soil organisms, which further weakens aggregates. Aggregate size distributions shift, such that microporosity increases at the expense of macroporosity, which results in decreased capability to absorb water. The degree to which tillage affects water capture is governed by complex interactions of tillage type and time, with an array of soil characteristics such as texture, original structure, and organic matter content. Overall, long-term tillage of any soil decreases aggregate resistance to physical disruptions such as raindrop impact and tillage operations of all sorts.

Water Retention Management

Weed control during time periods between crop growing seasons and weed control within the growing season are critical to retaining water for the cropping system. Tillage and herbicidal weed control are two extremes of a spectrum. Both can be effective in controlling weeds and preventing water use by weeds. They differ greatly, however, in terms of their effect on water evaporation from the soil surface. Herbicidal weed control minimizes water evaporation, while tillage maximizes evaporative losses: the reason being that all tillage events expose moist soil to the atmosphere, while herbicides kill weeds with no soil disturbance. Thus, water retention is maximized if the soil is not stirred in most environments.

Residue cover also reduces water evaporation rates because it reflects and absorbs heat energy, preventing it from warming the soil beneath. Note that the residue cover slows evaporation rate, but does not stop evaporation. Eventually the water in the surface soil beneath a residue cover will dry out if no rainfall occurs, but at a much slower rate than if no cover is present. Thus in climates where rain events occur frequently, the residue is more beneficial for water retention than in geographic regions where weeks and months pass with no rainfall. For the climates

illustrated in Figure 2, the summer precipitation regions benefit from residue cover in terms of evaporation control, whereas the winter precipitation areas with dry, hot, summers benefit less from residue cover. Residue cover is still valuable in winter precipitation areas for other reasons such as improved water capture and erosion control.

Efficient Water Use by Crop Plants

Once the water is captured and retained in the soil, the farmer needs to be sure that the best agronomic practices are used so that healthy plants are available to use the water. These practices include crop variety choice, proper planting dates, and adequate fertilization. Weak, stressed plants cannot develop root systems that will maximize crop yield response to the available soil water and precipitation that is received during the growing season.

Conclusion

Managing cover, the primary dryland-farming management technique, affects water capture, water retention, and even efficiency of water use. Cover absorbs raindrop impact energy, catches snow in winter seasons, slows runoff water, decreases water evaporation, and improves nutrient management for the plants. Furthermore, it provides protection against erosion by wind and water. None of these roles is independent of the others. Dryland farming will always play a key role in feeding the world's population, and as that population grows, the need for improved management techniques will also increase.

Further Reading

Brengle KG (1982) *Principles and Practices of Dryland Farming*. Boulder, CO: Colorado Associated University Press.

Dregne HE and Willis WO (1983) *Dryland Agriculture*. Agronomy No. 23. Madison, WI: American Society of Agronomy.

Farahani HJ, Peterson GA, and Westfall DG (1998) Dryland cropping intensification: a fundamental solution to efficient use of precipitation. *Advances in Agronomy* 64: 197–223.

Skujiņš J (1991) *Semiarid Lands and Deserts: Soil Resource and Reclamation*. New York, NY: Marcel Dekker.

E

Earthworms *See* **Fauna**

EDAPHOLOGY

A L Ulery, New Mexico State University, Las Cruces, NM, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Edaphology is the science or study of soil, especially with respect to plant growth. The root of the word is 'édaphos,' Greek for 'foundation,' 'soil,' 'ground,' or 'land.' 'Soil science' is the term more commonly used today and includes the study of soil as a natural body in the landscape (pedology), as well as a medium to be managed for optimum crop and rangeland productivity, environmental waste disposal, and construction. Pedology focuses on the description and classification of soil and how it is formed. While both pedology and edaphology refer to the study of soil, the latter deals specifically with soil as a medium for plant growth. (*See* **Pedology:** Basic Principles.)

Edaphologists study the soil by examining a variety of factors or soil properties that govern plant (flora) and animal (fauna) life in the soil. Edaphic factors have been described as those soil conditions that affect the organisms living in a particular area. These factors are distinguished from climatic or physiographic factors that can also affect plant growth largely through the influence of water and temperature by concentrating on the root system and soil. The study of edaphic factors is an important part of ecology.

Ecosystems are thrown out of balance when major disturbances occur such as the overgrazing of rangeland or the conversion of native forests or grassland to agriculture. Such ecosystems are no longer in equilibrium and cannot be adequately described by traditional theories of plant succession or climax models. An edaphic climax is a localized vegetative community that differs from the surrounding vegetation because of different soil types and properties, including water-holding capacity, drainage class, soil depth, and soil fertility. New state and transition models for the study of nonequilibrium ecology and vegetation dynamics recognize that the interactions between the soil resource and the associated vegetative community determine the functional status of ecologic processes. This theory also states that the condition of the soil is directly connected to aboveground vegetation.

The organisms that live in a soil influence soil formation and are in turn affected by the edaphic properties that govern water, air, and nutrient supply. Thus, a continual feedback relationship is established between the soil flora and fauna, and the physical and chemical characteristics of the soil. Edaphology is the study of that relationship and how it affects plant growth. One approach to edaphology is agronomic and involves cultivating the soil and adding water and nutrients for maximum plant yield. Irrigation and soil-fertility management are examples of this approach. Ecological studies that correlate plants to specific soil types or properties in addition to landscape position and climate information are another example of edaphology. Evaluating or predicting the effect of management on soil–plant interactions and edaphic factors is often the objective of studies on ecosystem health, soil quality, and land degradation.

Edaphic Properties

Edaphic factors include those physical, chemical, and biological properties of a soil that influence plant growth. The various properties that are important for plant growth are interrelated but can be distilled to a few simple requirements. Essentially, what plants need from the soil are water, air, nutrients, and physical support. Sunlight is required for photosynthesis

but is provided aboveground and may be affected by aspect, latitude, and other physiographic and climatic factors. Some of the soil or edaphic properties that affect plant requirements are listed in Tables 1–3.

Edaphic factors often resemble soil-quality indicators. 'Soil quality' is a somewhat contentious term in the soil science community, but it refers to some very useful concepts concerning the ability of a soil to sustain biological production and diversity, regulate and partition water, filter and buffer contaminants, store and cycle nutrients, and provide structural support. Soils of high quality are fertile and productive, resist erosion, and help maintain good air and water quality. They are also resilient to natural and anthropogenic stresses and are therefore resistant to degradation. Soil-quality indicators are those physical, chemical, and biological properties that measure a soil attribute or function and are sensitive to changes in environment and management (Tables 1–3).

Table 1 Some physical edaphic properties and soil-quality indicators that affect plant growth by influencing water and air movement, or rooting depth in soil

Soil property	*Water and air movement*	*Root support or impedance*
Texture	X	X
Structure	X	X
Soil or solum depth	X	X
Depth to impermeable layer	X	X
Infiltration	X	
Hydraulic conductivity	X	
Porosity and macropores	X	X
Soil-water potential	X	
Water-holding capacity	X	
Drainage class	X	X
Bulk density	X	X

Table 2 Some chemical edaphic properties and soil-quality indicators that affect plant growth by influencing water and air movement, nutrients, or rooting depth in soil

Chemical property	*Water and air movement*	*Nutrient availability*	*Root support or impedance*
pH		X	
Cation exchange capacity		X	
Organic matter content	X	X	X
Nitrate nitrogen		X	
Organic nitrogen		X	
Total or soluble phosphorus		X	
Total or soluble potassium		X	
Micronutrients		X	
Clay mineralogy	X	X	X
Salinity or electrical conductivity	X	X	
Toxic ions or heavy metals		X	

Table 3 Some biological edaphic properties or soil-quality indicators that affect plant growth by influencing water and air movement, nutrients, or rooting depth in soil

Biological property	*Water and air movement*	*Nutrient availability*	*Root support or impedance*
Earthworm number	X	X	X
Respiration rate (measure of microbial population and activity)		X	
Microbial diversity		X	
Composition of organic matter	X	X	X
Plant root depth	X	X	X

Edaphic Factors Affecting Water and Air

In general, soil physical properties have the greatest effect on water and air movement into and through the soil. The role of water is so important in plant growth that some definitions of edaphology refer to it as the science of physics applied to soil and water. Water and air are considered together because they use the same pathways to enter and move through the soil. The water-holding capacity of a soil refers to the amount of water contained in a soil and is determined largely by texture, structure, porosity, soil depth, and organic matter content.

'Texture' refers to the particle-size distribution of the soil material. Particles range from sand size (2 mm) to clay size, which is less than 2 μm (1000 times smaller). Finer-textured or clayey soils hold more total water than sandy soils, but all of that water is not available to the plants, because it is held so tightly to the clay particles. Sandy soils have large, interconnected pores that fill and drain easily. While the texture refers specifically to the distribution of particles less than 2 mm in diameter, another valuable piece of information is the percentage of gravel or amount of particles greater than 2 mm. This factor affects soil packing, porosity, bulk density, nutrient status, and root growth.

Porosity results from the arrangement of soil particles into aggregates with inter- and intraaggregate pores, as well as from root growth and faunal activity. The pores within and around the aggregates conduct water and air through the soil. Macropores can form after roots decay or as a result of burrowing soil fauna. One of the most dominant groups of soil animals are earthworms, which may number from approximately 50 to 300 worms m^{-2} in agricultural

soils and more than 500 worms m^{-2} in forest and grassland soils. Earthworms shred and consume organic matter and excrete wastes as casts, which help create stable soil aggregates and enhance soil porosity. Their burrows are persistent over time and provide a major conduit for drainage and preferential water movement through the soil, but also minimize surface water erosion by increasing infiltration. The presence and abundance of earthworms often indicate healthy, organic-rich soils.

Bulk density is the dry soil weight per volume of soil and is a good measure of porosity and compaction. Pore space contributes to the volume of a soil and thus, as porosity increases, the dry weight decreases, resulting in a lower bulk density. Compacted soils have higher bulk densities and tend to restrict water and air flow as well as root growth. In addition, texture and organic matter both affect bulk density. Sandy soils have a higher bulk density than clayey soils because of lower porosity. Organic matter weighs very little, takes up a large volume, and so creates lower bulk density. Some minerals have a high particle density and can contribute to a higher soil bulk density if they are dominant in the soil. Soils with a lower bulk density contain more water than soils with a high bulk density, but they tend to drain too quickly and become droughty.

Organic matter contributes to water-holding capacity in several ways. It aggregates the soil and produces a crumb-like structure at the surface that facilitates infiltration, percolation, storage, and diffusion of water and air. Sticky substances such as waxes and tannins left behind when soil organisms decompose organic residues help bind soil particles together into porous clusters that hold and release water, nutrients, and humus. Soil organic matter or humus holds as many as 20 times its weight in water and has many beneficial physical, chemical, and biological properties. The O and A horizons in soil profiles have the highest amounts of organic matter and are often measured in field studies to evaluate their influence on nutrient supply or relationship to vegetation. Organic carbon or total organic matter is also commonly measured in edaphic studies.

The depth of the soil and the water-holding capacity also directly determine the amount of water in a soil. Many trees and desert shrubs have root systems that exploit soil water and nutrients to depths below 2 m, but most agronomic plant roots are concentrated in the upper meter of the soil. The bottom of a soil profile may be an impermeable layer or hardpan, bedrock, or a shallow groundwater table. A change in texture – for example, a sandy layer under a clay loam, or vice versa – can restrict water flow through the soil. Shallow, eroded soils hold less water than deep, uniformly textured soils. Gleying and mottles are evidence of standing water and poor aeration or anaerobic conditions.

Mineralogy, specifically the kind and amount of certain clay minerals, may also affect water and air movement through the soil. Some soil minerals such as montmorillonite, a member of the smectite family, absorb water and swell to fill up cracks and macropores, resulting in restricted hydraulic conductivity. This is particularly problematic in sodic soils or alkali-affected areas. Sodium is a cation with low charge density that tends to neutralize and disperse negatively charged clay and the aggregates formed from clay and organic matter. When aggregates disperse they form an impermeable seal at the soil surface and along cracks that limits the movement of water and air into the soil. The sodium adsorption ratio or exchangeable sodium percentage are two measures of the sodium content affecting plant growth, especially in arid regions where the rainfall is insufficient to leach salts out of the soil profile. Sodic soils are also a problem in closed basins with shallow groundwater that is often saline and detrimental to plant growth.

Edaphic Factors Affecting Plant Nutrients

For healthy growth, plants require many inorganic nutrients in varying amounts and forms. Macronutrients have concentrations of at least 500 mg kg^{-1} in plants, while micronutrients are required in lower amounts, usually less than 100 mg kg^{-1}. Nitrogen (N), phosphorus (P), and potassium (K) are the most commonly measured and applied macronutrients by farmers and gardeners. Inorganic fertilizers are labeled with their N–P–K amounts and should be applied to the soil only when soil test results indicate that the nutrients are below optimum levels. Specific edaphic factors that may be measured to monitor soil fertility include nitrate-N, available P, and soluble K (Table 2). Other commonly measured nutrients include calcium (Ca), magnesium (Mg), sulfate (SO_4^{2-}), and iron (Fe). Sometimes it is not just the total amount of each nutrient or compound that affects plant growth and distribution on a landscape, but the relative proportions of nutrients as well. For example, the calcium-to-magnesium ratio may be used to explain the occurrence of some endemic plants growing on serpentine soils in California, USA.

The soil reaction or pH is related to nutrient availability and optimum plant growth. This is one of the most common soil chemical properties measured. Some plants prefer acidic soils, while others are best suited for basic or alkaline soils. There are few plants that will survive in extremely acid or alkaline soils, however, as most flora and fauna prefer conditions between pH 5.5 and 8.5.

Soil pH also affects the microbial population of the soil, which is necessary to convert soil organic matter to inorganic nutrient forms. Most microbes prefer warm, moist, near-neutral conditions to mineralize organic matter and wastes. Soil microbes are responsible for decomposing plant residues and wastes and converting them to their final inorganic products of carbon dioxide, water, and nutrients. Where inorganic fertilizers are unavailable or too expensive, or when slow release of nutrients is desired, organic materials such as manure, compost, and plant residues are the best nutrient source for plants. In addition to carbon, these materials contain nitrogen, phosphorus, sulfur, and other nutrients. Soils rich in organic matter are darker and tend to warm up faster in the spring. The combined physical, chemical, and biological benefits of organic matter or humus make it one of the most important edaphic factors affecting plant growth.

Soil organic matter and clay minerals are usually negatively charged because of their composition. This diffuse negative charge allows them to attract and retain positively charged cations from the soil solution. The ability of a soil to adsorb or hold cations is called the cation exchange capacity (CEC). This is an important feature of soil materials because so many plant nutrients are cations (e.g., K^+, Ca^{2+}, Mg^{2+}, ammonium (NH_4^+) and many micronutrients). Cations are held loosely and temporarily by negatively charged exchange sites in and around soil particles so that they are not leached away with every rainfall or irrigation event. However, their attraction is not permanent and the cations are easily available to plants and microbes as they are needed. Soils that have abundant humus and smectite or vermiculite clays are higher in CEC than sandy soils low in organic matter. Organic matter is also valuable in complexing or chelating some metals to make them more soluble and plant-available. In addition, soil organic matter has been shown to negate the toxicity of aluminum (Al) in some soils.

In addition to aluminum, several other elements are toxic to plants and may require characterization to determine whether they are affecting plant growth and distribution on the landscape (Table 4). Nickel (Ni), manganese (Mn), lead (Pb), copper (Cu), chromium (Cr), and cobalt (Co) are among the metals that have been analyzed and correlated with vegetation growth around mines or smelters. Some elements are plant micronutrients at low concentrations, but become toxic at high concentrations in the soil or solution. Examples of these elements are boron (B), copper, and zinc (Zn).

Soil mineralogy is also an important edaphic factor, influencing the availability of some important nutrients such as phosphate and sulfate. Allophane or clay materials derived from volcanic glass can fix or permanently adsorb phosphate and sulfate anions. Iron and aluminum oxides are prevalent in leached, acidic soils of humid zones and can also combine with phosphate to make insoluble phosphate minerals that are unavailable to plants. In arid regions, where calcareous soils are common, the calcium combines with and binds the phosphate. Phosphorus is generally most plant-available at near-neutral pH and in soils not dominated by allophane, iron or aluminum oxides, or calcium carbonate. Available or extractable phosphorus is measured to distinguish it from organic or mineral forms that are not accessible to plants.

Potassium and ammonium also become fixed or unavailable in soils high in vermiculite. The size of these ions and their charge density allow them to become permanently sandwiched in the interlayer region of high-charge vermiculites that collapse around them in response to opposite electrical charges. Thus, soluble potassium and/or clay mineralogy is measured to determine the amount of potassium available for plant uptake.

Table 4 Toxic elements that may affect plant growth and soil quality

Trace element	*Maximum concentration allowed in irrigation water*[a] *(mg^{-1})*	*Essential to plants at low concentrations?*[b]
Aluminum (Al)	5.0	Possibly
Arsenic (As)	0.10	Possibly
Beryllium (Be)	0.10	No
Bismuth (Bi)	NA	No
Boron (B)	NA	Yes
Cadmium (Cd)	0.01	No
Chromium (Cr)	0.10	Possibly
Cobalt (Co)	0.05	Possibly
Copper (Cu)	0.20	Yes
Fluorine (F)	1.0	Possibly
Iron (Fe)	5.0	Yes
Lead (Pb)	5.0	Possibly
Lithium (Li)	2.5	Possibly
Manganese (Mn)	0.20	Yes
Molybdenum (Mo)	0.01	Yes
Nickel (Ni)	0.20	Yes
Selenium (Se)	0.02	Possibly
Silver (Ag)	NA	No
Tin (Sn)	NA	No
Vanadium (V)	0.10	Possibly
Zinc (Zn)	2.0	Yes

NA, no maximum concentration has been designated.
[a]Criteria established by the Food and Agricultural Organization (FAO) of the United Nations.
[b]Some elements are required by plants at low concentrations but are toxic at higher levels (therefore 'Yes'), and some elements are still under investigation or only partially accepted as being essential, thus the term 'Possibly.'

Salinity or electrical conductivity (EC) is often measured in arid regions or areas under irrigation. EC increases as the salt content increases. Salinity can be a problem in closed basins with shallow water, tables. Salts can affect plant growth in two ways, either directly as a toxic specific ion or indirectly by lowering the osmotic potential of the soil water, which hinders the plants' ability to uptake water. Increased salts lower the osmotic potential, which in turn lowers the total soil-water potential, making plants work harder (expend more energy) to extract water from the soil. Specific ion toxicity can be a problem when boron, chloride (Cl^-), selenium (Se), or sodium (Na) concentrations are excessive. Toxicities of certain ions are plant-dependent; however, high sodium concentrations may also disperse the soil particles and create physical impediments to water and air flow. Inorganic fertilizers are also salts, designed to dissolve easily in water. Excess fertilizer may contribute to soil EC and saline stress.

Edaphic Factors Affecting Root Growth

Many of the factors already discussed also affect plant root growth and development. Compaction and high bulk density limit the root system if the roots cannot force their way through the soil. However, good soil–root contact is important for nutrient and water uptake, so moderate compaction may actually enhance plant growth. Repeated cultivation by heavy machinery can cause a plow pan or hard layer to develop that can impede root growth and water movement. This layer is usually found at the bottom of the Ap horizon, and the thickness of the Ap horizon is thus a useful edaphic factor to measure (Figure 1).

Other horizons of interest in a soil profile include the B horizon, a zone of accumulation of clay (Bt), organic matter (Bh), calcium carbonate (Bk), iron oxides (Bs), or salts (Bz) (Figure 1). These horizons may affect rooting depth and plant growth either physically by impeding root growth and water flow or chemically by fixing nutrients or supplying toxic concentrations of salts and other compounds. Bt horizons can also enhance the water-holding capacity of a soil. The solum depth (the combined thickness of the A and B horizons) is sometimes analogous to soil depth, when the C horizon is considered to be unweathered or little affected by pedogenic processes (Figure 2).

Shallow soils often result from erosion, which is a major problem in many areas of the world. Good plant cover is one of the best ways to minimize soil loss, whether from wind or water erosion, and thus is an important component in soil conservation and management. Physiographic factors such as slope percentage, aspect, and location on the landscape are often recorded along with edaphic factors to evaluate plant–soil–landscape relations. The depth of the soil profile and the thicknesses of various horizons are pedological as well as edaphic factors; this is an example of where the distinction between edaphology and pedology becomes blurred.

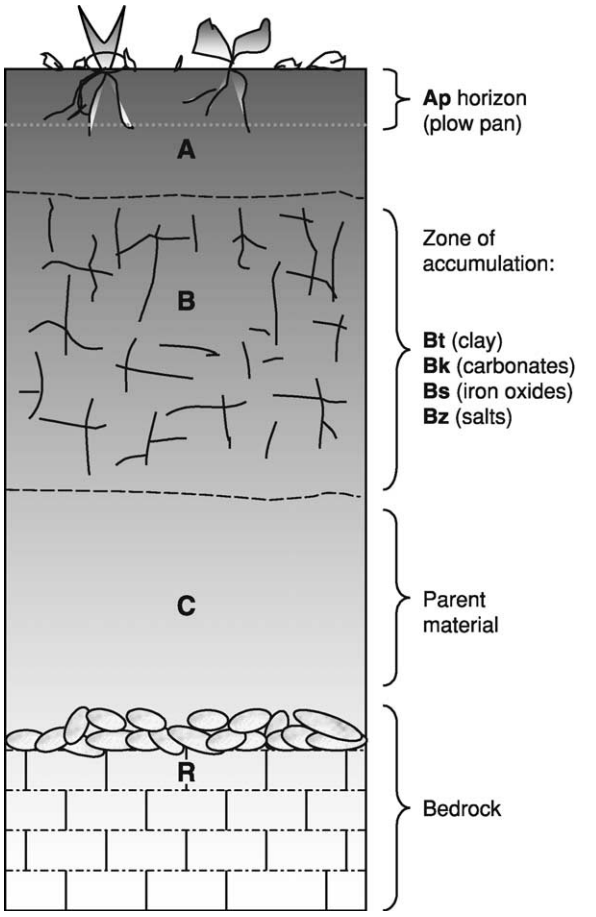


Figure 1 A soil profile, a vertical section of the soil through all its horizons (layers), extending into the C horizon (parent material). Every soil is individual and has its own unique characteristics and properties. Some of the master horizons and just a few of the many possible subhorizons that may form in a soil as it develops from the parent material in response to climate, vegetation, biota, topography, and time are shown. These include the Ap and various B horizons that may inhibit root growth and water movement through the profile if they are dense or cemented. Alternatively, the accumulation of clays and organic matter in the B horizon can enhance water-holding capacity and nutrient availability, and thus plant growth.

Temperature is an important edaphic factor, influencing plant growth and root development. Soil temperature is affected by several physical factors, including porosity, water content, soil color, organic matter content, bulk density, and soil depth. It is also a function of landscape placement and physiographic features such as aspect, elevation, and latitude. The temperature may be measured directly in the soil, or

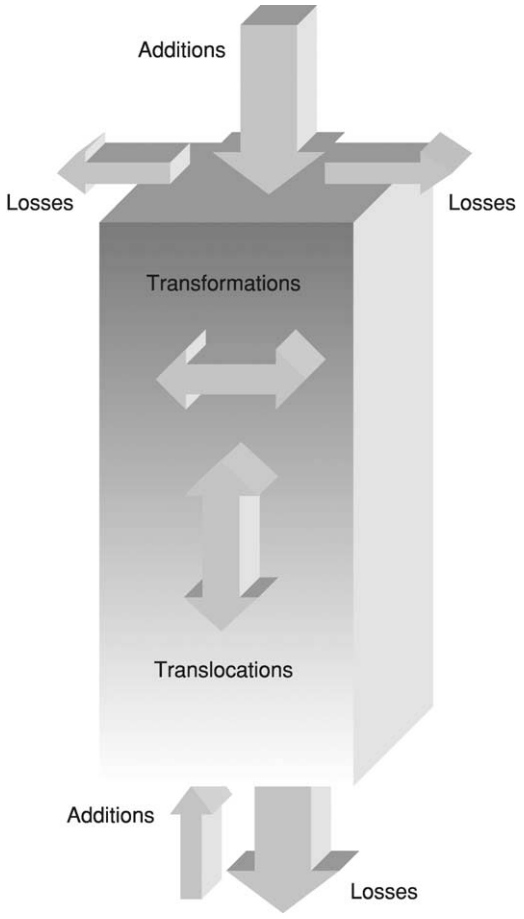


Figure 2 Pedogenic processes include the addition and removal of matter into and out of the profile as well as the chemical alteration (transformation) and movement of matter within the profile (translocation).

indirectly as a function of climatic factors such as air temperature and solar radiation, as well as the edaphic features mentioned above.

Research

In addition to the use of the word 'edaphic' instead of 'soil' to describe any condition or property pertaining to the soil, there are specific studies conducted by ecologists, plant scientists, agronomists, and soil scientists (or edaphologists) that are designed to measure, characterize, monitor, and assess the health of the soil as it serves as a medium for plant growth. The literature on edaphology research is largely divided into ecologic and agronomic studies. In the science of ecology, the spatial variation or distribution of vegetation as a function of soil type or edaphic property is important to characterize an ecosystem. The information gleaned from these studies provides a baseline of native vegetation in the landscape and helps identify changes due to anthropogenic and climatic effects. The long-term goal of this kind of study is to manage or protect native or desired vegetation by understanding the relationships among plant communities, the underlying soil, disturbance regimes, and physiographic features.

The other area of edaphology research is concerned with sustainable agriculture and long-term cropping practices to maintain soil fertility and productivity. The field of agroecology deals with the management and cultivation of agronomic ecosystems (or agroecosystems). Intensive conventional agriculture can deplete a soil and lower its quality, reducing the soil's capacity to function in an ecosystem or to resist degradation. By monitoring soil-quality indicators and various edaphic factors, scientists can assess the sustainability of adopted management practices. The two broad areas of ecological and agricultural edaphology have also been combined to allow comparison of intensively managed agroecosystems with virgin soils and native vegetation.

Soil is dynamic, teeming with organisms, and is an integral, interactive part of the environment. It is the foundation upon which our civilization stands and it must be used wisely and conserved if we are to continue benefiting from its numerous functions, including food and fiber production, environmental protection, water and nutrient storage, and engineering materials.

List of Technical Nomenclature

Agroecology	The science of applying ecological concepts and principles to the design and management of sustainable agroecosystems
Ca^{2+}	Calcium ion
CEC	Cation exchange capacity
Climax	The most advanced successional community of plants capable of development under, and in dynamic equilibrium with, the prevailing environment
EC	Electrical conductivity, related to the concentration of salts in solution
Ecology	The science of the relationship between organisms and their environment
Edaphology	The science that deals with the influence of soils on living things; particularly plants, including human use of the land for plant growth
Gleyed/gleying	A condition resulting from prolonged soil saturation and reducing conditions, indicated by the presence of bluish or greenish colors from ferrous iron in the soil mass or as mottles

Humus	Organic soil compounds exclusive of undecayed plant and animal tissues, their 'partial decomposition' products, and the soil biomass. The term is often used synonymously with soil organic matter
K^+	Potassium ion
Macronutrient	A plant nutrient found at relatively high concentrations (greater than 500 mg kg^{-1}) in plants; usually refers to nitrogen, phosphorus, and potassium, but may also include calcium, magnesium, and sulfur
Mg^{2+}	Magnesium ion
mg kg^{-1}	Milligrams per kilogram of dry soil or plant material. Similar to parts per million (ppm)
Micronutrient	A plant nutrient found in relatively small amounts (less than 100 mg kg^{-1}) in plants. These include boron, chloride, copper, iron, manganese, molybdenum, nickel, cobalt, and zinc
mm	Millimeters or 10^{-3} meters. Used to define the length or diameter of soil particles, roots, etc
Mottles	Spots or blotches of different color or shades of color interspersed with the dominant soil color
NH_4^+	Ammonium ion
Sodium adsorption ratio (SAR)	A relation between soluble sodium and soluble divalent cations that can be used to predict the exchangeable sodium fraction of soil equilibrated with a given solution. It is defined as the concentration of sodium divided by the square root of the sum of calcium and magnesium concentrations
Worms m^{-2}	The number of earthworms per square meter of soil

See also: **Pedology:** Basic Principles; **Quality of Soil**

Further Reading

Brady NC and Weil RR (2002) *The Nature and Properties of Soils*, 13th edn. Upper Saddle River, NJ: Prentice-Hall.

Doran JW, Coleman DC, Bezdicek DF, and Stewart BA (eds) (1994) *Defining Soil Quality for a Sustainable Environment*. SSSA Special Publication No. 35. Madison, WI: Soil Science Society of America.

Gershuny G and Smillie J (1999) *The Soul of the Soil: A Guide to Ecological Soil Management*. White River Junction, VT: Chelsea Green.

Gliessman SR (1998) *Agroecology: Ecological Processes in Sustainable Agriculture*. Chelsea, MI: Ann Arbor Press.

Hillel D (1998) *Environmental Soil Physics*. San Diego, CA: Academic Press.

Miller RW and Gardiner DT (2001) *Soils in Our Environment*, 9th edn. Upper Saddle River, NJ: Prentice-Hall.

Plaisance G and Cailleux A (1981) *Dictionary of Soils*. Paris, France: La Maison Raustique Publication (Translated from French and published for the USDA and NSF: New Delhi, India: Amerind Publishing Company, PVT, Ltd).

Porazinska DL and Wall DH (2001) Soil conservation. In: Levin SA (ed.) *Encyclopedia of Biodiversity,* vol. 5, pp. 315–326. San Diego, CA: Academic Press.

Singer MJ and Munns DN (2002) *Soils: An Introduction*, 5th edn. Upper Saddle River, NJ: Prentice-Hall.

Sparks DL (2003) *Environmental Soil Chemistry*, 2nd edn. San Diego, CA: Academic Press.

Tugel AJ and Lewandowski AM (eds) (1999) *Soil Biology Primer.* Ames, IA: NRCS Soil Quality Institute.

Electron Paramagnetic Resonance

See **Electron-Spin Resonance Spectroscopy**

ELECTRON-SPIN RESONANCE SPECTROSCOPY

N Senesi and G S Senesi, Università di Bari, Bari, Italy

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Electron-spin resonance (ESR) spectroscopy, otherwise known as electron paramagnetic resonance (EPR) spectroscopy, is a nondestructive, noninvasive, highly sensitive and accurate analytical technique that can detect and characterize chemical species possessing unpaired electrons, i.e., paramagnetic. Soil species able to produce ESR spectra include organic free-radical moieties in humic substances (HS) and organic- and mineral-associated paramagnetic metal ions. Further, ESR labeling, trapping, and metal-probing methods are used to study the dynamics of macromolecules in soil solution, HS complexation chemistry, and ion sorption on mineral surfaces.

Basic Principles

The Resonance Phenomenon

The basic physical phenomenon underlying ESR spectroscopy is referred to as the 'Zeeman effect,' which consists of the interaction between the magnetic moment of an unpaired electron and an external magnetic field that produces the splitting of the energy levels of the unpaired electron. If a static magnetic field of strength H is applied in the z-direction, i.e., parallel to the z-axis, the Zeeman interaction leads to an energy, E, for an unpaired electron, given by:

$$E = -\mu_z H = -g\beta H M_z \qquad [1]$$

where μ_z is the electron magnetic moment in the direction of the conventional z-axis; g is a dimensionless constant called the spectroscopic electronic splitting factor or g-value or g-tensor; β is the electron Bohr magneton; and M_z is the component of the electron-spin angular momentum in the direction of the applied magnetic field H (z-axis). The values that M_z may assume are $+1/2$ and $-1/2$ depending on the alignment of the electron spin, S, either parallel (high-energy) or antiparallel (low-energy) to the magnetic field direction. Thus, two energy levels exist with an energy difference that increases linearly with the magnitude of H (Figure 1).

In a sample containing unpaired electrons in the thermodynamic equilibrium in a magnetic field of value H_0, a population difference exists between the two energy levels, with an excess population in the lower level. If an incident electromagnetic radiation is supplied to the sample, e.g., by applying an alternating magnetic field of frequency ν_0 perpendicular to the static magnetic field H_0, an absorption of energy, ΔE, occurs provided that ν_0 satisfies the following equation:

$$h\nu_0 = \Delta E = g\beta H_0 \qquad [2]$$

where h is the Planck constant. This is known as the 'resonance condition.' The measurement of this energy absorption, recorded as its first derivative, is the ESR signal (Figure 1), i.e., the basis of ESR spectroscopy.

Spectral Parameters

The position and the shape of the ESR signal depend on the environmental conditions in the vicinity of the electron, which may result in spectral patterns more complicated (Figure 1b) than the single-line spectrum (Figure 1a). The most important effects that influence the position and pattern of the ESR spectrum are the electron Zeeman, nuclear hyperfine, ligand superhyperfine, and nuclear quadrupole interactions. These effects are related to spectral parameters that include the g-factor and nuclear hyperfine and superhyperfine coupling constants.

At a given magnetic field and for a particular microwave frequency, the electron Zeeman effect shifts the resonance position of the free electron from a value of $g = 2.00232$ to a g-value (eqn [2]) that is dependent on the molecular properties of the paramagnetic species. The g-value of organic free radicals is generally close to 2.00 and does not distinguish well between different radicals, whereas for paramagnetic metal ions it is often typical of a particular ion and its valence state. For most paramagnetic ions, the g-value is anisotropic, i.e., dependent on the orientation of the molecule relative to the external magnetic field, and is completely described by the three components g_{xx}, g_{yy}, and g_{zz}, along the three axes x, y, z (x-, y-, z-system) that are generally coincident with molecular symmetry axes. These three components differ from each other ($g_{xx} \neq g_{yy} \neq g_{zz}$) for paramagnetic species that have no principal axis of symmetry, whereas a single isotropic g-factor, g_{iso} or $g_0 = 1/3$ ($g_{xx} = g_{yy} = g_{zz}$) is exhibited by octahedral, tetrahedral, or cubic symmetries. Species with axial symmetry, such as Cu^{2+} and V^{4+}, have one principal or threefold axis of symmetry, conventionally the z-axis, and equivalent x- and y-axes.

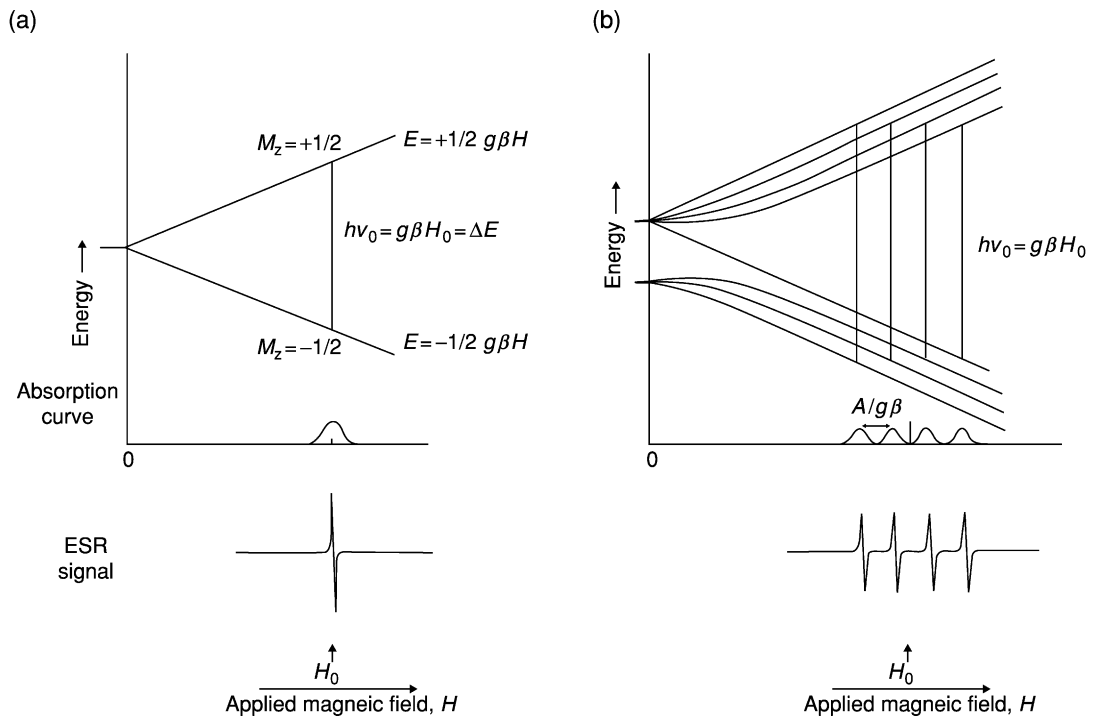


Figure 1 Effect of an applied magnetic field, H, on the energy levels (E) of the two spin states of an electron ($M_z = \pm 1/2$). Electron-spin resonance (ESR) transition(s) at $\nu_0 = g\beta\ H_0/h$, actual absorption curve, and experimental first-derivative ESR spectrum for the cases of no nuclear interaction (a) and interaction with a nucleus having $I = 3/2$ (b), e.g., Cu^{2+}. The nuclear hyperfine splitting is given by $A/g\beta$.

These species feature so-called rigid-limit spectra and exhibit two g-values, usually labeled $g_{\|}$ ($= g_{zz}$, the g-value parallel to the symmetry axis or z-axis), and $g\perp$ ($= g_{xx} = g_{yy}$, the g-value perpendicular to the z-axis in the x,y-plane). The g-values can provide information on the nature of the paramagnetic species and the symmetry of the metal ion in the sample matrix.

The 'hyperfine splitting or structure' arises from the 'nuclear hyperfine' interaction of the unpaired electron with its nucleus if it features a nonzero spin ($I \neq 0$), such as Cu ($I = 3/2$), Mn ($I = 5/2$), or V ($I = 7/2$), and is independent of the applied field H_0. Permanent local fields arising from magnetic nuclei split each electron spin level into $2I + 1$ components, which results in a set of $2I + 1$ equally spaced hyperfine-structure lines replacing the single line resonance in the ESR spectrum. For example, four lines appear for Cu (**Figure 1b**), six for Mn, and eight for V. The hyperfine splitting is generally approximated by $A/g\beta$, where A is the magnitude of the nuclear hyperfine interaction, the so-called hyperfine coupling constant. The parameter A, like g, may also exhibit an orientation-dependence (i.e., anisotropy), thus it may provide useful information on the nature and molecular symmetry of paramagnetic species possessing magnetic nuclei.

The 'superhyperfine splitting or structure' arises when an interaction occurs between the unpaired electron and ligand nuclei having nonzero nuclear spin in paramagnetic metal–ligand complexes. The most common nucleus giving rise to superhyperfine structures in ESR spectra is ^{14}N, which has a nuclear spin of $I = 1$. Thus, each hyperfine line is split into three approximately equally spaced components of equal intensity. Very complex spectra are obtained, however, when more than one N-ligand, especially if not equivalent, is involved in metal complexation.

The 'nuclear quadrupole' interaction arises from the nuclear quadrupole moment (for example, of the Mn nucleus, $I = 5/2$) and the electric field gradient. The interaction, besides modifying the hyperfine splitting, also 'mixes' the hyperfine levels, so that 'forbidden' transitions can occur. Generally, quadrupole effects are very small and rarely observed in powder and frozen solution spectra.

Instrumentation and Methodology

The Instrument and the Experiment

The basic components of a typical ESR spectrometer operating in the X-band (microwave) frequency

region of the electromagnetic spectrum are (Figure 2): (1) an electromagnet supplying a static, or direct current (DC), continuous magnetic field that can be varied in strength, usually in the range 0–1 tesla ($1\,T = 10^4$ gauss); (2) the source of microwaves, usually a vacuum tube oscillator named 'klystron' that provides a monochromatic coherent source of electromagnetic radiation with a frequency typically around 9.5 GHz that is held constant to within 1 part in 10^6; (3) a resonant or 'microwave' cavity, which is a hollow-conducting box generally cuboid where the sample is placed and irradiated; (4) a waveguide that couples the klystron to the resonant cavity that is locked electronically to the resonance frequency; (5) a circulator incorporated between the klystron and the cavity, which ensures optimum transfer of the microwave power from the klystron to the cavity and from the cavity to the detector; (6) a small iris diaphragm that interfaces the cavity to the waveguide and whose effective diameter can be varied until all the incident microwaves are absorbed by the cavity so that it is said to be 'critically coupled'; (7) a magnetic field modulator that increases the intensity of the reflected microwave at low frequency (usually 100 kHz) by superimposing upon the external magnetic field H_0 a second magnetic field of a few tens of millitesla; and (8) a phase-sensitive detector locked to the magnetic field modulation frequency.

In the practical ESR experiment, the microwave frequency is usually fixed and the external applied magnetic field is continuously swept until the resonance condition (eqn [2]) is met. Most ESR spectrometers allow for repetitive sweeping of the magnetic field through the resonance and employ a computer to average transients, thus improving sensitivity. The output of the detector–amplifier system, that is, the ESR spectrum, is displayed as the first derivative of the absorption peak and recorded on a chart recorder as a function of the applied external magnetic field, H (Figure 1). Occasionally, the second derivative is plotted to improve the resolution of closely lying lines.

Sample Preparation

The ESR technique can analyze samples in any form, that is, liquid, solid powder, amorphous polymer, frozen glass, or single crystal. Generally, solid powders are packed into an ESR tube with an inside diameter of a few millimeters. Aqueous solutions or suspensions are placed in capillary tubes or specially designed flat cells to reduce dielectric absorption of microwave radiation by water molecules. Tubes or cells are made of highly pure quartz with no defects, which gives no ESR signal and has a very small dielectric loss. Clay films can also be used mounted on flat quartz 'tissue cells.'

Removal of paramagnetic O_2 molecules that can broaden ESR spectra by dipole–dipole interaction is sometimes required before ESR analysis. If the concentration of paramagnetic species in a sample is too large, the sample must be magnetically diluted in a diamagnetic matrix to minimize spin–spin coupling that may broaden the spectrum.

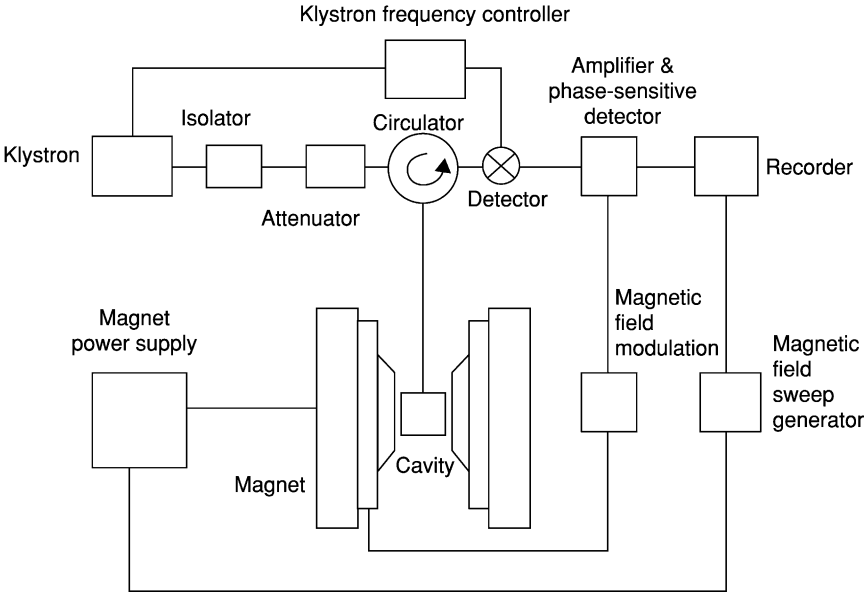


Figure 2 Block diagram of a typical X-band ESR spectrometer.

Sensitivity and Resolution

The sensitivity of the technique depends primarily upon the number of spins present in the sample. Typically, a modern X-band spectrometer can detect 10^{15}–10^{16} spins, which implies a sensitivity of 10^{-7}–10^{-8} moles. The sample must be correctly placed in the cavity that generally has a detection region about 2 cm high, with the greatest sensitivity at its center. Sensitivity is generally improved by setting the modulation amplitude at larger values that, however, should not exceed a fraction of the peak-to-peak resonance linewidth, otherwise distortion of the spectral line shape can occur. According to the Curie law, maximum sensitivity for paramagnetic species is attained at the smallest possible sample temperature. Although organic free-radical species can generally be measured easily at room temperature (RT), for most paramagnetic metal ions cooling at liquid N_2 (bp 77 K) or liquid He (bp 4.2 K), a higher temperature is often required. The extremely large dielectric loss of water at microwave frequencies prevents working at RT with aqueous systems and requires ESR measurements made at liquid N_2 or liquid He temperature on frozen aqueous solutions. However, large metal–organic molecules cannot tumble rapidly in solution, thus the ESR spectrum at RT is often similar to that observed for powders or frozen solutions.

The major limitation of an ESR experiment is the resolution of the signal linewidths that may overlap to such an extent that information is lost. The resonance linewidth is affected by two mechanisms, namely, homogeneous and inhomogeneous broadening. The most important homogeneous broadening effects are: (1) 'spin-lattice relaxation,' by which an electron at a higher energy level returns to the ground state by losing energy to its environment (i.e., the 'lattice'), either in a short time, thus producing linewidth broadening, or in a long time, which results in narrower linewidths; and (2) microwave-power saturation, which may arise if a sufficiently large microwave power is applied which tends to equalize the electron populations of the two levels, thus decreasing signal intensity and increasing linewidth. In the absence of saturation, the ESR signal intensity should increase as the square root of the microwave power. Inhomogeneous broadening arises from nonuniformities in the magnetic field throughout the sample resulting from other neighboring paramagnetic species or from neighboring magnetic nuclei, or from dipolar interactions between unlike spins. These effects, often referred to as spin–spin interactions, are random in direction and result in a merging of individual resonant lines or spin packets into a single overall line or envelope, with loss of line resolution and related information.

Two techniques are potentially useful to overcome this limitation: (1) the electron nuclear double magnetic resonance (ENDOR) spectroscopy, which is a combination of ESR and nuclear magnetic resonance (NMR) techniques; and (2) the electron spin echo (ESE) spectroscopy, which is a time-domain electron magnetic resonance technique. ENDOR and ESE have not yet been applied to soil chemistry, thus major scientific activity is expected to occur in this area.

Determination and Interpretation of Spectral Parameters

Once the ESR spectrum is measured, values of spectral parameters can be determined accurately and rigorously by comparing it with a computer-simulated spectrum for the species being studied, which is calculated with the use of trial parameters and the expressions for the magnetic fields at which transitions occur. Then, the procedure is repeated until satisfactory agreement is found between the two spectra. In practice, the elaborate computer simulation procedure may be avoided by the relatively simple, even though not rigorous, computation of spectral parameters directly from the experimental ESR spectrum and from spectrometer measurement settings using the equations described in detail below.

Once the ESR parameters are obtained, they can be related to the nature of paramagnetic species, metal oxidation state, and type and site symmetry of metal binding by using rigorous physicochemical approaches and complex mathematical elaborations. This procedure can be avoided by empirical comparison and correlation of experimentally obtained ESR spectral parameters with ESR data available on similar model, synthetic, or natural systems, which can provide the information of chemical and structural interest to the soil scientist.

Whole Soil

The ESR spectra of whole-soil samples commonly show a very broad and intense ferromagnetic resonance peak centered at approximately $g = 2.00$, which can be attributed to strongly magnetic materials consisting of randomly oriented crystals and/or polycrystalline powdered particles of different shapes and random orientation, such as iron oxyihydroxides and minerals rich in Mn, Ti, and other transition metals. High porosity and crystal defects and/or impurities can further increase ESR signal linewidth. As a result of these effects, ESR of whole soil can provide very limited chemical and structural information, thus soil components, i.e., HS and their most important fractions, humic acid (HA) and fulvic acid (FA),

metal–HS complexes, clay minerals and metal–clay associations, must be isolated from the whole soil and studied separately.

Organic Free Radicals in Soil Humic Substances

ESR Spectra and Parameters

Humic substances are the most abundant and most chemically and biologically active fraction of the non-living organic matter in soil, whose macromolecules typically contain indigenous free radical moieties. The ESR spectrum of HS free radicals is usually obtained at RT on solid or water-dissolved samples of unfractionated HS, and HA and FA fractions and subfractions. The magnetic field is swept over a relatively narrow scan range (generally 5–10 mT) through the field at which the free electron resonates until the resonance condition is met, which generally occurs at a field of about 340 mT, which corresponds to a *g*-value close to that of the free electron. Typical spectrometer settings and operating conditions used in the measurement of ESR spectra of HS free radicals are: microwave frequency close to 9.5 GHz; microwave attenuation, 13 dB, corresponding to a microwave power of about 10 mW, at which the signal is generally least saturated; and modulation amplitude, 0.63 mT. A typical ESR spectrum of HA and FA free radicals features a single-line resonance, devoid of any structure (Figure 3a), while a partially resolved hyperfine structure is rarely observed (Figure 3b).

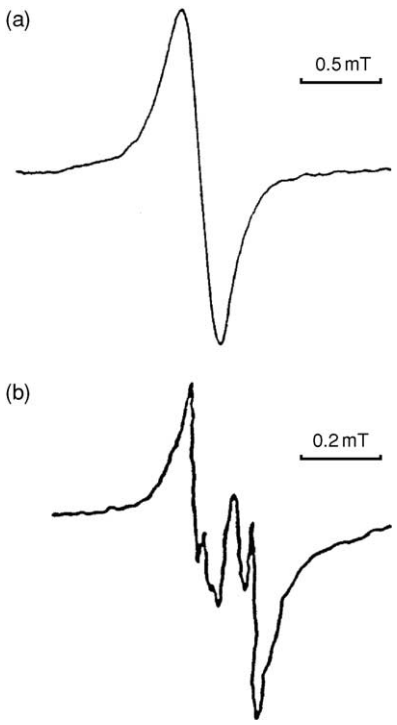


Figure 3 Typical single-line electron-spin resonance (ESR) spectrum of organic free radicals in soil humic acids and fulvic acids (a) and four-line ESR spectrum observed for some acid-boiled humic acids (b). (Adapted with permission from Atherton NM, Cranwell PA, Floyd AJ, and Haworth RD (1967) Humic acid. I. ESR spectra of humic acid. *Tetrahedron* 23: 1653–1667.)

At most, three spectral parameters of importance for the characterization of the nature, origin, and properties of HS free radicals can be derived from ESR spectra, which are the *g*-value, the width of the absorption line, i.e., the linewidth, and, rarely, the hyperfine structure. The *g*-value can be accurately approximated from the magnitudes of the magnetic field at which the resonance occurs for the sample and for a standard of known *g*-value, usually N,N-diphenylpycrylhydrazil (DPPH) diluted in powdered KCl ($g_{DPPH} = 2.0036$). A small amount of standard contained in a capillary tube can be taped to the sample tube and the signal of the standard is used as a field 'marker.' Since ν_0 is identical for both, the sample and standard, g is readily calculated from the ratio of field positions using the relationship derived from eqn [2]:

$$g_u = g_k H_k / H_u \qquad [3]$$

where 'u' (unknown) and 'k' (known) refer to the sample and standard, respectively, and H_k can be calculated from eqn [2] using ν_0 and g_k. The width of the resonance line, or linewidth, ΔH, is generally measured in tesla or in gauss ($1\,\text{gauss} = 10^{-4}\,\text{T}$), as the peak-to-peak separation of the first derivative ESR signal. The hyperfine splitting is measured as the separation (in gauss or in tesla) between the hyperfine lines.

The concentration of unpaired electrons, i.e., of organic free radicals in HS can be determined by comparing its signal area with that of a standard chart containing a known content of paramagnetic centers, e.g., the 'strong pitch' supplied by the manufacturer, or DPPH. Rigorously, a double integration of the first derivative curves should be performed to obtain the corresponding areas. However, since the line shape of the sample and standard are generally both of the Lorentzian type, the double integration can be avoided by simply calculating each area as the product of the height (h) and square of the width (w) of the first derivative signal. If hyperfine splitting of the sample occurs, the area under the hyperfine lines must be summed to obtain the total area to be considered. The spin concentration, expressed in spins per gram, can then be calculated by:

$$(\text{spin g}^{-1})_u = [(\text{spin g}^{-1})_k (hw^2)_u (\text{gain})_k] / [(hw^2)_k (\text{gain})_u q_u] \qquad [4]$$

where 'u' and 'k' are as above (eqn [3]), gain is the signal amplifier gain used, and q_u is the weight (in grams, moisture- and ash-free) of the sample.

Since this procedure requires that the sample and standard be measured in an identical environment, i.e., the same matrix, geometry, volume, temperature, and spectrometer settings, either a double-cavity instrument or concentric sample tubes should be used. However, more often, separate tubes for the sample and standard are used, placing them alternatively in exactly the same position in the resonant cavity.

Interpretation of ESR Parameters

The *g*-values commonly measured for organic free radicals in soil HAs and FAs range between 2.0030 and 2.0050, which are consistent with indigenous semiquinone radical moieties conjugated to aromatic rings (e.g., $g = 2.0041$ for 9,10-anthraquinone), although a contribution from methoxybenzene radicals (*g*-values range, 2.0035–2.0040) and N- or S-associated radicals ($g = 2.0031$–2.0037) cannot be excluded.

The linewidths of the ESR signal of HA and FA free radicals generally range between 3.5 and 7.5×10^{-4} T for solid samples and between 2.0 and 3.0×10^{-4} T for the corresponding water solutions. ESR linewidths do not show any particular dependence on the nature and origin of FAs and HAs, but they are generally slightly greater for FA than for HA from the same source. Factors affecting ESR signal linewidth are free-radical concentration and aggregation state of the sample, interactions with solvent and/or metal ions, power-saturation effects, and temperature.

The ESR spectra of HA and FA free radicals rarely show a hyperfine structure. For example, some acid-boiled HAs feature a four-line hyperfine structure (**Figure 3b**) attributed to the interaction of the unpaired electron with two nonequivalent H nuclei, whereas an oxidized soil FA exhibits a three-line spectrum ascribed to the interaction of the unpaired electron of a semiquinone O atom with two adjacent equivalent H nuclei.

The concentration of organic free radicals in soil HAs and FAs depends on their nature and origin, and generally ranges between 10^{16} and 10^{18} spins g^{-1}. FAs usually show one-third to one-fifth the spin content of HAs from the same source. The spin concentration of FAs measured in solution at neutral and alkaline pHs is always greater than that measured in the solid state. Several factors affect the spin content measurement, including power-saturation effects, the presence of pronounced shoulders associated with a broad signal, and, in the case of solution samples, solvent, pH, and time. Caution should thus be used in evaluating and comparing the spin contents of HS. A reasonable reliability and comparability can be expected, however, in the evaluation of changes in spin concentrations occurring in a certain sample subjected to variations of various physical and chemical conditions.

Factors Affecting ESR Parameters of Humic Substances

The changes of concentration, *g*-value, and linewidth of free radicals in HS subjected to variations of factors such as pH, ionic strength, state of aggregation, hydrolysis, alkylation, redox potential, irradiation, temperature, and humidity can provide valuable insights into their molecular features.

Increase in pH, chemical reduction, acid hydrolysis, or visible- and ultraviolet (UV)-light irradiation enhances the free radical concentration of HS, while leaving almost unaltered the *g*-value and the signal linewidth. However, in all cases except acid hydrolysis, the spin content increase is not sustained in time, i.e., it returns gradually, and almost reversibly, to a value similar to that before the treatment. These results suggest that short-lived, 'transient' free radicals chemically similar to stable native radicals are produced in HS upon any of the mentioned treatments.

Increase in temperature up to 450°C or exposure to high-energy irradiation (gamma rays) causes a marked increase in spin content, broadening of linewidth, and decrease in *g*-value in solid HAs. These changes, which are accompanied by loss of O and increase in C content, can be ascribed to homolytic bond-cleavage and subsequent delocalization and stabilization of newly produced free electrons from semiquinonic O atoms to aromatic C atoms.

Chemical or electrochemical oxidation generally produces a time- and pH-dependent decrease in the free radical content of water-dissolved HAs and FAs, which is reversed by a subsequent reduction treatment. A decrease in spin content is also observed upon increase of neutral electrolyte concentration in FA and HA solutions at pH close to neutrality. However, no changes in *g*-values or linewidths are observed upon any of these treatments. A decrease in both organic free-radical concentration and ESR signal linewidth is observed with increasing humidity of solid HA samples, which is almost reversible upon redrying of the samples. Contrasting results are obtained for the effect of methylation on the free radical content of HS, i.e., it decreases dramatically for podzolic soil HAs and FAs, whereas it increases for several other soil HAs and FAs.

Structural Implications

The free radical concentration is correlated to other compositional and structural parameters of HS, either positively, e.g., to the E_4/E_6 ratio (ratio of

absorbance at 465 nm and 665 nm), absorbance at 465 nm, atomic C/H and O/C ratios, O percentage, phenolic OH content, and aromaticity, or negatively, e.g., to H percentage and aliphaticity. These results imply that the free radical content is directly related to the dark color, aromatic or aliphatic character, molecular complexity, and particle size of HS. A good correlation also exists between the free radical concentration and the degree of humification of HS in peat soils and in organic-amended soils.

Accumulated ESR evidence strongly supports an indigenous quinone (Q)-hydroquinone (H_2Q) system as the principal responsible for the generation and stabilization of semiquinone radicals ($HQ^{\bullet}$) and radical anions ($Q^{\bullet -}$) in HS macromolecules, which is based on the simple electron donor–acceptor (or charge-transfer) reversible mechanism:

$$Q + H_2Q \underset{2OH^-}{\overset{2H^+}{\rightleftarrows}} 2HQ^{\bullet} \rightleftarrows 2Q^{\bullet -} \qquad [5]$$

The large decrease in spin content observed in podzolic HAs and FAs as a consequence of selective blocking of phenolic OH groups by methylation confirms these groups as the most important electron donors responsible for the formation and existence of organic free radicals in HS. A quinone content of 0.5 mmol g^{-1}, typical of most HAs, would theoretically yield up to 6×10^{20} spins g^{-1} of semiquinone anion at alkaline pH and in the presence of a donor group such as hydroquinone. This effect can increase from two to five orders of magnitude the amount of stable free radicals in HS, with relevant implications in their chemical and biochemical reactivity. Further, semiquinone radicals can be generated by reduction or photoirradiation of quinones in solution in the presence of electron donors.

Interactions of Humic Free Radicals with Organic Chemicals and Metal Ions

The ESR technique can usefully be applied for evaluating the role of organic free radicals in the interaction of HS with organic chemicals, such as pesticides. For example, the increase in free radical concentration measured in the interaction products between HAs and *s*-triazine or substituted urea herbicides provides evidence of the occurrence of an electron donor–acceptor mechanism, with formation of charge-transfer complexes. Differently, the products of interaction of chlorophenoxyalkanoic herbicides with HA show a considerable decrease in free radical concentration, which suggests the occurrence of homolytic cross-coupling reactions between HA free radicals and free radical intermediates generated in the preliminary chemical, photochemical, and/or biological degradation of the herbicide.

The ESR spectroscopy can also be used to study the effect of some metal ions on HS free radicals. In general, addition of diamagnetic metal ions such as Na^+, Zn^{2+}, and Al^{3+} to HA or FA solutions does not affect their free radical content, whereas addition of paramagnetic metal ions such as Fe^{3+}, Cu^{2+}, Mn^{2+}, VO^{2+}, Ni^{2+}, and Co^{2+} causes a marked decrease in HS free radical content as a function of the metal species and HS origin. In studies of podzolization processes, ESR data suggest that recently formed, low-molecular-weight HS cannot undergo further polymerization because their free radicals combine with Fe^{3+} and move to the Bh horizon. In another study, ESR spectroscopy reveals the formation of semiquinone radicals by single electron transfers occurring at goethite and manganese oxide surfaces during the oxidation of hydroquinone. Further, ESR spectroscopy can be used to characterize the nature of HS formed by oxidative polymerization of phenolic and other organic compounds in the presence of natural clays, soils, soil oxides, and other metal-ion catalysts.

Paramagnetic Transition Metal Ions in Soil Constituents

The ESR technique can be successfully applied to study natural complexes of paramagnetic metal ions of importance in soil, such as Fe^{3+}, Cu^{2+}, Mn^{2+}, and V^{4+}, with HAs, FAs, and plant litters, and structural paramagnetic metal ions such as Fe^{3+} in soil minerals. In particular, ESR analysis can provide unique information on the identity and oxidation states of metal ions, metal binding site(s) including ligand types, coordination and symmetry, and stability of metal–organic complexes.

ESR Spectra and ESR Parameters

The ESR spectra of paramagnetic metals in soil organic and mineral constituents are usually obtained at either RT on powder samples or at liquid N_2 temperature (77 K) on powders or frozen (77 K) solution samples. Usually, the magnetic field is initially scanned tentatively over a wide range (from 0 to 1 T), then an enlarged spectrum is recorded over a narrower field (0.2 or 0.1 T) comprising the region where resonances appear to allow for their detailed analysis and interpretation. Measurement conditions commonly employed are: microwave frequency, approximately 9.2 GHz (spectra at 77 K) or 9.8 GHz (spectra at RT); microwave attenuation, 13 dB (microwave power oscillates between 9.0 and 9.2 mW); and modulation amplitude, between 0.63 and 4 mT,

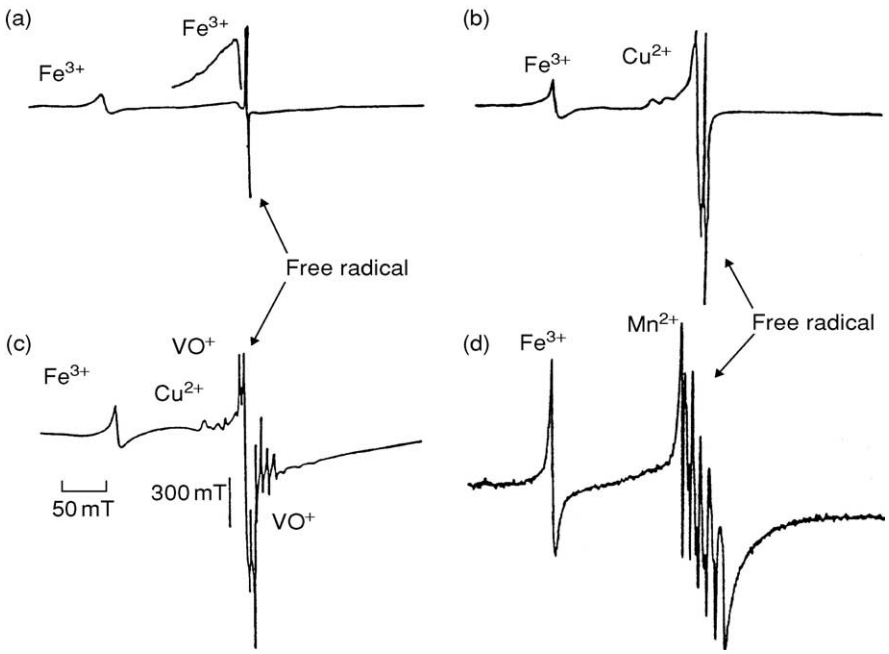


Figure 4 Representative wide scan range (800 mT) electron-spin resonance spectra at 77 K of a Mollisol humic acid from the IHSS Reference and Standard collection (a) (inset, higher gain); a Paleosol humic acid (b); a loam soil humic acid (c); and an aqueous extract of decomposing leaf litter from a forest soil (d). (Reproduced with permission from: Senesi N (1996) Electron spin (or paramagnetic) resonance spectroscopy. In: Sparks DL (ed.) *Methods of Soil Analysis*: *Chemical Methods*, pp. 323–356. Book series no. 5. Madison, WI: Soil Science Society of America/American Society of Agronomy.

according to the metal ion being measured. Four examples of typical ESR spectra of metal–organic complexes naturally occurring in isolated soil organic fractions are shown in Figures 4 and 5.

Three types of spectral parameters can be obtained from ESR spectra of metal–organic complexes: (1) the g-value(s) of the metal(s) present in the sample; (2) the hyperfine coupling constant(s), A, if the metal nucleus has a nonzero spin; and (3) the ligand superhyperfine splitting(s) if the unpaired electron of the metal ion is delocalized on to magnetic nuclei of surrounding ligands. Rarely, resonance lines due to 'forbidden transitions' are observed, e.g., for Mn^{2+}.

The g-values and hyperfine and superhyperfine constants, A, can be calculated from experimental spectral data and spectrometer setting values used according to the standard equations:

$$g = h\,\upsilon_0/\beta H_0 = 0.714484\,\upsilon_0/H_0 \qquad [6]$$

$$A(\mathrm{cm}^{-1}) = A(\mathrm{MHz})/c = 2.80247\,(a\,g)/(c\,g_e) = 0.469766\ 10^{-4}\,a\,g \qquad [7]$$

where υ_0 (megahertz) is the microwave frequency value at the resonance condition; H_0 is the value of the magnetic field at which the resonance is centered (on calibrated chart paper); g_e (2.00232) is the g-value of the free electron; g is calculated by eqn [6]; a is the hyperfine splitting measured as the peak-to-peak separation (in 10^{-4} T) between the hyperfine lines in the experimental spectrum; and c is the speed of light in a vacuum. Accuracy of magnetic field calibration of chart paper is checked by using a suitable standard.

Ferric iron The ESR spectra of HAs, FAs, and decomposing litter layers generally exhibit an asymmetrical isotropic resonance line at about $g = 4.2$, which is consistent with high-spin (five unpaired d-electrons) Fe^{3+} ions held in tetrahedral or octahedral sites of low-symmetry (rhombic) ligand field possibly by carboxylic and/or phenolic hydroxyl groups (Figure 4). This form of Fe exhibits considerable resistance to proton and metal exchange, and to chemical reduction, which suggests that Fe^{3+} is strongly bound and protected in inner-sphere complexes in HS.

A very broad signal near $g = 2$ is often exhibited by HS (Figure 4a, inset), which possibly arises from extended spin–spin coupling of various neighboring Fe^{3+} ions. Iron in such sites is easily reduced by chemical agents and easily extracted by complexing agents, thus suggesting that it is weakly bound on external surfaces of HS. Two weak resonances near $g = 9$ and $g = 6$ are sometimes observed in ESR spectra of HS, which possibly arise, respectively, from Fe^{3+} in sites with near orthorhombic symmetry and from high-spin Fe^{3+} in largely distorted, axially symmetric crystal fields.

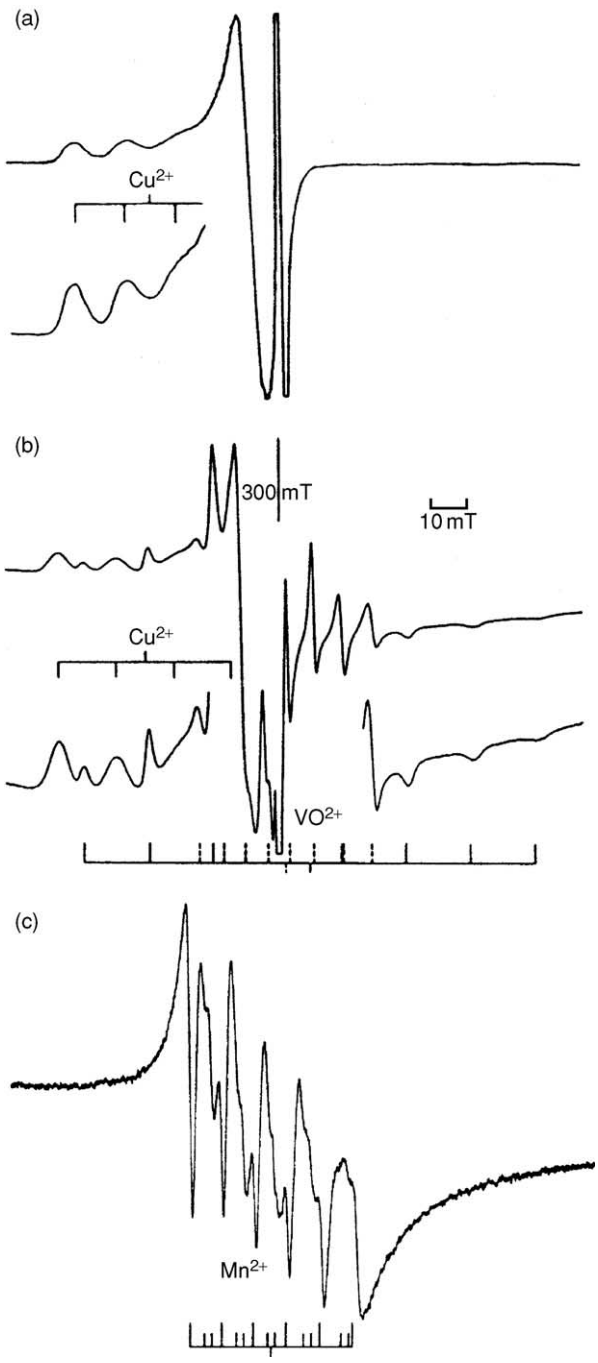


Figure 5 (a)–(c) Same electron-spin resonance spectra as in **Figure 4b,c,d**, respectively, but recorded on an enlarged scan range (200 mT). (Reproduced with permission from: Senesi N (1996) Electron spin (or paramagnetic) resonance spectroscopy. In: Sparks DL (ed.) *Methods of Soil Analysis*: *Chemical Methods*, pp. 323–356. Book series no. 5. Madison, WI: Soil Science Society of America/American Society of Agronomy.

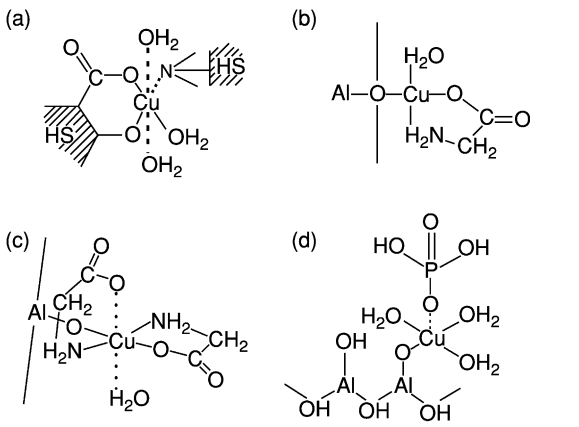


Figure 6 Models for Cu^{2+} complexes in humic substances (a), on surfaces of gibbsite (b), and boehmite (c) in the presence of glycin, and on aluminum oxides and allophanes in the presence of phosphate (d).

An ESR signal at $g = 4.3$ can be observed in layer silicates such as kaolinites, vermiculites, micas, and smectites, which are generally composed of an anisotropic and an isotropic g-factor, and possibly arise from structural octahedral $FeO_4(OH)_2$ groups. Soil clays such as vermiculites and weathered phlogopite often exhibit a strong ferrimagnetic resonance at $g = 2$ that can be attributed to ferrimagnetic clusters of structural Fe^{3+} and/or ferric oxides and tends to obscure other spectral details. No direct ESR evidence of Fe^{2+} species is obtained in soil mineral and organic components.

Divalent copper Soil HS often exhibits an anisotropic rigid-limit spectrum of the axial type in the $g = 2$ region (**Figures 4b,c, and 5a,b**), which is ascribed to Cu^{2+} ions. The nuclear spin of Cu is $I = 3/2$, thus its ESR spectrum should be split into four (i.e., $2I + 1$) features at both $g_{\|}$ and $g_{\perp}$. However, only the component at $g_{\|}$ is generally resolved partially into a quadruplet (**Figure 5a** (and inset), **b** (and inset)), while the splitting of the $g_{\perp}$ component is not resolved. The ESR parameters of this spectrum are consistent with a $d_{x^2-y^2}$ groundstate for Cu^{2+} ions held in square planar (distorted octahedral) coordination sites (tetragonal symmetry) as inner-sphere complexes with either only O ligands, such as carboxyls, phenolic hydroxyls, carbonyls, and, often, water molecules, or both O and N ligands (**Figure 6a**), or even only N ligands (i.e., a tetraporphyrin site). The resolved pattern observed in some cases at $g_{\perp}$ is ascribed to superhyperfine coupling of the Cu^{2+} unpaired electron to N ligand nuclei ($I = 1$).

The values of spectral parameters also provide evidence of a variable covalent bond contribution (i.e., delocalization of the unpaired electron toward the ligands) for Cu^{2+} in HS. At small Cu^{2+} loading, covalent bonding is favored and complexation to amine-N groups is preferred to O ligands; whereas large Cu^{2+} loading in HS determines the formation of small covalency binding of Cu^{2+}, largely to

O-containing ligands. The accurate analysis of Cu^{2+} ESR patterns at $g_{\|}$, possibly with the aid of a computer-simulated spectrum, often reveals the presence of two (or more) superimposed quadruplets, which proves the existence of different binding sites for Cu^{2+} in HS.

Vanadyl ion Soil HAs and FAs may feature a richly structured but relatively well-resolved pattern at about $g = 2$ (Figures 4c and 5b), comprising two distinct, overlapping rigid-limit spectra of the axial type. The one is the typical pattern of Cu^{2+} described previously, and the other consists of two superimposed hyperfine octuplets corresponding to the $g_{\|}$ and $g_{\perp}$ components of VO^{2+} ions (nuclear spin of V, $I = 7/2$) rigidly bound in square planar coordination sites as inner-sphere complexes with phenolate and possibly N ligands (large covalency, tightly bound forms) and/or surface carboxylate groups and water molecules (small covalency, relatively labile and exchangeable forms).

Divalent manganese The ESR spectra obtained for decomposing leaf litters (Figures 4d and 5c) and some soil HAs and FAs feature a well-resolved isotropic pattern in the region $g = 2$, which consists of six almost equally spaced principal lines and, possibly, 10 secondary lines (corresponding to forbidden transitions) of lesser intensity. The ESR parameters of such spectra are consistent with high-spin hexahydrated Mn^{2+} ($I = 5/2$) ions in outer-sphere complexes, bound by electrostatic forces to six O atoms of carboxylate and phenolate groups in a distorted octahedral environment.

Spin-Derivatization Studies

Spin-Labeling and Spin-Trapping

The spin-labeling technique, which consists in the attachment of a simple and stable paramagnetic species, such as the nitroxide radical, to the compound of interest, is a powerful ESR method for evaluating the dynamics of HS macromolecules in solution, and, in turn, for providing information on the aggregation state, molecular conformation, and micellar character of HS. Spin labels can also be used to investigate anisotropic molecular motion on mineral surfaces, which provides information on the nature of adsorption processes not obtainable by other methods.

An interesting illustration of this technique is the use of two organic anionic nitroxide spin labels, provided with either a carboxylate or an organophosphate functional group, to study the nature of bonding to noncrystalline alumina, boehmite, and gibbsite in aqueous suspensions. The ESR analysis reveals that both species are rapidly adsorbed on to the large surface area of alumina and boehmite, whereas only the organophosphate is adsorbed on gibbsite, with a loss in rotational motion, especially of the carboxylate. The values of motional restriction of organic radical anions on surfaces suggest that boehmite adsorbs carboxylate largely by ligand exchange with a single-surface OH and organophosphate by bidentate binding, whereas noncrystalline alumina binds weakly carboxylate by nonspecific electrostatic forces in addition to ligand exchange of surface OH.

The ESR spectra of the spin-label probe 5-SASL (stearic acid spin-label with nitroxide free radical in position 5 of the hydrocarbon chain) in HA suspensions suggest its bonding with surface hydrophobic sites of HA below pH 5, whereas these sites appear not available for bonding above pH 5. Kinetics of release of the nitroxide spin probes, neutral Tempol and cationic Tempamine, from Ca-hectorite aggregates and/or pastes measured by ESR suggest that the probes are not durably sequestered. The use of spin-labeled xenobiotics, e.g., pesticides, may represent a promising application of ESR spectroscopy to study their distribution, fate, and availability in soils.

The technique of spin-trapping involves the reaction of the 'spin trap,' such as nitrosobenzene, 2-methyl-nitrosopropene, 5,5-dimethylpyrroline 1-oxide (DMPO), or phenyl-*N*-*t*-butylnitrone (PBN), with a short-lived radical, e.g., produced in HS by light irradiation, reduction, or raising the pH, to yield a relatively stable radical having an ESR spectrum that allows identification and quantification of the short-lived radical trapped.

Metal Spin Probes

Complexation chemistry of humic substances The ESR analysis of synthetic complexes obtained by reaction of a paramagnetic metal 'probe' such as Cu^{2+}, Mn^{2+}, or VO^{2+}, with natural soil HS can provide useful information on the chemistry of 'residual' complexing capacity of HS toward metal ions and on the stability of formed complexes. For example, when an FA is doped with 5.5–50.1% Fe^{3+}, the $g = 2$ signal is enhanced relative to the $g = 4$ signal, which suggests that Fe^{3+} is preferentially bound in easily exchangeable forms to surface sites of FA. Evidence of inner-sphere complexes of Mn^{2+} coordinated octahedrally, possibly with carboxyl, phenolic hydroxyl, and/or a carbonyl group, is obtained for some soil HAs doped with Mn^{2+} at high pH (>8) or high temperature (>50°C). Rotational correlation times obtained by ESR analysis combined with gel-filtration chromatography data allow measurement of the dynamics of motion and stoichiometry of FA–VO^{2+} complexes in

solution, and give information on their molecular conformation and aggregation properties. Quantitative ESR spectroscopy can also be used as a more sensitive, convenient, and faster technique to determine weighted-average equilibrium constants (K_c) for water-soluble Mn^{2+}–FA complexes, which are in excellent agreement with K_c values determined by ion-exchange methods.

Ion exchange on layer silicates ESR spectroscopy confirms that metal ions such as Cu^{2+}, Mn^{2+}, VO^{2+}, and Cr^{3+} retain their inner hydration sphere and a great degree of rotational mobility on exchange sites of layer silicate clays. For example, ESR spectra of exchangeable Cu^{2+} and Mn^{2+} on kaolinite indicate the presence of planar $Cu(H_2O)_4^{2+}$ ions oriented parallel to the surface and possessing a great degree of mobility. In contrast, an orientation-dependent, rigid-limit ESR spectrum of Cu^{2+} is obtained for hydrated Cu^{2+} ions adsorbed on hectorite when the interlamellar spacing on the clay is limited to the equivalent of one or two molecular layers of water, which suggests a great degree of motional restriction for Cu^{2+} ions.

Chemisorption on mineral surfaces ESR analysis proves that paramagnetic metal ions at trace levels are bound rigidly in chemisorbed, nonexchangeable forms at isolated surface sites of soil oxides, hydroxides, and aluminosilicates, whereas they are sorbed in exchangeable forms to layer silicate clays. For example, Cu^{2+} is proven to chemisorb at isolated sites on noncrystalline $Al(OH)_3$ and microcrystalline AlOOH, possibly by formation of one or two direct bonds between surface Al-O groups and Cu^{2+}, which is favored at high pH. At low pH, gibbsite is able to chemisorb small amounts ($<0.5\,\text{mmol}\,100\,\text{g}^{-1}$) of monomeric Cu^{2+} that is oriented with its z-axis perpendicular to the (001) planes of the mineral. At pH >5, the appearance of a broad, featureless resonance and the reduction of the rigid-limit spectrum of Cu^{2+} suggest the presence of different, possibly hydrolyzed and polymerized forms of Cu^{2+} on the surface. The large decrease in g-values and increase in $|A_{\|}|$-values of the low-pH, gibbsite-chemisorbed Cu^{2+} upon exposure to NH_3 vapors suggest that Cu^{2+} may form a complex with NH_3 while remaining rigidly bound to the oxide surface. ESR data indicate that Cu^{2+} ions in the presence of the chelating ligand glycine absorb on microcrystalline gibbsite and boehmite in the form of ternary complexes in which Cu^{2+} coordinates simultaneously with one surface hydroxyl and one (on gibbsite) or two (on boehmite) glycine molecules, with the orientation of the Cu^{2+} z-axis normal to the (001) sheets of the mineral (Figure 6b,c).

ESR studies reveal that Cu^{2+} can be adsorbed by a direct bond (chemisorption) in nonexchangeable forms on allophanes and imogolite both by a preferential binuclear mechanism involving two adjacent AlOH groups and on a weaker type of binding site probably involving isolated AlOH or SiOH groups. Here also NH_3 is able to readily displace H_2O and OH-ligands from chemisorbed Cu^{2+}, leading to the formation of ternary Cu^{2+}–ammonia–surface complexes. At low pH, Cu^{2+} is also shown to chemisorb strongly to bidentate sites of titanium dioxide, whereas at higher pH a weaker complex is formed involving single Ti-OH groups.

The ESR technique shows that Cu^{2+} is adsorbed on aluminum oxides and allophanes in the presence of phosphate, forming a ternary complex with the phosphate coordinated to the axial position of a surface-bound Cu^{2+} (Figure 6d). Large amounts of phosphate suppress Cu^{2+} adsorption, apparently by blocking the coordination of Cu^{2+} to surface AlOH groups. ESR evidence is also provided that VO^{2+} can be coadsorbed with phosphate on boehmite and aluminosilicates as an inner-sphere complex. Finally, ESR proves that Cr^{3+} in the presence of selenite, phosphate, and fluoride is coadsorbed on hectorite and montmorillonite, with formation of ternary surface complexes.

Thermodynamic constants from ESR parameters Metal spin probes can be used to estimate thermodynamic stability constants of metal–surface complexes from the ESR parameter $g_{\|}$. For example, a linear relationship is found between the $g_{\|}$-value and the corresponding formation constants of square planar Cu^{2+} complexes with hydrous Al_2O_3, TiO_2, and some silicas in the presence of bidendate ligands. A decrease in the $g_{\|}$ value by 0.1 units corresponds to an increase in thermodynamic stability of about eight orders of magnitude. On these bases, a major revision of current concepts of cation adsorption on layer silicates is possible.

List of Technical Nomenclature

β	Electron Bohr magneton
ΔE	Absorption of energy by the electron
ΔH	Width of the resonance line, or linewidth
μ_z	Electron magnetic moment
A	Hyperfine coupling constant
DC	Direct current
DMPO	5,5-Dimethylpyrroline-1-oxide
DPPH	N,N-Diphenylpycrylhydrazil

E	Energy for the unpaired electron
ENDOR	Electron nuclear double resonance
EPR	Electron paramagnetic resonance
ESE	Electron spin echo
ESR	Electron-spin resonance
FA	Fulvic acid
g	Electronic-splitting factor, or g-value, or g-tensor
$g_{\parallel}$	Component of the g-tensor parallel to the symmetry axis
$g_{\perp}$	Component of the g-tensor perpendicular to the z-axis
g_{iso} or g_0	Isotropic g-factor
g_{xx}, g_{yy}, and g_{zz}	Components of the g-tensor along the axes x, y, z
h	Planck constant
HA	Humic acid
H, H_o	Strength of the static magnetic field
H_2Q	Hydroquinone
$HQ^{\bullet}$	Semiquinone radical
HS	Humic substances
I	Nuclear spin
K_c	Weighted-average equilibrium constant
M_z	Electron-spin angular momentum
NMR	Nuclear magnetic resonance
PBN	Phenyl-*N*-*t*-butylnitrone
Q	Quinone
$Q^{\bullet -}$	Semiquinone radical anion
q_u	Weight of the sample
RT	Room temperature
S	Electron spin
5-SASL	Stearic acid spin-label with nitroxide free radical in position 5 of the hydrocarbon chain
T	tesla
υ_0	Frequency of the alternating magnetic field

See also: **Clay Minerals**; **Fourier Transform Infrared Spectroscopy**; **Heavy Metals**; **Organic Matter:** Principles and Processes; Interactions with Metals; **Pollutants:** Persistent Organic (POPs); **Sorption:** Metals

Further Reading

Abragam A and Bleaney B (1970) *Electron Paramagnetic Resonance of Transition Ions*. Oxford, UK: Clarendon Press.

Berliner LJ (1976) *Spin Labelling: Theory and Applications*. New York: Academic Press.

Gordy W (1980) *Theory and Application of Electron Spin Resonance*. New York: Wiley-Interscience.

McBride MB (1986) Magnetic methods. In: Klute A (ed.) *Methods of Soil Analysis*, 2nd edn, part 1, pp. 219–270. Book Series No. 5. Madison, WI: Soil Science Society of America/American Society of Agronomy.

Poole CP (1983) *Electron Spin Resonance: Comprehensive Treatise on Experimental Techniques*, 2nd edn. New York: John Wiley.

Senesi N (1990) Application of electron spin resonance (ESR) spectroscopy in soil chemistry. In: Stewart BA (ed.) *Advances in Soil Science*, vol. 14, pp. 77–130. New York: Springer-Verlag.

Senesi N (1990) Molecular and quantitative aspects of the chemistry of fulvic acid and its interactions with metal ions and organic chemicals, part I, The electron spin resonance approach. *Analitica Chimica Acta* 232: 51–75.

Senesi N (1996) Electron spin (or paramagnetic) resonance spectroscopy. In: Sparks DL (ed.) *Methods of Soil Analysis: Chemical Methods*, pp. 323–356. Book Series no. 5. Madison, WI: Soil Science Society of America/American Society of Agronomy.

Senesi N and Steelink C (1989) Application of ESR spectroscopy to the study of humic substances. In: Hayes MHB, MacCarthy P, Malcolm RL, and Swift RS (eds) *Humic Substances. II. In Search of Structure*, vol. 2, pp. 373–407. Chichester, UK: John Wiley.

Thomson AJ (1990) Electron paramagnetic resonance and electron nuclear double resonance spectroscopy. In: Andrews DL (ed.) *Perspectives in Modern Chemical Spectroscopy*, pp. 295–320. Berlin, Germany: Springer-Verlag.

Weil JA, Bolton JR, and Wertz JE (1994) *Electron Paramagnetic Resonance: Elementary Theory and Practical Applications*. New York: John Wiley.

Electrostatic Double-Layer *See* **Cation Exchange**

ENERGY BALANCE

M Fuchs, Agricultural Research Organization, Bet Dagan, Israel

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

The genesis of soils is the transformation of the lithosphere in contact with the atmosphere. It involves physical, chemical, and biological processes that depend on meteorological conditions. For this reason, the geographic distribution of soil types is related to the distribution of climatic zones. The variables characterizing the climate, temperature, humidity, precipitation, wind, and radiation take on the values that equilibrate the energy balance of the surface. The driving energy source is solar radiation. The atmosphere, water in its three phases, and life determine how components of the energy balance derive from the solar energy input. Temporal variation of incoming solar energy causes heating or cooling of the atmosphere, heating or cooling of the ground, evaporation or condensation of water, thawing or freezing of water, and photochemical reduction or oxidation. Spatial variation of the sun's energy input drives the active atmospheric circulation.

The Energy Input From the Sun

The Earth's surface intercepts permanently from the Sun a mean radiant energy flux density of $342\,\mathrm{W\,m^{-2}}$. This power input undergoes a predictable annual fluctuation of $\pm 3.3\%$, because the Earth revolves around the Sun along an elliptic track. It reaches its maximum by 3 January and its minimum on 5 July. Unpredictable changes of solar activity cause additional variations amounting to less than $\pm 1\%$.

The annual energy density reaching the top of the atmosphere has a latitudinal distribution. It decreases from $13.17\,\mathrm{GJ\,m^{-2}}$ (mean flux density $= 434\,\mathrm{W\,m^{-2}}$) at the equator down to $5.45\,\mathrm{GJ\,m^{-2}}$ (mean flux density $= 173\,\mathrm{W\,m^{-2}}$) at the poles. The latitudinal distribution of the daily amount varies in the course of the year, because the rotation axis of the Earth has a fixed declination of 23.5° on the plane formed by the Earth's track around the Sun. The instantaneous flux density fluctuates in the course of a day because of the Earth's rotation. The range of variation is between the extraterrestrial mean maximum of $1368\,\mathrm{W\,m^{-2}}$ and zero.

Energy Balance of a Planet Without Atmosphere

To illustrate the impact of the atmosphere on the energy budget of the Earth's surface, it is instructive to consider the energy balance of a planet on the same track around the Sun, but deprived of its atmosphere. In this case the only energy-exchange processes are by radiation:

$$(1 - r_S)R_S - \epsilon_G \sigma T_G^4 = 0 \qquad [1]$$

where r_S is the mean albedo or solar radiation reflection coefficient of the planet ground surface, R_S is the incoming radiation originating from the Sun ($342\,\mathrm{W\,m^{-2}}$ in the wavelength band from 0.2 to $3\,\mu\mathrm{m}$), σ is the Stefan–Boltzmann constant ($5.67 \times 10^{-8}\,\mathrm{W\,m^{-2}\,K^{-4}}$), and ϵ_G is the emissivity of the planet's surface ($\cong 0.98$). The radiation emitted by the planet $\epsilon_G \sigma T_G^4$ is in the wavelength band between 5 and $100\,\mu\mathrm{m}$. It balances the solar radiation absorption of the surface. Setting the albedo of the planet's surface to be equal to that of the Earth as viewed from space, $r_S = 0.37$, the planetary radiation loss of the surface $\epsilon_G \sigma T_G^4 = (1 - r_S)\ R_S = 216\,\mathrm{W\,m^{-2}}$. The resulting mean Kelvin surface temperature T_G of the planet is 249.6 K or −23.6°C.

An increase of $1\,\mathrm{W\,m^{-2}}$ in solar activity believed to have occurred over the past 300 years would have produced a warming of 0.2°C in a planet without atmosphere. This is the probable temperature rise of the Earth that occurred during the 200 years preceding the acceleration of the warming trend observed since the early 1900s.

Energy Balance of the Earth's Surface

The presence of an atmosphere transforms the energy balance, because the molecules of gases and the solid particles in suspension absorb and scatter some of the incoming solar radiation. The absorbed energy heats gases and particles. In turn, the heated gases and particles emit long-wave radiation toward the surface and to outer space. According to Kirchhoff's law, the emissivity is equal to the absorptivity of radiation at the same wavelength, so the atmosphere absorbs a fraction of the radiation emitted by the Earth's surface, thereby preventing its escape to outer space. The absorbed radiation is partly radiated back to the ground. Turbulent convection carries some of the heat absorbed by the ground to the atmosphere and

contributes to the temperature elevation of the atmospheric gases and of the particles held in suspension. Part of this heat is included in the back-radiation. The remainder is lost by radiation into space. Heat transport from the ground surface to the atmosphere implies that the ground is warmer than the atmosphere. As radiative emission is proportional to the fourth power of the Kelvin temperature, radiation from the atmosphere to outer space is less than the heat emitted from the ground. To compensate for this decreased global emission, the ground elevates its surface temperature to emit more radiation.

Early information on the energy balance of the Earth was derived from the analysis of meteorological and hydrological ground observations. Satellite measurements of the radiative components of the atmosphere, modeling of the radiative transfer through gases, and of atmospheric circulation are the sources of the current state of knowledge.

Figure 1 shows annual mean values of the main components of the Earth's energy balance. Of the $342\,W\,m^{-2}$ of solar radiation reaching the top of the atmosphere, $205\,W\,m^{-2}$ meets clouds that transmit $88\,W\,m^{-2}$ to the Earth's surface and $77\,W\,m^{-2}$ is reflected into space. One hundred and thirty-seven watts per square meter impinge on clear atmosphere and $110\,W\,m^{-2}$ is transmitted to the surface. The mean atmospheric transmissivity τ_A for solar radiation is 0.58. The Earth's surface reflects $30\,W\,m^{-2}$ into space; thus the net absorption of solar radiation is $168\,W\,m^{-2}$. The spectrum of solar radiation reaching the Earth's surface is referred to as short-wave radiation and includes ultraviolet, visible, and near-infrared radiation.

The presence of water on the ground surface and in the atmosphere has a major impact on the energy balance of the Earth. The 60% mean cloud cover enveloping the Earth reflects back to space 22% of the incoming solar radiation. Water vapor in the cloudless part of the atmosphere absorbs two-thirds of the $27\,W\,m^{-2}$ of solar radiation absorbed by atmospheric gases; clouds absorb $40\,W\,m^{-2}$.

Terrestrial matter emits radiation in the spectral waveband from 5 to 100 μm. It is referred to as 'long-wave radiation' or 'far-infrared radiation.' The radiative emission of the atmosphere including clouds is $324\,W\,m^{-2}$. The Earth's surface absorbs $\alpha_G\ R_A = 318\,W\,m^{-2}$ of this flux density (where $\alpha_G = 0.98$ is the long-wave absorptivity of the Earth surface) and reflects $6\,W\,m^{-2}$. The surface emission is

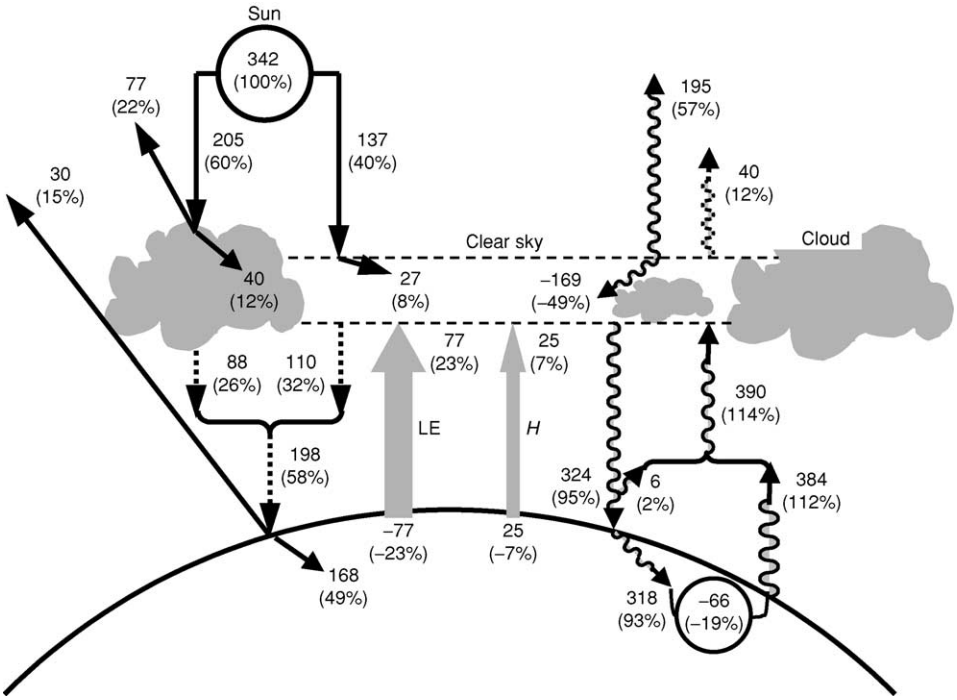


Figure 1 Mean annual values of the Earth's energy balance (main terms expressed in $W\,m^{-2}$). The numbers enclosed in parentheses are percentages of $342\,W\,m^{-2}$ extraterrestrial solar radiation as 100. The straight-tailed arrows symbolize flux densities in the solar wave band (0.2–3 μm). Arrows with a wavy tail represent radiation emitted by the Earth. The wavelength range of this terrestrial radiation is from 5 to 100 μm. Vertical arrows show downward flux densities. Slanted upward arrows represent reflected flux densities. Slanted downward arrows indicate absorbed flux densities. Dotted-tailed arrows mark flux densities transmitted either upward or downward according to the direction of the arrowhead. The gray arrows indicate latent heat for evapotranspiration (LE) from water surfaces or water held in the soil, and *H* is the sensible heat convected from the surface to the atmosphere.

$\epsilon_G \sigma T_G^4 = 384\,W\,m^{-2}$, leading to a mean temperature of the Earth: $T_G = 288.3$ K or. 15.1°C. The total long-wave radiation loss from the surface is $390\,W\,m^{-2}$. The resulting net long-wave radiative loss of the Earth surface is $-66\,W\,m^{-2}$.

As Kirchhoff's law states $\alpha_G = \epsilon_G = 0.98$, the summarized format of the Earth's surface radiation heat balance R_n reduces to:

$$R_n = \tau_A R_S + \epsilon_G (R_A - \sigma T_G^4) \qquad [2]$$

Comparison between T_G of the Earth (15.1°C) with T_G (−23.7°C) of the planet without an atmosphere gives an estimate of the global temperature elevation due to the radiative shielding by the atmosphere: gases and suspended liquid − solid particles. The radiative blanketing effect is best illustrated by comparing the net long-wave loss of $-216\,W\,m^{-2}$ from the planet without atmosphere with the net long-wave loss of $-66\,W\,m^{-2}$ from the Earth's surface. This radiative heat gain is commonly referred to as the 'greenhouse effect.' The wording is misleading, because the heating of a greenhouse exposed to solar radiation results from preventing turbulent convective heat transport and not from radiation trapping.

Gases and suspended particles emit $195\,W\,m^{-2}$ of long-wave radiation to space. As the atmosphere receives more radiation from the Earth surface than it emits to space, its net long-wave radiation deficit is only $-169\,W\,m^{-2}$. An additional $40\,W\,m^{-2}$ is lost to space by transmission through the radiative window of the atmosphere. The release of radiatively active gases such as CO_2 and aerosols into the atmosphere diminishes this transmission. Yet, for a proper perspective on the significance of the possible effects, one should keep in mind that water vapor absorbs 60% of long-wave radiation emission blocked by the atmosphere as compared with only 26% absorbed by CO_2.

The mean value of R_n is $102\,W\,m^{-2}$, but consensus among scientists on its value has not yet been reached. Most publications agree within a few percent on the solar radiation term $\tau_A R_S$, but estimates of the net long-wave radiative loss vary between −72 and $-40\,W\,m^{-2}$.

The net radiation dissipates as turbulent sensible heat convection into the air (H) and latent heat for evapotranspiration (LE):

$$R_n + H + LE = 0 \qquad [3]$$

The partition between H and LE depends on the availability of water at the ground surface. As oceans cover more than 70% of the Earth, water is readily available for evaporation.

The instantaneous energy balance for a local site also involves conductive heat flow G through the surface into or out of the ground subsurface. The heat flow is inward when the surface is warmer than the subsurface and outward for the opposite condition. Its magnitude depends on the thermal conductivity and the volumetric heat capacity of the ground. The depth of the subsurface layer affected by the diurnal change of the solar radiation on land is approximately 1 m. The corresponding annual variation penetrates to a depth of 20 m ($\approx 1\,m \times \sqrt{365}$). The heat penetration in oceans is at least one order of magnitude larger than on continental areas. As during the annual cycle heating and cooling of the Earth's surface have symmetrical amplitudes, the term G vanishes in eqn [3] applied to a local site. The term G also vanishes in eqn [3] when it is applied to the entire planet, because the distribution of heating and cooling on the rotating spherical Earth is symmetrical.

Based on observations of the hydrologic balance (worldwide distribution of precipitation and runoff from rivers flowing into the oceans) current estimates set mean evapotranspiration expressed as latent heat flux density to $-77\,W\,m^{-2}$. Sensible heat flux density of $-25\,W\,m^{-2}$ is obtained as the residual term of eqn [3] (**Table 1**). Clearly, evapotranspiration is an important constituent of the surface energy balance that cools the Earth's surface by dissipating a large proportion of the absorbed radiative heat.

Table 1 shows that the net radiation of oceans $R_n = 118\,W\,m^{-2}$ is more than that of the Earth. This is because the albedo of water is only 0.06, whereas the mean Earth-surface albedo is 0.15. Mean evaporation from oceans amounts to a latent heat flux density of $-94\,W\,m^{-2}$. The theoretical equilibrium latent heat flux density into a vapor-saturated atmosphere is $-73\,W\,m^{-2}$. The higher evaporation is an indication that the average water vapor content of the atmosphere above the oceans is below saturation, vapor being removed by condensation and advection to the continental areas. As latent heat release equivalent of the rainfall values over oceans is only $86\,W\,m^{-2}$, some of the evaporation from the oceans condenses and precipitates on continents. Indeed, the hydrologic balance of continents shows that mean

Table 1 Partition of the global surface energy balance ($W\,m^{-2}$)

	R_n	*LE*	H
Oceans	118	−94	−24
Land	64	−37	−27
Earth	102	−77	−25

R_n, the Earth's surface radiation balance; LE, latent heat for evapotranspiration; *H*, heat convection.

precipitation over land releases 56 $W m^{-2}$ as latent heat into the atmosphere and exceeds the 37 $W m^{-2}$ of mean latent heat flux density for evapotranspiration from land locked water. The balance is closed by runoff of water from the rivers flowing into the oceans. Continental evapotranspiration includes direct evaporation of water from the soil and transpiration of the soil water taken up by vegetation. Desiccation of soil surfaces forms a barrier that retards direct evaporation of water held in the deeper layers. Plant roots can extract water held in the soil below the surface and transport it to the leaves where evaporation occurs. For this reason, the presence and vitality of a vegetation cover have a major role on latent heat dissipation from land areas.

Eqn [3] also summarizes the energy balance of the atmosphere. Its net radiation R_n is the sum of the total solar radiation absorption, 67 $W m^{-2}$ (40 + 27) and the net long-wave radiation loss -168 $W m^{-2}$ amounting to -102 $W m^{-2}$. This flux density is balanced by the 77 $W m^{-2}$ of latent heat released by the condensation of the water vapor produced by evapotranspiration and the 25 $W m^{-2}$ of direct convective heating.

Anthropogenic Change of the Energy Balance

The previous sections show that the long-wave radiative effects of clouds, aerosols, and of gases elevate the Earth's surface temperature. The -66 Wm^{-2} long-wave radiation loss by the Earth surface are partly reabsorbed in the atmosphere, but 40 Wm^{-2} escape through a radiative window of the atmosphere to space. Gases with long-wave absorption bands such as CO_2, CH_4, and NO_X diminish the transparency of this radiative window. The anthropogenic release of these gases elevates their presence in the atmosphere and increases long-wave radiation absorption. Part of this absorbed radiation is remitted toward space, but part augments the long-wave radiative load of the ground surface. As 5.5 Wm^{-2} is required to raise the temperature by 1 K, the obstruction of the radiative window could increase the surface temperature by several degrees.

Air pollution is also reducing the transparency of the atmosphere for solar radiation on a global scale. A recent study shows that solar radiation reaching the Earth's surface has decreased by 20 Wm^{-2} over the past 30 years. The study does not show how much of the missing solar radiation is absorbed or reflected to outer space. Yet, the lowering of the primary energy input reaching surface should cool the Earth's surface or at least retard the heating trend due to long-wave absorbing gases.

Further Reading

Budyko MI (1958) *The Heat Balance of the Earth's Surface*. (translated by Stepanova NA). US Department of Commerce. Washington, DC: Government Publishing Office.

Budyko MI (1982) *The Earth's Climate: Past and Future*. London, UK: Academic Press.

Chahine MT (1992) The hydrological cycle and its influence on climate. *Nature* 359: 373–380.

Kiehl JT and Trenberth KE (1997) Earth's annual global mean energy budget. *Bulletin of the American Meteorological Society* 78: 197–208.

Paltridge GW and Platt CMR (1976) *Radiative Processes in Meteorology and Climatology*. Amsterdam, the Netherlands: Elsevier Scientific.

Peixoto JP and Oort AH (1992) *Physics of Climate*. New York: American Institute of Physics.

Rind D (2002) The Sun's role in climate variations. *Science* 296: 673–677.

Stanhill G and Cohen S (2001) Global dimming: a review of the evidence for a widespread and significant reduction in global radiation with discussion of its probable causes and possible agricultural consequences. *Agricultural and Forest Meteorology* 107: 255–278.

ENVIRONMENTAL MONITORING

P J Loveland and P H Bellamy, Cranfield University, Silsoe, UK

© 2005, Elsevier Ltd. All Rights Reserved.

> The pains that come from the necessities of nature, are monitors to us to beware of greater mischiefs, which they are the forerunners of; and therefore they must not be wholly neglected, nor strain'd too far. (John Locke, *Some Thoughts Concerning Education* §107, 1693)

Introduction

Soil monitoring (SM) is often driven by the desire to understand soil quality and changes in it. This usually requires the establishment and population of soil-quality indicators (*See* **Quality of Soil**). This can,

and does, have a profound influence on what SM is carried out, how a program is designed and executed, and what it costs. Effective SM implies long-term commitment and very tight control of potential sources of variability. Decisions taken at the planning stage are of extreme importance, and careful attention to them can save much money and trouble. Two essential points are:

- There is no universal system of soil monitoring; the means must fit the purpose and *vice versa*.
- There is a clear distinction between a soil-inventory program and a soil-monitoring program.

The first point might seem obvious, but is often overlooked. What is appropriate for an industrial site and its surroundings, where potential hazards to human health might be the driver, can be wholly inappropriate in an area where maintenance of a particular soil ecosystem is the target.

The second point is that an inventory is a collection of information about a place. For soils, it might be the features observed at the location of a soil profile and the properties of the profile horizons. 'Monitoring' implies temporal change, i.e., observation, sampling, and measurement repeated over time. There is thus a different intent and commitment in soil monitoring compared with an inventory, although the latter might be the starting point (baseline) of a soil-monitoring program. It is stressed that a single set of measurements in time does not constitute soil monitoring.

Most commonly, soil-monitoring targets change in a specific property or properties, e.g., pH, metal content, biomass. However, it can be equally important to study a process, e.g., soil erosion. Increasingly, there is a demand to understand potential soil functions. Linking soil monitoring to other environmental monitoring programs can maximize long-term benefits from the data collected. Planning SM needs to take these different objectives into account.

Finally, although many of the points discussed below might also apply to small parcels of land, this article is mostly concerned with the concept of soil monitoring applied at the regional or national scale.

Why Monitor Soils?

In recent years, there has been growing acceptance that not enough is known about soils and the changes that they are, or might be, undergoing, especially under increased pressure from man. This contrasts with the effort that has gone into the assessment of air and water. There are concerns that pressures on soils, usually driven by economic needs, will lead to irreversible decline in their ability to maintain ecosystem diversity, food, fiber and timber production and this will lead, indirectly, to serious problems with water and air quality. Thus, the whole framework of sustainable development would be compromised. There have been many statements over the last few decades pointing to the need to consider soils in terms of the:

- establishment of a set of principles for the rational use of soils and protection of them against irreversible degradation;
- pursuit of programs of soil conservation and reclamation and appropriate use of soils;
- recognition of soils as a fragile and essentially irreplaceable resource;
- importance of specific soil properties as the determiners of best use and land management practices;
- need for inventories and monitoring programs to establish baseline soil properties and changes in these over time.

Against this background, soil monitoring can be instigated for a number of reasons:

A. Basic science: are soil properties changing and, if so, at what rate and in what direction, and what does this tell us about our understanding of the world? For example, establishment of the magnitude of change in soil carbon content could improve inputs into models of the global carbon cycle;

B. Political: the need for information from which to formulate, test or improve policy, e.g., conventions on transboundary pollution require reductions in the acid load to soils; this could be demonstrated by showing that soil pH is remaining above a given value over a given time;

C. Legal: whether change has occurred to a soil that is in conflict with legislative obligations, e.g., demonstrate that loadings of so-called heavy metals are within statutory limits;

D. Financial: are the properties of a soil changing (or likely to change) over a given time-frame such that the soil is, or will become, more or less valuable in a given context, e.g., is there a potential pollution problem with otherwise very valuable development land?

Much of this thinking is increasingly considered in terms of the ability of soils to perform several functions, often simultaneously, i.e., food and fiber production, filtering, buffering, acting as a pool of genetic resource, support for diversity of above- and belowground ecosystems, keeper of the archeological record, support for the built environment. Thus SM is under pressure to provide information on the performance (or potential performance) of these functions and, often, to guide choices as to which function

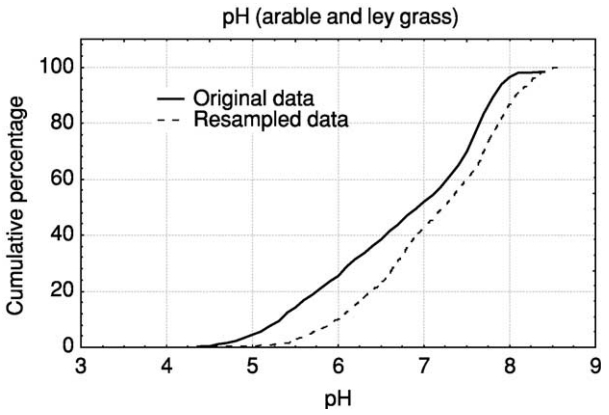


Figure 1 Change in topsoil pH between 1980 and 1995 under an arable/ley grassland rotation in England and Wales ($n = 900$). (Loveland PJ (1990) The National Soil Inventory: survey design and sampling strategies. In: Lieth H and Markert B (eds) *Element Concentration Cadasters in Ecosystems*, pp. 73–80. Weinheim, Germany: VCH Verlagsgesellschaft.) (Data from the National Soil Inventory of England and Wales.)

should be 'allocated' to a parcel of land in preference to another, e.g., build here but not there. Thus, simple measurement of simple properties is often no longer the absolute *raison d'être* of soil monitoring, although it is often the baseline. Figure 1 shows that after 15 years the pH of topsoils (0–15 cm) has increased under arable/ley grass rotation in England and Wales, i.e., the soils are less acid. This finding has been interpreted as a mixture of a reduction in acid deposition (policy response), better targeting of the use of agricultural lime (improved agronomic practice), and deep plowing into less acid subsoil because of larger, more powerful farm machinery (economic driver). All these reasons can be viewed as useful outcomes of SM and could lead to more focused targeting of future environmental research.

The Starting Point

Whatever the driver for SM, some fundamental questions need addressing:

A. What information is already available about soil type(s), properties, scale of observations, etc.?
B. Is this information internally consistent, e.g., are, or were, all the data obtained in the same way, or in ways comparable enough for the purpose intended?
C. What information is not available and does this matter?
D. Is further information required at the same spatial scale as existing data, at a smaller scale (fewer points), or a larger scale (more points)?
E. Is information required simply as values of a property, or in ways that describe soil functions; if so, which function is of greatest interest?
F. What is the target, e.g., all soils in an area, the common soils, rare soils, a specific ecosystem, a specific soil process?
G. What magnitude of change is of interest, over what time scale, and is this a realistic target?
H. Is the property measurable or is further research needed?
I. Is it necessary to collect more information, or can the need be met from surrogate measurements, pedotransfer functions, extrapolation and interpolation?

Organizational and practical questions arise:

1. How can soil monitoring yield the greatest amount of information for the greatest number of people for the least cost?
2. At what time interval is the monitoring to take place?
3. Is this interval suitable for all soil properties of interest?
4. Can the needs of the users be satisfied with one soil-monitoring design, i.e., number and layout of sites; if not, what modifications are needed?
5. Will one organization do everything, or is it to be a consortium (if so, who leads?)
6. Does the lead organization have the ability, infrastructure and will to keep that role for a sufficient period? It is no use setting up a 10-year program if those involved drop out after 5 years;
7. Are there agreed protocols for the design of the monitoring network, for the location, layout, and description of sites, for the soil sampling, for the preparation and storage of samples, for laboratory work, for quality control; if not, who will orchestrate this?
8. Who will keep the samples, the raw data, the processed data, and be responsible for them into the future?
9. Who 'owns' the data, i.e., are there issues of copyright, intellectual property rights, confidentiality, and controls on the release of data?
10. Is there an agreed publication policy for the output from the soil monitoring, given that many years might pass before sufficient data are available to allow robust statements to be made about change?
11. What will the various options cost, bearing in mind that the costs of very long term site maintenance, quality control, and storage can be substantial.

Many of these points might seem trivial. However, the time scale over which SM must run before large amounts of meaningful data are acquired has to be seen in decades. Thus the objectives, responsibilities, targets, and costs need to be thought through with great care and appropriate commitments obtained.

Soil-Monitoring Networks

Many countries have conventional soil survey maps showing the spatial distribution of soil types. Although the scale of the map controls the smallest area of any soil type that can be shown, even small-scale maps, e.g., 1:250 000 scale, can indicate the relative abundance of soil types. At a very early stage, it has to be decided whether the network is to be capable of sampling rare soils, i.e., those occupying only a small proportion of the landscape. As a guide, a network that is expected to 'capture' a soil occupying 5% of the landscape with 95% confidence would need several thousand sampling locations (Table 1). (*See* **Spatial Variation, Soil Properties**.) An example is the National Soil Inventory (NSI) for England and Wales (UK; approx. 150 000 km^2), which sampled almost 5700 points, but only captured about 400 of the 700 soil series that are known to occur in the two countries. If no soil maps exist, then the problem becomes even more pressing.

Numerous designs have been proposed for national or regional soil-monitoring networks. The NSI network is based on sampling at the intersects of a 5×5-km-square grid. The Forest Soil Monitoring Network in Europe is based on a regular 16×16-km grid, and there are many other examples of the use of square grids. The most efficient regular geometric grid is triangular, but the square grid has the advantage that it can be co-located easily with a national geographic grid, thus making it easy to find points on the ground using conventional maps. This proviso may become less important with the increasing use of global positioning systems (GPS) (but these currently do not work equally well everywhere). One caution often expressed about square grids is that the periodicity of the grid might match that of some periodicity of the soil itself. However, whilst this might be true at the field scale, e.g., rows of crops in orchards and olive groves, 'tramlines' in arable crops, etc., no evidence has yet been presented that such periodicity exists for very large areas of land. On a practical note, if a square grid is used, it is best to locate its origin such that the sampling points do not fall on the edges or corners of map sheets; a 1-km offset is usually adequate. Alternative networks can be designed based on a simple random selection of points, stratified random selections, and a stratified, systematic unaligned selection of points (*See* **Spatial Variation, Soil Properties**). An alternative approach is weighted or purposive sampling, i.e., the sampling points are located on the basis of, for example, an existing network of sites such as experimental farms, sites for other forms of monitoring, selected soil–landscape combinations or, possibly, perceived soil functions. The principal problems with this approach are:

- It might not yield enough points to allow the spatial structure of the data to be estimated reliably by geostatistical methods – too much of the landscape is omitted;
- There is a high risk of introducing bias into the design.

However, such schemes can yield reliable baseline data for particular soil properties, e.g., lead content, so long as the limitations are appreciated. Well-conducted monitoring networks of this kind are being run, for example, in Switzerland, Austria, and the Netherlands among others (see also the Quality Control section, below).

Table 1 Size of sample required (n) to estimate property (P) within limits $P \pm 0.1P$ for three levels of confidence, calculated from the normal approximation

		n		
P *or 1*−P	*Confidence limits*	*80%*	*90%*	*95%*
0.5	±0.05	164	269	384
0.4	±0.04	245	331	576
0.3	±0.03	382	627	896
0.2	±0.02	655	1075	1536
0.1	±0.01	1474	2420	3456
0.05	±0.005	3112	5109	7296

Adapted from Webster R and Oliver MA (1990) *Statistical Methods in Soil and Land Resource Survey*. New York: Oxford University Press, with permission.

Protocols

Whatever is done and by whom, it is absolutely essential to have protocols in place for all aspects of the work: site selection, accuracy of location, sampling strategy at each site, sampling, sample treatment, analytical procedures, data storage and handling, and data access. Uncontrolled sources of potential variability are the greatest enemy of long-term monitoring. Peoples' memories are not reliable over long timescales, nor can it be relied on that those people will be available some years later to answer any questions. Information must be written down, multiple copies kept, photographs taken, and proper recording and training undertaken. Above all, it is vital that information is shared between investigators.

Monitoring Sites

Methods for establishing all of these types of networks are described in detail in the literature.

Whatever is chosen, it is extremely important to avoid bias in the location of sites, i.e., the person on the ground should not be allowed to pick and choose where to locate a site and how to sample. It is also strongly recommended that the network should have sufficient points to be robust, i.e., to buffer against loss or major change at some of the sites. Whilst it is obvious that the larger the number of sites (n), the more one is likely to capture the range of values of a variable, it is often forgotten that n should be sufficiently large to allow adequate resampling to establish a given level of change. For example, in England and Wales, the 0- to 15-cm layer of *c*. 2500 arable/ley grassland sites (this counts as a single land use in the UK) was sampled in 1980, and these samples had a mean organic carbon content of *c*. 3.2%. In 1995, these sites were resampled in sufficient numbers to be able to detect a change of 0.2% (absolute) in this mean at 95% (this level of change represents just under 4 t organic C ha^{-1}). This was achieved by sampling *c*. 900 of the original 2500 sites at random (Table 2). It is worth using existing data, or reasonable synthetic data, however limited, to test out some of these sampling strategies at different degrees of detection of change as part of the exercise in determining how many sampling points might be required.

Table 2 Size of sample required to detect different levels of change in mean organic carbon content in arable/ley grassland sites (at 95% confidence level)

Change in % mean organic carbon	*Sample size*
1.0	48
0.5	180
0.2	820
0.1	1668

Adapted from National Soil Inventory of England and Wales.

Table 3 shows the effect on the data collected for zinc during the NSI program if smaller numbers of sites are visited at regular intervals across the original square grid or by selecting similar numbers of points randomly from that grid. One obvious result of taking fewer samples is that some of the extreme (high or low) values are missed, so the observed range becomes smaller. Hence the spatial variability of values in the region may appear less than it is.

Although it might be possible to control activities at a small number of key sites, it is uncommon to be able to tie land owners or managers to particular land-use agreements for decades. Thus, sites can be built upon, become dumps for waste, become polluted and thus unacceptably dangerous to work on, or the land use might change so drastically that the site is no longer fit for monitoring purposes. The latter is always difficult to assess. Some would argue that if a soil is converted from arable to woodland monitoring the progress of that change is a matter of considerable interest. Others might argue differently because they want to know how cultivated land behaves in the long term. Whatever is decided, the situation whereby the loss of a relatively small number of sites compromises the whole network needs to be avoided. Again, to use the example from England and Wales (UK), approximately 3% of the *c*. 900 sites under arable/ley grassland in 1980 had been lost to agriculture by 1995. This change was interesting in itself, but the 'loss' of these sites did not seriously alter the conclusions to be drawn from the data for the remaining sites (approximately 870).

The simplest site is a sampling point, a spot on the ground located to a given accuracy. Points need to be located geographically according to a protocol that fits the purpose of the network. Experience has shown that people with some background of field work, using good topographic maps

Table 3 The effect of reducing the sample size on parameters for total zinc (mg kg^{-1})

No. of samples	*Mean*	*Median*	*Minimum*	*Maximum*	*Lower quartile*	*Upper quartile*	*Quartile range*	*SD*
On regular 5 × 5-km grid								
5676	96.78	82.0	5.0	3648.0	59.0	108.0	49.0	108.85
1434	94.95	83.0	7.0	1985.0	59.0	108.0	49.0	89.40
363	92.66	81.0	8.0	765.0	57.0	108.0	51.0	69.85
94	95.56	83.5	14.0	434.0	61.0	110.0	49.0	65.10
15	74.94	72.0	28.0	126.0	52.0	95.0	43.0	26.05
Selected randomly								
5676	96.78	82.0	5.0	3648.0	59.0	108.0	49.0	108.85
1438	92.78	81.0	6.0	2017.0	58.0	107.0	49.0	82.50
367	90.30	81.0	10.0	753.0	56.0	103.0	47.0	66.39
97	95.26	85.0	10.0	753.0	54.0	110.0	56.0	85.26
17	105.76	87.0	37.0	270.0	67.0	118.0	51.0	62.77

Adapted from National Soil Inventory of England and Wales.

(1:10 000 or 1:25 000 scale) and/or aerial photographs, can easily relocate a point to within 10 m of a target point. With GPS, relocation to within 2–3 m is possible, although nearby buildings, trees, and hilly ground can seriously distort signals. It can be helpful to mark these points to aid future relocation. A vandal-proof approach is the burial of either a metal plate (a sheet of aluminum alloy approximately 25 cm square and 5 mm thick) or of a small but powerful permanent magnet (wrapped in a plastic bag) at approximately 50 cm depth, i.e., well below plow depth. These can easily be found again after many years by appropriate hand-held, electromagnetic detectors.

There are many alternative arrangements of kinds of sites. Two of the most common are:

- Area-based sites, in which it is decided that it is insufficient for the purposes of the monitoring that a given area of land shall be sampled in a consistent manner. This approach is quite common in what might be termed 'agricultural monitoring,' where the argument is that fields or other parcels of land are subject to such intense management that the whole parcel becomes the site;
- Key sites or control sites that are specifically managed so as to be available for revisiting in the long term. An area of relatively uniform soil is identified (often of a hectare or more) and sampling plots are laid out across it in sufficient number to allow resampling according to a strict protocol that avoids resampling of previous plots for many years. These kinds of sites can be extended to allow investigation of large-scale processes such as soil erosion or the effects of afforestation, etc., but it is axiomatic that large sites, with their (partial) dedication to the long-term study of specific soil processes are expensive to run. In most countries, therefore, that have followed this thinking, the number of such sites is small (10–15 is not uncommon).

There is a growing belief that a small number of such key sites should be a component of any soil-monitoring network as they can form a framework by which the quality control of a much larger network can be achieved.

Soil Sampling

There are two key questions to be asked when considering soil sampling:

- Which soil layers are to be sampled, e.g., soil horizons (the people doing the sampling need to be able to recognize these) or fixed-depth sampling, and are litter layers to be included?
- Is the sample to be taken at a point, e.g., from the face of a soil pit, or is it to be a bulked sample taken using the geographic point as a centroid?

In practical terms, often a combination of these approaches is used. Surface layers are readily sampled around a centroid, e.g., 25 subsamples at 4-m intervals on a 20 × 20-m grid, a 'random' walk across a parcel of stated extent, subsampling along the radii of a circle, or a numbered layout of sufficient sampling plots to allow a considerable number of resamplings without the risk of revisiting disturbed soil. There is much to be said for this 'bulked' approach as at least a partial answer to dealing with the short-range variability of soils. Others argue for a formal, geostatistical approach to each site, so that its variability is known and an appropriate sampling scheme drawn up. This can, however, have large cost implications if it is to be applied to hundreds of sites. Whatever is done should not compromise future sampling – the temptation to remove the whole of the topsoil layer at the centroid should be resisted; similarly, soil pits should be small.

Finally, the sample must be big enough, because there are usually several people who want some of it. At least 2 kg on an air-dried, less than 2 mm basis is a useful target. This has implications for the volume of organic (peat) soils sampled. Sample storage needs considerable prior discussion with all who might have an interest, e.g., the chemists might be happy with dried material, but what do the biologists want? Long-term 'fresh' sample storage implies access to specialized facilities – Do they exist? Can they handle the number of samples (this year, next year, every *X* years, etc.)? Storage of frozen samples can pose similar problems. Air-dried material can be stored in screw-cap glass bottles with Teflon lid liners (but these are not cheap). The amount of shelf space required for the next 10, 20, 50, etc. years needs to be estimated, as well as the cumulative weight of the material. An intelligible sample-identification system is required so that samples can be retrieved, and a mechanism is necessary that ensures no one arbitrarily moves the samples or throws them away.

If the physical properties of soils are of interest, e.g., bulk density, water-retention characteristics, and hydraulic conductivity, then other sampling techniques are required (e.g., undisturbed cores or samples of known volume) and are best discussed with the appropriate specialists. It might be necessary to designate a separate area of the site for this kind of work, as considerable disturbance can be involved. Above all, it is necessary to allow for sufficient sampling space to accommodate all likely measurements without compromising any of the others.

Quality Control

This is not just a laboratory matter, but applies equally well to field operations. Whichever is under consideration, protocols, testing, and training are paramount; slipshod procedures at any stage can undermine the whole enterprise.

In the field, there is much to be said for having a formal recording system – the site location, local land characteristics, aspect, slope, etc. being recorded in standard ways, and (permitted) deviations from these should be recorded also. For example, it is common to find that the specific geographic point is not accessible (e.g., because it is covered by a road or building). Rather than abandon the site, limited, structured deviation is permitted, e.g., move 100 m north, then east, then south, etc., the details of the deviation being recorded. Such information can be recorded on proformas on portable computers or on printed proformas (synthetic papers are now widely available that will withstand any kind of weather).

Long-term quality control requires that a proportion of the samples are analyzed in duplicate or triplicate during the current program and during every subsequent resampling. This is to detect long-term drift. The proportion is a matter for discussion, but, as the monitoring program continues, the amount of 'repeat' analysis becomes the largest component of the laboratory work, with implications for costs. If samples are to be sent to distant laboratories, a system has to be established for this. For example: Are field refrigerators or 'cool boxes' required? How quickly should the samples be sent (overnight for biological monitoring)? Does the laboratory know what to do when the samples arrive? Protocols have to be set up for all these matters and tested before the monitoring commences, e.g., can the protocol deal with samples arriving at the start of a public holiday?

Ideally, all the determinations should be made in one laboratory but, even where this is not possible, thought should be given to dividing the work into blocks, e.g., all the metals are determined in one laboratory, all the organics in another, all the biology at a third. Are the chosen laboratories up to standard and are they familiar with soils? Do they have formal QA procedures in place and are they accredited and audited? Do they use widely recognized standard methods (International (ISO) standards should be used wherever possible)? Are they viable in the long term? Cheapest is not necessarily best.

Similar approaches are necessary with regard to output from the soil-monitoring program. The data need to be held securely in a system that has a good prospect of long-term stability in terms of software and staffing. Data ownership, intellectual property, and related issues need to be clarified at an early stage, as does a policy over publication of results.

Summary

Soil-monitoring programs need to be fit for purpose. At the very start, a clear answer to the question 'Why are we doing this?' is needed. Considerable planning is required to achieve success. A large number of issues need to be considered and decided at an early stage and written into robust protocols. Above all, soil monitoring has considerable spatial, temporal, and cost implications, and there has to be a commitment at the decadal scale.

See also: **Acid Rain and Soil Acidification**; **Aggregation:** Physical Aspects; **Carbon Cycle in Soils:** Dynamics and Management; **pH**; **Quality of Soil**; **Spatial Variation, Soil Properties**

Further Reading

Acton DF and Gregorich LJ (eds) (1995) *The Health of Our Soils: Towards Sustainable Agriculture in Canada*. Ottawa, Canada: Agriculture and Agri-food.

Bak J, Jensen J, Larsen MM, Pritzl G, and Scott-Fordsmand J (1997) A heavy metal monitoring programme in Denmark. *The Science of the Total Environment* 207: 179–186.

Blum WEH, Brandstetter A, Riedler C, and Wenzel WW (1996) *Bodendauerbeobachtung: Empfehlung für eine einheitliche Vorgangsweise in Österreich*. Wien, Austria: Bundesministerium für Umwelt, Jugend und Familie.

Council of Europe (1992) *Soil Protection*. Recommendation to the Council of Ministers R(92)8, 18 May 1992. Strasbourg, France: Council of Europe Publications.

Davidson DA (2000) Soil quality assessment: recent advances and controversies. *Progress in Environmental Science* 2: 342–350.

Dumanski J (ed.) (2000) Indicators of land quality and sustainable land management. *Agriculture, Ecosystems and Environment* 81 (Special Issue).

Eriksson J, Andersson A, and Andersson R (1997) *Tillståndet i svensk åkermark (Current Status of Swedish Arable Soils)*. Rapport 4778. Stockholm, Sweden: Naturvårdsverket. (In Swedish with extensive English Summary.)

European Commission (2002) *Soil Protection for Sustainable Development: The Soil Protection Communication*. Environment Directorate General. Brussels, Belgium: European Commission Publications.

FAO (1994) *Harmonisation of Soil Conservation Monitoring Systems*. Proceedings of an International Workshop, 14–17 September 1993. Research Institute for Soil Science and Agrochemistry. Budapest, Hungary. Geneva, Switzerland: FAO.

Federal Republic of Germany (1998) *Bundesbodenschutzgesetz*. (Federal Soil Protection Act), 17 March 1998. Bonn, Germany: (Available in English translation.)

Huber S, Freudenschuss A, and Stärk U (eds) (2001) *European Soil Monitoring and Assessment Framework*. Technical Report No. 67. Copenhagen, Denmark: European Environment Agency.

Lorenz M, Becher G, Fischer R, and Seidling (2000) *Forest Condition in Europe – 2000: Technical Report*. Geneva, Switzerland: UNECE.

Loveland PJ (1990) The National Soil Inventory: survey design and sampling strategies. In: Lieth H and Markert B (eds) *Element Concentration Cadasters in Ecosystems*, pp. 73–80. Weinheim, Germany: VCH Verlagsgesellschaft.

McGrath SP and Loveland PJ (1992) *The Soil Geochemical Atlas of England and Wales*. Glasgow, Scotland: Blackie Academic.

Office for Economic Co-operation and Development (1998) *Agriculture and the Environment: Issues and Policies*. Paris, France: OECD Publications.

Schipper LA and Sparling GP (2000) Performance of soil condition indicators across taxonomic groups and land. *Soil Science Society of America Journal* 64: 300–311.

Schulin R, Desaules A, Webster R, and von Steiger B (eds) (1993) *Soil Monitoring: Early Detection of Soil Contamination and Degradation*. Basel, Switzerland: Birkhäuser-Verlag.

Sykes JM and Lane AJM (eds) (1996) *The United Kingdom Environmental Change Network: Protocols for Standard Measurements at Terrestrial Sites*. London, UK: The Stationery Office.

Wagner G, Quevauviller P, Desaules A, Muntau H, and Theocharopoulos S (2001) Comparative evaluation of European methods for sampling and sample preparation of soils. *The Science of the Total Environment* 264(1–2): 3–204.

Webster R and Oliver MA (1990) *Statistical Methods in Soil and Land Resource Survey*. New York: Oxford University Press.

ENZYMES IN SOILS

R P Dick, Ohio State University, Columbus, OH, USA
E Kandeler, University of Hohenheim, Stuttgart, Germany

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Biological and biochemically mediated processes in soils are fundamental to terrestrial ecosystem function. Ultimately, members of all trophic levels in ecosystems are dependent on the soil as a source of nutrients and energy, and for degradation and cycling of complex organic compounds. Soil enzymes are central to these processes by catalyzing innumerable reactions in soils that have global biogeochemical significance.

Enzymes are proteins that act as catalysts by accelerating rates of reaction without undergoing permanent change. Enzymes are specific activators because they combine with their substrates in stereospecific fashion that decreases the stability of certain susceptible bonds (i.e., changes electronic configuration), which reduces the energy of activation of reactions (the amount of energy required for a reaction to proceed).

In soils, enzymes can exist intracellularly (inside the cytoplasmic membrane), which is of course important in cellular life processes. In addition, enzymes can exist outside the cytoplasmic membrane, in the periplasmic space or cell surface, and as extracellullar enzymes in soil solution or stabilized in the soil matrix. The latter two categories are known as the abiontic form. Both intracellular and extracellular enzymes may be involved in biogeochemical processes, as discussed below, but it is difficult to quantify the contribution of each group of enzymes in performing a given reaction.

It is generally assumed that soil enzymes are largely of microbial origin, but it is also possible that animals and plants may contribute enzymes to soils. However, studies on the fate of root enzymes have shown that they are rapidly degraded and/or denatured, suggesting they do not contribute much to the overall enzyme activity of soils. It is difficult to conclusively discriminate between sources of enzymes in soils, and thus evidence for the primary role of microbes as a source of soil enzymes is from indirect evidence such as sterilization of soils using radiation, which kills the organisms but leaves the enzymes catalytic.

Soil fauna are probably a limited source of enzyme activity in soils but earthworm (*Lumbricus terrestris*) casts stimulate enzyme activities on localized basis. There is increased enzymatic activity in rhizosphere soil compared with nonrhizosphere soil for many enzymes because roots can excrete extracellular enzymes and there is increased microbial activity due to root exudates and sloughing of cellular debris. This elevated activity in the rhizosphere is probably due to roots stimulating microbial activity than from enzymes exuded by roots.

The activity of approximately 100 enzymes has been identified in soils. Undoubtedly the number in soils is far greater, but techniques for determining the presence or activity of other enzymes have not yet been developed for soils. The soil enzymes most often studied are oxidoreductases, transferases, and hydrolases. The oxidoreductase dehydrogenase has been widely studied in soils partly because of its apparent role in the oxidation of organic matter, where it transfers hydrogen from substrates to acceptors. Catalase activity is based on the rate of release of oxygen from added hydrogen peroxide or the amount of recovered hydrogen peroxide. Some hydrolases and transferases have been extensively studied because of their role in decomposition of various organic compounds and thus are important in nutrient cycling and formation of soil organic matter. These include enzymes involved in: the C cycle, i.e., amylase, cellulase, xylanase, glucosidase, and invertase; the N cycle, i.e., protease, amidase, urease, and deaminase; the P cycle, i.e., phosphatase; and the S cycle, i.e., arylsulfatase. Lyase activity has been found in soils but relatively few studies have been conducted on this group of enzymes.

The functional role of abiontic soil enzymes has yet to be conclusively shown experimentally. First there has been limited success in extracting enzymes from soils. This is because the enzymes complexed in the soil matrix often lose their integrity during the extraction process. As a result soil enzymes have largely been studied by measuring activities, which makes separation of abiontic enzymatic activities indistinguishable from activities associated with the living organisms.

Nonetheless, it has been hypothesized that abiotic enzymes may have important relationships with soil organisms. For some microorganisms, substrates or their breakdown products may be useful for the organism but, due to substrate size or insolubility, unavailable for direct microbial uptake. It may be advantageous for a microbial cell to be located on the surface of a humic colloid containing a number of enzyme molecules. Indeed, for some species their successful survival in a hostile soil environment may depend on an association with humus–enzyme complexes.

Spatial Distribution of Enzymes in Soils

Extracellular Enzyme Stabilization in Soil Matrix

Enzymes have been categorized according to their location in the soil. Three enzyme categories (termed 'biotic enzymes') are associated with viable proliferating cells located: (1) intracellularly in cell cytoplasm; (2) in the periplasmic space; and (3) at the outer cell surfaces. The remaining categories are broadly characterized as abiontic, from the Greek words '*a*,' meaning 'removal or absence of a quality,' and '*bios*,' 'having a form of life.' Abiontic enzymes are those exclusive of live cells that include enzymes: excreted by living cells during cell growth and division; attached to cell debris and dead cells; and leaked into soil solution from extant cells or lysed cells, but whose original functional location was on or within the cell. Additionally, abiontic enzymes can exist as stabilized enzymes in two locations: adsorbed to internal or external clay surfaces; and complexed with humic colloids through adsorption, entrapment, or copolymerization during humic matter genesis.

Microscale Distribution

The functionality and activity of enzymes in soils are closely controlled by their location with enzyme–substrate interactions occurring at pico- and nanoscales. It is difficult to study extracellular enzymes directly, as the recovery of an enzyme is generally less than 20% of the total present in soil. In addition, humic–enzyme complexes may be modified during the extraction procedure. Microscale investigations have concentrated on different size aggregates after physical separation or by studying microhabitats of the high-turnover, particulate organic matter (mainly undecomposed plant material or microbial debris).

Generally, soil organic matter chemistry ranges from minimally decomposed and course plant debris with a high C/N ratio to highly processed humic substances as organomineral complexes with a narrow C/N ratio associated with clay fractions. These spatial and physical size distribution patterns of C sources control the location of microbial community members and associated enzymes. This variability can be related to microhabitat hot spots of microbial activity on organic material associated with the rhizospheres (roots), drilospheres, and detritus (decaying organic matter). Characterization of rhizosphere soil with a root mat–soil interface technique has shown that the abundance of microorganisms and their activities decreases to levels similar to bulk soil within a distance of several millimeters from root surfaces. This spatial relationship is closely related to the levels in the rhizosphere of easily degradable root exudates, and mass flow and diffusion distance of dissolved organic substrates used by soil microorganisms. The soil–litter interface distance of influence is similar to root surfaces with enzyme activities, being greater up to 1.3 mm from the litter interface.

The spatial distribution of various microbial groups or species, besides substrate availability, is related to protection from predation from larger organisms. Commonly, much of the soil microbial biomass and enzyme activities are associated with the smaller-sized fractions (fine silt and clay). Studies have found that the clay fraction is dominated by bacterial biomass, with invertase, phosphatase, and protease being greater in the fine-textured fraction, but fungi and xylanase activity dominating in the coarse sand fraction that contains particulate organic matter, suggesting xylanase is of fungal origin. Enzyme activities of the coarse-sized particle fractions are strongly influenced by tillage practices; the enzyme activities of these particle fractions are therefore a valuable indicator for organic matter input and management changes.

Macroscale Distribution

Enzyme activity declines with soil depth and generally is correlated with organic C and microbial biomass distribution in the soil profile. This has been shown in many studies and is to be expected, as most of the biological activity and organic matter is in the surface soil horizons.

Plot-scale investigations have determined that organic matter turnover rates, microbial biomass, and enzyme activities of soil samples vary with vegetation and soil type. Relatively little is known about the topographic, pedogenic, soil mineralogy, and other properties that control microbial and enzyme distribution at landscape levels. The characterization of these interactions is essential to achieve a better understanding of complex ecosystem processes. This latter scale is of particular importance for developing a soil-quality indicator, because enzyme activities can vary more as a function of soil type than the differences caused by soil management.

Fuzzy operations (which use logic inference procedures by allowing individuals, or soil properties in this case, to be assigned into continuous class membership values instead of exact hard classes in order to make interpretations based on several or many soil properties) and multivariate analysis (factor analysis of many soil properties simultaneously; in this case to see which ones are most important in explaining soil variation in space or between management systems) have shown that land use is the strongest factor and soil contamination the weakest factor governing the level of soil enzyme activities at the ecosystem level in Central Europe. Soil type is an important site factor, as it summarizes climatic, topographic, and geologic conditions, acidification, and vegetation influence on soil biology and enzyme activity.

Methods of Studying Enzyme Activities in Soils

The discussions above indicate that, for most soil enzymes studied, enzymes exist as extracellular or abiontic enzymes that can retain their activities outside living cells and in viable soil organisms. The proportion of total activity coming from extracellular activity for a given enzyme is generally unknown and probably varies from enzyme to enzyme. This offers two ways to measure enzymes in soils; direct extraction of pure enzyme protein or measurement of enzyme activity.

Direct extraction and purification of enzymes in soils has been largely unsuccessful. This is evident when total enzyme activity of the soil is used to estimate how much enzyme protein is in the soil, and the amount extracted is always much lower than the estimate based on activity. The protein cannot be separated easily from mineral or organic colloids or it is denatured during the extraction process. This difficulty with enzyme extraction from soils means that soil enzymes are studied by measuring enzyme activity. A large number of assays have been developed to measure enzyme activities. For the assays to be effective, it is imperative that the reaction can be stopped completely after a specified incubation period and that the product is quantitatively extracted from the soil.

Because measuring activity is the primary means of studying enzymes in soils, there are a number of implications and limitations for understanding the spatial distribution, biochemistry, microbial ecology, and other factors relative to soil enzymes. The first consideration is the assay itself. Enzyme assays are done at substrate concentrations exceeding enzyme saturation on a known amount of soil and under a strict set of conditions that include temperature, buffer pH, and ionic strength. Thus, the results are operationally defined and any change in these conditions will change the measured activity. This means that enzyme assays measure the potential activity under optimal conditions and not the *in situ* activity, because the assay provides an environment quite different from the original soil.

A second implication of soil enzyme assays is that enzymatic activity of the abiontic or extracellular enzymes cannot be separated from that of living cells. Enzyme assays, particularly those with long incubations, generally include antiseptics such as toluene to inhibit growth and metabolism during the assay (shorter assays, 1–2 h, may not need growth inhibitors because the time is too short for significant microbial growth). As a result enzyme activity is not necessarily related to microbial activity (although

many assays can be highly correlated with microbial biomass or respiration, but this relationship for the same enzyme can vary widely due to soil type, soil management, chemical and physical manipulations of soils, and climatic factors).

A majority of enzymes assays developed for soils are hydrolases, because there are a wide number of hydrolases in soils that are important in soil functions. The activity of most hydrolases is investigated using artificial substrates that do not occur naturally in soil. One of the most common types are simple esters that combine the functional group of the substrate, e.g., phosphate (for phosphatase) or glucose (for β-glucosidase) with a chromophore, such as *p*-nitrophenol (PNP), which is easily extractable from soils (except very high organic matter soils where there may be some adsorption of PNP). After a specified incubation period, a $CaCl_2$-alkaline solution is added, which stops the reaction, enables quantitative extraction of PNP, and causes PNP to turn yellow for colorimetric determination of the PNP product.

Other hydrolytic enzymes that work on major polymeric substrates include cellulases, chitinases, and lipases. Important enzymes in the N cycle are those that hydrolyze proteins, and peptides, and release NH_4 from amino acids. These enzymes are largely extracellular, because the substrates have a large molecular weight and/or are insoluble, making it difficult or impossible for their direct uptake across microbial membranes.

Oxidoreductase enzymes that have been detected in soils are dehydrogenase (which is actually a number of different intracellular enzymes that can perform the same reaction), glucose oxidase, catalase (peroxidase), and polyphenol oxidases. These latter two enzymes are important in lignin degradation. Dehydrogenase activity has been of particular interest because it is a component of the electron transport system of oxygen metabolism and requires intracellular environment of viable cells to express its activity. Therefore, it is not likely to exist in an extracellular form and thus should be a good indicator of physiologically active microorganisms, unlike nearly all other enzymes. However, it has turned out to be a poor indicator of microbial activity. This may be due to unsuitable assay conditions or the presence of extracellular phenol oxidases in the soil that cause the same reaction, or there can be common soil constituents that act as electron acceptors (e.g., nitrate or humic acids), all of which cause an overestimation of dehydrogenase activity (which is one reason why this assay may not correlate with microbial activity). Furthermore, elevated levels of Cu can interfere with the assay procedure. Other oxidoreductases of note are the polyphenol oxidases because of their role in humification of soil organic matter.

Other assays that have been developed include lysases (e.g., glutamate decarboxylase, which hydrolyzes apartic acid, and L-histidine ammonia and tyrosine decarboxylase, which are both involved in N mineralization) and transferases (e.g., dextransucrase, which hydrolyzes sucrose, releasing glucose and fructose and thiosulfate S-transferase, which oxidizes elemental S). Fluorescein diacetate hydrolysis is a broad-spectrum hydrolytic enzyme assay; this reaction can be carried out by proteases, lipases, and esterases.

Many of the enzyme assays utilize spectrophotometric analysis of reaction products. These methods are advantageous because they are, for the most part, rapid and accurate, and they have low equipment costs. Titrimetric methods are commonly used for N-cycling enzymes, where the hydrolysis product from organic N compounds is NH_4. Other methods include fluorimetric techniques, which can be very sensitive, radioisotopic detection of products (rarely used), manometric methods, typically used for oxidases and decarboxylation assays, ion-specific electrodes, and various chromatographic methods (e.g., high-performance liquid and gas chromatography), available for measuring a wide array of enzyme products.

Enzyme Kinetics

Enzyme reaction kinetics can be described by various equations. Enzyme-catalyzed reaction velocity is generally described by a rectangular hyperbola where enzyme concentration is held constant and substrate concentration is varied over a wide range. At low substrate concentrations, this reaction is a first-order reaction, which shifts to a zero-order reaction at high concentrations. The classic mathematical derivation for the rate equation was first developed by Michaelis and Menten in 1913 as follows:

$$S + E \underset{k_2}{\overset{k}{\rightleftharpoons}} ES \overset{k_3}{\rightleftharpoons} E + P$$

when S is substrate, E is enzyme, ES is enzyme–substrate complex and P is product. The reaction rate (v) at any moment is equal to k_3 and when substrate concentration reaches infinity the rate of reaction approaches a maximum (V_{max}). For this to hold there are a number of assumptions that include: the substrate concentration remains nearly constant during the time of measurement; and if substrate concentration is much greater than enzyme concentration, only a negligible amount of substrate can accumulate in the ES intermediate complex. The K_m value is the Michaelis–Menten constant where substrate concentration is at one-half of the maximum velocity (moles per liter). The Michaelis–Menten

constants (V_{max} and K_m) are specific properties of individual enzymes.

Traditionally the Michaelis–Menten constants were estimated by doing various linear transformations. The constants can then be determined by the slopes and intercepts based on these transformations of the data. With the advent of widely available, high-speed computers, it is now preferable to use nonlinear fitting of model equations to estimate K_m and V_{max}, because these have the least-biased fit of the data.

The Michaelis–Menten constants are used to characterize a particular enzyme. The V_{max} can be used to determine the concentration of the enzyme if the molecular weight is known; however, because it is difficult to extract enzymes from soils, the molecular weight is not normally known. Therefore, V_{max} can be considered as an index of the amount of enzyme in the soil but its unit of measurement is activity.

The K_m constant is an index of the affinity of the enzyme for the substrate (the smaller the K_m the greater the affinity). In soils this is referred to as the 'apparent' K_m, because there are a range of biological, chemical, and physical factors that can affect enzyme affinity. First, in all likelihood a soil has many isoenzymes that can act on a particular substrate, because most enzymes come from a variety of soil organisms or even from plant roots. Secondly, some of the activity for most enzymes originates from a component stabilized on mineral surfaces or complexed in soil organic matter, and there are cofactors in soils or other environmental conditions that can affect the reactivity of the enzyme. These latter physical and chemical factors may change the conformation of the enzyme or some other enzymatic property, which will affect affinity.

Michaelis–Menten constants can be used to provide insights into how soil management affects enzymatic properties. Long-term cultivation of soils causes the V_{max} to decrease and K_m to increase, the latter indicating that there was less affinity of the enzyme for the substrate. Differences in K_m values between these systems suggest that there is a different suite of isoenzymes and/or differences in abiontic forms of the enzymes in soils.

Soil-Quality Indicators and Technologies

Sensitivity to Soil Management and Ecosystem Stress

Soil is the final arbitrator of the impact of human activity on the ecosystem's health and function. Consequently, there has been interest in developing indicators of soil quality to reflect the impacts of pollution, and agricultural and forestry activity on the ability of soils to perform important biogeochemical ecosystem processes.

The rationale for soil enzyme activity as a soil-quality indicator is that enzyme activities:

1. Are often closely related to important soil-quality parameters such as organic matter, soil physical properties, and microbial activity or biomass;
2. Can begin to change much sooner (1–2 years) than other properties (e.g., soil organic C), thus providing an early indication of the trajectory of soil quality with changes in soil management;
3. Can be an integrative soil biological index of past soil management;
4. Utilize procedures that are relatively simple compared with other important soil-quality properties (e.g., physical and some biological measurements) and therefore have the potential to be done routinely by soil-testing or environmental-analysis laboratories.

A conceptual model for enzyme assays as integrative biological indexes is based on the observation that activity originates both from viable cells and abiontic components. The abiontic component provides an indicator of semipermanent changes in soil quality, because such enzymes are probably complexed and protected in the soil humic- or clay-complexes and change over a period of years. The interpretation of this characteristic from a soil quality perspective is that soil management that promotes stabilization of organic matter and associated structural properties (e.g., aggregation and porosity) would also promote stabilization of enzymes in the soil matrix.

Field-scale studies have shown that enzymes are differenentially affected on a temporal basis due to soil-management activities. Table 1 summarizes activities of enzymes involved in carbon-, nitrogen-, phosphorus- and sulfur-cycling. These studies have shown that soil enzyme activities vary seasonally and over the long term, which has improved the understanding of the functioning and dynamics of soil. In many cases, enzyme activities are early predictors of the effects of soil management on soil quality and how rapidly these changes are expected to occur.

Numerous studies have shown, in side-by-side comparisons, that enzyme activities are sensitive to soil-management effects such as crop rotation, tillage versus no-tillage, and forest management. A few studies have shown that enzyme activities can be temporally sensitive within the first 2 years of changes in soil management, long before there are measurable differences in soil organic matter. Figure 1 shows such an early effect, with a distinct separation after 3 years between soils managed with or without winter

cover crops without measurable changes in soil organic matter. However, not all enzymes can detect soil-management effects, and activities naturally vary as a function of soil type.

A few enzymes have potential as indicators of viable soil microbial biomass or activity. Most notable is dehydrogenase, but for the problems mentioned in the previous section it has not worked well as an indicator of microbial activity. Furthermore, enzymes that correlate closely with microbial activity may be less suited to predict long-term changes or trajectory in soil quality because they reflect recent management or seasonal (climatic) effects that may be transitory.

The effect of air-drying on enzyme activity varies with the enzyme and can cause an increase in activity, but for most enzymes activity is reduced by 40–60%. Screening of a range of enzymes has shown that, although activity may decrease with air-drying, the relative change (or rank according to field management) in activity for some enzymes within the same

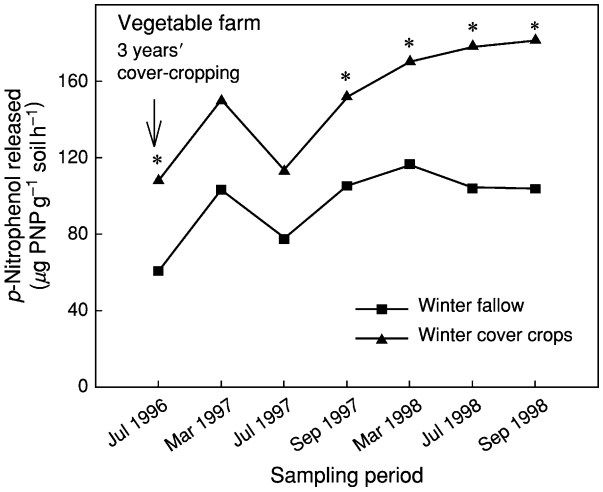


Figure 1 β-Glucosidase activity after 3 years of winter-fallowing or cover-cropping following summer vegetable crops in western Oregon, USA. (Adapted from Ndiaye EL, Sandeno JM, McGrath D, and Dick RP (2000) Integrative biological indicators for detecting change in soil quality. *American Journal of Alternative Agriculture* 15: 26–36. CABI Publishing, with permission.)

Table 1 The response of enzyme activities to the type of vegetation and soil

Soil enzyme activity	*Range of activities*	*Vegetation/soil type*
Xylanase activity (mg glucose g^{-1} per 24 h)	13–24	Spruce forest/n.d.
	0.28–8.0	Beech forest/n.d.
	1.8–3.0	Grassland/orthic luvisol
	0.24–1.83	Agricultural land/haplic luvisol, entisol
β-Glucosidase (μg *p*-nitrophenol $g^{-1} h^{-1}$)	20–55	Grassland/pachic arguistoll
	36–160	Forest/Haplohumult
	130–310	Crop rotation/hapludalf
	71–86	Crop rotation/pachic ultic argixerolls
	41–253	Crop, manured soil, pasture/typic haploxeroll
Protease activity (μg tyrosine g^{-1} per 2 h)	150–520	Agricultural land/haplic chernozem
	224–514	Pasture/typic dystrochrept
	120–430	Wheat seeds/loamy sand
	198–288	Crop rotation/haplic luvisol
Arginine deaminase activity (μg N $g^{-1} h^{-1}$)	2.5–5.0	Grassland/pachic arguistoll
	1.7–2.0	Crop rotation/phaeozem, lithosol, cambisol
	4.0–11.0	Forest/sandy soils
	0.1–1.3	Crop rotation/fluventic ustochrept
Arylsulfatase activity (μg *p*-nitrophenol $g^{-1} h^{-1}$)	30–50	Grassland/pachic arguistoll
	115–340	Agricultural land/hapludoll
	6.9–213	Pasture/typic dystrochrept
	21–49	Forest/podzol
	12–58	Crop, manured soil, pasture/typic haploxeroll
Alkaline phosphatase (μg *p*-nitrophenol $g^{-1} h^{-1}$)	40–80	Grassland/pachic arguistoll
	40–790	Agricultural land/aeric vertic epiaqualfs
	100–500	Crop rotation/hapludalf
	181–225	Crop rotation/ustochrept
Dehydrogenase (μg TPF g^{-1} 24 h)	114–155	Crop rotation/haplumbrepts, hapludalfs
	0.6–0.9	Crop rotation/fluvisol
	68–97	Crop rotation/ustochrept
	148–207	Crop rotation/fluventic xerochrept

Adapted from Kandeler E, Tscherko D, Stemmer M, Schwarz S, and Gerzabek MH (2001) Organic matter and soil microorganisms – investigations from the micro- to the macro-scale. *Die Bodenkultur* 52: 117–131.
n.k., soil type not known.

soil type remains the same. Air-drying stabilizes enzyme activity, greatly facilitates sample handling, and allows for timely analysis (unlike most other microbial measures that must be done as soon as possible after sampling). Combining this with the relative simplicity of many enzyme assays makes it possible to run a large number of samples on a routine basis and enables adoption by commercial soil-testing laboratories.

In general, hydrolytic enzymes are good choices as soil-quality indexes because it is likely that organic residue-decomposing organisms are the major contributors to soil enzyme activity. However, there is evidence that some of these enzymes can be confounded by long-term applications of fertilizers or liming. Phosphatase activity can be depressed by phosphate fertilizers and is also affected by pH, which is independent of other factors of soil quality such as organic matter content. Also, there is limited evidence that N fertilizers can have a similar effect on certain enzymes involved in the N cycle (e.g., urease and amidase).

Enzymes involved in the C cycle are thought to be better choices as a soil quality indicator than enzymes involved in cycling of nutrients that are heavily fertilized in agricultural systems. It seems likely C-cycling enzymes are more closely related to organic matter inputs, soil organic matter, and disturbance, all of which are related to soil quality. One such enzyme that that has been sensitive to soil management in many studies in β-glucosidase, which releases C for energy. Also, arylsulfatase has been an effective discriminator of soil-management effects and has been correlated with fungal biomass.

Enzyme assays have been used to evaluate the degree of highly degraded soils and polluted soil. Highly degraded soils can result from excessive erosion in agricultural or forest soils or from strip-mining activities. Enzyme expression has been closely related to the degree of recovery and corresponding microbial life and plant productivity in highly disturbed soils in the first years after remediation is initiated. Carbon hydrolytic enzymes appear to be well correlated with plant productivity in early stages of soil recovery, which is probably related to decomposition and organic matter accumulation being closely coupled with soil quality.

Enzyme assays on soils polluted with heavy or crude oil fractions have shown that relatively high rates of oil must be applied before enzyme activity decreases, whereas lighter petroleum products do not generally inhibit enzyme activity. Consequently, enzyme activity does not appear to be an appropriate technology for characterizing hydrocarbon-polluted soils. Pesticides have had no effect, stimulatory, or only short-term effects on soil enzyme activities even at pesticides rates far in excess of recommended rates for pathogen or weed control. However, in the case of heavy metal-contaminated soils, soil enzyme assays have potential for practical applications to assess bioavailability of metals. However, each metal often may require a specific enzyme assay to reflect the bioavailability of metals to soil microorganisms or plant availability. These assays are advantageous over total or even extractable metal content, because they better reflect toxicity to biological organisms, which for metals varies widely as a function of soil characteristics such as organic matter content/chemistry, textural distribution, and mineralogy.

It is important to recognize that soil enzyme activity is operationally defined. If the conditions of the assay (e.g., temperature, buffer pH, buffer type, ionic strength) are altered, results will also change. Therefore, to make meaningful comparisons among studies or over different time periods, it is important that the exact same protocol be followed for each enzyme assay.

Systematic studies across soil types, environments, and soil-management systems are still needed to fully determine the potential of soil enzyme activity to characterize soil quality and develop calibration data to interpret enzyme activities. Therefore, enzyme activities should be interpreted with caution and be measured along with other soil properties to assess soil quality.

Ecological Dose Value

The ecological dose value (ED_{50}) is analogous to LD_{50} (lethal dose at 50% kill rate) used for assessing the toxicity of substances on animal and human life. Applying this to enzyme activity means that an ED_{50} value would be the pollutant (inhibitor) concentration required to cause a 50% inhibition of enzyme activity in soil. Two general models have been used: the sigmoidal dose–response curve; and the Michaelis–Menten kinetic model.

Thus far, this approach has only been used to assess heavy metal pollution of soils; but presumably this could be used on a relative basis to quantify the impact of other soil pollutants or of highly disturbed landscapes (e.g., strip mines) on soil enzyme activities.

Detoxification of Polluted Soils

Enzyme technologies are emerging as a means to remediate polluted soils. Common organic contaminants include pesticides, volatile hydrocarbons from industrial and automobile-related compounds (e.g., benzene, toluene, trichloroethylene), polycyclic aromatic hydrocarbons from fossil fuel wastes,

polychlorinated biphenyls (PCBs) from electrical insulation, and chloroderivatives, chlorophenols, and chlorobenzene from paper mill effluents. A range of metals also commonly contaminates soils from industrial activity and mismanagement of irrigated agricultural land.

The goal of remediation is to transform pollutants into innocuous products. There are a number of potential ways that enzyme remediation processes can be invoked through microbial stimulation or introduction of a specific organism(s) capable of degrading a particular pollutant. Another approach is to add an enzyme or a suite of enzymes appropriate to degrade and detoxify a specific pollutant or to add substrates to soil that upon enzymatic hydrolysis release products that remediate soils.

Enzymes added to soils in a soluble form have been less effective, because the enzyme in this state is more susceptible to degradation by microorganisms, adsorption on mineral surfaces, and complexation in organic matter, and is not likely to be reusable. Consequently, the preferred method has been to stabilize the enzyme on a solid support by various chemical and physical immobilization techniques such as entrapment, encapsulation, covalent bonding, adsorption, and cross-linking or cocross-linking with bifunctional agents. Although immobilization typically reduces activity, the net effect is greatly increased potential for detoxification of soils by increasing stability and reusability of the enzyme. Supports have included clays (surface adsorption) and entrapment in gels, porous glass, or silica beads.

Summary

Soil enzymes are central to ecosystem processes because they catalyze innumerable reactions in soils that have biogeochemical significance. Catalytic soil enzymes can exist internally or on surface membranes of viable cells, be excreted into soil solution, or be complexed in the soil matrix or microbial debris. Extracellular enzymes may play an ecologic role for some microbial community members by hydrolyzing substrates that are too large or insoluble for direct absorption by microbial cells. More than 100 enzymes have been characterized in soils. With the exception of dehydrogenase and possibly a few other enzymes, which only exist in viable cells, nearly all other enzymes exist in both viable and complexed forms, independent of viable cells, stabilized in the soil matrix.

Research on soil enzymes provides insights into biogeochemical cycling of C and other nutrients and on microbial community functions in space and time. A relatively small amount of any given enzyme can be directly extracted from soil; therefore enzymes are mainly studied by measuring activity. The activity of enzymes varies temporally (seasonal), which often corresponds to microbial community responses to the environment, vertically (decreasing from the surface), at microscales, according to microbial community distribution, and at landscape level, where soil type is a major controlling factor (particularly textural distribution and organic matter).

Soil enzyme assays are emerging as technological tools for various applications in environmental and ecosystems management. Several enzymes have shown sensitivity in reflecting early changes (1–3 years) in soil quality due to soil management long before there are measurable changes in total organic C levels. This holds potential to guide ecosystem management for long-term sustainability. Enzyme assays can detect the level of degradation and recovery of soils in highly disturbed landscapes such as reclaimed strip-mine landscapes. The bioavailability of certain heavy metals in soils can be reflected with enzyme-activity measurements. Enzymes stabilized on colloid surfaces and incorporated into soils have been shown to degrade certain contaminants in soils.

There is still considerably more information needed to understand the ecology and function of extracellular enzymes in soils because of the diversity and complexity of the soil physical and chemical environment and microbial communities. Partly because of this, utilization of enzyme technologies requires careful consideration for interpretation and application. This is particularly true for assessing soil quality, where soil enzyme activities should be used in conjunction with other key soil measurements. Further research is needed to develop mechanisms for calibrating and interpreting soil enzyme technologies that are independent of soil type.

See also: **Pollutants:** Persistent Organic (POPs)

Further Reading

Burns RG and Dick RP (eds) (2002) *Enzymes in the Environment: Activity, Ecology, and Applications*. New York: Marcel Dekker, Inc.

Dick RP, Breakwill D, and Turco R (1996) Soil enzyme activities and biodiversity measurements as integrating biological indicators. Doran JW, Jones AJ *et al.* (eds) Handbook of Methods for Assessment of Soil Quality, pp. 247–272. SSSA Special Publication 49. Madison, WI: Soil Science Society of America, Inc.

Kandeler E, Tscherko D, Stemmer M, Schwarz S, and Gerzabek MH (2001) Organic matter and soil microorganisms – investigations from the micro- to the macroscale. *Die Bodenkultur* 52: 117–131.

Kiss I, Dragan-Bularda M, and Radulescu D (1975) Biological significance of enzymes in soil. *Advances in Agronomy* 27: 25–87.

Ladd JN (1978) Origin and range of enzymes in soil. In: Burns RG (ed.) *Soil Enzymes*, pp. 51–96. London, UK: Academic Press.

Ndiaye EL, Sandeno JM, McGrath D, and Dick RP (2000) Integrative biological indicators for detecting change in soil quality. *American Journal of Alternative Agriculture* 15: 26–36.

Skujinš J (1978) History of abiontic soil enzyme research. In: Burns RG (ed.) *Soil Enzymes*, pp. 1–49. London, UK: Academic Press.

Tabatabai MA (1994) Soil enzymes. In: Weaver RW, Angle S, Bottomley P *et al.* (eds) *Methods of Soil Analysis*, part 2, *Microbiological and Biochemical Properties*, No. 5. Madison, WI: Soil Science Society of America, Inc.

EROSION

Contents

Irrigation-Induced
Water-Induced
Wind-Induced

Irrigation-Induced

G A Lehrsch, D L Bjorneberg, and R E Sojka,
USDA Agricultural Research Service, Kimberly, ID, USA

Published by Elsevier Ltd.

Introduction

Soil erosion is caused by wind, tillage, precipitation, or irrigation. Erosion caused by irrigation, usually termed 'irrigation-induced erosion,' can be the most damaging because it affects many of the most productive soils in the world. These are the soils of arid irrigated regions, which typically have thin A horizons, little organic matter, and weak structure, making them highly erodible. Moreover, these soils, once degraded, recover very slowly. Irrigation-induced erosion occurs as an unintended consequence of irrigation for improved crop production.

To produce food and fiber worldwide, irrigation is vital. Irrigation enables crops to be produced in many areas where they could not otherwise be grown. In other drought-prone areas, irrigation on average doubles crop yield and nearly triples crop value, while improving production reliability and commodity quality. According to the Food and Agriculture Organization (FAO) of the United Nations, irrigation is practiced on only approximately 5% of the world's food-producing land, which includes rangeland and permanent cropland. That irrigated land, however, produces approximately 30% of the world's food. Similarly in the USA, only 15% of harvested cropland is irrigated, yet that land produces 40% of the nation's total crop value.

Three basic types of irrigation are drip, surface, and sprinkler. Drip irrigation supplies water to growing plants at very small rates, wetting relatively small soil volumes either at or below the soil surface. Properly designed and operated drip systems produce neither erosion nor runoff. In contrast, surface (or gravity-flow) irrigation requires water flow across the soil surface and is often designed to produce runoff to improve irrigation uniformity. With overland flow, however, comes erosion. In surface irrigation, the soil surface is the conduit used to deliver and distribute water. Surface irrigation that occurs (1) on sloping areas includes graded furrows (small ditches parallel to crop rows) and border strips, and (2) on relatively flat areas includes level or contour basins, terraces, and wild flooding. Sprinkler irrigation practices, too, can produce both runoff and erosion if not designed and managed properly. In sprinkler irrigation, water droplets are distributed through the air to the soil. Sprinkler irrigation includes: (1) moving lateral systems, including center-pivot, lateral-move, and big-gun systems; and (2) stationary systems, including solid-set and side-roll systems.

Irrigation-induced erosion from sprinkler irrigation resembles that from rainfall in many ways. In both cases, water droplet impact can deteriorate surface soil structure by fracturing soil aggregates, thereby producing aggregate fragments, primary particles, or both that can obstruct surface pores leading to surface sealing and increased runoff. Water that does not infiltrate into the profile accumulates on the surface and, once surface depression storage is

satisfied, runs off, often transporting detached soil downslope or off-site. Water droplet impact not only detaches soil but also increases turbulence in shallow flow, increasing the amount of sediment the flow can transport.

There are, however, notable differences between erosion from rainfall and from sprinkler irrigation. For sprinkler irrigation: (1) only a portion of the field receives water at any given time, (2) water droplet characteristics vary from system to system, and (3) irrigation is controlled and managed to apply water only when the growing crop needs more soil water or in preparation for planting, tillage, or harvest. An area's rainfall is usually very low in total dissolved solids (TDS) and its chemical composition changes little. In contrast, irrigation water contains TDS and can vary chemically as a function of water source.

The differences between erosion from surface irrigation and from rainfall are even more distinct. The key difference is surface irrigation's lack of water droplet kinetic energy, which affects surface soil structure and thus infiltration, runoff, and erosion. Also absent is the additional turbulence in overland and rill flow caused by droplet impact. In furrow irrigation, water is applied to only a small portion of the soil surface. Erosion from surface irrigation most often occurs during a number of small events rather than one or two large events, characteristic of erosion from precipitation. Water temperature, affecting water viscosity, is more likely to change during a 12- or 24-h irrigation under cloudless skies than during a rainstorm. The hydraulics of rill flow from rain also differ from those from surface irrigation. In rainfall rills, flow volume increases as water accumulates downslope. In furrow irrigation, the flow rate and volume decrease with distance down the furrow but increase with time as the soil's infiltration rate decreases. These processes gradually change the furrow stream's sediment detachment and transport capacities with both time and distance from the furrow inlet. As the irrigation proceeds, upper furrow ends often become deeper and narrower owing to detachment and transport from relatively large inflows, while the lower furrow reaches become shallower and wider owing to deposition from reduced flow. The duration of inflow, often 12 h or more, is much longer than the runoff from most rainfall events.

Sediment concentration in runoff tends to decrease with time during a furrow irrigation, but not necessarily during a rainstorm. In a furrow during irrigation, many factors change, which, in combination, may explain this phenomenon. Loose soil, frequently positioned in the furrow by recent tillage or cultivation, is often flushed from the furrow early in the irrigation. At the furrow head, coarser, more erosion-resistant fragments may armor the furrow bottom. As soil in the furrow becomes wetter, there is less tendency for the rapid aggregate disintegration that is common during the initial wetting of hot, dry soil. In the lower furrow reaches, deposition can cause the furrow to widen, thereby decreasing its flow depth and reducing shear.

The chemical composition of irrigation water affects irrigation-induced erosion, whether from sprinkler or surface irrigation. High sodium concentrations or sodium adsorption ratios (SAR) and low electrical conductivity (EC) in irrigation water allow the diffuse double layers of 2:1 clay domains to thicken, dispersing clays and weakening or fracturing aggregates. Primary particles, released from aggregates as clay disperses, and aggregate subunits obstruct surface pores, increasing both runoff and soil loss. In addition, small aggregates or fragments, rather than large ones, are more easily transported in overland flow, once they are detached. Moreover, irrigation-water chemistry can change markedly with water sources and sometimes through the irrigation season, as water sources change or as upstream return flow is mixed in changing proportions with surface water.

Significance

Furrow irrigation is an inherently erosive process. It is exacerbated by the need for long fields to increase farming efficiency and for clean tillage to ensure uniform and steady flow of water down the furrow. Soil erosion from irrigation occurs across entire fields as a consequence of overland flow and, from sprinkler irrigation, droplet impact. Soil or sediment loss, in contrast, is a measure of the sediment entrained in runoff that leaves a furrow or field at its outlet. Measured soil loss is often much less than the field total of eroded soil, predominantly from upper furrow reaches, because much sediment is redistributed and, as flow rates decrease, often deposited on to lower furrow reaches before it can leave the field in runoff. Annual soil loss from surface-irrigated fields can vary from less than 1 Mg ha^{-1} to more than 100 Mg ha^{-1}, depending on crop type, field slope, soil properties, and water management, particularly flow rate. A single 24-h furrow irrigation of erodible soil on slopes of more than 2% has caused more than 50 Mg ha^{-1} of soil loss in runoff. Little erosion occurs from level fields, surface-irrigated pastures, or fields producing forage. In contrast, much erosion occurs from row crops grown on fields with steeper slopes, generally those exceeding 2%.

The magnitude of sprinkler irrigation-induced erosion is not well documented for at least two

reasons. First, it is difficult to measure, particularly so because it varies widely across time and space. Second, it tends to be an on-field problem occurring only in the area being irrigated at that time. Although sprinkler irrigation is normally regarded as a less-erosive alternative to surface irrigation, problems sometimes occur, particularly where systems are improperly designed or poorly operated. Farmers may irrigate excessively steep slopes with sprinklers, creating erosion problems because they have exceeded their irrigation system's design limits. Where center pivots with high-volume end guns are placed on rolling topography, the combination of high application rates, variable sloping land, and tower-wheel tracks can produce severe erosion in a single irrigation or in one season.

Erosion, whether occurring from sprinkler or surface irrigation, is caused by humans. Consequently, with an understanding of the processes involved, properly designed irrigation systems, and enlightened, skillfull management, irrigation-induced erosion can be nearly eliminated in many cases or at least adequately controlled.

Erosion Under Sprinkler Irrigation

Processes Causing Soil Loss

Soil erosion from water, whether caused by precipitation or irrigation, can be described in terms of three components or processes: detachment, transport, and deposition. Detachment is the release of soil aggregates, aggregate fragments, or primary particles from the soil surface as a consequence of energy input, usually from droplet impact or shear from runoff flow. Transport occurs as detached soil, that is, sediment or bedload, is splashed about and carried downslope in overland flow. Deposition occurs as sediment settles out of the flow as the water's carrying capacity for sediment is exceeded. Depending upon flow hydraulics, deposition may occur within a few meters of the detachment point or may not occur until the sediment is transported off-site.

When properly designed and carefully operated, stationary sprinkler systems, especially solid-set systems with a grid of simultaneously operating sprinklers, apply water for lengthy periods at a relatively low rate (e.g., $3\,\mathrm{mm\,h^{-1}}$). The soil's infiltration rate is seldom exceeded, so little (if any) runoff or erosion occurs. In contrast, center-pivot systems, with a moving lateral that pivots around a fixed point, apply water at higher rates (e.g., $80\,\mathrm{mm\,h^{-1}}$) to smaller areas (e.g., 5 to 20-m-wide strips) than solid-set systems. With center-pivot irrigation, the irrigated area per unit length of lateral must increase with distance from the pivot point. Consequently, the outer spans of pivots have relatively high discharge rates per unit lateral length (e.g., $15\,\mathrm{l\,min^{-1}\,m^{-1}}$) and high instantaneous application rates per unit wetted area. This greatly increases the potential to exceed a soil's infiltration rate, causing runoff and erosion.

Soil erosion from sprinkler irrigation is directly proportional to the application rate in the wetted area which, in turn, is affected by sprinkler type. Low-pressure-type spray heads, which are relatively economical to operate and thus have become popular, have reduced pattern widths and increased application rates relative to other sprinkler types. Again, high application rates can lead to erosion, runoff, and soil loss.

Water-drop impact, or more specifically droplet kinetic energy, detaches surface soil particles and splashes the detached soil in all directions. Some of the soil entrained in the infiltrating water obstructs surface pores. Droplet energy also compacts surface soil. The increased bulk density and obstructed pores reduce infiltration. Droplet kinetic energy also causes turbulence in shallow surface flow, increasing the flow's carrying capacity for sediment. An irrigation's total kinetic energy is a function of its droplet size distribution; the larger the droplet, the greater the kinetic energy. Droplet size distributions can be altered within limits by modifying nozzle pressure, nozzle size, and spray-head deflector plate. Sprinkler irrigation system designers must often balance desired design parameters with environmental and economic constraints.

Slope and topography also affect erosion processes from moving lateral sprinkler systems, particularly center-pivot systems. Depending upon the slope and the pivot's direction of travel, runoff can move on to dry soil, with relatively large infiltration rates, or previously wetted soil, with much smaller infiltration rates. In the first case, runoff rates decrease rapidly, fortunately because the dry soil is easily eroded. In the second, runoff accumulates and concentrates in rills or larger, ephemeral gullies, increasing in rate, erosivity, and sediment-carrying capacity. In the special case where the pivot lateral is parallel to the slope direction, the effective wetted slope length is long and erosion can be particularly severe. Under both center-pivot and lateral-move systems, the tower-wheel tracks are relatively large flow paths 40–50 m apart, with smeared and sealed surfaces underlain by compacted soil. Runoff is common in wheel tracks where they are parallel to the slope direction. Even where the tower-wheel tracks cross the slope, the tracks cause problems, because they collect and concentrate overland flow.

Practices Controlling Soil Loss

Irrigation practices New irrigation systems must be properly designed. Central to the design is an accurate estimate, preferably based upon measurements, of the soil's infiltration characteristics, particularly the infiltration decrease with time. Both new and existing systems must be operated in accordance with (1) design parameters such as nozzle diameter and pressure, and (2) operational guidelines such as set times and travel speed.

To control irrigation-induced erosion, one must minimize runoff. Without runoff, there will be no sediment transport apart from splash at the point of detachment. To minimize runoff, irrigators should schedule irrigations using scientific techniques and apply no more water than is needed for maximum economic yield. From an erosion-control standpoint, no runoff should be the goal.

Modifying the sprinkler type, nozzle pressure, and nozzle diameter alters both the application rate and wetted area. Changes that decrease the sprinkler flow rate, decrease the application rate, or increase the wetted area minimize erosion, runoff, and soil loss. For spray heads, changing the nozzle and deflector plate changes the drop size distribution. Shifting the distribution to smaller and fewer large droplets reduces total droplet kinetic energy striking the soil, thus reducing detachment. Disadvantages of such a size distribution change are that smaller droplets evaporate more readily and are more susceptible to wind drift, which distorts the spray pattern, decreasing both irrigation uniformity and efficiency. Another disadvantage of small droplets is that they travel relatively short distances, giving the spray head a small wetted diameter and high application rate.

A goal of irrigation system design and operation is to match the system's application rate to the soil's infiltration rate (to minimize runoff), both spatially and temporally. This goal is difficult to achieve, however. One relatively new technique with promise for moving lateral systems is to use variable-rate sprinklers that can be programmed to operate on a site-specific basis. Appropriately programmed, the sprinklers could change their discharge rate, depending upon field slope, soil-infiltration differences, presence of rock outcrops, or other factors.

Soil and crop management practices One effective way to help reduce erosion caused by early-season irrigations is to eliminate unneeded seedbed-preparing tillage. In the spring, surface soil aggregates of many soils are structurally weak and susceptible to breakdown from tillage or droplet impact. Unnecessary springtime tillage weakens or breaks particle-to-particle bonds within aggregates, often fracturing them. Aggregate fragments and primary particles are more easily transported than are larger, intact aggregates. Moreover, such tillage buries crop residue and indirectly destroys soil organic matter, further weakening aggregates.

Some tillage practices, on the other hand, instead of contributing to soil erosion can help control it. One such practice, paratilling, uses broad, angled subsoiling shanks to partially lift and laterally shatter soil, increasing the tilled soil's infiltration rate, often substantially, thereby decreasing runoff and soil loss. Another tillage practice that decreases runoff is reservoir tillage. In this postplant operation, small water-storage basins or pits are formed at intervals across a field's surface. Those basins increase surface-depression storage by collecting and temporarily holding water, allowing the water to infiltrate rather than run off. Reservoir tillage reduces runoff, even when an irrigation system's application rate somewhat exceeds the soil's infiltration rate. This practice is particularly effective where performed under the outer spans of center pivots, where application rates often exceed soil infiltration rates.

No-till and conservation tillage are other tillage practices that reduce irrigation-induced erosion. These practices leave crop residues on the soil surface as mulch. Surface mulch absorbs droplet kinetic energy, protects soil structure, and maintains surface roughness, thereby minimizing the decrease in the soil's infiltration rate with time. No-till or conservation tillage also keep soil surfaces rougher, increasing both depression storage and the tilled soil's initial infiltration rate. Within limits, crops in a rotation can be sequenced to produce crop residue regularly throughout a multiyear rotation. A canopy of growing vegetation also absorbs droplet energy, reducing energy input directly to surface soil. Production practices that hasten canopy coverage can reduce erosion from droplet impact, and may reduce erosion from overland flow by shading surface aggregates and keeping them moist and less susceptible to slaking. Vegetation on the soil surface also slows runoff and absorbs overland flow shear.

Another management practice that helps to control runoff, thus soil loss, on slightly sloping surfaces is to till or plant so that the final tillage or planting marks are perpendicular, rather than parallel, to the slope direction. On rolling topography, one should practice contour tillage, in which both tillage and planting operations are performed on the contour, as much as possible. These practices slow runoff, allowing more time for water to infiltrate into the soil.

Erosion with Surface Irrigation

Processes Causing Soil Loss

In surface irrigation, as water flows across a soil's surface or advances down a furrow, it quickly wets relatively dry aggregates or clods in its path. As a consequence of the small matric potential in the dry soil, water will quickly enter the aggregate from all directions, causing 2:1 clay domains to swell, displacing O_2 and N_2 from particle surfaces, and often compressing those gases and air within the aggregate. As the compressed air finally escapes, the force it exerts often fractures interparticle bonds within the aggregate, or the aggregate itself, liberating aggregate fragments and primary particles. This process, in which an air-dry aggregate breaks into subunits or fragments when quickly wetted or immersed in low-electrolyte water, is termed 'slaking.' It contributes substantial amounts of soil for transport in the furrow stream, accounting in large part for the relatively great sediment concentrations often observed early in an irrigation.

Water must flow across the soil during surface irrigation. This flowing water exerts shear along the wetted perimeter, detaching soil once the imposed shear exceeds a threshold, termed the 'critical shear stress.' In a furrow, this critical shear varies both spatially and temporally. In addition to detaching soil, the flowing water transports detached soil downslope, further contributing to the erosion process. Level-basin irrigation systems may have no runoff, thus no soil loss from the basin. Other surface systems on sloping fields, in contrast, have runoff. To ensure adequate wetting of the soil near their field or furrow outlet, those surface irrigation systems are designed and operated so that 20–40% of the added water runs off. Thus, without proper precautions and management, soil loss will occur from many surface-irrigated areas.

Competing processes affect the erosivity and hydraulics of the flowing irrigation water. Infiltration through the wetted perimeter reduces the furrow flow rate with distance from the furrow inlet. This decrease in flow rate with distance reduces the furrow stream's shear and carrying capacity, at times leading to sediment deposition. As time passes, however, the soil's infiltration rate decreases and, with no change in the inflow rate, the furrow flow rate increases. Increasing the flow rate increases the shear and carrying capacity. Also, as much of the slaked and easily eroded soil is flushed from the furrow early in the irrigation, the sediment concentration in the furrow stream often decreases. This decreasing sediment concentration with time (and with increasing flow rate) increases the furrow stream's transport capacity.

Practices Controlling Soil Loss

Irrigation practices One of the best ways to control erosion of surface-irrigated land is to convert to a well-designed sprinkler irrigation system, with its higher efficiency, better application uniformity, minimal runoff, and often reduced labor needs. Sprinkler irrigation does require, however, more energy, a larger capital investment, and a greater level of management than surface irrigation. Sprinkler irrigation can also encourage disease and may not meet peak crop water demand. Thus, such conversion is not possible or practical in every situation, and other practices must be used to control erosion under surface irrigation.

As mentioned above, one must minimize runoff to minimize soil loss from irrigated fields. With surface irrigation, this goal is more difficult to achieve, because runoff is usually necessary to assure adequate application uniformity. None the less, irrigation should still be performed to produce no more runoff than is needed. Scientific irrigation scheduling, good water control, and close monitoring of ongoing irrigations help to minimize both runoff and soil loss.

In some areas, irrigators may be able to shorten furrow lengths. This reduces erosion, because the inflow rate can be reduced yet still allow the furrow stream to advance to the outlet in a reasonable length of time, termed 'advance time,' usually 25–40% of the total set time. Reducing inflow rates is desirable because much detachment and transport occurs near furrow inlets, where furrow flow rates are highest. On some fields, furrow length can be halved by adding a midfield gated pipe to supply the needed inflow. Shortening furrow lengths, however, may increase runoff and soil loss from the entire field (because twice as many furrows are producing runoff) and always requires more labor. For example, adding a midfield pipe doubles the number of furrows that need to be set and the pipe itself must be moved when performing field operations. If field size is reduced to shorten furrow lengths, then more time will be required to plant, till, and harvest those smaller fields. In many areas, furrow lengths cannot be shortened due to existing return-flow channels.

In some situations, furrows may be oriented to cross the slope slightly, rather than run parallel to the slope direction. This repositioning reduces the furrow's slope, reducing the flowing water's shear on the soil along the wetted perimeter, thus reducing both sediment detachment and transport capacity. Repositioning furrows may lead, however, to increased erosion of the now-steeper tailwater collection ditch.

Another means of reducing erosion is to manage furrow inflow rates and advance times appropriately. Inflow rates must be large enough for the furrow

stream to reach the outlet, but, once runoff begins, the inflow rate can be reduced ('cut back') to minimize erosion near the furrow inlet as well as runoff at the furrow outlet. Monitoring is required, however, because if a furrow's inflow is reduced too much, its outflow may cease, greatly reducing the uniformity of water application in that furrow. Also, to minimize differences in intake opportunity time from furrow inlet to outlet, irrigators would like advance times to be relatively small. However, a tradeoff must be made, since smaller advance times require greater inflow rates, yet those greater rates increase erosion near furrow inlets.

Some producers use surge irrigation to improve application uniformity. Surge irrigation is a technique wherein flow is applied intermittently ('surged') during a single irrigation set to overcome initially high infiltration rates near furrow inlets. While surge irrigation helps infiltration to be more uniform from furrow inlet to outlet, it must be used carefully or erosion near furrow inlets can be greater with its intermittent inflow than with continuous inflow if inflow rates are higher when surging than when not.

Irrigation water quality can be changed to reduce soil loss. In some areas, one may be able to mix water sources or otherwise add electrolytes to alter inflow water chemistry, principally by increasing the water's Ca^{2+} concentration. Increasing the concentration of divalent cations in the irrigation water reduces the thickness of 2:1 clay domains' diffuse double layer. This minimizes clay dispersion and enables aggregates to remain intact, less susceptible to transport downslope in the furrow stream. Since the divalent cations stabilize soil structure along furrow-wetted perimeters, they also lessen infiltration decreases with time that make furrow-irrigation management difficult. Gypsum is commonly added to water with very low EC or high SAR to improve its suitability for irrigation.

Runoff management practices Runoff can also be managed to minimize, or at least control, soil loss under surface irrigation. One technique is to use pump-back runoff reuse systems, in which all runoff and sediment are collected in a reservoir at the field end, then pumped back to the inlet, where the runoff is reintroduced during the same irrigation as inflow to the field. While incurring energy and equipment costs for pumping, pump-back return systems offer many benefits. Reintroduced inflow that contains some sediment reduces furrow-stream sediment-carrying capacity. Depending upon flow hydraulics, sediment eroded from the field may be redeposited on to the field near its origin. Where irrigation return-flow water-quality regulations are stringent, irrigators with pump-back systems will have no off-farm (or off-site) discharge of sediment, fertilizer, pesticides, weed seeds, or microbes.

To collect or retain soil eroded from irrigated fields, settling basins varying in size and shape may be constructed along runoff collection channels, often at field ends. These basins collect much of the runoff and, under quiescent conditions, allow soil particles from the runoff to settle. Some basins are large (for collecting runoff from 20 ha or more); some are small (for runoff from only a few furrows). After draining the basins at the season's end, the collected sediment can be returned to the field. While offering this advantage, settling basins suffer from many disadvantages. Erosion still occurs in the field. Clay-sized soil, containing most of the P, other plant nutrients, and agricultural chemicals, does not fully settle out but is largely lost in the basin's outflow during the irrigation season. A settling basin's sediment collection efficiency declines as it fills with sediment, reducing residence time in the basin. Land area is taken out of production. Settling basins also require weed control, can be safety hazards, and can be the source of flying insect pests. Energy, time, and, for bigger basins, heavy equipment not common on farms are required to remove sediment from the basins and redistribute the sediment on to the field or another area. In spite of these disadvantages, settling basins have their place, particularly when used in combination with other erosion-control practices.

Buried drains with standpipes are a special type of settling basin. In a field's tail ditch, plastic, corrugated pipe is placed in a trench as a drain. Standpipes that extend vertically from the drain to just above the soil surface are installed every 5–10 furrows along the drain's length, then the trench is backfilled. Earthen dams are then constructed across the tail ditch, just downstream of each standpipe's inlet, thus forming a small basin at each standpipe. In operation, each dam forces runoff to pond, allowing some sediment to settle, before the runoff enters the standpipe's inlet and drains from the field. With appropriate management, this special drainage system can eliminate excessive erosion that often occurs at field ends where furrow slope increases sharply as runoff drains into a deep tail ditch. These drainage systems: (1) increase yields from field ends, (2) bring additional land into production, (3) ease farm equipment's ingress and egress across the lower field boundary, and (4) reduce weed problems common in and near wet tail ditches. Unfortunately, buried drains do not control erosion from the bulk of the field, and still allow some sediment to enter the drain and be transported from the field in the drainage water.

Soil and crop-management practices Placing mulch or maintaining crop residues in irrigation furrows effectively reduces both erosion and soil loss. Mulch in the furrow absorbs shear and slows furrow-stream velocity, thus reducing both sediment detachment and transport. By reducing flow velocity, the mulch can reduce overland flow by allowing the added water more time to infiltrate. If available, previous crop residues should be used as mulch, but in some rotations and areas, straw from off-site works well. While effective at controlling furrow erosion and often increasing crop yields, if not properly managed, the mulch tends to float downstream and obstruct the channel, damming the water which breaks over into adjacent furrows, increasing their flow while reducing the flow in the obstructed furrow. By increasing infiltration, mulch can increase erosion from upper furrow reaches if the mulched furrows require greater inflows. Mulch placed in level basins can hinder the even spreading of water, at times channeling it to erode some areas and underirrigate others. Instead of placing mulch in an irrigation furrow, one can establish semipermanent vegetation, such as turf, along the furrow's wetted perimeter, much like a grassed waterway. Turf, once established, nearly eliminates furrow erosion but complicates field management and can reduce crop yield. Turf-covered furrows are a viable practice only for rare cropping patterns and on very steep slopes.

Narrow or twin-row plantings also reduce erosion. By positioning crop rows on bed shoulders, close to an intervening irrigated furrow, the plant root systems stabilize soil along the furrow's wetted perimeter, while overhanging vegetation drooping into the furrow and plant debris reduces furrow-stream velocity, minimizing both detachment and transport.

Filter strips, often seeded to small grain or forage, can also be placed perpendicular to the furrow direction at the downstream end of row-crop fields to trap sediment that would otherwise leave the field in the furrow outflow. By slowing and spreading the flow as it progresses through the 3- to 6-m-wide strip, the furrow stream's carrying capacity is greatly decreased, with much sediment being deposited within the strip. Filter strips do not, however, prevent erosion from occurring upslope, nor do they produce much marketable yield from the crop seeded in the strip.

A recently developed, highly effective erosion-control practice is the adding of certain types of synthetic organic polymers to surface irrigation water. These polymers, high-molecular-weight, moderately anionic polyacrylamides (PAM), are added to inflow water to be present at dilute concentrations of approximately $10\,\mathrm{mg\,l^{-1}}$. When evenly distributed throughout the inflow early in an irrigation, PAM stabilizes soil along furrow-wetted perimeters and flocculates sediment that may be present in the flow. PAM-treated water also reduces seal formation in the furrow, thus slowing the decrease in the soil's infiltration rate with time. All told, their use reduces furrow soil loss by approximately 95%, economically (e.g., less than US\$$40\,\mathrm{ha^{-1}}$) and with minimal additional management. PAM also reduces erosion and increases infiltration under sprinkler irrigation, but its use there requires specialized equipment and is not yet user-friendly.

Effective furrow-erosion control is also possible using whey, a natural organic by-product of cheese manufacture, at times viewed as a food-processing waste. When added without running off to newly formed furrows early in an irrigation season, it too stabilizes soil along wetted perimeters, in part owing to greatly enhanced microbial activity that leads to aggregate formation and stabilization at and below the wetted perimeter. Soil loss from subsequent irrigations of whey-treated furrows is reduced by 50–98% and infiltration increased by 50–60%.

A combination of practices can be particularly effective. PAM and/or conservation tillage can be used to reduce on-field erosion, while filter strips and small settling basins remove additional sediment before the runoff leaves the field. Larger settling basins and wetlands in return-flow streams can further reduce the runoff's sediment load before the runoff reaches receiving waters.

Summary

Controlling erosion on and soil loss from irrigated lands is critical to sustain agricultural production. Protecting and stabilizing the soil surface will minimize sediment detachment; slowing or reducing overland flow will minimize sediment transport. Reducing or managing runoff is the key to controlling soil loss wherever sprinkler irrigation or surface irrigation is practiced. Erosion caused by sprinkler irrigation is similar to that caused by rainfall, with many erosion-control practices effective for both. Techniques that protect the soil surface from raindrop or sprinkler-drop impact are effective in maintaining infiltration rates, reducing overland flow, and controlling both detachment and transport. Erosion processes with surface irrigation are quite different from those with rainfall, due to the absence of droplet kinetic energy input to the soil surface, and thus require different control strategies. Controlling erosion from surface irrigation is a challenge, due to the requirement for overland flow and runoff, and to varying flow regimes and soil infiltration rates.

For both sprinkler and surface irrigation, off-site soil loss is often least where combinations of control practices are employed. For surface irrigation, PAM use is not only economical but probably offers the most promise for effective erosion control for most furrow-irrigated production systems.

Prospects for Future Control

In the USA, surface irrigation is practiced on about 50% of the irrigated land; worldwide, however, more than 95% is surface-irrigated. Wherever surface irrigation is practiced, improved irrigation scheduling and better water control can reduce erosion and soil loss while minimizing off-site environmental damage. In furrow-irrigated areas where labor is available and relatively inexpensive, changing management practices to reduce runoff by shortening furrow lengths, reorienting furrows to reduce furrow slopes, and/or managing inflows will help reduce on-field erosion and off-site soil loss. In more industrialized areas, with established surface water quality standards, pump-back return systems offer the most comprehensive control of both runoff and soil loss. Filter strips and buried drains with standpipes can minimize future off-site soil loss. Without doubt, though, the use of PAM in surface irrigation holds the greatest potential for cost-effective erosion control.

Effective sprinkler erosion-control techniques already exist and more are on the horizon. Variable-rate sprinklers on center pivots will probably prove cost-effective for site-specific soil and water management to increase yields, improve water-use efficiency, and decrease water requirements while simultaneously reducing runoff and attendant soil loss. Engineering hindrances to PAM use in center pivots will probably be overcome, enabling PAM's erosion-controlling and infiltration-enhancing benefits to be extended to sprinkler-irrigated lands also. PAM's other environmental benefits, such as minimizing off-site discharge of sediment, weed seeds, plant disease agents, and microbes (including possible human pathogens), will become more important with stricter environmental regulations, spurring ever greater PAM use under irrigated conditions.

See also: **Erosion:** Water-Induced; **Irrigation:** Environmental Effects; **Overland Flow**

Further Reading

Bjorneberg DL, Kincaid DC, Lentz RD, Sojka RE, and Trout TJ (2000) Unique aspects of modeling irrigation-induced soil erosion. *International Journal of Sediment Research* 15: 245–252.

Carter DL (1990) Soil erosion on irrigated lands. In: Stewart BA and Nielsen DR (eds) *Irrigation of Agricultural Crops*, pp. 1143–1171. Agronomy Monograph 30. Madison, WI: American Society of Agronomy.

Kincaid DC and Lehrsch GA (2001) The WEPP model for runoff and erosion prediction under center pivot irrigation. In: Ascough JC and Flanagan DC (eds) *Soil Erosion Research for the 21st Century*, pp. 115–118. St. Joseph, MI: American Society of Agricultural Engineers.

Lehrsch GA and Robbins CW (1996) Cheese whey effects on surface soil hydraulic properties. *Soil Use and Management* 12: 205–208.

Lehrsch GA, Sojka RE, and Westermann DT (2001) Furrow irrigation and N management strategies to protect water quality. *Communications in Soil Science and Plant Analysis* 32: 1029–1050.

Sojka RE (1998) Understanding and managing irrigation-induced erosion. In: Pierce FJ and Frye WW (eds) *Advances in Soil and Water Conservation*, pp. 21–37. Chelsea, MI: Ann Arbor Press.

Sojka RE and Bjorneberg DL (2002) Erosion, controlling irrigation-induced. In: Lal R (ed.) *Encyclopedia of Soil Science*, pp. 411–414. New York: Marcel Dekker.

Trout TJ, Sojka RE, and Okafor LI (1990) Soil management. In: Hoffman GJ, Howell TA, and Solomon KH (eds) *Management of Farm Irrigation Systems*, pp. 872–896. St. Joseph, MI: American Society of Agricultural Engineers.

Withers B and Vipond S (1980) *Irrigation: Design and Practice*, 2nd edn. Ithaca, NY: Cornell University Press.

Water-Induced

J E Gilley, USDA Agricultural Research Service, Lincoln, NE, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Water erosion is caused by the detachment and transport of soil by rainfall, runoff, melting snow or ice, and irrigation. Excessive erosion can threaten the production of agricultural and forest products. Erosion may also impact water conveyance and storage structures, and contribute to pollution from land surfaces. Water erosion may occur within rills, interrill areas (the regions between rills), gullies, ephemeral gullies, stream channels, forest areas, and construction sites. Rainfall characteristics, soil factors, topography, climate, and land use are important elements affecting soil erosion. Conservation measures that have been effectively used to reduce soil erosion on agricultural areas include contouring, strip cropping, conservation tillage, terraces, buffer strips, and use of polyacrylamide on irrigated areas. Specialized

erosion control practices have been developed for use within stream channels, forest areas, and construction sites. One of the most effective means of reducing erosion is to maintain a vegetative or residue cover on the soil surface.

Impacts of Erosion

Erosion is a natural process. Topographic features such as canyons, stream channels, and valleys are created by long-term geologic erosion. Geologic erosion influences soil formation and distribution. Accelerated erosion results from the removal of natural vegetation by human activities such as farming, ranching, forestry, and construction.

The production of agricultural and forest products can be affected by excessive erosion. Erosion causes a breakdown of soil aggregates and accelerates the removal of organic and mineral materials. The loss of surface soil is critical because the exposed subsoil remaining following excessive erosion usually has reduced infiltration capacity, water storage, and nutrient characteristics. The exposed subsoil is usually finer-textured, making seedbed preparation and crop production more difficult. Smaller-size particles are more easily detached and transported by overland flow. Thus, sandy soils are made even coarser by erosion.

Sedimentation resulting from erosion may significantly reduce the effectiveness of water-conveyance and storage structures. The capacity and functional life of lakes, reservoirs, and streams can decrease as a result of sedimentation. The suitability of streams and rivers as a biological habitat and effective water supply can be affected by excessive amounts of sediment. The existence of sediment can also impair the use of streams and rivers as fish-spawning areas.

Nutrients, pesticides, and pathogens transported by sediment can contribute to pollution of streams and lakes, thus reducing their suitability for aquatic organisms and their use as water supplies. A large nutrient concentration in streams and lakes can also cause excessive vegetative growth, resulting in seasonal oxygen deficiencies. The type of fertilizer that is used, application rate, and nutrient content of the soil influence nutrient transport.

Types of Erosion

Water erosion can be separated into individual categories, each with distinct characteristics. Rills are small channels that form as runoff rate increases. The regions between rills are defined as interrill areas. Gully erosion occurs when concentrated flow is large enough to form large channels that cannot be crossed during normal tillage operations. Ephemeral gullies appear at the same position on the landscape each year, but they are small enough to be filled in by tillage operations. Stream-channel erosion may take place within a water-course that usually has continuous flow. Each of the erosion types may occur on croplands, rangelands and pastures, forest areas, and construction sites.

Rill Erosion

As overland flow moves downslope, it concentrates due to surface microtopography. Small channels or rills may form as the runoff velocity of overland flow increases. Rills often occur between crop rows or along tillage marks. The hydraulic shear of flowing water and soil properties influence rill erosion. Normal tillage operations usually remove rills. The soil materials detached within rills and sediment delivered from interrill areas are transported by rill flow. Once rills have formed, substantial amounts of erosion may occur, resulting in a loss of soil productivity.

Interrill Erosion

Raindrops impacting the soil surface detach soil particles on interrill areas. The detached soil particles may then be transported to rills by shallow overland flow. Soil properties, rainfall intensity, and slope all influence interrill erosion. Interrill erosion is often most apparent on the light-colored upper portions of convex slopes where tillage mixes surface soil and subsoil.

Gully Erosion

Deep channels larger than rills that cannot be removed by tillage are classified as gullies. Gullies usually form near the upper portion of intermittent streams or where trails, paths, or roads cause runoff to concentrate. In tropical areas, gullies may develop following deforestation and cultivation. The runoff-generating characteristics of the watershed influence gully formation and development. Once they form, gullies become a permanent part of the landscape. Gullies that separate portions of fields or pastures are a nuisance.

A gully continues to develop and move upslope where there is a water overfall. As water moves through the channel during large runoff events, a gully may rapidly expand and deepen. Large runoff events may also remove sections of the exposed banks that were previously undercut and had fallen into the channel. Gullies are described as active if their walls are free of vegetation and as inactive when they are stabilized by vegetation.

Ephemeral Gully Erosion

The topography of many fields causes runoff to collect and concentrate in a few natural waterways before leaving the field. These channels or ephemeral gullies serve as the primary drainage system for a field and form in the same locations each year. However, ephemeral gullies are transitory rather than permanent since they are small enough to be filled in by tillage operations. Ephemeral gully erosion occurs when soil particles are detached and transported from the concentrated flow channel.

Stream Channel Erosion

The removal of soil from the banks or beds of streams results in stream channel erosion. Only relatively small areas are affected by stream channel erosion, but the impacts on highly productive soils can be severe. Stream channels erode as water flows over the side of the stream bank or scours below the water surface, especially during severe floods. Stream channel erosion is usually greatest on the outside of bends. Meandering may also cause erosion along stream banks. A reduction in sediment delivery from upland areas through control practices or the capture of sediment in water-storage facilities may increase stream channel erosion.

Factors Affecting Erosion

Erosion is influenced by a variety of elements including rainfall characteristics, soil factors, topography, climate, and land use. Erosion models can be used to estimate the interrelated effects of these factors on erosion and to predict the impact of various land uses on runoff and soil loss.

Rainfall Characteristics

For erosion to occur, runoff must first be present. Runoff is defined as that portion of rainfall that does not infiltrate nor accumulate on the soil surface but moves downslope. Rainfall rate and duration both influence runoff and erosion. Runoff occurs only when rainfall intensity exceeds soil infiltration rate. Infiltration will decrease with time during the initial stages of a storm. Thus, no runoff may occur from a storm of short duration, while a storm of the same intensity but of longer duration may result in substantial runoff.

Rainfall intensity influences both the rate and volume of runoff. During a high-intensity storm, infiltration capacity is exceeded by a greater margin than during a less intense rainfall event. Thus, even though the precipitation amount may be similar for two events, a high-intensity storm will produce a greater volume of runoff.

Soil Factors

The susceptibility of soil particles and aggregates to detachment is influenced by soil characteristics. Soil texture, organic matter content, structure, and permeability have been shown to influence soil erodibility. Because they lack cohesiveness, sand particles are relatively easy to detach but they are difficult to transport by overland flow because of their large mass. In contrast, clay particles are difficult to detach since they readily bond together but are easily transported once they are separated from the soil mass. Silty soils are usually well-aggregated, but the aggregates break down rapidly when wetted, allowing nonaggregated soil particles to be easily transported. Soils containing large stable aggregates are difficult to detach and transport and usually have greater infiltration rates.

As soil organic matter content increases, individual aggregates become more stable, soil structure improves, and infiltration rate becomes greater. Cultivated soils recently removed from native vegetation, pasture, or meadow usually have excellent structure and stable aggregates. The incorporation of organic materials into the soil profile helps to increase aggregate stability. Cultivation without the addition of organic materials causes a reduction in aggregate stability and organic matter content, and an increase in soil erodibility.

More runoff usually occurs from fine-textured than from sandy soils because of differences in infiltration. Maintaining high infiltration rates is one of the most effective means of reducing erosion. Soil surface sealing caused by aggregate destruction and plugging of pores with soil particles may substantially reduce infiltration. High infiltration rates can be maintained when vegetative material protects the soil surface from sealing, when soil structure is preserved, and soil compaction is minimized.

Topography

Erosion is influenced by slope gradient, length of slope, and size and shape of the watershed. The velocity of flowing water becomes greater as slope gradient increases. A larger runoff velocity allows flowing water to detach and transport additional soil material. An increased accumulation of overland flow on longer slopes results in greater rill erosion. Convex slopes with a larger gradient at the bottom of a hillslope are more erosive than concave land surfaces. On concave surfaces, deposition frequently occurs at the bottom of the slope because of reduced transport capacity of flow. Crops growing in productive areas at the bottom of a hillslope may be covered with water and sediment during extreme precipitation events.

Climate

The total amount and intensity of rainfall influence the quantity of erosion within a region. The dense vegetation found on areas that receive substantial rainfall reduces erosion potential. Significant erosion may occur in regions with low rainfall and limited vegetation when high-intensity rainstorms do occur. Maintaining high infiltration rates not only reduces runoff and erosion, but also helps to maintain soil water supplies needed by vegetation.

Runoff from melted snow and ice moving over a thin layer of freshly thawed soil can cause substantial erosion. In colder climates, more erosion may result from snowmelt than from rainfall events. Frozen soil is not subject to erosion during several months of the year. However, significant runoff may result from the rapid melting that occurs when air temperatures rapidly rise or rain falls upon a snow-covered surface.

Land Use

Land use is the only factor affecting erosion that can be modified to reduce soil loss potential. Erosion results from different land-use conditions on cropland areas, rangelands and pastures, and forest areas.

Cropland areas Soil erosion potential can substantially increase as natural forest or rangeland is converted to cropland. Cropland areas least susceptible to erosion have a complete ground cover throughout the year. The amount of surface cover maintained on a particular site is influenced by cropping and management conditions. The greatest erosion potential on croplands exists after planting when residue cover is usually at a minimum and high-intensity rains frequently occur.

Cultivated land left fallow with no vegetative cover is particularly vulnerable to runoff and erosion (Table 1). Areas with steep slopes on which row crops such as corn and cotton are grown continuously may also be of concern. The dense surface cover resulting when row crops are planted in rotation with grasses and legumes may substantially reduce erosion potential.

On areas that receive sufficient precipitation, interseeding row crops with a legume can be an effective conservation measure. The legume provides a protective cover during the critical planting period. Following the cropping season, the legume may also serve as a supplemental nitrogen source.

Irrigation is used on some agricultural areas. Both irrigation and natural precipitation events may result in erosion. Because of the increased quantities of water introduced through irrigation, the potential for runoff may be greater on irrigated areas.

Table 1 Mean annual runoff and erosion under different land uses with mean annual rainfall of 1400 mm

Land use	*Runoff (mm)*	*Soil loss ($mg\,ha^{-1}$)*
Rotation; coastal Bermuda grass and crimson clover after corn	70	5
Corn grown continuously	180	27
Cotton grown continuously	250	49
Fallow; cultivated with no vegetative cover	470	135

After Carreker JR, Wilkinson SR, Barnett AP, and Box JE (1978) *Soil and Water Systems for Sloping Lands*. USDA, ARS-S-160.

Rangelands and pastures In humid areas, the dense sod found in pastures reduces erosion. When adequate surface cover is maintained on rangelands and pastures, erosion is usually minimal. In regions with limited rainfall, severe erosion may still occur from the exposed areas between bunches of grasses during intense storms in regions with limited rainfall. Bare soils lack a protective surface cover and root mass and are subject to physical abrasion from livestock hooves. The reduction in vegetative cover caused by excessive grazing may increase the potential for both water and wind erosion.

As soils erode, they become less resistant to further degradation. Soil aggregation, water infiltration, and water retention are diminished as surface soil is eroded and subjected to trampling. The proportion of grazing-resistant species that have shorter and shallower roots often increases on eroded sites. Grazing management systems should be selected to provide a productive balance of plant species that maintain root mass.

The areas adjacent to stream banks often contain productive soils and have greater amounts of soil water because they receive runoff from upland regions. Because of the relatively large amounts of vegetation on these areas, they are prime sites for grazing. Preventing cattle from grazing next to stream banks has been found to result in a substantial reduction in stream-bank erosion and sediment delivery to streams.

Gravel and cobble materials are found throughout the soil profile on some rangeland areas. Since shallow overland flow cannot easily transport gravel and cobble materials, they often remain on the surface of eroded soils. Thus, a surface armoring process is established that reduces further soil loss from the eroded rangeland sites.

Forest areas The vegetative cover continuously maintained on the surface of undisturbed forests substantially reduces soil loss. On an undisturbed

forest soil, most erosion occurs from channel banks and adjacent steep slopes. Road construction, timber harvesting, or fires may substantially increase erosion rates. The potential for erosion may be severe on forest areas when a fire destroys plant material and the litter layer. A condition called hydrophobicity occurs when soils repel water due to the intense heat from a fire. Erosion potential is influenced by the intensity of the fire and the extent of the burned area within a watershed.

Construction sites Improper construction practices can accelerate soil erosion. On-site damage in urban areas may result from the loss of topsoil by heavy equipment during construction activities or by erosion. If disturbed sites are not properly protected, runoff and erosion can adversely affect the surrounding environment. Although the area that is disturbed may be relatively small, the amount of soil eroded from construction sites can be substantial. Preventing sediment from leaving a construction site can reduce serious off-site impacts.

Erosion Control Measures

A variety of measures are available for controlling erosion on agricultural areas including contouring, strip cropping, conservation tillage, terraces, buffer strips, grassed waterways, and the use of polyacrylamide on irrigated areas. Specialized control measures are also available for reducing erosion from forest areas, construction sites, and stream channels. Depending upon the severity of the problem, it may be necessary to use a combination of control measures to reduce erosion to reasonable limits.

Erosion Control Measures on Agricultural Areas

Contouring Performing tillage and planting crops along the contour of the land can be an effective conservation measure. Rill development is reduced when surface runoff is impounded in small depressions. Contour farming not only minimizes erosion but also reduces runoff by storing rainfall behind ridges. The storage capacity of furrows is significantly increased with ridge tillage systems. Row crops are planted on top of the same furrow each year to maintain furrow storage capacity.

The effectiveness of ridges in trapping runoff and reducing soil loss decreases as slope gradient becomes greater. When contouring is performed on steep slopes, furrow storage capacity may be exceeded during high-intensity rainfall events. Water previously stored in the furrow is then released. As the volume of water increases with each succeeding row, small gullies may form. To prevent runoff from large precipitation events from overtopping ridges, a small slope gradient along the row is desirable. Field boundaries should be located on the contour or moved to eliminate odd-shaped fields with short rows.

Strip cropping Under strip-cropping conditions, alternate parcels of different crops are grown on the same field. The strips with the greatest surface vegetative cover capture soil eroded from upslope areas. Strip widths are dictated by farm implement requirements. To improve erosion control, the strips are usually planted on the contour in a rotation that shifts crops annually from one strip to the next. The most effective strip-cropping rotations include perennial grasses and legumes that alternate with grain and row crops. In arid and semiarid regions, strips may be placed perpendicular to the prevailing wind direction for wind erosion control.

Conservation tillage On cropland areas, erosion potential is substantially reduced when residue mulch from the previous crop is left on the soil surface (Figure 1). Residue mulch serves to protect the soil surface by adsorbing and dissipating raindrop energy. The percentage of residue cover maintained on the soil surface influences soil loss potential. Substantial reductions in erosion can result from small amounts of residue cover. Any tillage or planting system that leaves at least 30% of the soil surface covered with residue after planting has been defined as conservation tillage. It can be seen from Figure 1 that a 30% cover of wheat or corn residue can reduce soil loss by approximately 62% and 97%, respectively. The type of residue material influences the amount of erosion protection provided for a given surface cover. During runoff events, small impoundments may form above residue elements. The impoundments found above the larger-diameter residue elements such as corn have greater volumes and are therefore more effective in trapping sediment and reducing soil loss.

Excessive tillage can destroy soil structure, resulting in surface sealing and decreased infiltration. With the increased availability of herbicides, the use of tillage for weed control has diminished. When tillage is performed, existing crop residues are maintained by using special implements that cause only minimal disturbance to the soil surface. To maintain sufficient residue cover to control erosion, no-tillage is used before planting for some row crops such as soybeans.

Terraces On steep land, terraces or broad channels are built perpendicular to the slope to reduce rill erosion by decreasing overland flow length. Sediment settles from overland flow as runoff travels at relatively low velocities along the gentle grades used

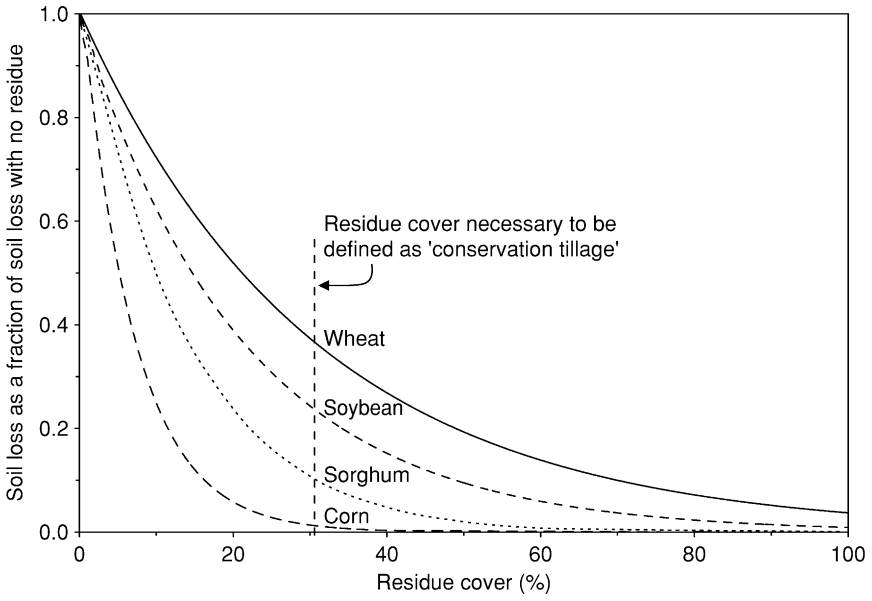


Figure 1 Ratio of soil loss for a given residue cover to soil loss with no cover. After Colvin TS and Gilley JE (1987) Crop residue–soil erosion combatant. *Crops and Soils* 39(7): 7–9, with permission.

in terraces. Therefore, the quality of surface water leaving terraced fields is improved. Gully formation is prevented since the terraces usually empty on to grassed waterways or into underground pipes. Since terraces retain runoff they also increase the amount of soil water available for crop production.

Important considerations in terrace design include soil characteristics, cropping and management practices, and climatic conditions. Contouring is included as a conservation practice on terraced fields since the crop rows are usually planted parallel to the terrace channel. Since they are expensive to construct, cause some inconvenience to the farming operation, and require periodic maintenance, terraces should only be used when other erosion-control measures do not provide adequate protection.

Buffer strips Buffer strips are designed to intercept runoff using permanent vegetation. Other erosion-control practices are usually employed in association with buffer strips. As an integral part of a planned conservation system, buffer strips may be located at a variety of locations within a landscape. To maintain buffer-strip performance, periodic maintenance is required. Contour buffer strips, filter strips, and grassed waterways are frequently used types of buffer strips.

Perennial grasses are usually planted along steep slopes within contour buffer strips. Sediment is removed as overland flow approaches and enters the grass strips. Site-specific conditions dictate the types of vegetation and spacing of contour buffer strips. As a result of sedimentation, a narrow terrace may form along the upslope portion of the grass strip. Contour buffer strips are much less expensive to establish than terraces.

Filter strips provide increased infiltration and remove sediment from overland flow. However, they do not interfere significantly with normal farming operations since they are located at the edge of fields or adjacent to streams, ponds, or wetlands. Areas with gentle slopes where rilling is not a problem are the best locations for filter strips.

Grassed waterways Runoff from terraces or other concentrated flow areas can be conveyed using grassed waterways, thus preventing channel erosion and gully formation. Costly downstream sedimentation is reduced because the sediment transported by overland flow is deposited in the grassed waterways. A stable outlet located below the grassed waterway serves to disperse the flow before it enters a vegetative filter.

Grassed waterways reduce peak runoff rates and provide a stable channel that can easily handle the flow that remains. Channel stabilization is provided by modifying the cross-section and slope of the waterway to limit flow velocity and by establishing vegetative protection. The types of vegetation used in the channel are dictated by local soil and climatic conditions. In addition to grassed waterways, permanent gully control structures may be needed on areas with relatively large runoff volumes or steep slopes.

To prevent failure, the waterway should not be used as a road, stock trail, or pasture, especially

during wet conditions. Care should also be taken when farm machinery crosses the waterway. The waterway should be managed to stimulate new growth and control weeds, and an annual application of fertilizer is recommended.

Polyacrylamide Erosion on irrigated areas has been reduced by the use of polyacrylamide. Polyacrylamide is a long-chain synthetic polymer that serves as a strengthening agent to bind soil particles together. Larger, heavier particles are more difficult to detach and transport. Polyacrylamide is added to the irrigation flow at rates dictated by the irrigation system, soil type, and water source. Increased infiltration and reduced transport of pollutants have also been reported for soils on which polyacrylamide has been applied.

Erosion Control Measures on Forest Areas

Proper road and drainage design, and effective slope stabilization can reduce erosion potential during road construction on forest areas. Detailed planning of the timber-harvesting operation can minimize runoff and soil loss. Compaction of the soil surface occurs when logs are moved across the land surface by tractors or skidders. Approximately 25–35% of the harvested area is disturbed during the tractor-logging operation. In comparison, the highlead system where the ends of the logs are raised off the ground by a cable system, and the skyline system where the disturbed logs are entirely lifted off the ground reduce the disturbed harvested area to approximately 15% and 12%, respectively. The risk of erosion is reduced when damaged trees are felled to reduce the velocity of overland flow. Check dams can be established in drainages using straw bales. Straw can also be spread on burned areas to protect the soil and stabilize reseeded areas.

Erosion Control Measures on Construction Sites

The construction project should be conducted in phases so that only those sites under active development are exposed and those areas should be kept as small as possible. Management practices should be implemented to reduce the volume and velocity of runoff, and to retain sediment within the construction site. Small sediment basins, perimeter sediment fences, and straw bale or fabric check dams can be used to reduce off-site sediment transport. Runoff from adjacent areas should not be allowed to enter the disturbed site. Finally, a vegetative cover should be established as soon as possible to provide permanent protection for the construction site.

Erosion Control Measures in Stream Channels

Stream channel erosion can be reduced by the use of vegetation, mechanical, or a combination of vegetation and mechanical means. Grading of the stream bank to a less severe slope may be necessary, depending upon the size of the upstream drainage area. Grasses, shrubs, and trees can be successfully used to stabilize stream channels. Fast-flowing water can be diverted away from stream banks by dikes made of loose stone or rock piles placed within the stream channel. Materials such as a mechanical cover of stone or rocks may also serve as a protective cover. A mechanical cover can be employed to protect areas with the greatest erosion hazard such as the bottom of a stream, while vegetation is usually used to stabilize the upper portion of the stream banks.

See also: **Erosion:** Irrigation-Induced; Wind-Induced

Further Reading

Carreker JR, Wilkinson SR, Barnett AP, and Box JE (1978) *Soil and Water Systems for Sloping Lands*. USDA, ARS-S-160.

Haan CT, Barfield BJ, and Hayes JC (1994) *Design Hydrology and Sedimentology for Small Catchments*. San Diego, CA: Academic Press.

Hudson N (1995) *Soil Conservation*. Ames, IA: Iowa State University Press.

Laflen JM, Tian J, and Huang C (2000) *Soil Erosion and Dryland Farming*. Boca Raton, FL: CRC Press.

Lal R (1999) *Soil Quality and Soil Erosion*. Boca Raton, FL: CRC Press.

Lal R, Blum WH, Valentine C, and Stewart BA (1998) *Methods for Assessment of Soil Erosion*. Boca Raton, FL: CRC Press.

Napier TL, Napier SM, and Turdon J (2000) *Soil and Water Conservation Policies and Programs: Successes and Failures*. Boca Raton, FL: CRC Press.

Pierce FJ and Frye WW (1998) *Advances in Soil and Water Conservation*. Chelsa: Sleeping Bear Press.

Schwab GO, Fangmeier DD, Elliot WJ, and Frevert RK (1993) *Soil and Water Conservation Engineering*, 4th edn. Singapore: John Wiley.

Toy TJ, Foster GR, and Renard KG (2002) *Soil Erosion: Processes, Prediction, Measurement and Control*. New York: John Wiley.

Troeh FR, Hobbs JA, and Donahue RL (1999) *Soil and Water Conservation: Productivity and Environmental Protection*, 3rd edn. Upper Saddle River, NJ: Prentice-Hall.

Ward AD and Elliot WJ (1995) *Environmental Hydrology*. Boca Raton, FL: CRC Press.

Wind-Induced

T M Zobeck, USDA Agricultural Research Service, Lubbock, TX, USA
R S Van Pelt, USDA Agricultural Research Service, Big Spring, TX, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Wind erosion is a dynamic physical process leading to soil degradation that occurs when strong winds blow on loose, dry, bare soils. Fine, fertile soil particles are often removed during wind erosion (wind-induced particle movement), reducing soil productivity and causing significant on-site and off-site problems. The widespread social and economic hardships that occurred in the USA during the disastrous Dust Bowl days of the 1930s were caused primarily by wind erosion on cropland. Much progress has been made in reducing the effects of wind erosion in the USA through soil conservation efforts made by individual land-owners with the technical assistance of the US Department of Agriculture, Natural Resource Conservation Service (USDA, NRCS). However, wind erosion continues as a national and international problem. Dust clouds may still be seen in many parts of the USA (Figure 1). Studies by the NRCS in 1997 show that wind causes about 44% of the 1.9 billion tons per year of soil lost from US cropland by water and wind erosion. In this article we explore the causes, effects, and control of wind erosion.

Wind Profile

Wind erosion is a process that results from the interaction of the wind and the soil surface. The *Encyclopedia of Climatology* defines wind as "a stream of air flowing relative to the earth's surface, usually more or less parallel to the ground." The Earth's surface exerts a drag on the wind, resulting in a vertical profile of wind speed that is described by a semilogarithmic equation:

$$u(z) = \frac{u^*}{k} \ln\left(\frac{z}{z_0}\right) \quad [1]$$

where $u(z)$ ($\mathrm{m\,s^{-1}}$) is the wind speed at height z (m); u^* ($\mathrm{m\,s^{-1}}$) is the friction velocity, which is indicative of the amount of atmospheric turbulence (its value is independent of height for a given wind profile); k is the von Karman constant (0.4), a dimensionless number; and z_0 (m) is the aerodynamic roughness height, which is a measure of the roughness of the ground surface. From eqn [1], it is evident that there is a sharp decrease of mean horizontal wind speed as the surface is approached. This gradient in wind speed produces a shearing force as the result of surface roughness causing a drag on the airflow. The shearing force produces a tangential stress on the surface, called the shear stress, that is equal to the product of the square of the friction velocity and fluid density. The shear stress provides the interchange of momentum necessary for erosion to occur.

Figure 1 This dust storm was advancing in a thunderstorm outflow in west Texas in 1997. Reproduced with permission from Chen W and Fryrear DW (2002). Sedimentary characteristics of a haboob dust storm. *Atmospheric Research* 61: 75–85, Elsevier.

Modes of Transport

The minimum wind velocity initiating particle movement is known as the threshold velocity. Particle susceptibility to wind erosion is affected by size, density, and shape. The quartz sand particle diameter most susceptible to wind erosion is approximately 100 μm; this size particle begins to move with a wind velocity of approximately $14\,\mathrm{m\,s^{-1}}$ measured at a height of 2 m. Particles larger and smaller than this size require greater wind velocities to initiate movement from rest.

Wind-blown materials move in three modes: creep, saltation, and suspension. In general, the largest soil particles (1–2 mm) will roll or slide along the surface in the creep mode because they are too massive to leave the soil surface. Particles between 100 μm and 1 mm in size tend to move in saltation (bouncing) mode. The third mode of particulate transport, suspension, involves soil particles less than 100 μm in diameter. These materials may travel great distances before returning to Earth.

Fine materials carried in suspension are less susceptible to direct entrainment by the wind than are fine sand particles. Saltating particles return to the soil surface with a force that is a function of their mass and speed, resulting in the disruption of soil aggregates and surface crusts and the release of finer particles. As this fine particulate is produced, turbulent eddies carry it higher and higher into the air. In order

for a particle to be lifted to any height, the upward velocity of the eddy must be greater than the terminal velocity of the particle. Sand-sized quartz grains have significantly higher terminal velocities than silt- and clay-sized particles and soil particulate organic matter, and are rarely lifted to heights greater than a few meters. While in most cases sand-size particles are redeposited in the source field or nearby, in extreme events fine sands may enter true suspension mode and be transported great distances.

As saltating particles are ejected from the surface, they initially rise at an angle of about 25–50° and the force of the wind accelerates them constantly until gravity returns them to the surface. The force these returning particles exerts on the soil surface often results in the ejection of several sand grains from the surface, so the number of grains in saltation often increases exponentially in a short distance in sandy soils. This phenomenon has been called the avalanching effect. There is a limit to the number of grains that can be entrained for any given wind speed, however, as the accelerating sand grains exert a drag on the passing wind that results in transfer of the wind energy into the energy of the sand particles and heat of friction. There is an equilibrium known as the transport capacity where as much saltating material is being deposited as is eroded. This is generally not true for particles moving in the suspension mode, as the amount of suspended particles continues to increase in the atmosphere with increasing fetch across the eroding surface.

Soil Surface Conditions

Although wind drives the wind-erosion process, the condition of the soil surface often controls whether or not a soil is transported by the wind. Any factor that reduces the impact of the wind on the soil surface reduces wind erosion.

The surface soil texture (the amount of sand, silt, and clay) is a primary factor affecting the erodibility of a site. Sandy soils tend to have very low stability and are easily moved by wind. The NRCS has related soil texture to wind erodibility, classifying the soil into wind erodibility groups according to soil texture and carbonate content (Table 1). Calcareous soils (soils containing enough calcium carbonate that they effervesce in the presence of dilute acid) tend to be more erodible than similar noncalcareous soils. The surface soil texture and carbonate content are difficult to change and are generally considered intrinsic or static soil properties. Other surface soil properties are temporal in nature and may change annually, seasonally, or even daily. Thus, the effects of these properties on wind erosion are temporal in nature and may be subject to management practices.

Table 1 Relation of soil texture and soil erodibility

Soil texture[a]	*Predominant soil texture class of surface layer*	*Wind erodibility group (WEG)*	*Soil erodibility index (I) (mg ha^{-1} year^{-1})*[b]
C	Very fine sand, fine sand, sand, or coarse sand	1	694[c] 560 493 403 358
C	Loamy very fine sand, loamy fine sand, loamy sand, loamy coarse sand, or sapric organic soil materials	2	300
C	Very fine sandy loam, fine sandy loam, sandy loam, or coarse sandy loam	3	193
F	Clay, silty clay, noncalcareous clay loam, or silty clay loam with more than 35% clay	4	193
M	Calcareous loam and silt loam or calcareous clay loam and silty clay loam	4L	193
M	Noncalcareous loam and silt loam with less than 20% clay, or sandy clay loam, sandy clay, and hemic organic soil materials	5	125
M	Noncalcareous loam and silt loam with more than 20% clay, or noncalcareous clay loam with less than 35% clay	6	108
M	Silt, noncalcareous silty clay loam with less than 35% clay, and fibric organic soil material	7	85
–	Soils not susceptible to wind erosion due to coarse surface fragments or wetness	8	–

[a]C, coarse; M, medium; F, fine.
[b]The soil erodibility index is based on the relationship of dry soil aggregates greater than 0.84 mm to potential soil erosion.
[c]The *I* factors for WEG 1 vary from 160 for coarse sands to 310 for very fine sands. For coarse sand with gravel, use a low figure. For very fine sand without gravel, use a higher value. When unsure, use an *I* value of 220 as an average figure.
Adapted from the USDA, NRCS (1999) *National Agronomy Manual*. Title 190 Part 502. Wind Erosion, USDA Natural Resources Conservation Service.

Soil surface roughness, also called microrelief or microtopography, can modify the effect of the wind on the surface and physically protect part of the surface from abrasion caused by blowing particles. In tilled soils, the tillage tool creates an oriented roughness, or ridges, in the direction of tillage and random roughness due to the random orientation of soil aggregates and clods. The effects of tillage-induced roughness on wind erosion depends upon wind direction when ridges are present. Winds in the direction perpendicular to tillage are affected by both ridges and clods while wind in the direction parallel to tillage is only affected by the random roughness. The physical protection of the soil surface by ridges and clods is illustrated in Figure 2. More of the soil surface is protected from abrasion when the wind blows perpendicular to the ridges.

In agricultural fields, tillage not only creates surface roughness but also disrupts the soil and creates an unconsolidated surface layer of soil aggregates of various sizes. Some of the aggregates and clods are small enough to be eroded by wind. Aggregates of mineral soils <0.84 mm diameter are generally considered erodible. Larger particles of organic soils blow because they have much lower density than mineral soils due to the high content of organic material present. The estimated potential wind erosion is related to the amount of nonerodible aggregates, as indicated in Table 2. Practices that encourage the development of nonerodible aggregates reduce the amount of wind erosion in tilled fields.

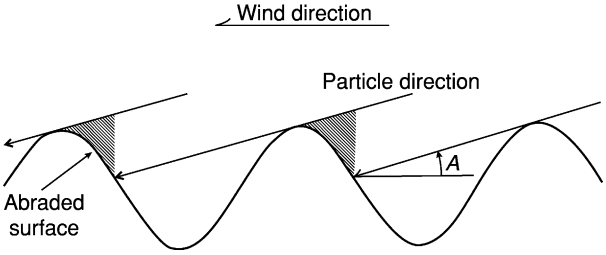


Figure 2 Schematic representation of a ridged field. Part of the surface is protected from abrasion. Adapted with permission from Zobeck TM (1991) Soil properties affecting wind erosion. *Journal of Soil and Water Conservation* 46: 112–118.

The ability of soil aggregates to resist breakdown due to abrasion caused by blowing sand grains or other applied forces is called aggregate stability. Fragile soils have low aggregate stability and are easily disrupted during wind erosion events, causing additional fine material to be added to the wind stream. Aggregate stability is quantified by measuring the amount of energy per unit mass needed to crush aggregates of a given size.

Rainfall often impacts the soil surface with enough force to disrupt aggregates and rearrange the particles on the surface of unconsolidated tilled soils to form a relatively thin consolidated layer, called a crust. The crust is more dense and resistant to abrading particles than the unconsolidated soil immediately below it. Very sandy deposits form very weak crusts that are easily destroyed by blowing sand particles. However, soils with even a small amount of clay and silt will bind together to form a layer that is 40–70 times as resistant to erosion as entirely erodible soil of the same texture.

In sandy soils, a thin layer of loose, highly erodible sand may form on the surface of a crust. This sand is highly susceptible to wind and starts blowing at relatively low wind speeds. The amount and distribution of this loose, erodible sand on the soil surface are related to the potential wind erosion of a site. If little

Table 2 Soil erodibility index (I) in units of mg ha^{-1} as determined by percentage of nonerodible soil

Aggregates >0.84 mm diameter (%)										
	Units (mg ha^{-1})									
Tens[a]	0	1	2	3	4	5	6	7	8	9
0	–	694	560	493	437	403	381	358	336	314
10	300	293	287	280	271	262	253	244	237	228
20	220	213	206	202	197	193	186	181	177	170
30	166	161	159	155	150	146	141	139	134	130
40	125	121	116	114	112	108	105	101	96	92
50	85	81	74	69	65	60	56	54	52	49
60	47	45	43	40	38	36	36	34	31	29
70	27	25	22	18	16	13	9	7	7	4
80	4	–	–	–	–	–	–	–	–	–

[a]The rows and columns represent the amount of nonerodible soil. For example, if nonerodible soil is 25%, first find the 20 in the Tens row and then the 5 in the Units column to find 193 mg ha^{-1}.

Adapted from the USDA, NRCS (1999) *National Agronomy Manual.* Title 190 Part 502. Wind Erosion, USDA Natural Resources Conservation Service.

or no loose, erodible sand is present, erosion generally does not occur.

Vegetation

Vegetation, whether growing or dead, standing or flat, reduces the susceptibility of the soil surface to erosion. Standing vegetative cover effectively raises the roughness height (z_0) and reduces the shear stress acting on the surface behind the barrier. Movement of the standing biomass absorbs some of the wind energy and eddies created in the wake of the standing biomass dissipate much of the rest. The effectiveness of standing biomass at reducing erosion increases with the height, density, and stiffness of the plants. Flat plant residue, while not greatly affecting the value of z_0, also serves to protect the soil surface from erosion. Dense flat residue and low-growing vegetation effectively form the surface against which the wind exerts its force, and thus protect the underlying soil surface. Where the flat residue is less dense, the residue acts much like random soil roughness elements. If saltation is initiated in a localized area of the field, the saltating grain may strike a relatively nonerodible plant part and settle to the ground, effectively breaking the chain reaction of saltation. Where adequate water from rainfall or irrigation is available, winter-cover crops are often planted as a preferred wind-erosion control technique.

Wind Erosion Modeling

Several wind erosion models have been used or proposed. The model most commonly used currently in the USA is the Wind Erosion Equation (WEQ), developed by the USDA. The WEQ was developed to predict long-term average annual soil loss from farm fields. The WEQ has been used by the USDA, NRCS to determine soil loss due to wind erosion for the US Natural Resource Inventory. The general form of the equation, $E = f(IKCLV)$, expresses soil loss (E) as functionally related to soil erodibility (I), ridge roughness (K), a climatic index (C), field length (L), and a vegetative cover factor (V). In practice, the I value is determined from a table similar to Tables 1 or 2 and a series of charts and nomographs are used to apply the other factors in the equation. A computer spreadsheet program is now often used to calculate soil loss using WEQ.

A modified version of the WEQ called the revised wind erosion equation (RWEQ) was recently released by USDA. The RWEQ uses some of the same concepts of the WEQ but incorporates new variables in the following equation: $Q = 109.8\ (WF \times EF \times SCF \times K' \times COG)$. The maximum soil loss determined by the field conditions, Q, is the product of several variables. The wind is described by a weather factor, WF, that accounts for the wind velocity, soil wetness, snow, air density, and gravity. The erodible fraction of soil (EF) is estimated based on intrinsic soil properties. The surface crust factor, SCF, reduces erosion based on clay and organic matter content. Soil roughness, K', is related to oriented and random soil roughness. And a combined crop factor, COG, accounts for the effects of crop canopy and flat and standing residue. RWEQ estimates soil loss due primarily to saltation, with no direct estimate of suspended load. In practice, soil and site factors and crop management are entered into a computer program that computes erosion for periods defined by the user.

A wind erosion model called the Wind Erosion Stochastic Simulator (WESS) is a single-event wind erosion submodel of the Environmental Policy Integrated Climate (EPIC) model of USDA. This model is similar to RWEQ and uses the following basic equation:

$$YW = (FI \times FR \times FV \times FD) \int_0^t \frac{YWR}{WL}\,\mathrm{dt} \qquad [2]$$

where YW is the wind erosion estimate, FI is a soil erodibility factor, FR is surface roughness, FV is vegetative cover, FD is a field length factor based upon the unsheltered field length WL, and YWR is the erosion rate based upon wind energy and soil properties.

A much more detailed Wind Erosion Prediction System (WEPS) is under development by USDA to make daily or shorter-time-interval estimates of wind erosion. WEPS is a modular computer-based prediction system that also requires significant soils, site, and crop management input. The model uses a series of complex submodels that account for the effects of weather, crop growth and decomposition, soil surface properties, soil hydrology, tillage, and erosion. In general, the model simulates the weather, grows a specified crop and determines when erosion occurs. The output includes estimates of saltation and suspended sediment.

Recent concern in global climate change has created interest in using models to estimate atmospheric dusts, produced primarily by deserts or semi-arid ecosystems. Advances in understanding of atmospheric transport of soil-derived dusts have been made by laboratories in France and Australia. Scientists at the Laboratoire Interuniversitaire des Systèmes Atmosphériques, Creteil, France have developed a model considering the size distribution of erodible particles and surface roughness as controls on dust emissions. Scientists with Commonwealth Scientific and Industrial Research Organization (CSIRO) in Australia have developed a model accounting for the effects of soil, climate,

and vegetation to estimate sand drift intensity and dust emission. The model has been coupled with a geographic information system to estimate dust emission over a large area of Australia.

Wind Erosion Sampling

Measurement of wind erosion is necessary to determine the effects of control measures, validate models, and assess the intensity of erosive forces. A wide variety of devices has been developed for this purpose. The ideal sampler has a high sampling efficiency by collecting sediment without disturbing the wind stream or developing suction or back-pressure. Since about 50% of the blowing sediment is transported within 2 cm of the soil's surface, sampling close to the surface is desirable to make accurate estimates of transported material.

A simple method to measure creep material is to bury a jar level with the soil surface and measure the sediment that rolls into it. The center portions should be covered to exclude saltation material. A combination creep/saltation sampler has been developed by the USDA to sample very near the soil surface at heights of 0–3, 3–9, and 9–20 mm (Figure 3). The sampler rotates into the wind and each inlet has a separate collection pan. This sampler is used in combination with other samplers to analyze the entire saltation zone.

Several types of samplers have been developed to sample the saltation zone, which is about 1–2 m above the soil surface. Big Spring Number Eight (BSNE) samplers are wedge-shaped samplers with 60-mesh screens on the top or side to relieve back-pressure and 2-cm-wide and 5-cm-high inlets (Figure 4). Several BSNE samplers are mounted on a pole to sample several heights. The samplers have wind vanes to ensure they are oriented into the wind. A similar sampling scheme (Wilson-Cooke-type dust samplers) employs bottles mounted on a pole with glass tubes as inlets (Figure 5).

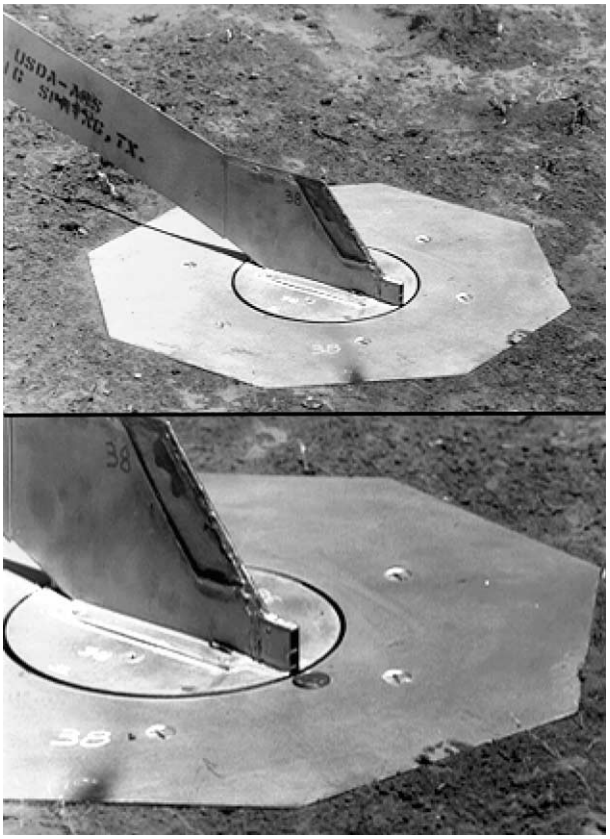


Figure 3 Creep/saltation sampler. Reproduced with permission from Zobeck TM (2002) Field measurement of erosion by wind. *Encyclopedia of Soil Science*, Marcel Dekker: New York.

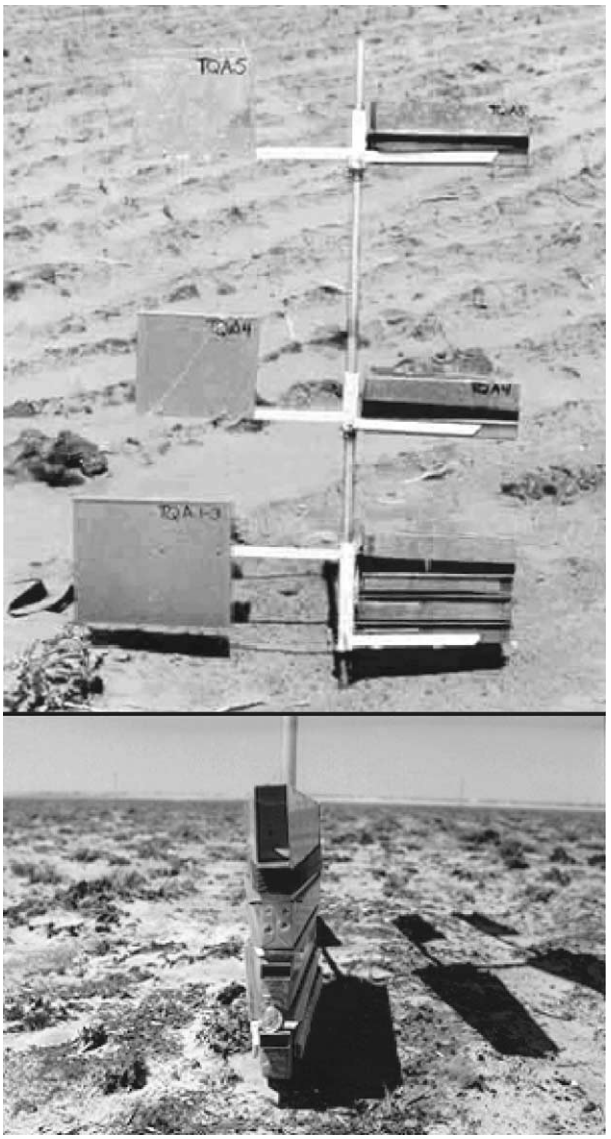


Figure 4 BSNE dust sampler. Reproduced with permission from Zobeck TM (2002) Field measurement of erosion by wind. *Encyclopedia of Soil Science*, Marcel Dekker: New York.

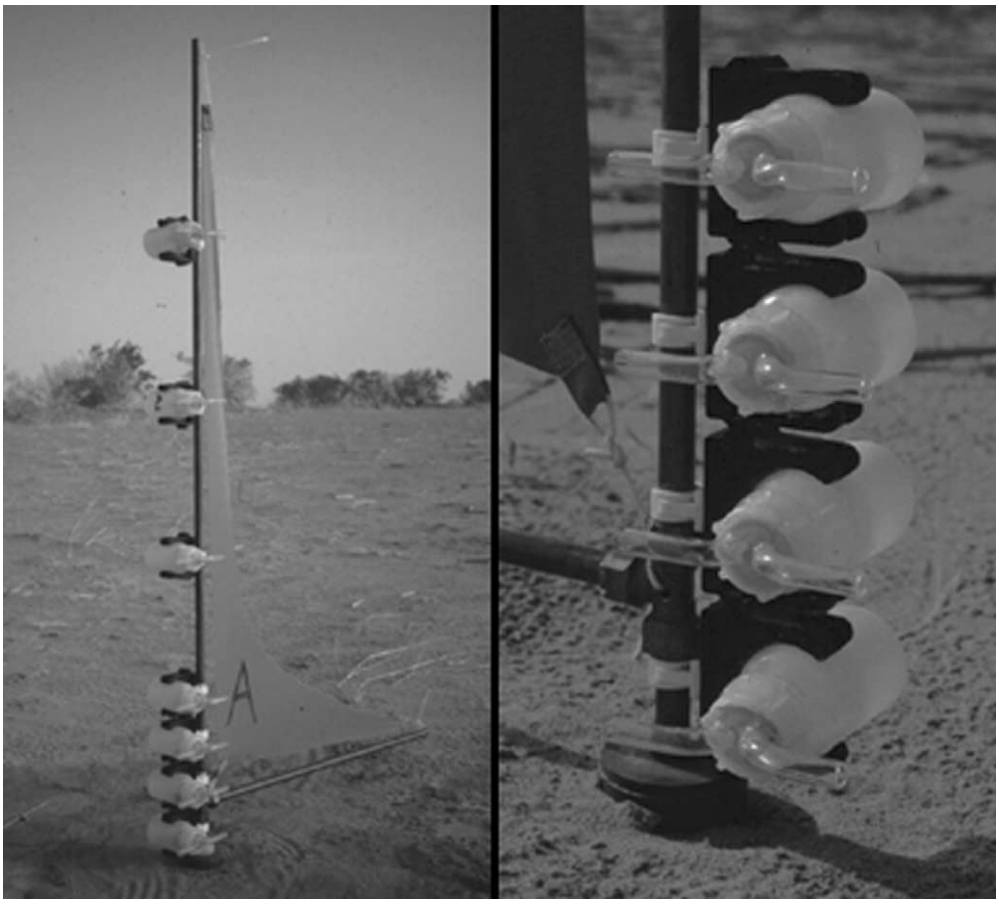


Figure 5 Wilson-Cooke-type dust sampler. Reproduced with permission from Zobeck TM (2002) Field measurement of erosion by wind. *Encyclopedia of Soil Science*, Marcel Dekker: New York.

Use of Anthropogenic Radioisotope Tracers

While wind-erosion measurement devices and samplers give good estimates of wind erosion for a given event, they are rarely employed for periods exceeding weeks or months. Wind erosion may occur at almost any time of the year and varies in intensity from year to year. From the time a field is put into cultivation, it is susceptible to erosion, and the effects of wind erosion are cumulative. Until the last 40 years, there has been no reliable way to estimate long-term wind erosion rates. Atmospheric testing of nuclear weapons in the 1950s and 1960s resulted in the release of radioactive isotopes that were distributed worldwide and were deposited, primarily by rainfall, on the soil surface. Many of these isotopes, including ^{90}strontium, ^{137}cesium, and $^{239/240}$plutonium, are strongly adsorbed to the surfaces of soil clays and organic matter. Strong pulses in 1959 and 1963–1964 created a very sharp rise in soilborne radioisotope tracers. Transport of these soilborne tracers is almost exclusively by actual physical movement of the particles to which they are adsorbed.

Radioisotope tracers, particularly ^{137}cesium, have been used for the last four decades to study soil redistribution from water erosion and for the last two decades to study soil redistribution from wind erosion. The greatest advantage to the use of radioisotope tracers is the ability to estimate long-term (since 1963) rates of wind erosion. Radioisotope activity of soil cores sampled in an undisturbed reference site such as native grassland is compared to the activity of soil cores sampled in a field or site for which the long-term estimate is desired. The loss or gain of radioisotope activity relative to the reference site has been shown to be proportional to the erosion or deposition of soil at the site of interest.

Effects of Wind Erosion

The on-site effects of wind erosion are generally considered to be those in or very near the eroding field. The on-site effects of wind erosion may be further

divided into long-term cumulative effects and immediate effects. The effects of wind erosion on the soil are cumulative and, where wind erosion is a problem, soil losses often exceed the rates of soil formation dramatically. In extreme situations, wind may erode the entire topsoil horizon and reveal the underlying, less fertile horizons. Wind erosion is a winnowing process that, while it moves all the surface soil in a field, tends to remove the finer, more fertile parts of the soil and transport these materials away from the field. The result is a soil with a lower percentage of silt and clay, lower cation exchange capacity, lower organic matter, and reduced water-holding capacity. All of these factors serve to reduce the productivity of the soil and to diminish crop yields.

Immediate on-site effects of wind erosion also result in economic losses to agriculture. During windstorms, sandblast injury from saltating sand grains striking and abrading plants damages crop stands, especially seedling stands. Depending on the intensity of the event, seedling stands may be retarded or stunted in later growth and development, damaged to the point that replanting is warranted, or completely destroyed. As the seedlings start to produce true leaves, and especially during exponential growth phases, tissue conversion to woody matter reduces the susceptibility of the plant to damage and the increase in vegetative ground cover tends to reduce wind erosion frequency and intensity. In general, small grains are less susceptible to sandblast injury than corn, soybeans, and sunflowers, which are more tolerant than cotton, cabbage-related crops, and legumes. The least sandblast-injury-tolerant crops are garden vegetables and flowers.

Studies by the USDA, Economic Research Service suggest that the off-site impacts from wind erosion may be much larger than the on-site impacts. The estimated off-site costs of wind erosion in the western USA range from $4 to $12 billion per year. Off-site impacts from wind erosion are caused primarily by the release of fugitive dust, which may travel long distances and which constitutes most of the airborne dust that settles in homes and on automobiles, imposing costs for cleaning, reduced recreational opportunities, and impaired health. Estimates of the annual flux of airborne dust deposition over land areas range from less than 10 to about 200 t km^{-2} $year^{-1}$. As expected, these estimates vary with respect to distance and direction from the major source regions. The source areas of suspended dust are primarily deserts, agricultural operations in arid and semiarid environments, unpaved roads, dry lake beds, braided streams, deltas, and glacial outwash. These areas can be located on any mid-latitude continent, and the literature is composed of studies of dust source regions on all continents.

Fugitive dust obscures visibility and negatively affects air quality. Reduced visibility results in the cancellation of flights and numerous automotive accidents annually. Reduced visibility also depresses tourism and recreation revenues. Fugitive dust fouls machinery and adversely impacts environmental and human health. Pathogens, both plant and animal, may be transported thousands of kilometers, even across oceans, on dust particles.

Wind Erosion Control

Any treatment that reduces the wind at the soil surface to below the threshold of particle movement will control wind erosion. A wide variety of practices have been developed.

Perhaps the best control is to establish and maintain adequate ground cover (Figure 6). The ground cover may be growing vegetation of some type of mulch material such as cotton-gin trash. The effect of such cover is shown in Figure 7.

Figure 6 This wheat cover crop protects the cotton from blowing dust in Texas. Courtesy of USDA Natural Resources Conservation Service.

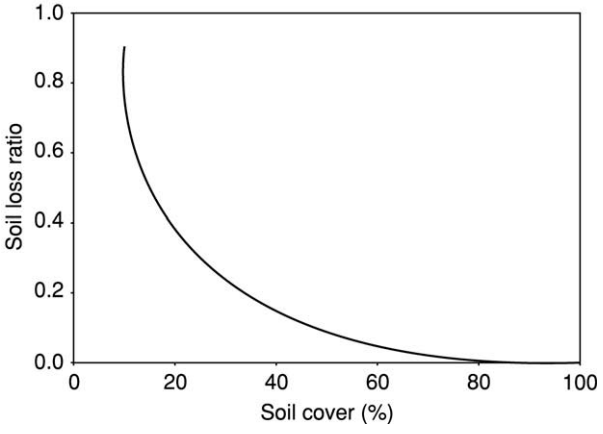


Figure 7 The effect of cover on soil loss. Reproduced with permission from Fryrear DW (1985) Soil cover and wind erosion. *Transactions of the American Society of Agricultural Engineers* 28(3): 781–784.

Since wind erosion tends to increase with field distance due to the avalanching effect, a reduction in field length is often accomplished using barriers. The barriers may consist of elaborate windbreaks called shelter belts that have several rows of vegetation, including trees and shrubs. The use of trees planted in shelter belts to control wind erosion is a technique that dates back to the US Dust Bowl days of the 1930s. Shelter belts were often planted along fence lines and around farmsteads to reduce the wind velocity and filter some of the suspended sediment from the air. In general, shelter belts reduce the wind velocity for a distance of 10 times their height in the downwind direction and three times their height in the upwind direction.

More recently, annual crops have been used as more closely spaced shelter belts within a field. Spacing of these 'wind strips' varies from two rows of wheat planted in the furrows between rows to 5–10-m-wide strips planted at 100–300-m intervals throughout the field (Figure 8). In emergency situations even physical barriers such as a snow fence may be used. The recommended porosity of barriers is about 40% for best results.

In some situations, it is not possible to maintain a vegetative cover on soil surface. In these circumstances, tillage may be used to maintain the soil in a cloddy condition (Figure 9). Care must be used to ensure tillage is done when the soil is moist enough to create clods. Tilling dry soils may exacerbate the problem by further pulverizing the soil and creating even more erodible particles.

Chemical treatments have also been used to control soil blowing. Chemical treatments tend to be expensive and are usually restricted to smaller or high-value areas. The chemical treatments bind the surface soil particles together. Any disturbance to the treated areas greatly reduces the effectiveness of the treatment.

Figure 8 Perennial grass barriers protect cucumbers in South Carolina. Courtesy of USDA Natural Resources Conservation Service.

Figure 9 Tillage maintains this field in a rough cloddy condition. Courtesy of USDA Natural Resources Conservation Service.

List of Technical Nomenclature

u^*	Friction velocity ($m\,s^{-1}$)
$u(z)$	Wind speed at height z ($m\,s^{-1}$)
z	Height (m)
z_0	Aerodynamic roughness height (m)

Further Reading

Bagnold RA (1943) *The Physics of Blown Sand and Desert Dunes*. New York, NY: William Morrow.

Chepil WS and Woodruff NP (1963) The physics of wind erosion and its control. *Advances in Agronomy* 15: 211–301.

Fryrear DW (1990) Wind erosion: mechanics, prediction, and control. *Advances in Soil Science* 13: 187–199.

Fryrear DW and Bilbro JD (1998) Mechanics, modeling, and controlling soil erosion by wind. In: Pierce FJ and Frye WW (eds) *Advances in Soil and Water Conservation*, pp. 39–49. Chelsea, MI: Ann Arbor Press.

Hagen LJ (1991) A wind erosion prediction system to meet users' needs. *Journal of Soil and Water Conservation* 46: 106–111.

Marticorena B and Bergametti G (1995) Modeling the atmospheric dust cycle: I. Design of a soil-derived dust emission scheme. *Journal of Geophysical Research* 100: 16 415–16 430.

Shao Y (2000) Physics and modelling of wind erosion. In: *Atmospheric and Oceanographic Science Library*, vol. 23. London: Kluwer Academic.

Skidmore EL (1994) Wind erosion. In: Lal R (ed.) *Soil Erosion Research Methods*, 2nd edn, pp. 265–293. Delray Beach, FL: St. Lucie Press.

USDA, NRCS (1999) *National Agronomy Manual*. Title 190 Part 502. Wind Erosion, USDA Natural Resources

Conservation Service. Available online at:www.weru.ksu.edu/nrcs.
Wiggs GFS (1997) Sediment mobilisation by the wind. In: Thomas DSG (ed.) *Arid Zone Geomorphology: Process, Form and Change in Drylands*, 2nd edn. Chichester: John Wiley.
Woodruff NP and Siddoway FH (1965) A wind erosion equation. *Soil Science Society of America Proceedings* 29: 602–608.
Zobeck TM (1991) Soil properties affecting wind erosion. *Journal of Soil and Water Conservation* 46: 112–118.

ESSENTIAL ELEMENTS

E A Kirkby, University of Leeds, Leeds, UK

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Studies over many years of cultivating plants in water and sand cultures supplied or lacking in particular mineral nutrients have established a list of elements that are essential for the growth of higher plants. The essential elements that are required in large amounts were established in the nineteenth century and most of the trace elements, those elements usually taken up and required in only small amounts, mostly in the first half of the twentieth century. Techniques for avoiding contamination have improved and the list has been extended to include Ni for higher plants and, since 1970, As, Cr, Co, F, I, Pb, Li, and Se for animals.

The list of essential elements for higher plants is based on three criteria of essentiality. These are, that a lack of the element makes it impossible for the plant to complete its life cycle, a lack of the element gives rise to specific deficiency symptoms, and that the element plays a specific role in the nutrition and metabolism of the plant. From this approach the following are now defined as essential elements:

Carbon
Hydrogen
Oxygen
Nitrogen
Sulfur
Phosphorus
Potassium
Calcium
Magnesium
Iron
Manganese
Zinc
Copper
Boron
Molybdenum
Nickel

Sodium and silicon have also been shown to be essential for some species of higher plants.

When uptake is restricted either by a lack of the nutrient or because it is in an unavailable form in the soil, plant growth is depressed and nutrient deficiency symptoms can ensue which are typical for the lacking nutrient.

Classification of Essential Elements

Essential elements for both plants and animals are often considered in terms of macro- and micronutrients (or 'trace elements'), the macronutrients, C, H, O, N, P, S, K, Ca, Mg, being present in much higher concentrations in plant tissues than the micronutrients, Fe Mn, Cu, Zn, Mo, B, Cl, Ni. This approach provides no indication of biochemical behavior nor physiological function. One such attempt at physiological classification is shown in Table 1.

The elements in the first group include N and S, which in reduced form are covalently bonded constituents of organic matter. Phosphorus, B, and Si constitute another group and show similarity in biochemical behavior in that they are absorbed as inorganic anions or acids and in the plant are bound by hydroxyl groups of sugars forming phosphate, borate, and silicate esters. The plant nutrient cations and anions K, Na, Ca, Mg, Mn, and Cl make up a third group and are taken up in ionic form; and in the plant they remain as such or are adsorbed to indiffusible anions. These nutrients have osmotic- and ion-balance roles as well as more specific functions in enzyme catalysis. The fourth group of elements, including Fe, Cu, Mo, and Zn, are present in plants as structural chelates or metalloproteins; and the first three participate in redox reactions. Mn in its role in photosystem II can also be added to this list.

Nutrient Deficiency Symptoms

Observed visual symptoms of chlorosis and necrosis in nutrient-deficient plants reflect impairment of

nutrient function and the degree of nutrient mobility within the plant (i.e., old leaves first showing symptoms indicate a mobile nutrient; young leaves presenting symptoms, indicate an immobile nutrient). A systematic approach to visual diagnosis can therefore be drawn up as shown in Table 2. It must be borne in mind, however, that symptoms only become clearly visible when deficiency is acute and the growth rate distinctly depressed. Additionally much natural vegetation adapted to nutrient-poor sites adjusts the growth rate to the most limiting nutrient so that visible nutrient deficiencies do not appear. Diagnoses can be especially difficult when more than one nutrient is involved either in deficient or toxic levels. Visual symptoms provide only a guide to nutrient status which must be supported by additional information, including soil and plant chemical analysis.

Nitrogen

Of all the plant nutrients, nitrogen is the most commonly deficient in soils. The two main forms of uptake are nitrate and ammonium ions, and maximum growth is favored by a combination of both. Ammonium N is assimilated in the root, whereas nitrate is mobile in the xylem and can be stored in the vacuoles of roots, shoots, and storage organs. Reduction of nitrate to ammonium via nitrate reductase occurs both in roots and shoots, and there are

Table 1 Classification of plant nutrients

Nutrient element	*Uptake*	*Biochemical functions*
1st group; C, H, O, N, S	In the form of $CO_2, HCO_3^-, H_2O, O_2, NO_3^-, NH_4^+, N_2, SO_4^{2-}, SO_2$. The ions from the soil solution, the gases from the atmosphere.	Major constituents of organic material. Essential elements of atomic groups which are involved in enzymic processes. Assimilation by oxidation-reduction reactions.
2nd group; P, B, Si	In the form of phosphates, boric acid or borate, silic acid from the soil solution.	Esterification with native alcohol groups in plants. The phosphate esters are involved in energy transfer reactions.
3rd group; K, Na, Ca, Mg, Mn, Cl	In the form of ions from the soil solution.	Non-specific functions establishing osmotic potentials. More specific reaction in which the ion brings about optimum conformation of an enzyme protein (enzyme activation). Bridging of the reaction partners. Balancing anions. Controlling membrane permeability and electrochemical potentials.
4th group; Fe, Cu, Zn, Mo	In the form of ions or chelates from the soil solution.	Present predominantly in a chelated form incorporated in prosthetic groups. Enable electron transport by valency change.

Reproduced from Mengel K and Kirkby EA (2001) *Principles of Plant Nutrition*, 5th edn. Dordrecht, the Netherlands: Kluwer Academic Publishers.

Table 2 Some principles of visual diagnosis of nutritional disorders

Plant part	*Prevailing symptom*		*Disorder*
			Deficiency
Old and mature leaf blades	Chlorosis	Uniform	N (S)
		Interveinal or blotched	Mg (Mn)
	Necrosis	Tip and marginal scorch	N
		Interveinal	Mg (Mn)
Young leaf blades and apex	Chlorosis	Uniform	Fe (S)
		Interveinal or blotched	Zn (Mn)
	Necrosis (chlorosis)		Ca, B, Cu
	Deformations		Mo (Zn, B)
			Toxicity
Old and mature leaf blades	Necrosis	Spots	Mn (B)
		Tip and marginal scorch	B, salt (spray injury)
	Chlorosis, necrosis		Nonspecific toxicity

Reproduced from Marschner H (1995) *Mineral Nutrition of Higher Plants*, 2nd edn. San Diego, CA: Academic Press.

striking differences in distribution of activity between species. With increasing intensity of nitrate supply, the shoot generally plays a greater role in the reduction. The NH_3 formed cannot be reoxidized and provides the source for the synthesis of the low molecular weight N compounds, including amino acids, which are further metabolized to high molecular weight cell constituents containing N.

Nitrogen makes up about 20–40 mg g^{-1} of the dry weight of the plant, most of which is present in reduced form as organic substances. These include low molecular weight compounds, including amino acids, amides, peptides, amines, and ureides, which are involved in intermediate metabolism and transport of N within the plant. Additionally N is a constituent of proteins, nucleic acids, and other compounds of high molecular weight such as phytohormones. As an enzyme constituent, N is involved in all reactions taking place in the cell, including energy metabolism, and, in the form of nucleic acids, is responsible for storage and transfer of genetic information.

Lack of nitrogen slows down the synthesis of protein and generally inhibits plant growth. Plants are small, with spindly stems, and senescence is enhanced, older leaves falling prematurely. Growth of roots is less depressed than that of shoots. Chloroplast development is disturbed and hence N-deficient leaves develop chlorosis, which is evenly distributed over the whole leaf. Since nitrogen is highly mobile within plants and is readily transported in reduced form from older to younger sites to meet demand for growth, older leaves first show the symptoms of chlorosis. The premature senescence associated with N deficiency particularly relates to a lack, in synthesis and translocation, of the N-containing phytohormones, cytokinins. When N is well supplied, not only is senescence delayed and growth stimulated, but shifts occur in the root-to-shoot ratio in favor of the shoot, all factors which appear to be related to induced changes in the phytohormone balance in favor of cytokinins.

Sulfur

The most important source of S to plants is sulfate taken up by the roots, although it can be absorbed through the leaves as sulfur dioxide. Uptake by the root appears to be sensitively controlled by shoot-to-root signals in the form of the tripeptide glutathione (GSH), the main long-distance transport form of reduced S. Sulphate is transported to the shoot and is reduced predominantly in the chloroplasts of mature leaves. Demand for S in the shoot is expressed by low concentrations of GSH in the phloem sap to the root, which favors the synthesis of SO_4^{2-} transporters and hence sulfate uptake.

Sulfur makes up between 1–5 mg g^{-1} of the dry weight of plants. The first stable S-containing product in the reduction of sulfate is the amino acid cysteine, from which methionine can be formed, both amino acids being building blocks for protein. One of the main functions of S in proteins and polypeptides is in the formation of disulfide bonds between polypeptide chains. These so-called S-S bridges contribute to the conformation of enzyme proteins. Other essential S containing compounds include the vitamins thiamine and biotin, as well as coenzyme A, which is essential for respiration and the synthesis and breakdown of fatty acids in animals as well as plants.

Since S is an essential constituent of protein, S deficiency inhibits protein synthesis, and within the plant shoot there is an accumulation of non-S-containing amino acids as well as nitrate–N. Growth rate is reduced and frequently the plants are brittle and thin, with shoots more affected than roots. Chloroplasts are decomposed during S deficiency and leaves become uniformly chlorotic. Since S is not readily mobilized from older tissues, it is the younger leaves that usually first show the deficiency.

Sulfur deficiency in the field is becoming increasingly common. Sulfur-containing fertilizers are now less frequently applied to soils even though there is a greater demand of crops as a consequence of the higher application of other nutrients and the resulting greater annual offtakes of sulfur resulting from higher yields. Additionally since the early 1970s, the emissions of SO_2 have fallen dramatically so that many areas of Europe and North America are at risk from S deficiency.

Phosphorus

Phosphorus limitation of plant growth is widespread in soils. This is largely because of the insoluble forms in which phosphate is present in soil and the extremely low mobility of phosphate in soils in comparison with other plant nutrients. The uptake mechanism of phosphate is not usually a limiting step, since plant roots are capable of absorbing phosphate against a very steep concentration gradient from extremely low external P concentrations well below those of the soil solution. Supply is restricted because acquisition from the soil is dependent on the process of diffusion, which in many cases accounts for most phosphate supplied to plant roots. This contrasts to other nutrients such as nitrate, Ca, and Mg for which the supply to the root surface is largely by mass flow of nutrients in the soil solution driven by transpiration. For phosphate, this form of transport cannot meet plant demand, because the concentration in the soil solution is too low.

Under these limiting conditions of P supply, plants can adapt by physical and chemical modification to enhance phosphate acquisition. Root surface can be extended, as, for example, by root hair formation, to allow a more efficient exploitation of the soil. Mycorrhizal fungi associated with the root also fulfills a similar function, the mycelium of the fungus extending into the soil to enhance phosphate uptake, with the host root supplying the fungus with carbohydrates. Rhizosphere acidification either by proton or organic acid excretion is also a widespread response to P deficiency which can induce P release from soil minerals. One such example is citric acid, which is excreted by the proteoid roots of white lupin at high rates and amounts saturating confined root zones. The acid, which is a strong chelator of Ca, Fe, and Al, is thus able to mobilize phosphate from sparingly soluble compounds in both acid and calcareous soils.

Phosphorus generally accounts for about 1–5 $mg\,g^{-1}$ of the dry matter of plants. Uptake by the roots appears to be regulated by a feedback signal of P-cycling in the phloem from shoot to root. Phosphate is at the center of metabolism. It is present in many sugar compounds involved in photosynthesis and respiration, and plays an essential role in energy metabolism as a component of ATP, ADP, AMP, and pyrophosphate (PPI). It is also a constituent of the phospholipids that occur in membranes and is an intergral component of the nucleotides RNA and DNA. In many seeds and fruits, phytate in the form of Ca and Mg inositol hexaphosphate acts as a phosphate reserve.

Plants showing P deficiency are typically stunted and often have a rigid, erect appearance. One of the earliest symptoms is a specific inhibition of leaf expansion and leaf surface area induced by restricted delivery of water. Older leaves often show a darkish green color by chlorophyll concentration and the stems may be a reddish color owing to enhanced formation of anthocyanins. Root growth is much less inhibited than shoot growth, and the roots act as a dominant sink for photosynthates and P from mature leaves. This response in favor of root growth at the expense of the shoot allows a greater exploitation of the soil for P as described above.

Potassium

Potassium is an essential element for all living organisms and is the most important nutrient cation not only in relation to its high concentration in plant tissues but also with respect to its physiological and biochemical functions. Unlike the elements so far discussed, K is taken up as a cation and remains in this form after uptake as the most abundant cation in the cytoplasm and vacuole, where it is required to neutralize organic acids and other anionic groups. Concentrations in plants range between 10–60 $mg\,g^{-1}$ K dry weight, with younger tissues showing higher values. Potassium is highly mobile in plants, showing the highest cation concentration in both the xylem and phloem saps, and there is evidence that K transport from shoot to root in the phloem regulates K uptake by the root.

Potassium is closely related to meristematic growth associated with acidification and loosening of the cell wall. Cell extension results from the accumulation of potassium needed to stabilize the pH of the cytoplasm and increase the osmotic potential of the vacuoles. Potassium and some phytohormones react synergistically in this respect.

High concentrations of K are needed for the active conformation of many enzymes of intermediate metabolism and biosynthesis, including protein synthesis. Potassium also plays a most important role in regulating the water status of plants by lowering water potential. Similarly it also accumulates in guard cells in particularly high concentrations, thereby regulating the opening and closing of the stomata. Potassium also functions to promote photosynthesis and the translocation of photosynthates.

Plants can suffer from potassium deficiency without showing major symptoms, with only a reduction in growth, so called 'hidden hunger.' As the deficiency progresses, older leaves are first to show symptoms, as these supply the younger leaves with potassium. In many plant species, chlorosis and necrosis begin at the leaf margins and tips, where the localized lack of potassium has disturbed the water relations. In some species such as clover, however, irregularly distributed, white necrotic spots appear on the leaves. In field-grown crops, K deficiency can often be recognized by the decreased turgor of the leaves under water stress which appear flaccid. Lignification of vascular bundles can be impaired by K deficiency, thus weakening stems and making crops prone to lodging, i.e. flattening of the crop by wind and rain as a consequence of weakening of the culms.

Calcium

Higher plants contain appreciable amounts of Ca, usually in the range of 5–30 $mg\,g^{-1}$ in the dry matter, but plant species and varieties differ greatly in their requirements. In order to obtain maximum growth, dicotyledons, particularly the legumes and herbaceous plants, have a much higher requirement than the cereals and other grasses. Calcium is taken up by plant roots in ionic form as the divalent ion and is depressed by competition from other mineral

cations such as K, NH_4, and Mg. In comparison with K, the mechanism of uptake is not particularly efficient. The concentration of Ca in the soil solution is often relatively high, and the uptake of Ca and, to a large extent, also that of translocation within the plant as a whole is passive and reflects the flow of water from the bulk soil to the atmosphere driven by transpiration.

Uptake is limited to the apoplastic or free-space pathway mainly accessible in nonsuberized young roots, which is in accord with the restriction of Ca from the living parts of the cell. This relatively large, divalent cation is bound or is in an exchangeable form in the cell wall or at the exterior surface of the plasma membrane. Substantial quantities are also sequestered and stored in the vacuoles. The maintenance of an extremely low concentration of Ca in the cytoplasm to usually less than 1 μmol l^{-1} Ca is essential in order to prevent major disturbance to metabolism by precipitation of inorganic P, or competition with Mg at binding sites, thereby inhibiting the activities of essential enzymes. It is also a prerequisite for the role of Ca as a second messenger in regulating cellular functions in response to many stimuli in plants which are indicated by changes in cytosolic Ca concentration. This role of Ca in signaling in plants is currently the subject of much research.

Since Ca is greatly restricted from the living parts of cells, it is present in only very low concentrations in the phloem sap. Thus once Ca has entered a leaf it remains there and cannot be translocated to meet the requirements of younger leaves or fruits or storage organs. Transport of Ca from shoot to root is virtually absent and, in order to meet plant demand and supply via the xylem, there must be a continuous supply of Ca to plant roots from the soil. When uptake is restricted, the growth of meristematic tissue in young roots and leaves is rapidly disturbed. Membrane function is impaired and cells become leaky. Root growth stops and young leaves are twisted and necrotic. An undersupply of Ca can occur to the storage tissues of many fruits and vegetables. Ca-related disorders include 'bitter pit,' where the flesh and surface of the apple are covered in brown necrotic spots, and 'blossom end rot' in tomatoes, where there is a cellular breakdown at the distal end of the fruit.

Magnesium

Magnesium in some respects is similar to Ca in that it is taken up as a divalent ion and is present in cell walls in a bound or exchangeable form. On the other hand, it occurs in the cytoplasm in high concentrations, where it activates many enzyme systems. It also forms a complex with ATP which can then bridge with an enzyme through Mg and a N atom of the enzyme protein to activate the enzyme. Mg thus participates in all cellular actions involving ATP. Also, unlike Ca, it is highly mobile throughout the plant.

Magnesium is generally present in plants in concentrations from 1–5 mg g^{-1} of the dry weight and plays a most important role in photosynthesis. As well as being present at the center of the porphyrin structure of the chlorophyll molecule, it is additionally required in the synthesis of this molecule. However, in leaves only about 25% of Mg as a maximum is bound as chlorophyll. It is also needed in ionic form to activate ribulose bisphosphate carboxylase in the primary fixation reaction of CO_2 in the stroma of the chloroplasts of C_3 plants. Magnesium is also essential in the maintenance of structure and conformation of nucleic acids.

Since magnesium is highly mobile in plants, visible symptoms of deficiency occur first in the older leaves, usually as interveinal chlorosis. Deficiency decreases size, as well as disturbing structure and function of chloroplasts of the parenchymatous cells of the leaf. Magnesium deficiency in crops is often encountered on acid, well-leached soils, and there is increasing evidence of deficiency in forest ecosystems in Central Europe associated with air pollution and soil acidification. In animal nutrition, intensive grassland management can lead to low availability of Mg in the herbage and the occurrence of the acute disease hypomagnesemia or grass staggers in dairy cows.

Iron

In well aerated soil Fe is mostly oxidized and is relatively insoluble with the concentration of complexed ferric ion in soil solution an order of magnitude lower than that required for optimal plant growth. Two distinct strategies exist by which plants lacking in Fe are able to increase the solublity and absorption of Fe from the soil. The first, Strategy −1, is found in all plants except the grass family (*Poaceae*). Uptake is in the form of Fe^{2+} the production of which is dependent on the activity of the plasma membrane bound Fe (III) reductase of root cells to reduce Fe^{III} complexes from the soil. Plants utilizing this strategy respond to a lack of Fe by changes in the physiology and anatomy of the roots which increase the ability of the plant to acquire Fe. These responses include enhanced ferric reduction capacity at the root surface, stimulation of H^+ efflux from the roots, induction of transfer cells and increased formation of root hairs. By contrast, plants using Strategy −11, the *Poaceae*, respond to a lack of Fe by releasing Fe-chelating substances of the mugeneic acid family of phytosiderophores.

(MAs) from the roots. These phytosiderophores solubilize soil inorganic Fe^{III} compounds by chelation to form Fe^{III}–MAs complexes which supposedly traverse the plasma membrane of the root cells by a specific transport system. The transport of Fe from root to shoot in the xylem has been shown in a number of plant species to be mainly as Fe^{III} dicitrate.

Iron undergoes alternate oxidation and reduction between Fe^{2+} and Fe^{3+} and is present in numerous enzymes and proteins transferring electrons in the photosynthetic and respiratory chains. Chloroplasts provide the location for about 80% of Fe in leaves, and lack of Fe primarily affects photosynthetic activity. Protein and lipid synthesis are impaired, resulting in a disturbance in chloroplast structure and development, particularly of the thylakoid membranes. Moreover iron is also specifically required in at least two steps in the biosynthesis of chlorophyll.

In accordance with a major function of Fe in chloroplasts, iron-deficient plants are characterized by interveinal chlorosis of the leaves similar to that of Mg deficiency but occurring first in the younger leaves because of the restricted mobility of Fe within the plant. In some cases the leaves become completely white and devoid of chlorophyll and at a later stage may develop necrosis. Some plant species, including members of the rose family and various fruit trees, are particularly sensitive. The deficiency is especially common in high-pH calcareous soils and known as lime-induced chlorosis, where HCO_3^- ions may restrict Fe uptake.

Manganese

Manganese occurs in soil in three oxidation states in soils, as Mn^{2+}, as Mn^{3+}, and Mn^{4+} in the form of insoluble oxides and is mainly taken up as Mn^{2+} released from the higher-valency oxides under reducing conditions. In its biochemical function, Mn very much resembles Mg, and both of these ions activate a number of enzymes, e.g., in the TCA cycle. Both are also able to complex with ATP and bridge it with an enzyme complex (phosphokinases and phosphotransferases). The most well known role of Mn is in the photolysis of water mediated by a Mn-containing enzyme complex attached to photosystem II. The water-splitting reaction is cyclic, whereby O_2 is evolved with the concomitant reduction of Mn^{IV} to Mn^{III}, which in turn is reoxidized on the transfer of electrons to photosystem II (Hill reaction). As expected, a lack of Mn first depresses chloroplast function. Oxygen evolution is depressed, then, as deficiency becomes more severe, chlorophyll formation is decreased and the ultrastructure of the thylakoid membranes drastically impaired. Plant response to deficiency is species-dependent. Dicotyledons show interveinal chlorosis of the younger leaves, whereas, in cereals, greenish gray spots occur on the more basal leaves (gray speck in oats). Soils associated with Mn deficiency include acid soils derived from parent materials low in Mn, as well as high-pH soils containing free carbonates, especially when rich in organic matter.

Zinc

Zinc is taken up by plants as the divalent ion Zn^{2+}. Although only a few Zn-containing enzymes are known in plants, there are numerous enzymes activated by Zn which makes it difficult to interpret some of the complex changes that can occur in Zn-deficient plants. For example the disturbance of protein synthesis in Zn-deficient plants can be caused by a marked decrease in ribonucleic acid (RNA) as a consequence of a lower activity of the Zn-containing RNA polymerase, or by a decrease in structural integrity of the ribosomes or by enhanced RNA degradation. Membrane integrity can be impaired by Zn deficiency since, in healthy plants, Zn preferentially binds to SH groups of the membrane to stabilize structure; but additionally Zn is required in the synthesis of the fatty acids in membrane lipids. The Zn-containing isoenzyme of SOD (Cu-Zn superoxide dismutase) plays an important role in detoxification of the superoxide radical (O_2^-), protecting membrane lipids and proteins against oxidation. Under Zn deficiency, Cu-Zn-SOD is reduced and, particularly when light intensity is high, elevated levels of superoxide radicals are produced which destroy the membranes. This is the cause of 'sunscald,' a disorder associated with Zn deficiency which affects many fruit and vegetable crops, where enhanced lipid oxidation in the leaves leads to stunted growth, the destruction of chlorophyll, and the appearance of necrosis. The lower concentrations of the growth hormone indole acetic acid (IAA) in Zn-deficient plants is also probably related very closely to the well-known deficiency symptom of 'little leaf' and 'rosette' of apples, which describes the reduction in growth of young leaves and stem internodes. Deficiency occurs on high-pH soils and in crop plants symptoms of interveinal chlorosis of the leaf are common. In the monocotyledons, maize and sorghum chlorotic bands appear on either side of the midrib which later become necrotic.

Copper

Copper is similar to Fe in that it is present in redox enzymes and undergoes a valency change (Cu^+ and

Cu^{2+}) which allows electron transfer. It plays a role in photosynthesis as a constituent of plastocyanin, which accounts for about 50% of copper localized in the chloroplasts and enables photosynthetic electron transport. Also, in the form of the enzyme Cu-Zn-SOD, mainly located in the stroma of the choroplasts, Cu is directly involved in the detoxification of superoxide radicals which are generated during photosynthesis. Most of the other Cu-containing enzymes, for example, cytochrome oxidase, the terminal electron acceptor in the respiratory chain, react with oxygen to reduce it to H_2O_2. Other oxidases are phenolase and laccase, involved in the synthesis of quinones, melanins, and lignin.

Typical Cu-deficiency symptoms are especially expressed in the young shoot tissue because Cu is relatively immobile in plants. Chlorosis, white tip, necrosis, leaf distortion, and dieback are common. Enhanced tiller formation in cereals is a consequence of necrosis of apical meristems. Lignification is impaired, leading to lodging in cereals. The formation of vegetative tissue is less affected than the spectacular reduction in seed and fruit yields from Cu deficiency-induced male sterility.

Molybdenum

Molybdenum differs from the other micronutrients in that it is taken up from the soil as an anion (MoO_4^{2-}). Required in minute amounts, it is an essential component of two major enzymes in higher plants, nitrate reductase and nitrogenase, which is present in nodulated legumes and required in the N fixation process. The effective mechanism in both these enzymes depends on valency change of the Mo. Because of its role in NO_3 reductase, plants supplied with NH_4 have a lower requirement and deficiency symptoms are less severe or even absent than in plants supplied with NO_3-N.

Molybdenum deficiency, unlike that of all the other micronutrients, occurs on acid soils or soils high in iron oxides, which can fix Mo_4^{2-}. Also unlike the other micronutrients, deficiency symptoms are not confined to the younger leaves, since Mo is mobile within the plant. Deficiency can resemble that of N, with older leaves first showing chlorotic symptoms but with leaf margins becoming necrotic where NO_3^- accumulates. 'Whiptail' in cauliflower is one of the most well known symptoms, where the leaf lamina is not formed and only the leaf rib is present.

Boron

Boron is taken up as an undissociated boric acid. Uptake has long been held to be by passive following water flow, but recent evidence suggests an active component at low concentration of supply. Within the plant, B forms very stable complexes with organic compounds, including various sugars and their derivatives, uronic acid, and some *o*-diphenols all of which are abundant in cell walls.

Boron is the least well understood of all the plant nutrients and, in contrast to most other nutrients, is not the constituent of an enzyme. One of the most rapid responses to B deficiency occur in a matter of hours; when B is withheld there is a cessation in root growth. Cell wall development of apical meristems is disturbed, a response in accord with a function for B in the ultrastructural arrangement of the cell wall component in the apoplast (established in the late 1990s), the B-containing pectic polysaccharide B–rhamnogalacturonan II. This complex is composed of boric acid and two chains of pectic polysaccharides cross-linked through the borate diester bonding and forms a network of pectic polsaccharides, which is probably the ubiquitous form of B binding in the cell walls of higher plants.

The involvement of B in cell membrane function may also be inferred from the rapid recovery of metabolically linked ion transport following the addition of B to deficient roots. Additionally, membrane-bound ATPase activity, which is low in the B-deficient roots, is restored within 1 h of B application to the levels in the B-sufficient roots. Boron appears to stabilize the structure of the plasma membrane by complexing with membrane constituents such as glycoproteins and glycolipids, thereby keeping channels or enzymes at optimum conformation.

Because of the concentration of B at the cell wall–plasma membrane interface, primary effects of B deficiency on the morphology and physiology at this site are attracting increasing attention, including studies of accumulation of certain phenolics, inhibition of lignin synthesis, enhanced IAA activity, and decreased levels of diffusible IAA in response to B deficiency.

Boron deficiency is widespread. It is easily leached from soils and, as B availability decreases with increasing pH, it is common on alkaline and calcareous soils. In most plant species, B is immobile within the plant so that symptoms appear first at the apical growing points. However, in a number of species, including some fruit trees and vegetable crops, this is not the case, because B is phloem-mobile as it complexes with polyols, which are the primary photosynthetic product of these species. Boron is therefore more evenly distributed in these plants. An important role of B is in pollen germination and pollen tube growth. Both processes are highly sensitive to a lack of B as is also the viability of pollen grains. At low B supply therefore, seed,

grain, and fruit production are very much more affected than is vegetative growth.

Chlorine

Chlorine is an unusual essential element as it is abundant in the environment and is taken up in high amounts by plants as Cl^-, whereas the concentrations required in its role as an essential element are minute. Indeed preventing contamination was particularly difficult in the first experiments to demonstrate the essentiality of Cl. In leaves Cl is accumulated in the chloroplasts and it is required as a cofactor to activate the Mn-containing water-splitting system in photosystem II. Chloride also functions in osmoregulation and together with K is often the main osmoticum in plant cells. At low concentrations, where Cl acts as a micronutrient, these osmoregulatory functions are restricted to specialized tissues and cells, including extension zones in roots and guard cells, where the Cl may be concentrated to much higher levels than in the bulk tissue.

Typical symptoms of deficiency include wilting of leaves, curling of leaflets, bronzing and chlorosis, and severe inhibition of root growth. Some plant species such as palm trees and kiwi have a high demand for Cl but it is questionable whether deficiency ever occurs in the field.

Nickel

Recently it has been recognized that Ni can be added to the list of essential elements. For about 25 years it has been known that Ni is an essential part of the enzyme urease which catalyzes the hydrolytic breakdown of urea to NH_4^+ and CO_2. Experiments with soybeans and cowpea have demonstrated that, regardless of the source of N-nutrition (urea, NH_4-N, NO_3-N, or N fixation), in the absence of Ni in the nutrient medium, urea accumulates in the leaves, producing severe symptoms of leaf-tip chlorosis. There is other supporting evidence that urea is a normal intermediate in nitrogen metabolism which, by the presence of Ni, is maintained at a low concentration, thereby preventing toxicity. Convincing evidence that Ni is essential for barley also comes from germination experiments using seeds obtained after growing plants for three generations deprived of Ni. The seeds were nonviable but viability could be restored by soaking in a Ni-containing solution.

See also: **Nutrient Availability**; **Nutrient Management**

Further Reading

Ae N, Arihara J, Okada K, and Srinnivasan A (eds) (2001) *Plant Nutrient Acquisition: New Perspectives*. Berlin, Germany: Springer-Verlag.

Bergmann W (1992) *Nutritional Disorders of Plants – Development, Visual and Analytical Diagnosis*. New York: Gustav Fischer-Verlag.

Brown PH, Welsh RM, and Carey EE (1987) Nickel: a micronutrient essential for higher plants. *Plant Physiology* 85: 801–803.

Clarkson DT and Hanson JB (1980) The mineral nutrition of higher plants. *Annual Review of Plant Physiology* 31: 239–298.

Epstein E (1999) Silicon. *Annual Review of Plant Physiology Plant Molecular Biology* 50: 641–664.

Fageria NK, Baligar VC, and Jones CA (1991) *Growth and Mineral Nutrition of Field Crops*. Marcel Dekker. Inc.: New York, Basel, Hong Kong.

Gerandas J and Sattelmacher B (1997) Significance of Ni supply for growth, urease activity and the contents of urea, amino acids and mineral nutrients of urea-grown plants. *Plant and Soil* 190: 153–162.

Marschner H (1995) *Mineral Nutrition of Higher Plants*, 2nd edn. San Diego, CA: Academic Press.

McDowell LR (1992) *Minerals in Human and Animal Nutrition*. San Diego, CA: Academic Press.

Mengel K and Kirkby EA (2001) *Principles of Plant Nutrition*, 5th edn. Dordrecht, the Netherlands: Kluwer Academic Publishers.

Mordvedt JJ, Cox RF, Shuman LM, and Welch RM (eds) (1991) *Micronutrients in Agriculture*, 2nd edn. Madison, WI: Soil Science Society of America.

Xu G, Magen H, Tarchitzky J, and Kafkafi U (2000) Advances in chloride nutrition of plants. *Advances in Agronomy* 68: 97–150.

EUTROPHICATION

A J Gold, University of Rhode Island, Kingston, RI, USA
J T Sims, University of Delaware, Newark, DE, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Eutrophication describes a cascade of processes that occur in aquatic and terrestrial ecosystems in response to an increase in nutrient inputs. Phosphorus (P) and nitrogen (N) are the nutrients that drive most eutrophication processes. In aquatic environments eutrophication can lead to excessive algal growth, low levels of dissolved oxygen, the death of fish, increased turbidity, and a loss of species diversity. These conditions threaten the long-term sustainability of fisheries and recreational uses of surface waters. In terrestrial environments, excess nutrient inputs to soils in managed ecosystems (e.g., agriculture, urban horticulture) can increase the risk of nutrient losses to groundwaters and surface waters. This can then lead to eutrophication of surface waters and also to human health problems if drinking waters are too enriched in nitrate-N or contaminated by toxic by-products produced through chlorination of eutrophic waters. Overloading natural ecosystems (e.g., forests) with nutrients can lead to soil acidification and changes in the composition and diversity of native plant species, both of which are usually ecologically undesirable.

Because N and P are essential nutrients for the growth and well-being of plants and animals – and because they are relatively inexpensive – many human activities result in the discharge of N and P into some sector of the environment. Both nutrients are discharged from point sources such as municipal wastewater treatment plants and concentrated animal-feeding operations (CAFOs). Years of effort and billons of dollars have been invested in reducing point-source nutrient pollution of surface waters by improving the nutrient-removal efficiency of wastewater treatment plants. N and P are also common nonpoint pollutants that are widely added to soils as soil amendments for crop production (e.g., fertilizers, animal manures, and other by-products such as composts, and municipal wastewaters and biosolids) or via soil-based wastewater treatment systems such as septic systems. Combustion of fossil fuels generates biologically available N and contributes to atmospheric contamination of many watersheds. Soil characteristics (physical, chemical, and biological) and soil-management practices profoundly influence the potential for nonpoint-source pollution of water bodies by nutrients and also determine the response of terrestrial biota to nutrient additions. Because of this, the implementation of 'best management practices' (BMPs) that reduce N and P losses from agricultural soils by processes such as erosion, runoff, and leaching, is a high priority worldwide today.

Aquatic Eutrophication

Causes

Additions of nutrients can generate remarkable changes in the primary productivity (the generation of biomass through photosynthesis) of aquatic systems. Because so many aspects of a freshwater ecosystem can be traced to nutrient supply, nutrient status is the basis for the widely used trophic (i.e., nutrition level) classification system applied to water bodies. Low-nutrient water bodies are classified as 'oligotrophic' (poorly nourished); high-nutrient water bodies are termed 'eutrophic' (highly nourished), and the intermediate state is referred to as 'mesotrophic' (Figure 1). Extremely nutrient-rich conditions do occur and these water bodies are classified as 'hypereutrophic.'

Limiting Nutrients

The nutrient status of a water body is not constant. Decreases can occur due to control of point-source pollution discharges and widespread implementation of BMPs to curb nutrient inputs from nonpoint pollution sources. More often, the nutrient supply within an aquatic ecosystem increases over time. This increase is known as eutrophication – it can occur naturally in response to slow increases in the stores of organic matter and sediment within the system. If the rate of nutrient increase is accelerated due to human activities, the process is known as 'anthropogenic eutrophication.'

In the nineteenth century, Justus von Liebig developed the 'law of the minimum' to indicate that growth of most organisms is controlled by the 'limiting nutrient,' the nutrient in least supply. At the beginning of the twentieth century, Brandt extended Liebig's insights to surface water, noting that plankton (microscopic, free-floating plant and animal organisms that form the base of the food chain in aquatic environments) abundance was correlated with nutrient concentrations in freshwater lakes in Germany. Later, A.C. Redfield pointed out that the Liebig law should be viewed in the context of the relative ratios of nutrients found within living algae. Redfield found

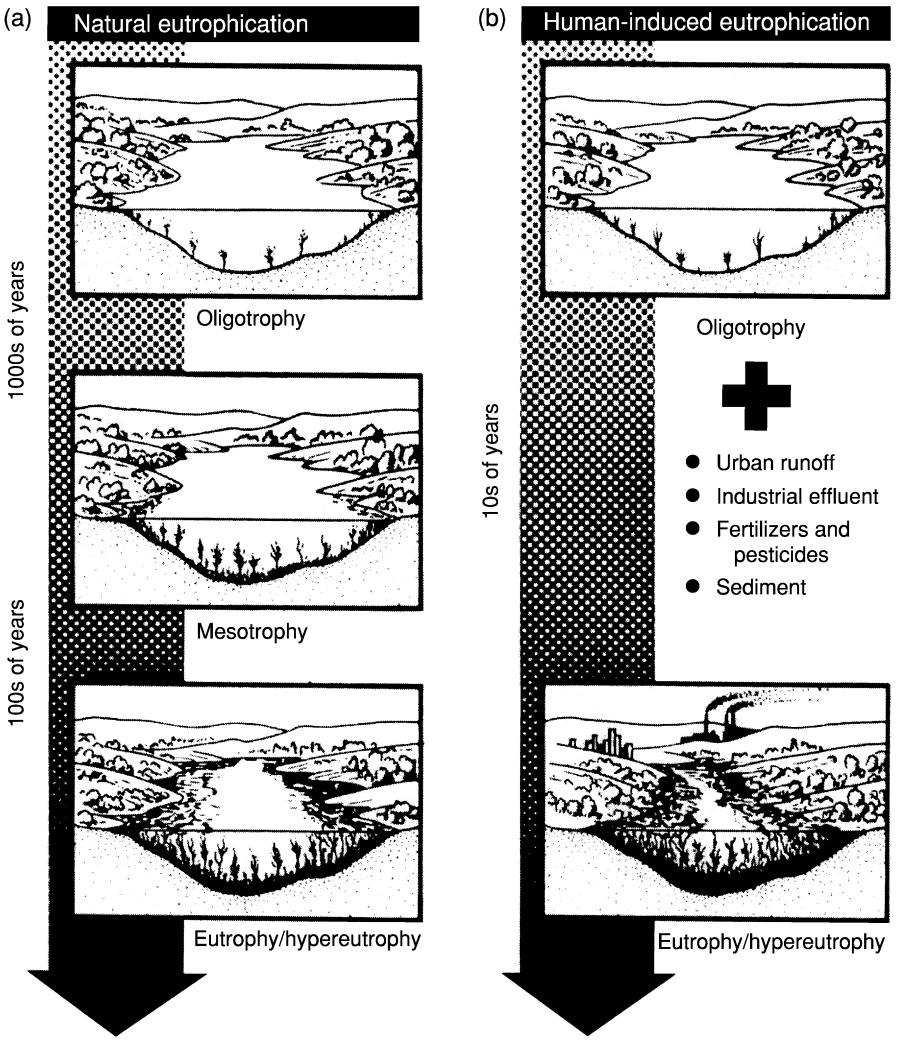


Figure 1 (a) Natural eutrophication describes the response of lakes to nutrient enrichment. Eutrophication increases primary productivity, leading to greater biomass and sedimentation. Oligotrophic lakes are nutrient-poor; eutrophic lakes are nutrient-enriched; (b) human activities induce anthropogenic eutrophication and generate rapid changes in nutrient enrichment and ecosystem characteristics. Reproduced with permission from NALMS (1990) *Lake and Reservoir Restoration Guidance Manual*. Madison, WI: North American Lake Management Society.

that aquatic algae and aquatic macrophytes are composed of fixed ratios of atoms:

106 C atoms:16 N atoms:1 P atom

By using the molecular weights of these atoms, we can translate these ratios into their mass ratios and express the composition in terms of the wet weight of living algae:

40 C:7 N:1 P per 500 wet weight of living algae

Thus, the law of the minimum suggests that additions of P will control primary production when the N:P mass ratio of available nutrients is greater than 7:1 and that the addition of 1 g of P can generate 500 g of algal biomass. If the N:P mass ratio is less than 7:1, N will be the limiting nutrient and additions of 1 g of N can be expected to stimulate 72 g of new algal biomass.

The limiting nutrient concept offers great practical value for the management of aquatic ecosystems. It suggests that decision-makers may be able to control ecosystem functions and values by managing inputs of a single nutrient. Today, most scientists agree that P is the limiting nutrient in freshwater ecosystems and that N limits primary production in coastal marine systems. C is rarely limiting, due to the presence of carbon dioxide in the atmosphere. These insights have emerged after heated debates within the scientific community, and exceptions to these 'rules' can be found in the literature.

Perhaps the most striking evidence for P limitation of lake eutrophication was demonstrated in a whole-lake study conducted in Canada by an international team in the 1970s (Figure 2). An oligotrophic lake with two similar basins was divided by a plastic curtain. One basin received additions of C and N, the other basin received C, N, and P. The basin receiving phosphorus developed intense blue-green algae blooms, while the other side remained clear and unchanged. The results of this experiment served to galvanize decision-makers around the globe to curtail P inputs into freshwater ecosystems.

A number of characteristics, such as rapid flushing rate, high depth-to-surface area ratio, or highly calcareous waters, can reduce the eutrophication response to nutrient inputs in freshwater lakes. Based on lake chemistry, morphology, and history of eutrophication, lakes can also develop internal stores of P that can continue to induce eutrophication even if external sources are controlled. P can accumulate on the bottom of lakes from the deposition of eroded soil particles transported to the lake by its tributaries. Accumulations of P in the form of organic sediments can also occur as eutrophication increases the quantity of organic matter (e.g., algae, macrophytes) in aquatic systems. Of equal importance, if aerobic conditions predominate in lake sediments and in the water overlying these sediments, dissolved P can chemically react with inorganic sediment constituents such as iron (Fe), manganese (Mn), and aluminum (Al) and form insoluble compounds that are much less bioavailable than soluble P. However, if anaerobic conditions develop in bottom waters and sediments, some of these compounds, especially iron phosphates, can dissolve and remobilize soluble P into the water column, where it will once again be available for uptake by algae and other aquatic species. In some cases, to mitigate the effects of these 'internal loads,' organic sediments are dredged to remove P physically from the lake, or chemically stabilized with additives such aluminum sulfate (alum) that precipitate soluble P into biologically unavailable forms.

Figure 2 Additions of small amounts of phosphorus to one section of Lake 226 in the Experimental Lakes Area of Ontario, Canada, caused extensive surface blooms of blue-green algae (top) and vividly demonstrated the importance of phosphorus as a cause of excessive algal growth or eutrophication. Photo from Fisheries and Oceans Canada, Experimental Lakes Area.

Both 'controlled addition' experiments and whole-ecosystem studies support the premise that N is the limiting nutrient to eutrophication in most coastal marine systems. Nitrogen additions in mesocosms simulating the Narragansett Bay of Rhode Island have generated increases in algal biomass. In addition, long-term data from Swedish estuaries show that changes in algal abundance are closely related to changes in N inputs rather than P inputs (Figure 3).

The amount of P present in a water body is partially controlled by internal factors that affect solubility and by the import of soil (erosion of particulate P) and water (dissolved P) from the watershed and any 'flushing' processes that subsequently export P and carry it further downstream. Biological processes do not create or reduce the mass of P within an aquatic ecosystem. In contrast, N can be continuously added or removed from the land and water by biological processes. N-fixing plants such as legumes (terrestrial) or blue-green algae (aquatic) transform atmospheric N_2 into organic forms of N, thus adding N to soils and/or waters. Denitrifying bacteria transform dissolved nitrate-N into gaseous forms (N_2O, N_2) and thus remove N from aquatic and terrestrial systems. Although N can accumulate in bottom sediments, the dynamics of the N cycle provide opportunities for N to leave aquatic systems in gaseous form, reducing stores of N within the system. Because gaseous losses of P do not occur, P tends to accumulate in sediments to a greater extent than N.

Several factors combine to produce low N:P ratios (i.e., below the Redfield mass ratio of 7:1) and N limitation in coastal marine waters versus higher N:P ratios and P limitation in freshwater lakes. First, the sources of nutrients to lakes and coastal marine waters differ. Both coastal and lake ecosystems receive nutrients from terrestrial and atmospheric sources. However, coastal systems also receive inputs from

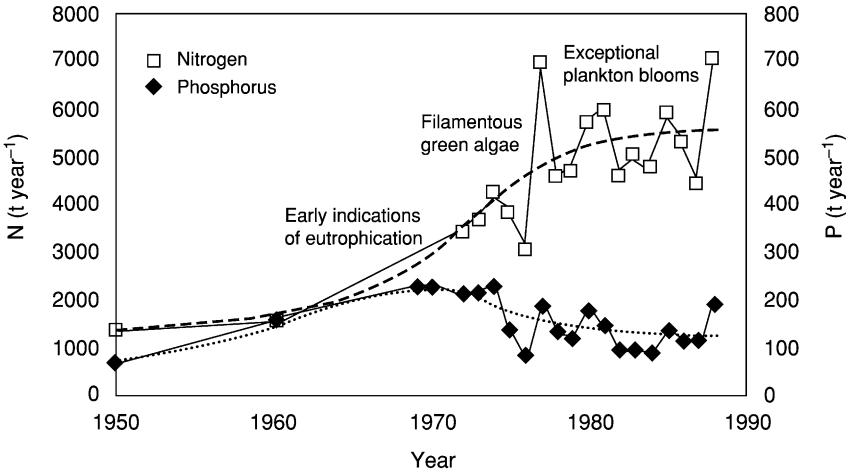


Figure 3 As evidenced by the biotic response of Laholm Bay in Sweden, nitrogen inputs generally control eutrophication in coastal marine waters. Reproduced with permission from Howarth RW, Anderson DM, Church TM *et al.* (2000) *Clean Coastal Waters*: *Understanding and Reducing the Effects of Nutrient Pollution*. Washington, DC: National Academy Press.

ocean waters with low N:P ratios due to denitrification on the continental shelf. Second, N fixation combined with slower flushing rates in lakes contributes to higher N:P ratios in lakes, whereas coastal marine waters experience little N fixation and any such additions are subject to rapid flushing from tides. Finally, sediment adsorption of P is often lower in coastal marine systems than in freshwater lakes, enhancing the relative supply of dissolved P in coastal waters.

Consequences of Aquatic Eutrophication

Stratification and Oxygen Depletion

The stimulation of primary production by nutrients can create algal blooms that reduce water clarity and lead to a number of other changes in aquatic ecosystems:

- Increased biomass of phytoplankton, suspended and attached algae;
- Decreased water column transparency;
- Shifts in phytoplankton composition to bloom-forming species, many of which may be toxic or inedible;
- Accumulation of carbon within the system;
- Changes in vascular plant production and species composition;
- Decrease in living aquatic habitats, including sea grasses and coral reefs;
- Depletion of deepwater oxygen concentration, resulting in hypoxia;
- Changes in fish species and fish production;
- Taste, odor, and water supply-filtration problems;
- Decrease in aesthetic values.

Oxygen is consumed as algae die and are decomposed by respiring microorganisms. The resulting oxygen depletion can be a major problem in stratified aquatic systems (Figure 4). Summer stratification occurs in freshwater lakes deeper than 3–5 m as the upper waters absorb solar energy and become warmer and less dense than the cool bottom waters. When stratified, a layer of rapid temperature change known as the 'thermocline' or 'metalimnion' effectively isolates the bottom waters and prevents oxygen diffusion from the atmosphere to waters below the thermocline. In estuarine systems, bottom waters can be isolated from the surface by both temperature gradients and a 'halocline,' a zone of rapidly changing salinity. Terrestrial water inputs create a low-salinity layer that floats above a heavier higher-salinity layer that is dominated by incoming waters from the open ocean. Regardless of the cause of stratification, dead and decaying algal biomass from surface waters will settle and, as decomposition proceeds, the oxygen supply of bottom waters becomes severely depleted. Oxygen depletion results in hypoxia (low oxygen) or anoxia (no oxygen). Both conditions directly kill many aquatic species. Anoxia can also alter sediment chemistry: reduced conditions can lead to sediment release of P, and production of hydrogen sulfide and methane, all contributing to water-quality deterioration. Hypoxia from eutrophication has damaged fisheries worldwide, from the Chesapeake Bay and Gulf of Mexico in North America, to the Baltic and Black Seas in Europe.

Changes in Species Composition

Eutrophication alters the composition and diversity of aquatic plants, affecting ecosystem structure and

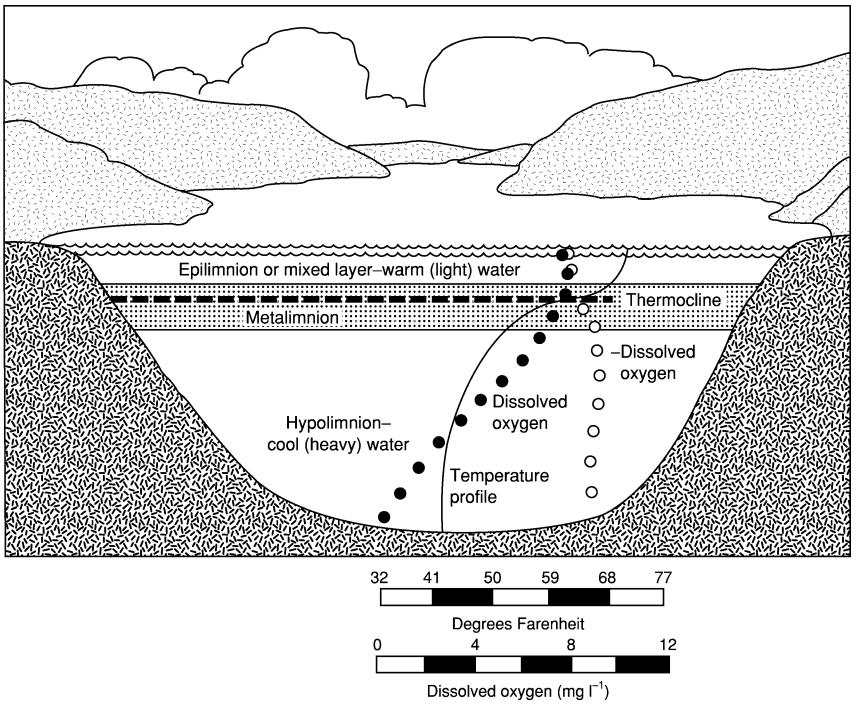


Figure 4 In temperate regions, most freshwater lakes deeper than 5 m undergo thermal stratification during the summer. Stratification separates a water body into distinct layers, isolating the cool bottom waters from atmospheric mixing and oxygen replenishment. Eutrophic lakes (solid circles) create substantial organic matter and its decomposition can deplete the stores of dissolved oxygen, leading to hypoxia or anoxia of the bottom waters. In contrast, oligotrophic waters generate little excess organic matter and minimal oxygen depletion occurs during stratification. Reproduced with permission from NALMS (1990) *Lake and Reservoir Restoration Guidance Manual*. Madison, WI: North American Lake Management Society.

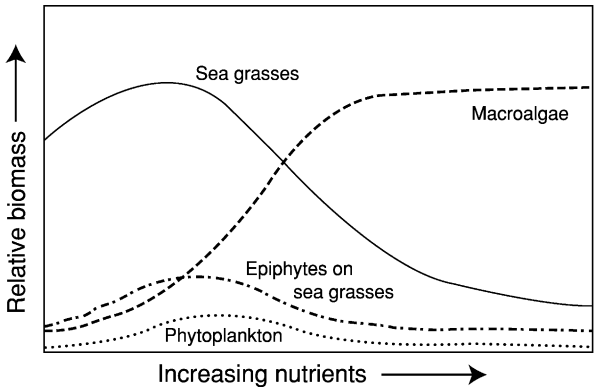


Figure 5 Increasing nutrients in shallow marine systems can shift aquatic plant communities from sea-grass beds that provide valuable habitats for marine organisms to nuisance macroalgae that cover the sediment with mats of rotting biomass. Nutrient enrichment stimulates the growth of phytoplankton in the water column and attached algae (epiphytes) on the sea grass, limiting light penetration below levels for sea-grass sustainability. Reproduced with permission from McComb AJ (ed.) (1995) *Eutrophic Shallow Estuaries and Lagoons*. Boca Raton, FL: CRC Press.

the food web (Figure 5). Increased inputs can shift algal composition in a freshwater lake from diatom-dominated systems, typical of oligotrophic lakes, to blue-green algae-dominated systems. Blue-green algae release toxins and are not readily ingested by secondary consumers. In addition many blue-green algae contain gas-filled vacuoles, causing the algae to float and accumulate on the water surface, effectively shading the lower waters and eliminating many important submerged plant species. Rotting masses of blue-green algae washed up on the shoreline of previously clear lakes is a discouraging sign that accelerated eutrophication has overtaken a lake's ecosystem. In coastal marine estuaries and bays, eutrophication has been linked to harmful algal blooms – often called 'red tides' – that cause widespread fatalities in fish and other marine organisms.

Sea-Grass Destruction

In shallow estuaries, increased nutrient inputs threaten the viability of sea grasses that serve as critical living spawning and nursery habitats. Sea-grass destruction rates rival the loss of tropical forests, creating serious problems in the Baltic Sea, Gulf of Mexico, and estuaries along the east coast of the USA. Sea-grass survival and growth are closely linked to light penetration. Elevated N inputs indirectly decrease light by stimulating both the growth of phytoplankton in the water column and attached algae (epiphytes) directly on

sea-grass leaves. Because sea grasses stabilize bottom sediments, declines in sea-grass beds can increase suspended sediments, further decreasing water clarity and hastening the loss of submerged aquatic vegetation. Thus, restoration of sea grasses requires control of N inputs and stabilization of bottom sediments.

Eutrophication can stimulate nuisance blooms of benthic macroalgae (seaweeds) that have lower light requirements than sea grasses. The macroalgae create thick extensive mats over sea-grass beds and the sediment surface, effectively reducing habitats for fish species that require mineral substrates to spawn. During periodic die-off cycles, macroalgae decay results in oxygen depletion, fish kills, and accumulation of rotting algae along the shoreline.

Degradation of Corals

Nutrient enrichment is also a threat to coral reefs, with both N and P implicated in coral reef degradation. Although coral reefs are among the most productive ecosystems on the planet, they thrive in high-clarity, nutrient-poor environments. Only slight nutrient enrichment of coral reefs is needed to stimulate the growth of attached macroalgae, which inhibits coral propagation. In addition, macroalgae blooms generate periodic hypoxia on the reef. Nutrient degradation of coral reefs is a global problem, with notable examples reported from the Great Barrier Reef, the Caribbean, and Hawaii.

Modeling Eutrophication

Phosphorus-Limited Freshwater Systems

Reducing nutrient inputs to aquatic systems can reverse the effects of eutrophication. Great progress has been made in our ability to predict the extent of lake eutrophication through models that link P loading (a term used to describe nutrient input: mass per time) to hydrologic and morphologic lake characteristics, such as hydraulic residence time (years), mean depth, and lake volume. These models derive from the work of R.A. Vollenweider.

Two different equations can be used to predict lake response to P inputs. The first predicts mean lake P concentrations in a lake:

$$P = P_i/(1 + T^{0.5})$$

This equation predicts that mean lake P concentration (P, in milligrams per cubic meter) will increase in proportion to the inflow concentration of phosphorus (P_i; milligrams per cubic meter), which is derived from P loading per outflow). P is expected to decrease with increasing hydraulic residence time (T, in years, obtained from lake volume per outflow).

The second predicts algal biomass (commonly quantified by the concentration of chlorophyll a (Chl a; in milligrams per cubic meter) – a photosynthetic pigment found in algae) – written as a function of mean P:

$$\text{Chl a} = 0.68P^{1.46}$$

The use of these equations can provide valuable insight into the required reduction of P loading needed for lake and watershed management efforts to improve water quality.

Nitrogen-Limited Coastal Marine Systems

Because of the diversity and complexity of coastal marine systems, no single modeling approach can capture the range of responses of marine ecosystems to nutrient loading. A variety of modeling approaches are now underway and reflect basic differences in characteristics such as physical setting (i.e., fjords, shallow estuarine lagoons, or river estuaries), biological communities (i.e., mangrove swamps, sea grasses, or planktonic systems), and hydrodynamics (i.e., watershed inputs, tidal flushing, water retention, and stratification patterns). Some of these models have been tested for selected locations and have proven quite useful for the prediction of sea-grass response and phytoplankton production to nutrient enrichment.

Terrestrial Eutrophication

Nutrient enrichment can also affect both managed and natural terrestrial systems. In managed systems such as agriculture, the most common cause of nutrient enrichment is overapplication of fertilizers, manures, or other organic by-products. Overapplication frequently occurs where there is an inadequate land base to assimilate the amount of organic by-product produced (e.g., sewage sludge generation in urban areas or manure production from CAFOs that are geographically concentrated on a small land base). As a result nutrient sources are applied at rates and times of the year that overwhelm the capacity for plant uptake. P accumulates in soils to values well above those needed for crop production. For both N and P, overapplication increases the potential for losses to ground and surface waters by leaching and runoff. In natural ecosystems such as forests, regular additions of nutrients in fertilizers and organic by-products are rarely made. However, because the atmosphere has become enriched with plant-available forms of N as a by-product of fossil fuel combustion, terrestrial eutrophication of natural ecosystems has become a global problem. Atmospheric loading of plant-available N can exceed

40 kg ha^{-1} per year within areas down-gradient of industrial zones, such as in southern California and northern Europe. These levels are more than 10 times greater than those found in much of the continental USA and represent a major increase over preindustrial times. High-density animal production systems also contribute to atmospheric loading of N. These locales can generate elevated emissions of gaseous ammonia-N that results in highly enriched atmospheric deposition of N to adjacent soils and surface waters. The long-term sustainability of natural ecosystems is threatened by these elevated inputs. In experiments conducted on grasslands in Europe and North America, increasing atmospheric deposition of plant-available N creates substantial declines in the species diversity of plants and insects.

Forested ecosystems can become 'N-saturated' as a result of eutrophication; atmospheric deposits of N can exceed the uptake capacity of the forest vegetation and soils. The excess N often converts to nitrate-N, a soluble anion that moves rapidly through soils and then travels to water bodies via groundwater discharge and surface runoff, thus contributing to eutrophication of aquatic systems. The transformation of nitrate from these external inputs also generates soil acidity that can contribute to depletion of essential cations in soils, particularly calcium (Ca) and magnesium (Mg), in turn diminishing the long-term productivity of the forest ecosystem. The increase in soil acidity can also mobilize Al ions and threaten aquatic life owing to the toxicity of Al to aquatic organisms.

Watershed Export and Models

The modern era is generating unprecedented levels of nutrient loading to surface waters. In the past 100 years, riverine discharge of N to large marine systems has increased four- to 10-fold, while P discharge has risen threefold. These increases in loading coincide with the expansion of human-generated nutrient inputs to the terrestrial environment. In particular, the global supply of available N has doubled since World War II, through production of synthetic N fertilizers, increased combustion of fossil fuels, and the expansion of land areas devoted to the cultivation of N-fixing leguminous crops.

Soils and soil management can play a profound role in the export of terrestrial nutrients to surface waters. Nutrient-balance studies conducted on watersheds ranging from local (i.e., less than 1000 ha) to regional scales (i.e., Baltic or Mississippi River Basin) demonstrate that rivers discharge only 10–30% of nutrient inputs to the watershed (Figure 6). Scientists now recognize the capacity of terrestrial environments to serve as nutrient 'sinks' that retain, remove, or transform nutrients and mitigate the effects of nutrient loading to surface waters. The soil N and P cycles contain a robust array of physical, chemical, and biological processes and transformations that control the fate of the terrestrial N and P inputs. Enormous quantities of N are immobilized by soil microbes and stored in soil organic matter. Agricultural soils retain 2000–5000 kg N ha^{-1}, and even more N can be retained in the organic matter of

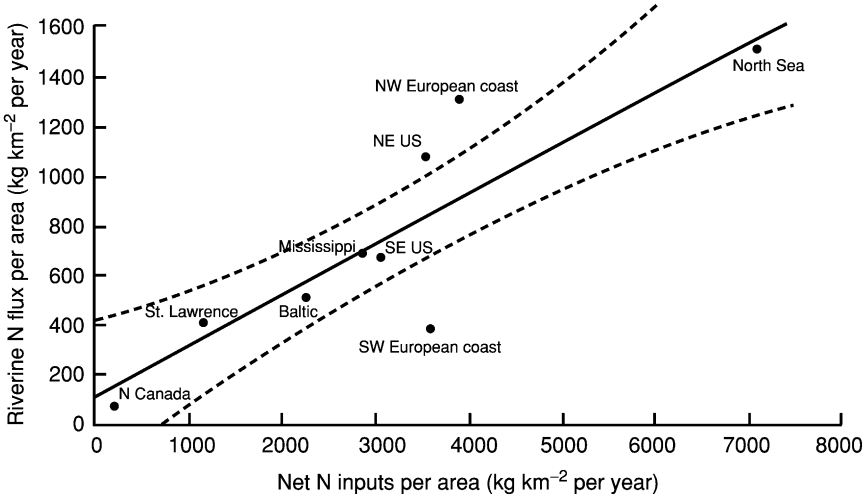


Figure 6 Nitrogen-balance studies conducted on large watersheds demonstrate that rivers discharge only 10–30% of nitrogen inputs. Transformations within the soil play an important role in reducing the effects of human-generated inputs to coastal marine waters. Reproduced with permission from Vitousek PM, Aber J, Howarth RW *et al.* (eds) (1997) Human alteration of the global nitrogen cycle: causes and consequences. *Issues in Ecology 1*.

wetland, forest, and grassland soils. Unfortunately, forest soils reaching N saturation will cease to function as N sinks, increasing the delivery of N to coastal waters.

Large quantities of P can be stored in soils by chemical fixation processes with soil constituents, such as Al, Ca, Fe, and organic matter. As with N, some soils in Europe and the USA are now becoming saturated with P due to long-term overapplication of fertilizers and animal manures, and may become sources of P in the future. Over time, human activity can alter the N stores in organic matter and the potential for losses of 'fixed' P in soils. Organic N can be mineralized and released through disturbance such as artificial drainage of wet soils, or conversion of forests and grasslands to routinely plowed croplands. P can be lost from soils as particulate P by erosion, a major problem worldwide, or as soluble P by leaching and surface runoff.

The control and reversal of aquatic eutrophication often rely on models to evaluate nutrient flux from forest, wetlands, riparian zones, and agricultural lands, in response to different management practices. Many modeling approaches are in use and reflect differences in the goals, setting, and scale of the simulation. Mathematical representations of soil hydrology, sediment transport, and biogeochemistry constitute the building blocks of these models – and the modeling approaches range from simple statistical relationships to highly parameterized mechanistic equations. Because of the inherent spatial variation in soil properties combined with the coarse scale of many available spatial databases, modelers recommend that indicators of uncertainty, such as confidence levels, be incorporated into estimates of nutrient export and that the complexity of the model match the available data and needs of decision-makers.

P flux models typically focus on soil processes and weather conditions that affect the movement of soluble and sediment-bound P forms in overland flow. In contrast, N export is usually dominated by groundwater flux, and models track nitrate losses by simulating N retention time, plant uptake, and microbial transformations within the root zone, and the timing and quantity of groundwater recharge out of the root zone.

Soil hydrologic models are central to nutrient export, because soil-moisture controls recharge to groundwater, rainfall/runoff relationships, and the extent of oxidized and reduced environments within the soil. Soil-moisture models typically generate a water balance based on infiltration, overland runoff, evapotranspiration, groundwater recharge, and changes in moisture storage within soil profiles in response to precipitation. Soil-moisture models range from mechanistic, highly parameterized schemes such as the Richards equation, to simple 'capacity' models that estimate soil-water flux in response to rules involving soil constants such as field capacity, permanent wilting point, and depth of the root zone. Intermediate, 'functional' models such as the SLIM model, developed by Addiscott, combine rate-based and capacity approaches and permit evaluation of nutrient dynamics in slowly draining soils. Once the soil-moisture status of the upper soil is established, overland runoff is often simulated through the curve number approach of the US Soil Conservation Service or through Green-Ampt techniques. Models such as AGNPS, developed by the USDA Agricultural Research Service, simulate runoff and sediment transport either in response to large 'design' storms or rely on daily soil-moisture balance approaches to estimate overland runoff.

Controlling Eutrophication: Future Trends

Natural and anthropogenic processes can mitigate nutrient losses and thus the potential for eutrophication. For example, advances in our understanding of soil denitrification can reduce the flux of nitrate from watersheds. This process generally occurs under anaerobic conditions where labile carbon is available to serve as the electron donor. Water table management through controlled drainage creates periodic drying and wetting cycles on cropland and shows promise for reducing nitrate export by stimulating nitrification (the microbial transformation of ammonium to nitrate that occurs in aerobic soils) and denitrification within the same location in soils. In addition, advances in tillage and crop residue management can take advantage of research that suggests that small fragments of easily decomposable organic matter can create localized 'hotspots' of respiration and denitrification, even in aerobic soils. In many watersheds denitrification in wetland soils receiving inputs of nitrate-laden groundwater is a major N removal process, particularly along corridors of undisturbed riparian (i.e., riverine) vegetation that intercept groundwater before it recharges streams. Thus, destruction of wetlands can result in disproportionate increases in nutrient loading due to the loss of the nutrient retention function of those locations. Conversely, reforestation of degraded lands can slowly sequester N. Efforts are under way to reduce P losses through the implementation of soil conservation BMPs (i.e., buffer strips, grassed waterways, terraces) that reduce soil erosion, the major source of P to surface waters. Clearly, comprehensive nutrient-management plans designed to prevent overapplication of manures and fertilizers is a critical aspect of

nutrient control. Equally important to the control of anthropogenic eutrophication is the recognition of the limits of soil-based disposal of organic by-products from agricultural activities – and the development of value-added products for those wastes, such as composts and pelletized manure-based fertilizers, or fuels for 'bioenergy' plants.

List of Technical Nomenclature

Chl a	Chlorophyll a (mg m^{-3})
P	Mean lake phosphorus concentration (mg m^{-3})
P_i	Inflow concentration of phosphorus (mg m^{-3})
T	Hydraulic residence time (years)

See also: **Denitrification**; **Macronutrients**; **Nitrogen in Soils:** Cycle; **Nutrient Management**; **Phosphorus in Soils:** Overview; Biological Interactions; **Pollution:** Groundwater; **Septic Systems**; **Watershed Management**

Further Reading

Aber JD and Melillo JM (1991) *Terrestrial Ecosystems*. Philadelphia, PA: Saunders College Publishing.

Addiscott TM, Whitmore AP, and Powlson DS (1991) *Farming, Fertilizers and the Nitrate Problem*. Wallingford, UK: CAB International.

Doering PH, Oviatt CA, Nowicki BL, Klos EG, and Reed LW (1995) Phosphorus and nitrogen limitation of primary production in a simulated estuarine gradient. *Marine Ecology Progress Series* 124: 271–287.

Harlin MM (1995) Changes in major plant groups following nutrient enrichment. In: McComb AJ (ed.) *Eutrophic Shallow Estuaries and Lagoons*, pp. 173–187. Boca Raton, FL: CRC Press.

Howarth RW, Anderson DM, Church TM *et al.* (2000) *Clean Coastal Waters: Understanding and Reducing the Effects of Nutrient Pollution*. Washington, DC: National Academy Press.

NALMS (1990) *Lake and Reservoir Restoration Guidance Manual*. Madison, WI: North American Lake Management Society.

Nixon SW (1995) Coastal marine eutrophication: a definition, social causes and future concerns. *Ophelia, International Journal of Marine Biology* 41: 199–219.

Smith VH (1998) Cultural eutrophication of inland, estuarine and coastal waters. In: Pace ML and Groffman PM (eds) *Successes, Limitations, and Frontiers in Ecosystem Science*, pp. 7–49. New York: Springer-Verlag.

Vallentyne JR (1974) *The Algal Bowl – Lakes and Man*. Miscellaneous Special Publication 22. Ottawa: Department of the Environment.

Ver LMB, Mackenzie FT, and Lerman A (1999) Biogeochemical responses of the carbon cycle to natural and human perturbations: past, present, and future. *American Journal of Science* 299: 762–801.

Vitousek PM, Aber J, Howarth RW *et al.* (eds) (1997) Human alteration of the global nitrogen cycle: causes and consequences. *Issues in Ecology 1*.

Wetzel RG (2001) *Limnology: Lake and River Ecosystems*, 3rd edn. New York: Academic Press.

EVAPORATION OF WATER FROM BARE SOIL

C W Boast and F W Simmons, University of Illinois, Urbana-Champaign, IL, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Evaporation of water from bare soil is often an important component of the soil water balance. If plants are living in the soil, evaporation of water from plant surfaces and evaporation from the soil are conceptually distinguished by calling the former transpiration and the latter evaporation, although both are evaporative processes – that is, both are characterized by a transformation of liquid water into water vapor. The combination, transpiration from plants and evaporation from soil, is called evapotranspiration. The bare soil surface of interest here can range from an entire plant-free (fallow) field to a small area between plants; for example, between trees in an orchard or between plant rows in a row-cropped field.

Evaporation Rate and Cumulative Evaporation

For an area of bare soil, it is often useful to know, at time t, the rate, $E(t)$ at which water is evaporating, per unit area of soil surface (quantity of water evaporating per square meter per second). The quantity of water is sometimes expressed by its mass, but more commonly by its volume, in which case the units of the evaporation flux density, $E(t)$, are (cubic meters per square meter per second = meters per second).

Alternatively, the evaporation process can be quantified by E_{cumul}, the cumulative amount of evaporation (per unit area of soil surface) which has occurred since some baseline time t_0. The relationships between $E(t)$ and E_{cumul} are:

$$E_{cumul} = \int_{t_0}^{t} E(t)\mathrm{d}t \quad \text{and} \quad E(t) = \frac{\mathrm{d}E_{cumul}}{\mathrm{d}t} \qquad [1]$$

If water volume is used to quantify the amount of water evaporating, that is, if $E(t)$ is expressed in meters per second, then cumulative evaporation is expressed in meters.

Processes Involved in Evaporation, and Latent Heat of Evaporation

In order for liquid water to change to water vapor, at a given location, three processes must occur: liquid water must be transported to, water vapor must be transported away from, and energy must be transported to the location. The relationship between the mass of water which is converted from liquid to vapor form and the amount of energy required is quantified by the latent heat of vaporization of water, denoted L_e. For pure, free water at 20°C, the value of L_e is $2.45 \times 10^6\,\mathrm{J\,kg^{-1}}$. Since L_e decreases only slightly with rising temperature, by approximately 0.1% per degree centigrade, L_e is often taken as constant over the range of temperatures encountered in natural systems.

The ultimate source of the energy which causes water to evaporate from a bare soil is electromagnetic radiation from the sun. The solar constant quantifies the average rate that solar (short-wave) radiation strikes the Earth's atmosphere: $1370\,\mathrm{W\,m^{-2}}$. The rate that short-wave radiation strikes the Earth's surface is denoted $J_s\downarrow$ (watts per square meter) and amounts to approximately one-half of the solar constant for a clear atmosphere, with the sun directly overhead. Integrated over a 24-h period, under more typical conditions, the daily amount of short-wave radiation reaching a bare soil surface can be as large as $20\text{–}30 \times 10^6\,\mathrm{J\,m^{-2}}$ or more. If this radiation load were to evaporate water at the rate $L_e = 2.45 \times 10^6\,\mathrm{J\,kg^{-1}}$, then, in order of magnitude the daily amount of water evaporated would be $10\,\mathrm{kg\,m^{-2}}$. Dividing by the density of water, $1000\,\mathrm{kg\,m^{-3}}$, gives this approximate upper limit of daily water evaporation as 0.01 m. This 1-cm daily evaporation serves as a rough benchmark for evaporation measurements. If the rate of evaporation E is expressed as kilograms per square meter per second, L_eE represents the rate that energy is consumed in the evaporation process (watts per square meter). On the other hand, if E is expressed as meters per second, and L_e is expressed as joules per cubic meter, then L_eE again represents watts per square meter.

Net Radiation

Seldom, even under ideal conditions, does daily evaporation amount to 0.01 m, because many other factors besides the rate, $J_s\downarrow$, of short-wave radiation striking the soil surface influence the actual amount of evaporation. First, a fraction, denoted α, of incoming short-wave radiation is reflected, and the rate that energy is reflected, $J_s\uparrow$ (watts per square meter) equals $\alpha J_s\downarrow$. The fraction α is called the soil surface's albedo, and, by definition, $0 \leq \alpha \leq 1$. Typical values of α range from 0.1, for wet, rough, dark-colored soil with the sun overhead, to 0.4, for dry, smooth, light-colored soil.

In addition to the downward-directed and upward-directed short-wave radiation, the net effect of electromagnetic radiation on the evaporation process includes two other terms: upward-directed long-wave radiation (emitted by the soil surface), whose rate is denoted by $J_l\uparrow$ (watts per square meter), and downward-directed long-wave radiation (mostly emitted by clouds), which is denoted $J_l\downarrow$ (watts per square meter). Combining all these terms gives the rate of net radiation, R_n (watts per square meter), as:

$$R_n = J_s\downarrow\,(1-\alpha) - (J_l\uparrow - J_l\downarrow) \qquad [2]$$

During the daytime, the short-wave portion of this expression dominates, making the net radiation (directed toward the soil surface) positive and providing energy that can cause evaporation. At night the short-wave portion is essentially zero and the upward-directed long-wave radiation is usually greater than the downward-directed long-wave radiation (because the Earth's surface is usually warmer than the temperature of clouds or bare sky), so the net radiation (R_n, directed downward toward the soil surface) is usually negative. That is, the electromagnetic radiation energy balance at night is generally away from the soil surface, which provides no energy for evaporation, but provides an energy sink for vapor condensation at the soil surface and within the soil.

Energy Balance

Three possible fates of electromagnetic energy which arrives at the soil surface during the day are to heat the air (at a rate H, in watts per square meter), to heat the soil (G, in watts per square meter), and/or to evaporate water (L_eE, in watts per square meter). The idea that these are the only three possibilities is illustrated in Figure 1a and is expressed as an energy balance at the soil surface:

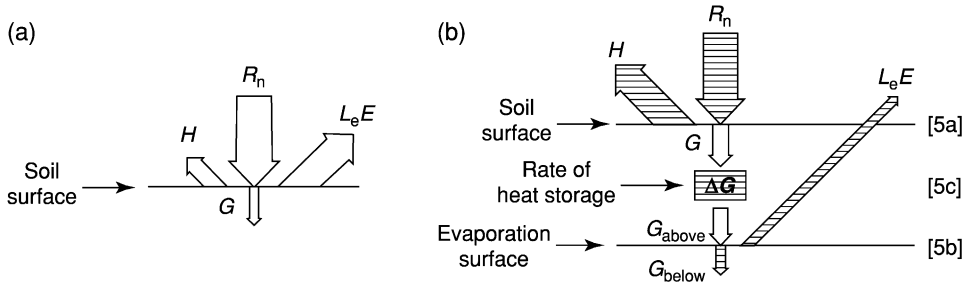


Figure 1 Energy balance (a) for the soil surface if all evaporation occurs there (Eqn [3]), and (b) (if all evaporation occurs at an 'evaporation surface' below the soil surface) energy balance for these two surfaces and the soil between them (Eqns [5] and (shaded) Eqn [6]).

$$R_n = H + G + L_eE \qquad [3]$$

The arrows in Figure 1 indicate the sign conventions, that is, an arrow points in the direction which is considered positive. During most of the daytime, each of the four terms in Eqn [3] is positive. At night, three of the terms, R_n, H, and G, are generally negative, reflecting the fact that heat is transferred toward the soil surface from air and soil, and that this heat is converted into outward-directed (long-wave) electromagnetic radiation. At night, E can be either positive (continued evaporation from the soil) or negative (dew formation), depending on the relative magnitudes of R_n, H, and G. Over any given 24-h period, the cumulative amounts of R_n, H, and E are usually positive, with day-long or multiday cumulative values of G tending to be positive during spring warming and negative during autumn cooling.

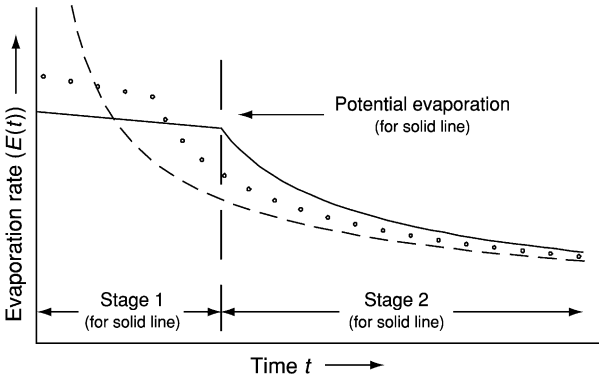


Figure 2 Schematics of (solid line) the two stages of evaporation from bare soil under conditions of constant potential evaporation, and (dashed line) soil evaporativity under conditions of unlimited potential evaporation. Dotted line is two-stage evaporation for slightly larger potential evaporation than solid line.

Stages of Evaporation

A striking feature of bare soil evaporation is illustrated by the solid line and curve in Figure 2. Under constant atmospheric and radiation conditions, evaporation from wet soil is observed to be nearly constant for a certain amount of time (stage 1, an almost horizontal line), then to decrease markedly (stage 2). The nearly constant rate of evaporation from bare soil during stage 1 is approximately equal to the rate of evaporation from an open body of water, or from a well-watered area with full plant cover, and this rate of evaporation is called the potential evaporation rate. On the other hand, atmospheric and radiation conditions vary over time, so the potential evaporation rate itself varies significantly, and evaporation from initially wet bare soil varies much more than is shown in Figure 2 during stage 1.

For both a constant and a varying potential evaporation rate, it has been observed that, after a period where evaporation of water from initially wet bare soil approximately equals the potential evaporation rate, the evaporation rate abruptly starts to fall below the potential evaporation rate. The time at which this occurs marks the boundary between stage 1 and stage 2 evaporation, and it is an important benchmark for describing the role of soil in the hydrologic cycle. One indirect way to detect when stage 2 begins is to measure the temperature of the soil surface. For a given set of atmospheric conditions, the daytime maximum surface temperature is generally higher, and the amplitude of the daily surface temperature fluctuation is larger during stage 2, than during stage 1. Also, for many soils there is a visible change in the color of the soil surface upon drying, and this change occurs at about the time when stage 1 evaporation ends and stage 2 begins. The change in color coincides with the disappearance of widespread water–air interfaces at the soil surface. (In other words, the lighter soil color indicates that the soil which is visible at the soil surface is 'dry.') Initially, the dry soil layer is very thin, and the dry layer becomes thicker over time.

Under certain circumstances a third stage of bare-soil evaporation occurs. For example, in the presence of a shallow water table whose depth is constant over

time, it is possible for a steady upward flow from the water table to the soil surface to occur. In this case, the evaporation rate decreases until it approaches this upward flow rate. Even if the approach is asymptotic, there may come a time after which, for all practical purposes, the evaporation rate is constant. This situation is sometimes called stage 3 evaporation. However, because the approach to constant evaporation rate is asymptotic, there is usually no clear-cut definition of the boundary between stage 2 and stage 3 evaporation.

If potential evaporation is a larger (but constant) value than that shown for the solid line in Figure 2, then initially the evaporation rate is larger (nearly horizontal dotted line). After a period of stage 1 evaporation (of shorter duration than for the solid line), the evaporation rate decreases (dotted curve) more rapidly than for the solid curve. Thus, depending on the potential evaporation rate, an entire family of curves can occur. For the idealized condition of an unlimited potential evaporation rate, represented by the dashed line in Figure 2, evaporation from initially wet bare soil starts out very fast and rapidly decreases.

Evaporation under Conditions of Unlimited Potential Evaporation

Evaporation of water from a bare soil, under conditions of unlimited potential evaporation, can be analyzed by considering what happens at the transition from stage 1 to stage 2 evaporation. Two events occur simultaneously, or nearly simultaneously: the evaporation rate suddenly starts to drop below the potential evaporation rate, and the soil surface water content changes precipitously from a wet value to a dry value. The larger the potential evaporation rate the sooner these events occur after the onset of evaporation. In the extreme, for an effectively infinite potential evaporation rate, it makes sense to visualize that stage 1 evaporation does not occur at all and that the surface soil water content changes from a wet value to a dry value at the onset of evaporation.

For an idealized, infinitely deep, soil with uniform initial ('wet') water content, θ_w, whose surface water content instantaneously changes to a dry value, θ_d, at $t_0 = 0$, this scenario is expressed as the following initial condition and boundary conditions:
Initial condition:

$$\theta = \theta_w \quad \text{at} \quad t = t_0 \quad \text{for} \quad 0 < z < \infty \qquad [4a]$$

Boundary conditions:

$$\theta = \theta_d \quad \text{at} \quad z = 0 \quad \text{for} \quad t_0 < t < \infty$$
$$\theta \to \theta_w \quad \text{as} \quad z \to \infty \quad \text{for} \quad t_0 < t < \infty \qquad [4b]$$

When the Richards equation, without its gravity term, is solved subject to these initial and boundary conditions, a remarkable result is obtained: regardless of the soil's hydraulic properties, the evaporation rate is inversely proportional to the square root of time, $E(t) = s(\theta_w, \theta_d)\, t^{-1/2}$. The proportionality constant, s, which is analogous to the sorptivity of infiltration theory, is a function (for a given soil) only of the initial water content, θ_w, and the surface water content, θ_d. This finding, although strictly applicable only (a) to a homogeneous soil, (b) of effectively infinite depth, (c) starting from uniform initial water content, and (d) uninfluenced by gravity, has been found to be applicable to situations which deviate greatly from these idealized conditions. And the seeming difficulty of an infinite initial evaporation rate ($t^{-1/2} \to \infty$ as $t \to 0$) is less troublesome when it is noted that (taking $t_0 = 0$, for simplicity of notation) the cumulative amount of evaporation after time t_0 is simply given by $2st^{1/2}$.

As a consequence of the simplicity of the square root of time behavior for evaporation under conditions of effectively unlimited potential evaporation, phase 2 evaporation has been modeled as a process whose rate equals $s'(t - t')^{-1/2}$, where s' and t' are parameters to account for the effect of stage 1 evaporation. One difficulty with this approach is that both s' and t' need to be chosen for each soil and each environmental situation, and it is very difficult to derive a theory which gives s' and t' as functions of measurable quantities.

Evaporation Below the Soil Surface

One of the salient features of evaporation from bare soil is the drying of the soil surface (first evident at about the time evaporation enters stage 2). As evaporation proceeds, a thicker and thicker layer of dry soil develops, and Eqn [3] ceases to be useful for determining the total evaporation rate. It can still represent the energy balance at the soil surface, but one of its terms, L_eE, represents only that part of the evaporation which occurs at the soil surface.

If the evaporation process is distributed over a range of depths, an exact statement of the energy balance must represent a range of depths. However, it has been observed that, at any given time, evaporation takes place over a quite narrow range of depths. Below this range of depths, water moves upward in liquid form, and above this depth range water moves upward as vapor. Because the range of depths is narrow, it is reasonable to approximate reality by supposing that all the evaporation occurs at a single depth (called 'the evaporation surface' or

'drying front' and shown in Figure 1b) and to write two energy balance equations:

$$\text{at the soil surface}: R_n = H + G \quad [5a]$$

$$\text{at the evaporation surface}: G_{above} = L_e E + G_{below} \quad [5b]$$

where G_{above} and G_{below} represent, respectively, the soil heat flux arriving at the evaporation surface, from above, and leaving the evaporation surface, to go below (both downward-directed). If ΔG represents the rate that heat is stored in the soil between the soil surface and the evaporation surface, then the equation:

$$G = \Delta G + G_{above} \quad [5c]$$

describes the relationship between G and G_{above}. These three equations contain seven variables, four more than the number of equations. This can be compared with four variables in Eqn [3], three more variables than the one equation in that case. Thus, the situation is more complicated than for Eqn [3]: there is one more unrelated variable. The three Eqns [5] can be combined to give:

$$R_n = H + \Delta G + L_e E + G_{below} \quad [6]$$

Two soil heat terms, ΔG and G_{below}, must be estimated in this equation, whereas only one such term, G, must be estimated in Eqn [3]. As is illustrated by the shaded arrows and the shaded rectangle in Figure 1b, Eqn [6] does not describe the energy balance at any particular location, because it contains terms at multiple locations; for example, R_n and H at the soil surface, and $L_e E$ and G_{below} at the evaporation surface. However, it does validly represent the relationship between energy terms at these two relevant locations. It is a tool for estimating the evaporation rate E (if effectively all evaporation occurs at an evaporation surface, below the soil surface), just as Eqn [3] is a tool for estimating E (if effectively all evaporation takes place at the soil surface).

Methods for Measuring Evaporation

Five strategies for measuring evaporation from bare soil are now described, three direct strategies and two approaches in which evaporation is indirectly measured by quantifying the amount of energy, $L_e E$, which goes into evaporating water (using either Eqn [3] or [6]). There are large equipment and labor costs for the direct methods, so a great deal of effort has gone into devising indirect methods to measure E.

Weigh the Loss of Water Mass in the Soil

One direct method is repeatedly to weigh a body of soil which is isolated from surrounding soil, but which is positioned such that the isolated soil is subjected to the same radiation and heat transport environment as the undisturbed soil. The soil plus its container is either called an evaporimeter or, to emphasize its separation from surrounding soil, a lysimeter. A microlysimeter consists of a cylinder (Figure 3a) that is small enough to be filled with undisturbed soil by pushing the cylinder into the soil (Figure 3b) and cutting off the soil at the bottom of the cylinder. To complete the installation, the bottom of a microlysimeter is covered (Figure 3c) such that the soil is hydrologically isolated from, but can be thermally connected to, the soil below, and the microlysimeter is weighed and then placed in a hole with the soil surface inside and outside the microlysimeter, as well as the cylinder rim, all at the same level (Figure 3d). Larger lysimeters, on the other hand, are often filled with disturbed soil that must be allowed to settle and reconsolidate, sometimes for years, before it can be considered representative of undisturbed soil.

The crucial property of an evaporimeter is that the rate of evaporation from the evaporimeter equal that from the nonisolated, undisturbed soil which the evaporimeter is supposed to mimic. The bottom of a microlysimeter blocks vertical flow of water, whether this is downward flow following rainfall or irrigation, or upward flow during a drying period. Thus the soil inside the microlysimeter can be either wetter or drier than nonisolated soil. Either way, the

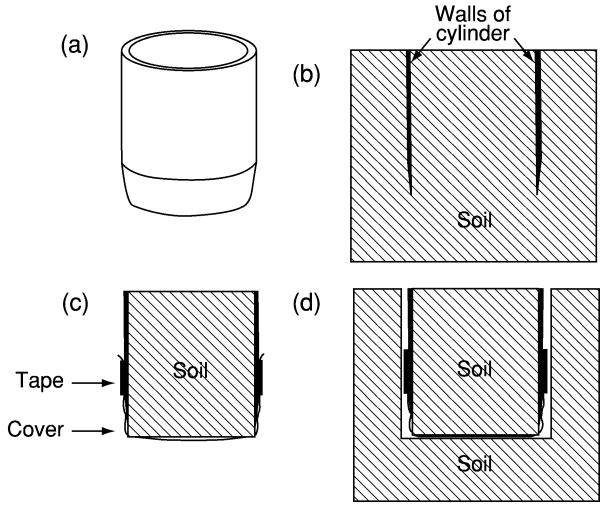


Figure 3 Procedure for microlysimeter determination of evaporation: (a) cylinder; (b) cylinder pushed into soil; (c) microlysimeter, partially taped, during preparation for first mass determination; (d) microlysimeter, in place, between mass determinations.

evaporation rate from the microlysimeter eventually deviates from what it is supposed to represent. And, in general, the shorter the microlysimeter the sooner this occurs. A rule of thumb is that the soil in a microlysimeter must be replaced (a) after any significant rainfall, and (b) almost every day during a drying period for a microlysimeter which is less than 0.1 m tall. The latter requirement can be relaxed to approximately weekly replacement for microlysimeters which are 0.3 m tall.

A variety of strategies have been employed to insure that a very large lysimeter, which is weighed *in situ*, does not deviate hydrologically from nonisolated soil. For example, if there is a shallow water table in the undisturbed soil, then the lysimeter can be equipped so that a water table is maintained at the same depth inside the lysimeter as the measured water table in intact soil. Once such a lysimeter is installed, it can be used indefinitely with a minimum of labor, whereas the labor cost of microlysimeters, which need to be periodically refilled with soil, can be very large. Microlysimeters have a low initial cost, and they can be used to measure evaporation for a small area of soil, for example to characterize bare soil evaporation in the space between plant rows for a row crop. A large lysimeter can characterize the combined effect of transpiration from plants and evaporation from soil, over an area which encompasses one or more row widths, vine spacings, or even a tree spacing, but large lysimeters generally cannot be used to determine separately water evaporation from soil and water transpiration from plants.

Detect the Arrival of Water Vapor Within the Air

Another direct method to quantify the rate of evaporation from bare soil is to measure the density of water vapor, ρ_v (kilograms per cubic meter) in a closed chamber containing air, above the soil surface, over a short interval of time. For example, if the evaporation rate is a rather low 0.06 mm h^{-1} (0.001 mm min^{-1}), then, for each square meter of soil surface area enclosed, 1 g of water vapor enters the air chamber per minute. The amount of water contained in air which is saturated with water vapor (at, for example, 15°C) is 12.7 g m^{-3}, so this rate of evaporation increases the relative humidity in, for example, a 0.8-m-tall chamber (which contains 10 g of water vapor per square meter at saturation) by 10% per minute. Since the evaporation rate can be 10 or more times as fast as in this example, the method can be quite sensitive for even a short closure time. Also, a reading must be completed fast enough so that the rising relative humidity of the air in the chamber does not reduce the evaporation rate.

Quantify the Upward Transport of Water Vapor in the Air

A somewhat less direct method to measure evaporation from a large field of bare soil is to use a model of vapor transport:

$$L_e E = L_e(\rho_{v,e} - \rho_{v,a})/r_v \qquad [7]$$

where E is expressed in kilograms per square meter per second, where $\rho_{v,e}$ and $\rho_{v,a}$ (kilograms per cubic meter) are the water vapor density at two positions, level 'a' in the air, and level 'e', either lower in the air than 'a' or at the evaporation surface, and where r_v (seconds per meter) is the resistance of the pathway from level 'e' to level 'a'. In Eqn [7], the parameter which is most difficult to quantify is the resistance factor, r_v. If 'e' represents the evaporation surface for stage 2 evaporation, the soil portion of r_v can be modeled as a resistance to diffusion in a dry soil layer between an evaporation surface and the soil surface. And a variety of methods have been devised to estimate the air portion of r_v; that is, to characterize the wind-driven and/or buoyancy-driven mechanisms for transport in the air near the Earth. Two key methods are: (1) to directly measure vertical air movements; and (2) to use equations which express r_v as a function of measured horizontal wind speed at one or more heights, sometimes with corrections for atmospheric stability conditions.

One disadvantage of Eqn [7], as well as the indirect methods described below, is that these approaches can be inaccurate for anything except very large, uniform fields. If they are attempted near the edge of a bare soil area, then the measurements can be strongly influenced by neighboring land use. For example, the required fetch distance, upwind of the measurement location, can be on the order of 100 times the neighboring plant height.

Measure Enough Elements in the Energy Balance to Calculate L_eE

In Eqn [3] or Eqn [6], R_n can be measured with a net radiometer, and the soil heat terms can be either measured or estimated. This leaves the sum $L_eE + H$ known (by difference, in either Eqn [3] or Eqn [6]), but the individual values of L_eE and H unknown. The most common scheme for breaking $L_eE + H$ into its two parts is practical only during stage 1 evaporation, but the first step in this scheme is more generally useful: the transport of heat into the air is modeled in a way which is very similar to how vapor transport is modeled in Eqn [7]:

$$H = \rho C_p(T_s - T_a)/r_h \qquad [8]$$

where T_s and T_a (degrees Kelvin) are the temperature at two levels with level 'a' in the air and with level 's' either lower in the air than level 'a' or at the soil surface (but not at the evaporation surface if this is below the soil surface), where ρ (kilograms per cubic meter) is the density of air, where C_p (joules per kilogram per degree Kelvin) is the specific heat of air at constant pressure, and where r_h (seconds per meter) is the resistance to heat transport from location 's' to location 'a'. As in Eqn [7], the entity which is most difficult to quantify is the resistance factor r_h.

One strategy for estimating evaporation is to define the Bowen ratio as $\beta = H/L_eE$, to replace H in Eqn [3] or Eqn [6] by βL_eE, and to solve for L_eE, giving, for Eqn [3], $L_eE = (R_n - G)/(1 + \beta)$, and giving a similar expression for Eqn [6]. If evaporation is occurring at the soil surface, during stage 1, then one strength of the Bowen ratio strategy is that (as portrayed in **Figure 1a**) H is usually smaller than L_eE, so β is small compared with 1, and even an approximate estimate of β can give a reasonably accurate estimate of the evaporation rate. A second strength of the Bowen ratio approach, for stage 1 evaporation, is that it is sometimes reasonable to assume that $r_h = r_v$. But only if level 's' in Eqn [8] is identical to level 'e' in Eqn [7]. Then β (the ratio of the right sides of Eqns [8] and [7]) can be determined from measurements or estimates of temperature and vapor density at the two levels.

These two strengths of the Bowen ratio approach are, for the most part, lost during stage 2 evaporation, in particular, when the factor r_v includes the resistance to vapor transport through a dry soil layer in addition to the resistance to transport through the air. Since heat transport, in the numerator of β, travels only through the air portion of this path, the factor r_h can be much smaller than r_v, and β can be large (or, as portrayed in **Figure 1b**, H can be much larger than L_eE, for dry soil). Thus, not only can one not assume that $r_h = r_v$ for stage 2 evaporation, the advantages which accrue when β is small, compared with 1, are lost.

Compare the Soil of Interest to a Completely Dry Soil

This method for measuring evaporation from drying bare soil requires establishment of a large body of dry soil, and taking enough measurements on the dry soil and on the drying soil to estimate the difference between their two energy balances. Combining Eqns [2] and [6] describes the energy balance of the drying soil:

$$H + \Delta G + L_eE + G_{below} = J_s\downarrow\,(1 - \alpha) - (J_l\uparrow - J_l\downarrow) \quad [9]$$

where it is assumed that all evaporation occurs at a single evaporation surface. (If the evaporation surface is at the soil surface, then ΔG equals zero and G_{below} could be replaced by G.) With the subscript 'zero' marking entities that are specific to the dry soil, the energy balance at the soil surface of the dry soil is:

$$H_0 + G_0 = J_s\downarrow\,(1 - \alpha_0) - (J_l\uparrow_0 - J_l\downarrow) \quad [10]$$

In this equation, there is no evaporation term, because there is no evaporation from the dry soil ($E_0 = 0$), and there are only two ways that the net radiation can be disposed of at the soil surface: heating air, at rate H_0, and heating soil, at rate G_0. Two terms in the right side of this equation are assumed to be the same for the dry soil as for the drying soil: the downward-directed short-wave and long-wave radiations, at rates $J_s\downarrow$, and $J_l\downarrow$.

Subtracting Eqn [10] from Eqn [9], and rearranging, gives:

$$L_eE = (H_0 - H) + (G_0 - \Delta G - G_{below}) + J_s\downarrow\,(\alpha_0 - \alpha) + \sigma(\epsilon_0 T_{s0}^4 - \epsilon T_s^4) \quad [11]$$

where the last term represents $J_l\uparrow_0 - J_l\uparrow$ (using the Stefan–Boltzmann law, with σ the Stefan–Boltzmann constant ($5.67 \times 10^{-8}\,W\,m^{-2}\,K^{-4}$), and with ϵ_0 and ϵ (dimensionless) the emissivities, and T_{s0} and T_s the soil surface temperatures of the dry soil and the drying soil, respectively). Because of evaporation, the surface temperature of the drying soil, T_s, is generally lower than that of the dry soil, T_{s0}, during the day. The dry soil's hotter surface causes more heating of the air there, so the first term in the right side of Eqn [11] is generally greater than zero during the day. Also, the solar reflection term is generally greater than or equal to zero, because the albedo of the dry soil, α_0, is greater than or equal to α for the drying soil. And, unless the emissivity of the dry soil is a lot smaller than that of the dry soil, the last term is greater than zero when $T_{s0} > T_s$. Generally, the only term in Eqn [11] whose sign is ambiguous during the daytime is the soil heat term. During the daytime the left side of Eqn [11] is positive, so, even if the soil heat term is negative, the other three terms in the right side of Eqn [11] more than compensate for this.

In general, two strategies have been employed to estimate the terms on the right side of Eqn [11]. In one strategy, the sum of the heat-flux term and the short-wave reflection term is assumed to be negligible compared with L_eE, and approximations are made in the other two terms so that their 24-h integrated value is proportional to the day's maximum $T_{s0} - T_s$ value. The advantage of this strategy lies in its simplicity, but these approximations are not

always valid. In the other strategy, each of the terms in Eqn [11] are estimated, and the net integrated effect is calculated.

In either strategy, the $H_0 - H$ term in Eqn [11] is estimated by use of Eqn [8] twice, once for the dry soil, and once for the drying soil, and, generally, r_h is determined from wind speed measurements. Two key simplifying assumptions are that the resistance factor, r_h, be the same for the dry and drying soils, and that the same T_a can be used for both soils. The former assumption is analogous to the assumption, $r_v = r_h$ (in Eqns [7] and [8]), in the previous method. The method is relatively new, and many methodological questions have not been thoroughly explored. However, Eqn [11] quantifies a key difference between a dry soil and a drying soil, namely, that heat which is consumed in evaporating water in the drying soil (left side of Eqn [11]) must go elsewhere for a dry soil (right side of Eqn [11]). Soil surface temperature plays an important role in all (except for albedo) of the processes during stage 2 evaporation, so measurement of the surface temperature can provide much information in this scheme, and other schemes, for estimating the evaporation rate.

List of Technical Nomenclature

α	Albedo of drying soil (dimensionless)
α_0	Albedo of dry soil (dimensionless)
β	Bowen ratio, H/L_eE (dimensionless)
ΔG	Rate of heat storage in zone between soil surface and evaporation surface ($W\,m^{-2}$)
ϵ	Emmissivity of drying soil (dimensionless)
ϵ_0	Emmissivity of dry soil (dimensionless)
θ	Volumetric soil water content (dimensionless)
θ_d	Volumetric soil water content at soil surface, during drying (dimensionless)
θ_w	Volumetric soil water content, before drying (dimensionless)
ρ	Density of air ($kg\,m^{-3}$)
ρ_v	Water vapor density ($kg\,m^{-3}$)
$\rho_{v,a}$	Water vapor density at level 'a' ($kg\,m^{-3}$)
$\rho_{v,e}$	Water vapor density at level 'e' ($kg\,m^{-3}$)
σ	Stefan–Boltzmann constant ($5.67 \times 10^{-8}\,W\,m^{-2}\,K^{-4}$)
C_p	Specific heat of air at constant pressure ($J\,kg^{-1}\,K^{-1}$)
E	Rate of water evaporation at soil or evaporation surface ($m^3\,m^{-2}\,s^{-1} = m\,s^{-1}$)
E_{cumul}	Cumulative water evaporation at soil or evaporation surface ($m^3\,m^{-2} = m$)
G	Rate of heat transfer from soil surface downward into drying soil ($W\,m^{-2}$)
G_0	Rate of heat transfer from soil surface downward into dry soil ($W\,m^{-2}$)
G_{above}	Rate of heat transfer arriving at evaporation surface from above ($W\,m^{-2}$)
G_{below}	Rate of heat transfer leaving evaporation surface downward ($W\,m^{-2}$)
H	Rate of heat transfer from soil surface of drying soil into air ($W\,m^{-2}$)
H_0	Rate of heat transfer from soil surface of dry soil into air ($W\,m^{-2}$)
$J_s\uparrow$	Upward-directed short-wave radiation ($W\,m^{-2}$)
$J_s\downarrow$	Downward-directed short-wave radiation ($W\,m^{-2}$)
$J_l\uparrow$	Upward-directed long-wave radiation from drying soil ($W\,m^{-2}$)
$J_l\downarrow$	Downward-directed long-wave radiation ($W\,m^{-2}$)
$J_l\uparrow_0$	Upward-directed long-wave radiation from dry soil ($W\,m^{-2}$)
L_e	Latent heat of vaporization of water ($2.45 \times 10^6\,J\,kg^{-1}$)
R_n	Net radiation ($W\,m^{-2}$)
r_h	Resistance to heat transport, from soil surface to position 'a' ($s\,m^{-1}$)
r_v	Resistance to vapor transport, from evaporation surface to position 'a' ($s\,m^{-1}$)
s	Proportionality constant for square root of time behavior during drying
s'	Parameter for stage-2-evaporation relationship
T_a	Temperature at level 'a' (K)
T_s	Temperature at surface of drying soil (K)
T_{s0}	Temperature at level 's' for dry soil (K)
t	Time (s)
t_0	Time at which evaporation commences (s)
t''	Parameter for stage-2-evaporation relationship
z	Depth (m)

See also: **Dryland Farming**; **Energy Balance**; **Evapotranspiration**; **Heat and Moisture Transport**; **Lysimetry**; **Radiation Balance**; **Thermal Properties and Processes**; **Thermodynamics of Soil Water**; **Water Cycle**

Further Reading

Ben-Asher J, Matthias AD, and Warrick AW (1983) Assessment of evaporation from bare soil by infrared thermometry. *Soil Science Society of America Journal* 47: 185–191.

Boast CW and Robertson TM (1982) A "micro-lysimeter" method for determining evaporation from bare soil: description and laboratory evaluation. *Soil Science Society of America Journal* 46: 689–696.

Brisson N and Perrier A (1991) A semiempirical model of bare soil evaporation for crop simulation models. *Water Resources Research* 27: 719–727.

Brutsaert W (1982) *Evaporation into the Atmosphere: Theory, History and Applications*. Norwell, MA: D. Reidel.

Evett SR, Matthias AD, and Warrick AW (1994) Energy balance model of spatially variable evaporation from bare soil. *Soil Science Society of America Journal* 58: 1604–1611.

Evett SR, Warrick AW, and Matthias AD (1995) Wall material and capping effects on microlysimeter temperatures and evaporation. *Soil Science Society of America Journal* 59: 329–336.

Hillel D (1998) *Environmental Soil Physics*. London, UK: Academic Press.

Lascano RJ and van Bavel CHM (1986) Simulation and measurement of evaporation from a bare soil. *Soil Science Society of America Journal* 50: 1127–1133.

Massman WJ (1992) A surface energy balance method for partitioning evapotranspiration data into plant and soil components for a surface with partial canopy cover. *Water Resources Research* 28: 1723–1732.

Monteith JL (1973) *Principles of Environmental Physics*. New York: American Elsevier.

Penman HL, Angus DE, and van Bavel CHM (1967) Microclimatic factors affecting evaporation and transpiration. In: Hagen RM, Haise HR, and Edminster TW (eds) *Irrigation of Agricultural Lands*, pp. 483–505. Monograph No. 11. Madison, WI: American Society of Agronomy.

Tanner CB (1967) Measurement of evapotranspiration. In: Hagen RM, Haise HR, and Edminster TW (eds) *Irrigation of Agricultural Lands*, pp. 534–574. Monograph No. 11. Madison, WI: American Society of Agronomy.

van Bavel CHM and Hillel DI (1976) Calculating potential and actual evaporation from a bare soil surface by simulation of concurrent flow of water and heat. *Agricultural Meteorology* 17: 453–476.

van de Griend AA and Owe M (1994) Bare soil surface resistance to evaporation by vapor diffusion under semiarid conditions. *Water Resources Research* 30: 181–188.

EVAPOTRANSPIRATION

G Stanhill, The Volcani Center, Bet Dagan, Israel

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Evapotranspiration is the total water loss to the atmosphere from a unit land surface area, usually expressed in units of depth; it includes the water vapor evaporating from the soil surface and from free water on plant surfaces as well as that transpired from within plant surfaces. The word 'actual' is sometimes added to distinguish it from 'potential' and 'reference evapotranspiration', terms used to describe the theoretical upper limit to total water loss from land surfaces under special experimental conditions.

Development of the Concept

'Evapotranspiration,' both actual and potential, was first defined by Thornthwaite in 1944, and the term became widely known and used following his 1948 publication in which potential evapotranspiration was calculated as a complex empirical function of air temperature and day length. However, the first published appearance of the word dates from 1937, albeit without explanation or definition and in a hyphenated form.

A major reason for the rapid and widespread adoption of the term was the success of the efforts made at this time, notably by Penman and later by Penman and Monteith, to estimate the total water loss to the atmosphere from standard climate measurements on a sound physical basis; another was the inadequacy of the methods then available to measure the different components of water loss under natural conditions.

However, Penman objected to the use of the word 'evapotranspiration' on the grounds that it was unnecessary, because 'evaporation' was an equally valid term for the separate components as well as for the total water loss to the atmosphere. There is historical

support for this position in that when the use of the word 'evaporation' was first recorded in the middle of the sixteenth century it also covered transpiration, water loss from plants. The converse is also true, and the broad and imprecise use of both these words has persisted, even in the scientific literature, for more than 400 years.

Within two decades, emphasis on the need to lump together all the components of evapotranspiration diminished and recognition of the need to split the components returned with the appreciation of the importance of biological factors in controlling transpiration and of the strong coupling between plant growth and crop yield and transpiration, but not evaporation. This linkage is a basic feature of most of the dynamic, climate-driven models now widely used to simulate plant growth and yield.

As the transpiration stream can now be continuously and accurately measured under field conditions by heat pulse tracing systems, it is possible to study the relationship between the two fluxes accurately, previously referred to as the transpiration ratio and later as water-use efficiency, in natural surroundings.

Methods of Measurement

Only three of the almost innumerable methods of estimating evapotranspiration that have been described can be regarded as representing direct measurement; all other approaches require values of the constantly changing water conductance characteristics of the land surface under study, which are very difficult to measure and therefore have to be estimated from other measurements.

As the flux of water in evapotranspiration from a land surface is between two and three orders of magnitude greater than the flux of carbon to the surface in dry-matter assimilation and similarly greater than the changes in the plant's water content, it follows that the rate of evapotranspiration can be directly measured with sufficient accuracy from changes in the mass of a sample land surface. Instruments for this purpose are known as weighing lysimeters, because measurement of drainage (the literal meaning of the word) is also required. Evapotranspiration measured with such an instrument is generally accepted as the standard against which other methods of measurement and estimation should be evaluated and calibrated. However, to serve as a standard, it is essential that the lysimeter contains a sample of the land surface whose water losses are fully representative of the surface of interest. To ensure that this is so requires that the depth and structure of the soil within the lysimeter, the height and density of the vegetation growing on it, and the microclimate above the lysimeter are indistinguishable from those of the surrounding surface it represents.

A range of weighing lysimeters meeting these demanding requirements has been described, including those with sufficient area and depth to support mature forest and orchard trees, those with undisturbed soil monoliths with water potential at their base matched to that of the surrounding soil, and instruments sufficiently sensitive to register the condensation of dew from the atmosphere. Unfortunately their cost, together with that of their installation and maintenance, has prevented the use of weighing lysimeters in sufficient numbers to measure the spatial variability of evapotranspiration.

Two meteorologic methods of directly measuring evapotranspiration are generally accepted as being of comparable accuracy with those obtained with weighing lysimeters; both of them have the advantages of portability.

The first to be used was the energy balance–Bowen ratio method, which requires measurements of the energy balance above the land surface of interest (i.e., the radiation balance above the vegetation surface and the heat flux below the soil surface) together with the ratio of the gradients of air temperature and vapor pressure above the land surface. The small size of these gradients, resulting from the need to confine them to within the shallow atmospheric layer whose microclimate is representative of the land surface of interest, is the major factor limiting the accuracy of this method of measuring evapotranspiration.

Recent technological advances have extended the use of the eddy–covariance flux method of measuring evapotranspiration which calculates the net eddy flux of water leaving the land surface from high-speed observations of the vertical component of wind speed and of humidity. However there are unresolved problems with this method, as shown by the frequent failure of energy-balance closure, i.e., the latent and sensible heat fluxes measured by this method are often less than the energy balance measured for the same land surface, suggesting that evapotranspiration is less accurately measured by the eddy–flux method than by the energy balance–Bowen ratio approach.

Magnitude of Evapotranspiration

Information on the magnitude of actual evapotranspiration is available from a number of water- and heat-balance studies of the Earth's land surfaces. In the water-balance studies, evapotranspiration is estimated as the difference between regionally averaged measurements of precipitation and the runoff from the same area: the use of long-term, regional average annual values enabling the changing storage term to

Table 1 Evapotranspiration from the Earth's land surfaces

	Water balance estimate[a] *(mm year^{-1})*	*Heat balance estimate*[b] *(mm year^{-1})*
All land surfaces	480	442
Northern hemisphere	435	398
Southern hemisphere	572	535
Africa	582	502
North America	403	393
South America	946	882
Asia	420	417
Australasia	534	403
Europe	362	375

[a]Adapted from Baumgartner A and Reichel E (1975) *The World Water Balance. Mean Annual Global Continental and Maritime Precipitation, Evaporation and Runoff.* Amsterdam, the Netherlands: Elsevier.
[b]Adapted from Henning D (1989) *Atlas of the Surface Heat Balance of the Continents. Components and Parameters Estimated from Climatological Data.* Berlin, Germany: Gebruder Borntraeger.

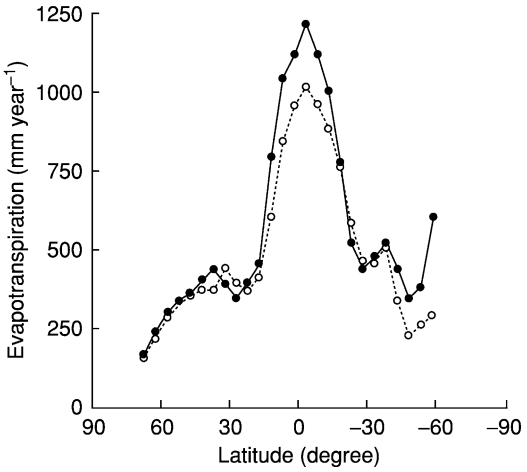


Figure 1 Evapotranspiration from the Earth's land surfaces, in millimeters per year. Empty circles, data source is Henning, 1989; full circles, data source is Baumgartner and Reichel, 1975. (Adapted from Henning D (1989) *Atlas of the Surface Heat Balance of the Continents. Components and Parameters Estimated from Climatological Data.* Berlin, Germany: Gebruder Borntraeger; Baumgartner A and Reichel E (1975) *The World Water Balance. Mean Annual Global Continental and Maritime Precipitation, Evaporation and Runoff.* Amsterdam, the Netherlands: Elsevier.)

be eliminated from the area's water balance. In the heat-balance approach, which can be used for periods less than a year, potential evapotranspiration is first calculated from climatologic measurements and then corrected to actual evapotranspiration using a function of measured precipitation.

The global, hemispheric, continental, and latitudinal values of evapotranspiration for the Earth's land surfaces presented in Table 1 and Figure 1 are taken from calculations using both the water-balance and heat-balance approaches. They agree to within 10% for the global, hemispheric, and most of the continental estimates; much of the larger differences found in the case of the latitudinal estimates can be attributed to the different values of precipitation used in the studies.

Maximum mean annual values of evapotranspiration reaching 1400 mm are found in the near-equatorial regions of South America and Southeast Asia, the specific areas coinciding with those with tropical forest cover and heavy rainfall. Maximum monthly values reaching 150 mm are found for these same areas of land cover in the mid-summer. On a daily basis, double this rate of evapotranspiration has been reported for tall vegetation subject to advective conditions, where latent energy is supplied from convective exchange with the passing air.

Control of Evapotranspiration

The major role of evapotranspiration in the Earth's water balance and humankind's growing shortage of water has focused attention on the possibility of controlling this so-called loss of water to the atmosphere: to do so in an environmentally responsible way requires an understanding of the processes controlling all the components of evapotranspiration, their interactions, and their biological significance.

In a purely physical sense, evapotranspiration can be understood as a transfer process in which water moves from the soil, a source of limited capacity and variable potential, and passes through parallel soil and plant pathways of variable conductance, into the atmosphere, a sink of variable potential but finite capacity. From this point of view evapotranspiration is controlled by either the source or sink strength, whichever is limiting. However, because the mechanisms controlling the rate of conductance of water through plants are biological and incompletely understood, this purely physical description represents a simplified view of the transpiration flux; and, when used to calculate this major component of evapotranspiration, empirical constants must be used.

The significance of the siting of the major biological control of conductance at the plant–atmosphere interface is that the plant is able to control its transpiration loss and so maintain its internal water balance at a favorable status and avoid dehydration. The major mechanism by which this is achieved is through the control of the size of the leaves' stomatal apertures: the plant's internal water status provides a

feedback signal for this control. Feedforward systems of stomatal control based on the water status of the air outside the canopy and of the soil surrounding the roots have also been described.

A second component of evapotranspiration of hydrologic importance in a number of forest ecosystems is the fraction of precipitation that is intercepted by the vegetation canopy and then evaporated directly to the atmosphere without reaching the soil surface. This modifies the canopy microclimate and so reduces transpiration by an amount which depends on the ratio of the surface resistance of the wetted to the dry vegetation, and to a lesser extent on air temperature. Resistance ratios in temperate climates range from values of ≈ 0.2, typical of coniferous forest, to ≈ 0.8, typical of a number of field crops, so that the evaporation rate of intercepted water can be expected to be, respectively, 5 and 1.25 times that of transpiration from the same canopies when dry. For evergreen forests in high-rainfall climates, evaporation of intercepted precipitation can form a very substantial proportion of total evapotranspiration and is not insignificant for deciduous forest canopies even during their leafless phase.

The effect of evaporation of intercepted water on plant growth and yield is less clear than is its hydrologic significance. The use of misting and fogging irrigation systems in protected agricultural systems under conditions of high evaporative demand suggests a wholly favorable influence of the reduction in transpiration, even though the coupling of the transpiration and dry-matter production processes implies that this will have a negative effect on growth.

Precision drip-irrigation systems have also been used to reduce evaporation from the soil surface, the third component of evapotranspiration, by confining water application to below the surface and the crop row. The resulting dry surface imposes a high resistance to upward water flow, so reducing evaporation. It also modifies the crop microclimate, increasing air temperature and reducing humidity within the plant canopy, and so presumably leads to greater transpiration rates which perhaps in turn affect crop growth and yield. In many agricultural systems, cultivation of the soil surface is widely practiced with the aim of reducing evaporation from the soil as well as eliminating transpiration from weeds.

With natural land covers, the type and density of the vegetation can significantly affect the different components of evapotranspiration and these influences can be important for land-use planning and management in water catchment areas which have more than one objective. The ways in which land use influences evapotranspiration is central to the long-debated and complex question of the effect of aforestation and deforestation on water supplies and, on a wider scale, climate and rainfall.

A number of attempts have been made in both the laboratory and the field to control evapotranspiration directly. The two methods used to reduce transpiration have been stomatal closure and the reduction of plant radiation balance through increased leaf reflection. Most of these experiments were unsuccessful and in the few cases where water loss was significantly reduced this was accompanied by a corresponding reduction of plant growth.

Progress in this potentially important field requires a better understanding of the biological mechanisms controlling transpiration and of the coupling of this process to that of plant growth. With such an understanding it may be possible, and even environmentally advantageous, to use biotechnologic techniques to develop plants with lower rates of transpiration and/or higher levels of plant water use efficiency. A major benefit of such an achievement would be a reduction in irrigation requirements, which currently use two-thirds of the Earth's renewable water resources.

Climate Change and Evapotranspiration

The impact of climate change on evapotranspiration has been studied to evaluate its effect on human water supplies and indirectly on plant growth and crop yields. However, attempts to simulate future rates of evapotranspiration under future climate regimes are subject to such large uncertainties that at present they provide plausible, possible scenarios rather than estimates of known certainty which could be used for water planning.

Some indication of the magnitude of evapotranspiration changes during the last century can be obtained on the basis of the mean climate changes measured over this period. The direct effects of the 70 ppm (20%) increase in the atmospheric CO_2 concentration are twofold and are in contrast: reduced stomatal conductance and increased vegetative growth. Experiments comparing plant growth and evapotranspiration from plant stands growing under the current and a doubled CO_2 concentration indicate that these two effects cancel each other, leaving canopy water loss substantially unchanged.

Other climate changes measured during the twentieth century, attributed to the increase in CO_2 and other radiatively active gases, include an increase in mean land air temperature of $0.6 \pm 0.15°C$, with the increase in minimum temperatures twice that of the maximum. A mean overall increase of 9 mm (1%) in annual precipitation over the Earth's land surfaces was observed over the period and a decrease in global

irradiance averaging $0.51 \pm 0.05\,W\,m^{-2}$ per year was recorded in the last 50 years. As the effect of the minor increase in air temperature would have been more than offset by the more substantial decrease in radiation, a decrease in potential evapotranspiration could be expected to have occurred. This has been verified from evaporation pan observations in the USA, the former Soviet Union, and India as well as from estimates in India and China based on the Penman equation.

There is no evidence that a corresponding change in actual evapotranspiration has occurred; this is shown by the absence of clear evidence for large, statistically significant, and widespread changes in the precipitation over the land surfaces of the Earth (the measured 1% increase is far below the error of its measurement) and in the annual streamflow and peak discharges of rivers. This lack of overall trend in the runoff component of the water balance has emerged from a major study of historical records of 142 major rivers throughout the world, possessing more than 50 years of data and representing drainage areas greater than $1000\,km^2$.

See also: **Energy Balance**; **Evaporation of Water from Bare Soil**; **Penman, Howard Latimer**; **Penman–Monteith Equation**

Further Reading

Baumgartner A and Reichel E (1975) *The World Water Balance. Mean Annual Global Continental and Maritime Precipitation, Evaporation and Runoff*. Amsterdam, The Netherlands: Elsevier.

Henning D (1989) *Atlas of the Surface Heat Balance of the Continents. Components and Parameters Estimated from Climatological Data*. Berlin, Germany: Gebruder Borntraeger.

Houghton JT, Meira Filho LG, Callander BA *et al.* (1996) *Climate Change 1995. The Science of Climate Change*. Cambridge, UK: Cambridge University Press.

Larcher W (1980) *Physiological Plant Ecology*, 2nd edn, pp. 258–267. Berlin, Germany: Springer-Verlag.

Molchanov AA (1963) *The Hydrological Role of Forests*. Jerusalem, Israel: Israel Program for Scientific Translations. (Translated from Russian *Gidrologicheskaya rol'lesa, Izdatel'stvo Akademii Nauk SSR Moskva 1960.)*

Monteith JL (1965) Evaporation and environment. In: Fogg GE (ed.) *The State and Movement of Water in Living Organisms*, vol. 19, pp. 205–235. Society of Experimental Biology Symposium. Cambridge, UK: Cambridge University Press.

Penman HL (1956) Evaporation: an introductory survey. *Netherlands Journal of Agricultural Science* 4: 9–29.

Penman HL (1963) *Vegetation and Hydrology*. Farnham Royal, UK: Commonwealth Agricultural Bureaux.

Rutter AJ (1975) The hydrological cycle in vegetation. In: Monteith JL (ed.) *Vegetation and the Atmosphere*, vol. 1, pp. 111–154. London, UK: Academic Press.

Stanhill G (1973) Evaporation, transpiration and evapotranspiration: a case for Ockham's razor. In: Hadas A, Swartzendruber D, Rijtema PE, Fuchs M, and Yaron B (eds) *Physical Aspects of Soil Water and Salts in Ecosystems. Ecological Studies – Analysis and Synthesis*, vol. 4, pp. 207–220. Heidelberg, Germany: Springer-Verlag.

Stanhill G (1986) Water use efficiency. In: Brady NC (ed.) *Advances in Agronomy*, vol. 39, pp. 53–85. Orlando, FL: Academic Press.

Tanner CB (1967) Measurement of evapotranspiration. In: Hagan RM, Haise HH, and Edminster TE (eds) *Irrigation of Agricultural Lands*, pp. 534–574. Agronomy Monograph No. 11. Madison, WI: American Society of Agronomy.

Waggoner PE (ed.) (1990) *Climate Change and US Water Resources*. Wiley Series in Climate and the Biosphere. New York: John Wiley.

F

FACTORS OF SOIL FORMATION

Contents
Biota
Climate
Human Impacts
Parent Material
Time

Biota

A H Jahren, Johns Hopkins University, Baltimore, MD, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

It is clear that organisms (such as plants, microorganisms, and soil invertebrates) that live in and on the soil both influence soil properties and are influenced by soil properties. A clear assessment of the isolated influence that a given type of biota has upon soil formation is best performed in the context of a biosequence, or series of sites in which all soil-forming factors are held constant, while the biota varies systematically. Biosequence studies are most often performed to evaluate the influence of differing plant communities, since vegetation at a site can be independent of climate, topography, parent material, and time. The most commonly performed biosequence studies assess the influence of grassland versus forest biota on soil-forming processes. Because grasses contribute a large portion of their tissue to the soil from within the soil (as below-ground tissue), grassland ecosystems result in significantly higher organic contributions to the soil, especially at depth, relative to forest ecosystems. Biosequence studies designed to evaluate the influence of deciduous versus conifer forests have also been performed, and highlight the unique role of the conifer needle litter layer as a discrete and chemically unique portion of the uppermost soil. The most recent studies have focused on the indirect effects of different vegetation on soil-forming processes, through the assessment of microbial populations or water infiltration characteristics along a biosequence.

The Biosequence

Efforts to explain soil characteristics in terms of the influence of biota are best facilitated by biosequence studies. These studies contain a series of soil profiles, across which the biotic soil-forming factor varies, while other soil-forming factors such as climate (cl), topography (r), parent material (p), and time of development (t) remain constant. Within the context of a biosequence, the effect of changing biotic factor (o) upon any soil property (s) can be assessed quantitatively, providing that all relevant factors and properties can be adequately described or assessed. The relationship between biota and soil properties can be formalized in order to highlight the functional influence that biota has on the soil, in the following definition of the biosequence:

$$s = \mathrm{f}(o)_{\mathrm{cl,r,p,t}} \qquad [1]$$

However, application of the biosequence study is problematic, for both conceptual and practical reasons. It is easy to imagine hundreds of ways in which organisms might affect soil characteristics: earthworms burrow through soil horizons, perhaps decreasing the bulk density; annual plants die and perhaps contribute organic matter to the soil profile; denitrifying bacteria might lower nitrate levels in soil solutions. However, acknowledging the influence of organisms on soil properties is fundamentally different from showing how organisms act as a

soil-forming factor. The constellation of organisms present in any soil environment is dependent upon soil properties. For example, moisture levels in a given soil environment might partially control the amount of plant biomass that grows at the site. Thus the amount of vegetation, which influences soil characteristics, is actually dependent upon the soil properties. This interdependence, which can be seen in most examples of soil biota, significantly complicates the application of biosequence theory, in which all soil-forming factors are assumed to be independent. In application, it may be difficult to find sites where the biotic factor varies, while climate is constant, since most organisms are intimately sensitive to the climate in which they live. This may be overcome by focusing upon 'patchy' communities of organisms – constellations of organism type and species that exhibit regular heterogenity within a small geographical area. However, this raises another potential complication, namely how stable are the patches over the time of soil formation, since the time scales of community change and soil formation may be very different within the same ecosystem? Finally, there is the example that, at a given soil, only some plant species, of all the seeds that arrive at the site, eventually grow to maturity. Thus the actual vegetation at such a site is a subset of the potential vegetation, and may be dependent upon soil properties, or other soil-forming factors.

In light of these issues, biosequence studies are best performed where kind and composition of species change across sites (in preference to changes in quantity or yield of organisms, which may be governed by soil properties). It is also best that these studies are performed within a geographic area restricted to that which might be thought to be subject to the same overall potential biota, as well as the same climate regime. Given that these conditions can be met, biosequence studies yield valuable information about the specific effects of the biotic soil-forming factor.

The Biotic Factor

The biotic factor of soil formation (o) is multifaceted, given the myriad of organisms that conduct their life activities in the context of soil. A pedologist interested in evaluating the influence of the biotic factor might wish to specify a group of organisms to focus upon, and to acknowledge the distinct processes associated with vegetation (o_v), microorganisms (o_m), animals (o_a), and human activities (o_h). At present, there is a paucity of systematic studies evaluating the biotic factor associated with both microbes and animals on soil formation, although significant interest exists in these areas. The effect of human activities on soil properties and soil formation is treated extensively elsewhere in this volume and will therefore not be included in the discussion to follow. Given the wide environmental tolerances of many common plants, vegetation is considered to be the facet of the biotic factor most independent of climate, topography, parent material, and time in biosequence studies. Many stable and adjacent, yet highly distinct, vegetation communities exist in both the Old and New Worlds, and have formed the basis of most biosequence studies.

Vegetation as a Soil-Forming Factor

The elemental composition of soil differs from that of geologic materials in its striking enrichment of carbon and nitrogen compounds, relative to most rocks. The ultimate sources of this carbon and nitrogen are organic compounds contributed to the soil by life processes, and the death, of soil organisms such as plants, animals, and microorganisms. Of these contributions, those from vegetation dominate: plants contribute organic compounds to the soil in a variety of ways, including the senescence or necrosis of tissue, exudation or respiration from the roots, and the liberation of reproductive tissues such as pollen, seeds, and fruit. There are reasons to expect that different kinds of vegetation would contribute organic compounds to soils in different ways: deciduous trees contribute leaf material to the soil in one large annual pulse, while evergreen trees contribute leaf (or needle) tissue continuously throughout the year, as their foliage slowly turns over. In addition, the biomass produced annually as a potential contribution to the soil varies by two orders of magnitude across different vegetative biomes (Table 1).

The distribution of mass within the vegetative communities is variable as well: in most forests, the

Table 1 Leaf biomass production across vegetation and community type

Vegetation or ecosystem type	*Annual biomass production (air-dry) (kg m^{-2})*
Deserts[a]	0.04–0.12
Alpine meadow[b]	0.05–0.09
Tall- and short-grass prairie[b]	0.16–0.50
Grasslands (North American)[a]	0.09–0.90
Deciduous forest leaves[a]	0.40–0.85
Conifer forest needles[a]	0.80–3.00

[a]Source: Webb WL, Lauenroth WK, Szarek SR, and Kinerson RS (1983) Primary production and abiotic controls in forests, grasslands, and desert ecosystems in the United States. *Ecology* 64: 134–151.
[b]Source: Clements and Weaver (1924) and Swederski (1931), as reported in Jenny H (1941) *Factors of Soil Formation: A System of Quantitative Pedology*. New York: McGraw Hill. Republished in 1994 by Dover Publications, Mineola, New York.

belowground mass is about one-quarter of the total mass of the vegetation, while in some grasslands, more biomass exists belowground than above.

When plant material is contributed to the soil, it becomes soil organic matter only after the action of decomposition. The process of decomposition is mediated by microorganisms, and these bacteria, protozoa, and fungi preferentially decompose plant material of high quality. The quality of plant contribution can be quantified by its carbon-to-nitrogen ratio; high nitrogen concentrations in organic compounds are initially taken as an indicator of readily decomposable material. Therefore a low carbon-to-nitrogen ratio in plant material suggests abundant nitrogen in the substrate, that would readily decompose, and thus affect pedogenesis by forming soil organic matter. The carbon-to-nitrogen ratio of plant tissues is extremely variable from species to species, and between tissues within an individual plant (Table 2).

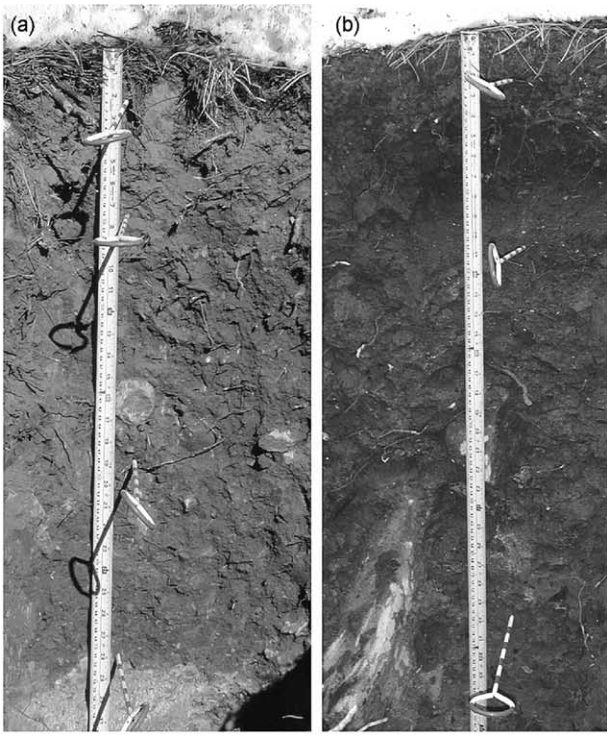
Conifer ecosystems contribute more tissue to the soil annually than do deciduous or grassland ecosystems, yet this tissue is of very low quality with respect to decomposition. In contrast, the A horizon of grassland soils exhibits organic matter with high quality, relative to deciduous or conifer systems. From this, we might hypothesize that a biosequence evaluating soils forming under conifer forests, deciduous forests, and grasslands would yield the following observations: (1) the conifer soil might have a thick undecomposed upper horizon, with relatively low levels of organic matter below; (2) the grassland might have well-decomposed and well-distributed organic matter throughout the soil profile; and (3) the deciduous system would exhibit properties intermediate to the conifer and grassland profiles.

In addition to the direct effect of organic contribution to the soil by vegetation, there are several

Table 2 Carbon-to-nitrogen ratio in plant tissue and resulting organic horizon

Vegetation or organic horizon type	*Carbon-to-nitrogen ratio (unitless)*
Shoot, root and leaf[a]	
Conifer, deciduous, grass and other	5–80
Pine needles (*Pinus coulteri*)[b]	57
Oak leaves	
Quercus pubescens[a]	45
Quercus dumosa[b]	39
Conifer forest A horizon	
Pine (*Pinus ponderosa*)[c]	33.5–54.9
Pine (*Pinus coulteri*)[b]	16.1
Fir (*Abies concolon*)[c]	26.0–42.3
Hardwood forest A-horizon	
Oak (*Quercus dumosa*)[b]	19.5
Mixed oak-hickory[c]	12.6
Bluestem grassland A-horizon[c]	11.7

[a]Source: Raschi A, Miglietta F, Tognetti R, and van Gardingen PR (1997) *Plant Responses to Elevated CO_2*. Cambridge: Cambridge University Press.
[b]Source: Quideau SA, Graham RC, Chadwick OA, and Wood HB (1998) Organic carbon sequestration under chaparral and pine after four decades of soil development. *Geoderma* 83: 227–242.
[c]Source: Bodman (1935) and Rost (1918), as reported in Jenny H (1941) *Factors of Soil Formation: A System of Quantitative Pedology*. New York: McGraw Hill. Republished in 1994 by Dover Publications, Mineola, New York.



(a) Fine-loamy smectitic mesic typic Haplustalf
Vegetation: pine, juniper, cactus, yucca, cedar
Parent material: volcanic (basalt)

Depth (cm)	*Horizon*	*Color*	*% Clay*	*Structure*	*pH*
0–6	A	7.5 YR 4/3	6	1fgr	6.5
6–19	Bt_1	7.5 YR 4/4	18	2fsbk	6.5
19–48	Bt_2	7.5 YR 3/3	25	3msbk	7.0
48–71	Bt_3	7.5 YR 3/4	35	3msbk	7.0
71–84+	C	7.5 YR 5/6		massive	7.0

(b) Very-fine smectitic mesic typic Haplustert
Vegetation: pine, juniper, yucca, various grasses
Parent material: volcanic (basalt)

Depth (cm)	*Horizon*	*Color*	*% Clay*	*Structure*	*pH*
0–5	A	7.5 YR 2.5/1	18	3mgr	7.0
5–26	ABtss	7.5 YR 3/1	50	3msbk	7.0
26–76	$Btss_1$	7.5 YR 3/2	80	3mabk	7.0
76–90+	$Btss_2$	7.5 YR 3/3	80	3mabk	7.0

Figure 1 Two soil profiles from the Kaibab national forest in Northern Arizona: (a) forest soil; (b) soil from a forest containing some grasses. The influence of grass vegetation can be seen as enhanced organic content at all depths in the profile.

indirect effects of vegetation that may influence soil properties. The decomposition of plant tissues produces humus, which contains many organic acid groups, including fulvic acids. The low pK of these acids may accelerate weathering processes and the production of secondary minerals from primary minerals. In this way the decomposition of plant tissues might facilitate pedogenic processes, such as the *in situ* production of secondary clay. The needles of conifer trees are well-known to result in low-pH soil conditions. The leachate of this evergreen foliage is usually ~2 units lower in pH value than leachate moving through deciduous litter, liberating organic acids that may participate in weathering processes. Vegetation type may indirectly influence the hydrologic properties of the soil: earthworm activity specific to oak vegetation results in water-stable aggregates that promote infiltration and aeration, and reduce erosion relative to nonaggregated soils. Also, the extensive fine-root systems present in grassland soils results in high percolation rates as water readily infiltrates subsurface horizons through a network of micropores.

Grassland–Forest Biosequence

The best-studied biosequence involves the 'Prairie-timber transition zone,' which is a coexisting grassland and forest ecosystem with abrupt transitions between the two vegetation types. A belt, or zone, of such transitions has been described in European Russia, Siberia, Canada, and the Midwestern USA. Within these zones, both grasslands and forests occur on the uplands as well as in the drainage channels, and rainfall is sufficient to support forest growth throughout the entire region. The sharp transition that forms the basis of the biosequence has puzzled ecologists and no satisfactory hypothesis has been put forward to explain the pedologically fortuitous configuration. Recent intruiging suggestions invoke long-term and long-range feedbacks between the grassland and the forests, perhaps via herbivores, acting to stabilize the discrete, but adjacent, ecosystems.

The influence of grasses on soil profile development can be partially isolated by examining soils under forests, and comparing them to those under forests containing some grasses near the grassland–forest transition (Figure 1).

Grass tissue is almost evenly distributed above- and belowground, therefore annual grasses readily contribute their tissue to the soil profile, not only at the surface, but from within the epipedon and subsurface horizons as well. Since increased organic matter results in dark soil colors, soils influenced by grasses in this way exhibit lower value and chroma than their forest counterparts, especially at the lower depths of the profile (Figure 1).

Various biosequence studies comparing grassland soils to forest soils reveal significantly higher levels of percentage carbon up to 1 m depth. The quality of this organic matter is higher under grasslands, with significantly lower carbon-to-nitrogen ratios than under forests, especially near the soil surface. Soil pH values are generally lower in forest soils, probably reflecting the influence of conifer litter, which results in acidic leachate. These differences in pH are often diminished with depth, however. Many studies have revealed more clay and enhanced structure under grassland vegetation, relative to forest soils, although

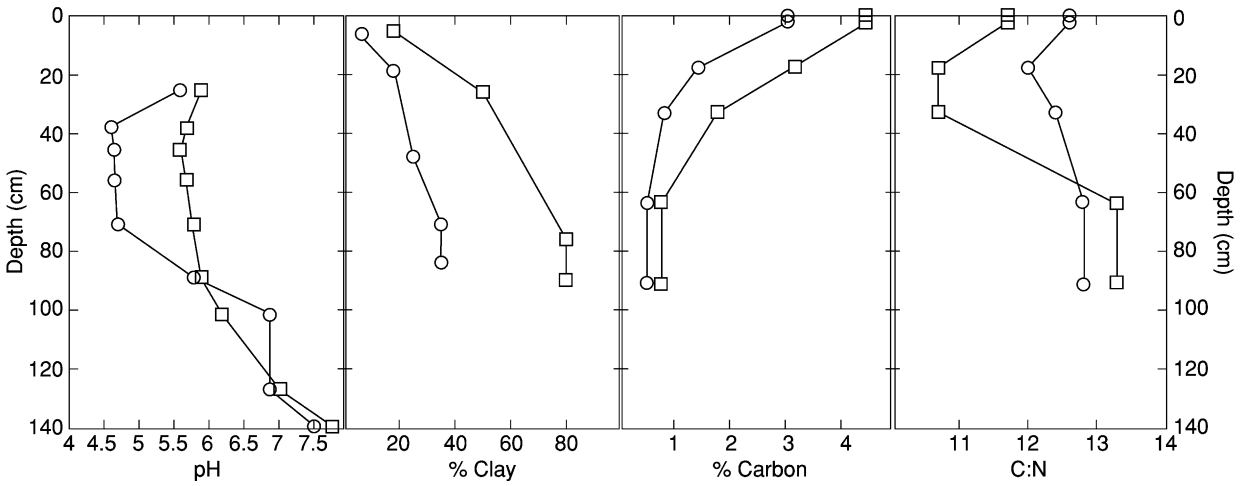


Figure 2 Comparison of soil characteristics below forest (circles) and grassland (squares) ecosystems, as reported in several biosequence field studies. pH after Jenny H (1941) percentage clay after Figure 1; percentage carbon and C:N after Rost (1918), as reported in Jenny H (1941) *Factors of Soil Formation: A System of Quantitative Pedology*. New York: McGraw Hill. Republished in 1994 by Dover Publications, Mineola, New York.

there is no agreement as to what mechanism acts to cause this relationship (Figure 2).

Conifer Forest–Deciduous Forest Biosequence

Recent studies focusing on soil development after the installation of lysimeters and using other long-term monitoring techniques have shed light on the effects of conifer versus deciduous forest ecosystems on soil development. An important example of such a biosequence resides within the San Dimas experimental forest in the San Gabriel mountains of California. Four decades of soil development have resulted in significantly higher soil carbon and nitrogen content under oak (deciduous) systems than under pine (conifer). However, such results may be highly species-specific: comparison of soils forming under beech (deciduous) versus spruce (conifer) vegetation along a biosequence in Denmark showed much higher levels of soil organic carbon under the conifer forest, at all soil depths (Figures 3 and 4).

The most notable result of deciduous versus conifer forest biosequence studies is that the conifer

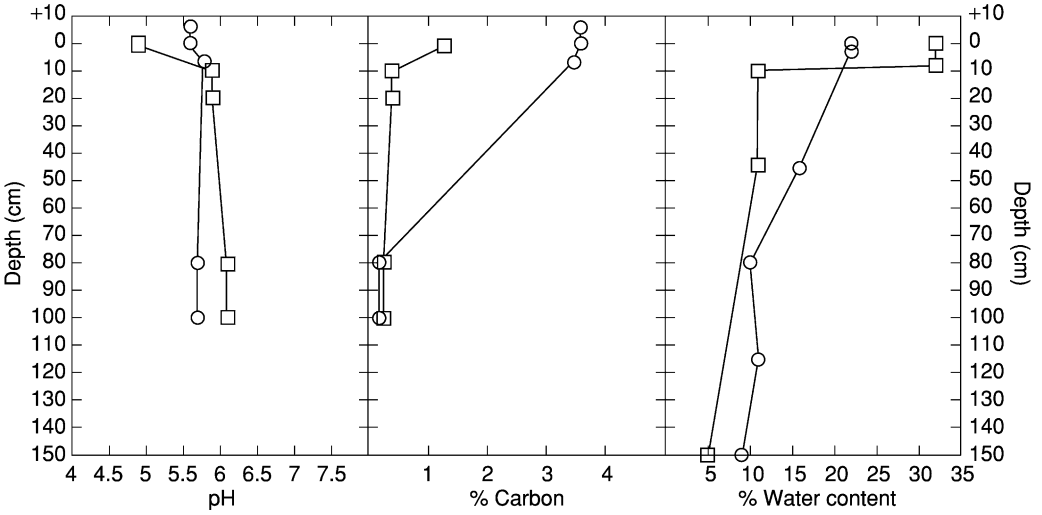


Figure 3 Comparison of soil characteristics below hardwood (circles) and conifer (squares) forests, as reported in several biosequence field studies. pH and percentage carbon after Graham RC, Ervin JO, and Wood HB (1995) Aggregate stability under oak and pine after four decades of soil development. *Soil Science Society of America Journal* 59: 1740–1744 (1995); percentage water content after Ekelund F, Rønn R, and Christensen S (2001) Distribution with depth of protozoa, bacteria and fungi in soil profiles from three Danish forest sites. *Soil Biology and Biochemistry* 33: 475–481.

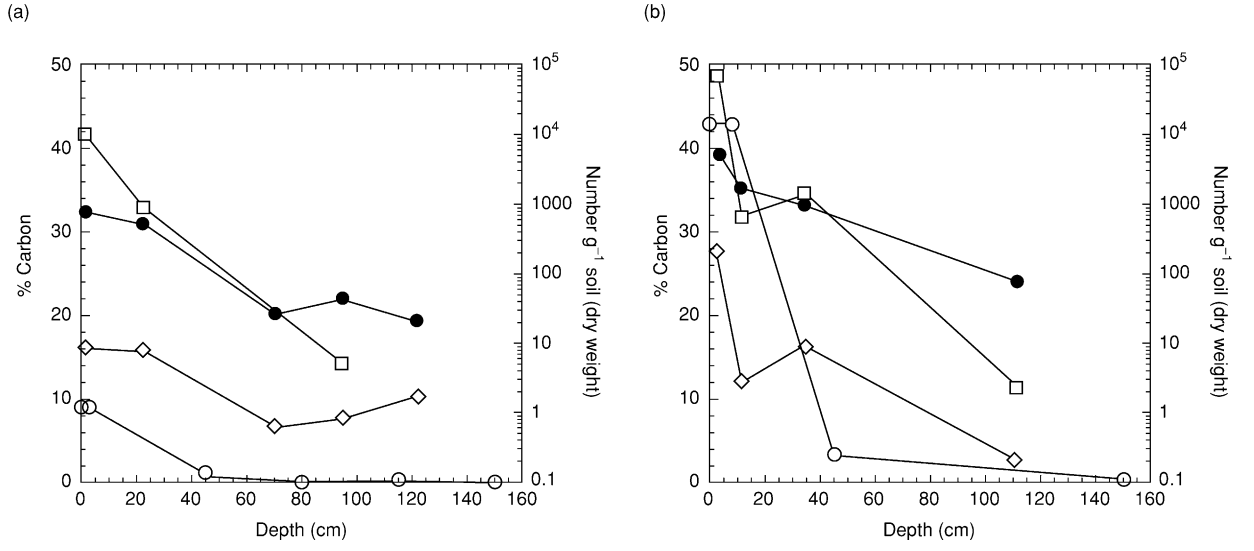


Figure 4 Comparison of microbial populations within (a) beech and (b) spruce forest soils. Open circles, percentage carbon; squares, protozoa; diamonds, bacteria ($\times 10^8$); filled circles, fungal hypha. Biosequence data after Ekelund F, Rønn R, and Christensen S (2001) Distribution with depth of protozoa, bacteria and fungi in soil profiles from three Danish forest sites. *Soil Biology and Biochemistry* 33: 475–481.

ecosystem dramatically influences the top ~10 cm of the soil, while the influence of the deciduous ecosystem may be more diffuse. A dramatic decrease in pH, increase in percentage carbon, and a large increase in water content, confined to the upper decimeter of soil, has been observed in soils forming under conifer forests by various biosequence studies (Figure 4). This may highlight the importance of the conifer needle litter layer as a stable and chemically unique portion of developing soils.

New areas of study have attempted to quantify microbial populations along biosequences and evaluate the effect of vegetation type on the microorganism community. Microorganisms control many soil processes, such as decomposition, and thus could be considered to affect soil processes indirectly as the result of changing vegetation along a biosequence. Deciduous versus conifer forest biosequences in Denmark show bacterial, fungal, and protozoan populations an entire order of magnitude higher under spruce (conifer) ecosystems, relative to beech (deciduous) forests. Although microbial populations decrease with profile depth under both types of forest, populations remain significantly higher throughout the profile forming beneath spruce (conifer) forests (Figure 4). Further study is needed to assess the different species composition and ecosystem significance of these differing microbial communities, but it is clear that vegetation biosequences show a wide range of both direct and indirect effects on soil formation that is highly variable across vegetation type.

List of Technical Nomenclature

% carbon	Carbon content (carbon g per 100 g of soil)
% clay	Clay content (total number per 100 particles that are clay-sized)
	'Clay-sized' ($<2\ \mu$m)
% water content	Water content (water g per 100 g of soil)
B	Biomass production (kg m^{-2})
C:N	Carbon–nitrogen ratio (unitless)
pH	$-$Log [H^+] (unitless)

See also: **Biodiversity**; **Classification of Soils**; **Factors of Soil Formation:** Climate; Human Impacts; Parent Material; Time; **Jenny, Hans**

Further Reading

Bicki TJ, Fenton TE, Luce HD, and Dewitt TA (1988) Comparison of percolation test results and estimated hydraulic conductivities for Mollisols and Alfisols. *Soil Science Society of America Journal* 52: 1708–1714.

Ekelund F, Rønn R, and Christensen S (2001) Distribution with depth of protozoa, bacteria and fungi in soil profiles from three Danish forest sites. *Soil Biology and Biochemistry* 33: 475–481.

Graham RC, Ervin JO, and Wood HB (1995) Aggregate stability under oak and pine after four decades of soil development. *Soil Science Society of America Journal* 59: 1740–1744.

Jenny H (1941) *Factors of Soil Formation: A System of Quantitative Pedology*. New York: McGraw-Hill. Republished in 1994 by Dover Publications, Mineola, New York.

Quideau SA, Graham RC, Chadwick OA, and Wood HB (1998) Organic carbon sequestration under chaparral and pine after four decades of soil development. *Geoderma* 83: 227–242.

Raschi A, Miglietta F, Tognetti R, and van Gardingen PR (1997) *Plant Responses to Elevated CO_2*. Cambridge: Cambridge University Press.

Roy J, Saugier B, and Mooney HA (2001) *Terrestrial Global Productivity*. London: Academic Press.

Schlesinger WH (1997) *Biogeochemistry: An Analysis of Global Change*, 2nd edn. London: Academic Press.

Webb WL, Lauenroth WK, Szarek SR, and Kinerson RS (1983) Primary production and abiotic controls in forests, grasslands, and desert ecosystems in the United States. *Ecology* 64: 134–151.

Climate

O C Spaargaren, ISRIC/WDC for Soils, Wageningen, The Netherlands
J A Deckers, Catholic University of Leuven, Leuven, Belgium

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Climate is a key factor in the formation and potential use of soils. It influences vegetation and soil fauna, transforms rock into regolith, and regulates chemical reaction rates and transport flow of substances in soils. Climate plays a crucial role in human land-use options, determining the length of growing seasons and the crops that can be grown. Soils, on the other hand, act as a climatological archive, recording the conditions and changes in the recent past. These examples show the intricate interaction between soils and climate.

Key climatic components in relation to soil are the seasonal amount of precipitation, the amount of evaporation, the temperature and its fluctuations, wind, and solar radiation. The amount and seasonal distribution of precipitation, together with

evaporation (or, combined with the influence of the vegetative cover, the evapotranspiration), determines the moisture content and the net flux of water in a soil. The temperature and the temperature variations (daily, seasonally, annually) largely regulate the rate of decomposition or alteration of organic and mineral substances in soils, and the faunal activity. Wind plays a key role in dry areas, and solar radiation is responsible for the necessary influx of energy.

Because climate has such an impact on soils, it is not surprising that the first realistic concepts of soil formation and soil distribution were linked to climatic conditions and climate belts (see example in Table 1). The concept of 'zonal soils,' i.e., soils that are in equilibrium with climate (and vegetation type), was first introduced by V.V. Dokuchaev during the 1880s in Russia. Rode describes Dokuchaev's concept of zonality of soils as:

> The zonal occurrence on the globe . . . consists in the fact that the areas occupied by soils of various types stretch in broad belts, on the whole latitudinally, like the climate zones. Such regularity is indeed the best proof of the very great role which climate plays in soil formation.

A very good correlation exists between the climatic, vegetation, and soil belts in Belarus, the Ukraine, and the European part of Russia. This is brought about by the fairly uniform parent material (somewhat sandier in the north, more loamy (loess) in the south), the regular increase in temperature from the north to the south, and a rainfall pattern showing an increase followed by a decrease also from north to south. It results in a latitudinal sequence of tundra soils (Gelisols), podzolic and forest soils (Spodosols, Alfisols), chernozemic and chestnut soils (Mollisols), desert soils (Aridisols), cinnamon soils (Ultisols), and Zheltozems (Oxisols), corresponding with arctic, boreal, steppic, desert, subtropical, and tropical climates.

Table 1 The relation between precipitation and the saturation deficit of the vapor pressure to correlate climate and soil type in Europe

Climate	*Soil type*	*Precipitation/ saturation deficit*[a] *(mm)*
Arid	Desert (Aridisols)	0–100
Semiarid	Chestnut brown (Mollisols)	100–275
Semihumid	Chernozem (Mollisols)	125–375
Humid	Brown forest (Alfisols, Ultisols)	275–500
Perhumid, cool	Podzol (Spodosols)	375–1200
Perhumid, cold	Tundra (Gelisols)	500–600

[a]The precipitation–saturation deficit quotient (in millimeters) is defined as the amount of precipitation divided by the saturation deficit, i.e., the saturated vapor pressure minus the actual vapor pressure.
Adapted from Meyer A (1926) Über einige Zusammenhänge zwischen Klima und Boden in Europa. *Chemie der Erde* 2.

The soil pattern in North America does not follow latitudinal belts. This is because, although the parent material is quite homogenous, at least in the interior part of the USA and Central Canada, the rainfall gradient is increasing from west to east, perpendicular to the temperature gradient, which is from north to south. Here Chernozems, Chestnut brown soils and Desert soils occur at the same latitude, unlike the pattern in Belarus, the Ukraine, and European Russia. Therefore, the concept of zonality has been difficult to apply in North America. This can be noticed from the soil maps published in the USA during the 1930s, and by the fact that, in developing the groupings of soils at that time, more consideration was given to soil characteristics and broad environmental conditions than climate alone. Only the highest category (Category VI, Solum Composition Groups), the distinction between the Pedocal and Pedalfer groups, being the presence or absence of accumulated secondary lime, respectively, reflects some climatic relationship (more evapotranspiration versus more leaching). This, however, only holds true if calcium carbonate is present in the soil system.

While in Russia the concept of zonality, with its close link to climate, has continued to govern the subsequent soil classification systems, the American Soil Taxonomy, introduced in 1975 and modified in 1999, focuses more on measurable soil properties and materials, and introduced soil climatic criteria (soil moisture regime, soil temperature regime) at the penultimate-highest (suborder) level. These soil climatic criteria still have some relationship to climatic conditions at the Earth's surface, concerning well-drained soils. For example, a xeric soil moisture regime (i.e., total soil moisture control section is, in normal years, dry for 45 or more consecutive days in the 4 months following the summer solstice and moist for 45 or more days in the 4 months following the winter solstice) is a typical moisture regime in areas of Mediterranean climates (moist and cool winters, dry and warm summers). Similarly, the (iso)hyperthermic soil temperature regime (i.e., mean annual soil temperature at 50 cm depth is 22°C or higher; 'iso-' indicating a temperature fluctuation of less than 6°C) is closely related to areas with a tropical climate.

As a basis for this discussion on the relation between soils and climate, the Köppen climate classification is used. This widely used, well-known, world-wide classification system combines temperature and precipitation into five main climatic zones, viz. tropical climates (A), dry climates (B), warm temperate climates (C), cool temperate climates (D),

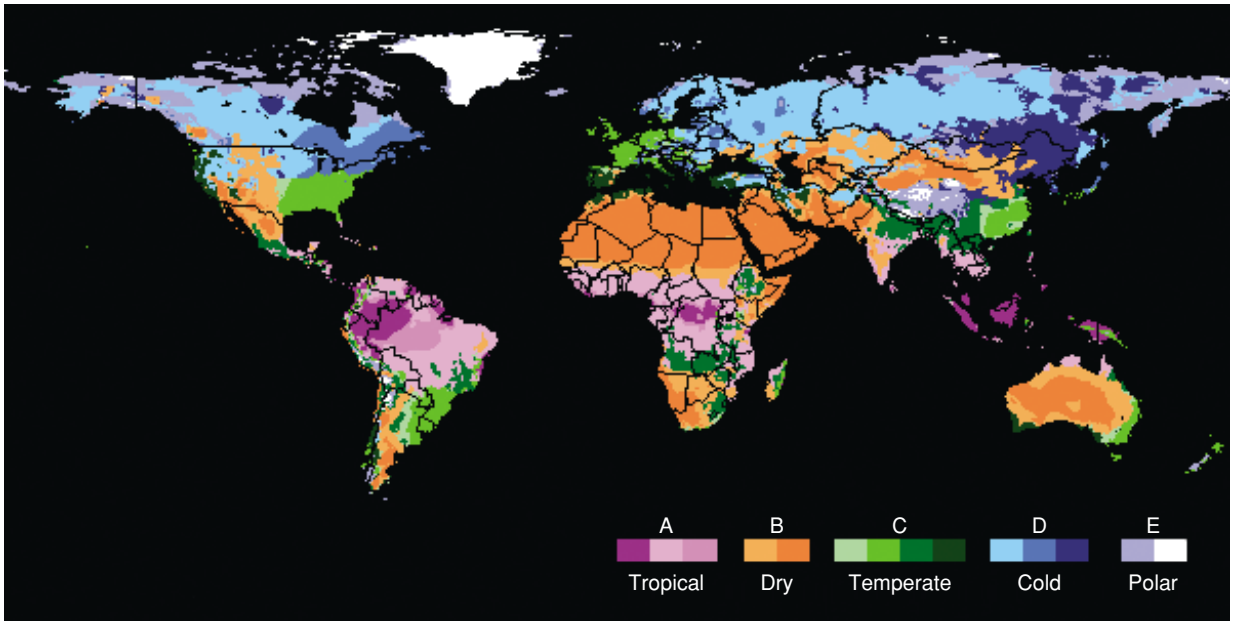


Figure 1 Koeppen's climate classification. Reproduced with permission from FAO Environment and Natural Resources Service, Sustainable Development Department (www.fao.org).

and polar climates (E) (Figure 1). Soils are classified according to both the second edition of *Soil Taxonomy* (ST) and the *World Reference Base for Soil Resources* (WRB). (*See* **Classification Systems:** FAO.)

Soils in Tropical Climates

The tropical climate in the Köppen system is defined as having an average temperature in every month of above 18°C. There is no winter season, and annual rainfall is large and exceeds the annual evaporation. Subdivision of this climate type is based on the absence or presence of a significant dry period, and the occurrence of monsoon rains (Af, tropical rainforest; Am, tropical rainforest with monsoon type of precipitation; Aw, tropical savannah).

The high temperature and precipitation of the almost-permanently moist tropical rainforest climates (Af and Am types of climate; (per)udic soil moisture and (iso)hyperthermic soil temperature regimes) yield an intense and deep chemical weathering and strong leaching of solutes. As a result primary minerals such as feldspars and micas are dissolved or transformed, leading to the loss of basic elements (Ca, Mg, K, Na) and silica, and the formation of such secondary minerals as kaolinite ($Al_2Si_2O_5 \cdot 2H_2O$), goethite (α-FeOOH), and gibbsite (α-$Al(OH)_3$). (*See* **Clay Minerals**.) On rare occasions also lepidocrocite (γ-FeOOH) may be formed. This process, known as ferralitization, is widespread in tropical regions, particularly on the older geomorphic surfaces. It gives rise to acid soils with low cation exchange capacity

Figure 2 Giant Podzol in Malaysia. (Reproduced with permission from the International Soil Reference and Information Center, ISRIC.)

(CEC $<16\ cmol_c^{-1}\ 100\ g^{-1}$ clay) due to the prevalence of low-activity clay, and even soils with variable charge if iron hydroxides and/or aluminum hydroxides are dominating. Such soils are known as Oxisols (ST) or Ferralsols (WRB), and kandic Great Groups of the Ultisols (ST), comparable with Acrisols and Nitisols (*pro parte*; WRB). On younger deposits, not yet as strongly leached as the soils on the older geomorphic surfaces, chemical weathering of primary minerals such as smectites, vermiculites, and chlorites produces large amounts of exchangeable aluminum, resulting in Paleudults (ST) or Alisols (WRB). In sandy deposits, the strong leaching leads to the formation of giant Spodosols (ST) or Podzols (WRB) (Figure 2).

Under tropical savannah climates, which have a pronounced dry season (Aw type; ustic soil moisture and hyperthermic temperature regimes), weathering and leaching are less intense compared with the (almost) permanently moist tropical climates described above. This is due to less rainfall, the upward movement of solutes during the dry period preceding the evaporation of the soil moisture, and frequent additions of airborne dust. The Aw type of climates often border the climates characterized as dry (B). Dust derived from the deserts are can be transported over long distances. Well known are the Harmatan trade winds in West Africa, frequently obscuring the sunlight during daytime, and the Sahara dust traveling across the Atlantic Ocean to the Caribbean islands. This results in annual addition of dust, frequently calcareous, to the soils surrounding the dry regions of the Earth.

At the same time, the process of desiccation plays an important role. During the long, dry period, soils dry out completely (characteristic for the ustic soil moisture regime), and hydrated forms of iron and aluminium hydroxides may lose at least part of their crystal water. Goethite (α-FeOOH) is transformed into hematite (α-Fe_2O_3), following the reaction $2FeOOH - H_2O \rightarrow Fe_2O_3$. Similarly, one would expect dehydration of gibbsite (α-$Al(OH)_3$) into boehmite (α-AlOOH), but this has not yet been demonstrated with certainty. (*See* **Metal Oxides**.)

The dehydration of goethite and the formation of hematite are responsible for the striking difference in color between the soils of the humid tropical regions and the soils of the savannah-type climates. Soils under humid tropical climates are dominantly yellow or yellowish-brown (the color of goethite), whereas the soils under tropical savannah climates are dominantly red (the color of hematite). This reddening is known as the process of rubefaction and occurs in all climates with a pronounced dry season.

Because of the less-intense weathering and leaching, and the addition of airborne dust, the soils under the tropical savannah climate are less acid and have higher base saturation compared with their counterparts under the tropical rainforest climate. Here we find, next to Oxisols (ST) or Ferralsols (WRB), Kandiustalfs or Lixisols (Figure 3) and Nitisols (*pro parte*) (WRB), and Paleustalfs and Haplustalfs (ST), or Luvisols (WRB). Other soils frequently encountered are Natrustalfs (ST) or Solonetzes (WRB), where sodium plays an important role; and Albaqualfs (ST) or Planosols (WRB), which have pronounced water stagnation at the surface during the rainy season. In level areas rich in shrink-swell clays (montmorillonite), Vertisols (ST and WRB) occur.

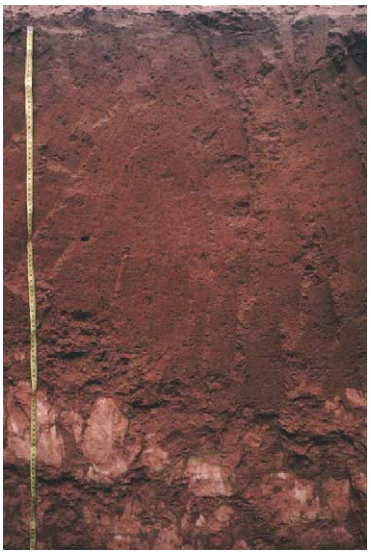


Figure 3 Deep red savannah soil over sandstone in northern Ghana. (Reproduced with permission from ISRIC.)

Soils in Dry Climates

Dry climates (B) in the Köppen climatic classification are distinguished from the more humid climates by the formula:

$$R = \frac{1}{2}T - \frac{1}{4}PW$$

in which R is rainfall (in inches), T is temperature (in degrees Fahrenheit) and PW is percentage annual rainfall in winter (half-year). In the dry climates potential evaporation exceeds precipitation on the average throughout the year. A distinction is made between a steppe climate (BS: ustic soil moisture regime, variable-temperature regimes) and a desert climate (BW; aridic or torric soil moisture regime, variable-temperature regimes), the first one being described by Köppen as semiarid, the latter one as arid. Half-R in the above formula designates the boundary between the two climates. In addition, both dry-hot (h) and dry-cold (k) climates are distinguished, with an additional option to indicate influence of fog (n) from cold ocean currents upon coastal deserts.

Evaporation exceeding precipitation leads to an upward movement of solutes in many soils of the dry climates. In steppe climates this leads to enrichment with secondary calcium carbonate ($CaCO_3$) and gypsum ($CaSO_4 \cdot 2H_2O$), whereas in desert climates even more soluble salts, such as mirabilite ($Na_2SO_4 \cdot 10H_2O$), thenardite (Na_2SO_4), hexahydrite ($MgSO_4 \cdot 6H_2O$), nahcolite ($NaHCO_3$), soda ($Na_2CO_3 \cdot 10H_2O$), trona ($Na_3CO_3HCO_3 \cdot 2H_2O$), and halite (NaCl), may precipitate. The accumulation of soluble salts is caused by the process known as 'salinization.'(*See* **Salination Processes**.)

The scarcity or even lack of vegetation means that wind action plays a significant role in these climates and has an important bearing on the soils. Wind speed not only regulates the amount of evaporation, also wind action is responsible for removal of fine particles (dust storms), and the accumulation of sands in dune complexes. Where fine particles are removed, often a gravely surface is left behind, giving rise to a so-called desert pavement (Figure 4). The gravels at the surface are shaped by scouring sand particles ('ventifacts') and have a shiny surface ('desert varnish').

Where cold ocean currents meet desert shores, such as the Benguela current along the coast of northwestern South Africa and Namibia, mist drifts inland, carrying moisture and solutes from the sea. This gives rise, over time, to deposition of, particularly, sulfates. The vast areas of accumulated gypsum, up to 150 km inland, in this region can largely be attributed to the influence of mist.

Soils in desert and steppe climates do not show much development apart from the accumulation of carbonates, gypsum, or salts. Under desert climates, various types of Aridisols (ST) occur, in particular Calcids (ST) or Calcisols (WRB), Gypsids (ST) or Gypsisols, and Durids (ST) or Durisols (WRB). Other soils are dominantly Entisols (ST), the sandy ones being Torripsamments (ST) or Arenosols (WRB), the others being Torriorthents (ST) or Regosols (WRB). Where salts have accumulated, Salids (ST) or Solonchaks (WRB) are found. Under steppe climates, where dominantly an ustic soil moisture regime prevails, Entisols (ST) or Regosols (WRB), Inceptisols (ST) or Cambisols (WRB), and Mollisols (ST) or Kastanozems and Chernozems (WRB) occur. Depending on the accumulated secondary material, Durustepts and Durustolls (ST) or Durisols (WRB), Calciustepts and Calciustolls (ST) or Calcisols and Kastanozems (WRB), and Argiustolls (ST) or Chernozems (WRB) occur. Other soils are Natrustolls and Natrustalfs (ST) or Solonetz (WRB), and Ustalfs (ST) or Luvisols (WRB).

Figure 4 Typical desert soil with pavement in Israel. (Reproduced with permission from ISRIC.)

Soils in Warm Temperate Climates

Warm temperate climates (C) are defined in the Köppen climate classification as having a coldest month with an average temperature below 18°C but above −3°C. Thus there is a distinct summer and winter season. Subdivision within the warm temperate climates is based on rainfall pattern (permanently moist (Cf), dry winter (Cw), or dry summer (Cs), and maximum temperatures during the summer season (a, hot summer (warmest month, above 22°C); b, warm summer (warmest month, below 22°C); c, cool, short summer (less than 4 months above 10°C). The permanently moist climate type (Cf) coincides roughly with the (per)udic soil moisture regime, the one with a dry summer (Cs) with the xeric soil moisture regime. Soil temperature regimes are usually mesic and thermic, but areas of the warm temperate climate with a cool, short summer (Cfc) may experience a frigid soil temperature regime.

Warm temperate climates have, either seasonally or permanently, an excess of rainfall. This gives rise to chemical weathering and leaching, temporarily or all year around. Due to the large fluctuation in temperature and the varying rainfall patterns, chemical weathering rates will be variable as well. However, a net transport of substances, either in solution or as colloids, from the upper parts of the soils to the lower parts or to the groundwater is the common denominator in most of the soils in warm temperate climates.

The chemical weathering of noncalcareous parent materials leads to the release of free iron hydroxides (amorphous iron hydroxide, $Fe(OH)_3$; goethite, α-FeOOH) and the formation of clay, particularly illite. This clay formation is mainly the result of microdivision from muscovite, accompanied by partial or total loss of fixed K^+ ions. The release of free iron hydroxides gives the soils in warm temperate climates their characteristic yellowish-brown color, a process also known as 'brunification.' In warm temperate climates with a distinct dry summer (Cs), dehydration of the hydroxides leads to more red colors (rubefaction) due to the formation of hematite (α-Fe_2O_3). The chemical weathering of calcareous parent materials starts with the dissolution and removal of the carbonates present. It is only after the process of decalcification that other processes, such as brunification, will start.

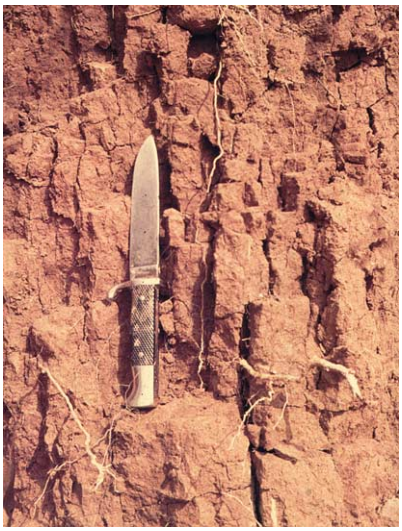


Figure 5 Red Mediterranean soil with well-developed argillic horizon in Greece. (Reproduced with permission from ISRIC.)

Of particular importance in these climates is physical translocation of clay particles by gravity water from the upper soil horizons to the lower soil horizons. This process, called clay illuviation, leads to clay-enriched, subsurface horizons, known as argillic (ST) or argic (WRB) horizons (Figure 5). The process is enhanced by a soil environment relatively poor in Ca^{2+} and Al^{3+} ions, both of which tend to flocculate the clay particles, thus inhibiting the dispersion necessary for the translocation of clay. It is further favored by desiccation of the soil, through which larger pores and cracks are created, and by the occurrence of heavy rain showers, creating the low-electrolyte environment needed for the translocation of clay. (*See* **Flocculation and Dispersion**.)

Soil types of the warm temperate climates are dominantly Alfisols, Inceptisols, and Spodosols (ST) or Luvisols, Cambisols, and Podzols (WRB). If dryness in the summer season prevails (climate types Csa and Csb), also soils with accumulation of calcium carbonate (Calcixerepts (ST) or Calcisols (WRB)), soils rich in sodium (Natrixeralfs (ST) or Solonetzes (WRB)), and shrink–swell clays (Vertisols (ST and WRB)) occur. In parts with high rainfall all year round, peat soils (Histosols (ST and WRB)) are found, together with soils having a thick, acid, and humus-rich surface horizon (Dystrudepts (ST) or Umbrisols (WRB)). Toward the boundary with cool temperate climates (D) Mollisols (ST) or Chernozems (WRB) are encountered.

Soils in Cool Temperate Climates

The cool temperate climate (D) in the Köppen climate classification is defined as having a coldest month with an average temperature below −3°C and an average temperature of the warmest month above 10°C. Distinction is made between a cool temperate climate, moist all-year round (Df) and climates with a dry winter (Dw). Further differentiation is based on maximum temperatures during the summer season: a, hot summer (warmest month, above 22°C); b, warm summer (warmest month, below 22°C); c, cool, short summer (less than 4 months above 10°C); d, very cold winter (coldest month, less than −38°C). The cool temperate climate approximately coincides with areas having frigid and (partly) cryic soil temperature regimes.

Large areas experiencing cool temperate climates have been formerly glaciated. This has resulted in deposition of glacial and periglacial deposits such as boulder clays, outwash material, fluvial deposits of braided river systems, lake sediments, loess, and cover sands. The heterogeneity of these sediments, in combination with the conditions of cool temperate climates, results in a variety of soil-forming processes.

Of importance in the cool temperate climates are the long, cold winters, during which the soils are frozen and covered by snow. In areas with slowly permeable parent materials (boulder clays, lake sediments, etc.) meltwater stagnates during springtime, causing waterlogging and temporary reducing conditions. This gives rise to dissolution and removal of iron, decline in soil structure in the surface horizons, and acidification. On better-drained parent materials (cover sands, braided river deposits, etc.), water stagnation may still be present due to the frozen subsoil, but not as long and as severe as on poorly drained parent materials.

The low temperature also hampers biological activity in the soils. Only during the warm summers will soil life be active and contribute to the process of homogenization. Consequently, humification and bioturbation are slow, particularly in the Dfc, Dfd, Dwc, and Dwd climates. On the other hand, where summers are sufficiently long and warm, and winters not very severe (Dfa, Dfb, Dwa, and Dwb climates), biological activity in the soils contributes greatly to the processes of homogenization and humification. This is especially so under grasslands with a high production of organic matter. (*See* **Grassland Soils**.) Burrowing soil animals homogenize the soils to a considerable depth, sometimes up to 2 m, and thoroughly mix the organic matter with the mineral soil constituents.

The processes described above lead, on the one hand, to well-differentiated soils with poor humification suffering strongly from water stagnation and, on the other hand, to well-homogenized and well-drained soils rich in organic matter. Dominant soils

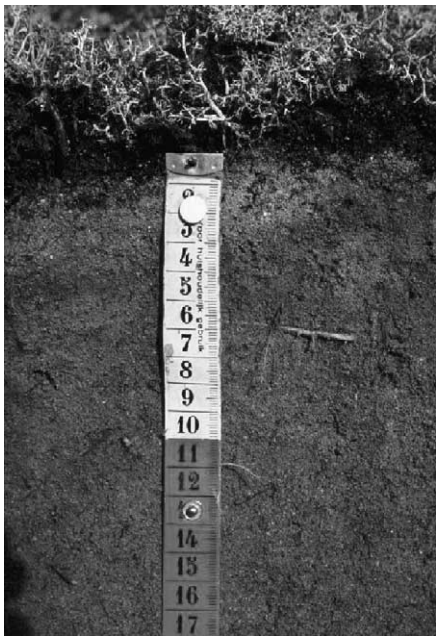


Figure 6 Mini-Podzol under tundra vegetation in northern Sweden. (Reproduced with permission from ISRIC.)

in the former group are Albaqualfs, Glossaqualfs, Epiaqualfs, Albaquults, and Epiaquults (ST) or Albeluvisols and Stagnic Luvisols (WRB). Mollisols (ST) or Chernozems (WRB) represent the group of well-homogenized and well-drained soils rich in organic matter.

The climate-induced coniferous forest in the colder part of the cool temperate climate contributes significantly to the development of acid soils, especially on sandy deposits. Here the process of cheluviation takes place, whereby organic acids, released from the vegetation debris, complex aluminium and iron that are subsequently leached down and deposited in lower parts of the soils. The result is the development of Spodosols (ST) or Podzols (WRB). Given the young age – geologically speaking – of the deposits and the slowness of the process because of the low temperatures, corresponding soils are often only weakly, though distinctly developed (Figure 6).

Soils in the Polar Climates

The polar climates (E) in the Köppen climate classification are defined as having an average temperature of the warmest month of less than 10°C. Two subclimates are recognized: the tundra climate (ET) and the climates of perpetual frost (EF). They coincide with the cryic soil temperature regime.

The low temperatures in regions with polar climates only permit physical weathering; chemical weathering is (almost) absent and, if any, is restricted

Figure 7 Horizon involutions in a Canadian soil. (Reproduced with permission from C. Tarnocai, Agriculture and Agri-Food, Canada.)

to the short thawing periods. In addition, biological activity is low due to the low temperature.

Characteristic for the soils in the polar climates are the effects of freezing and thawing and, in many cases, the presence of permafrost (permanently frozen subsoil with ice segregation). The annual freezing and thawing of the soils result in thermal contraction and expansion of minerals in rocks, thus slowly weakening the rock structure and permitting water to enter in small cracks. The freezing of water further widens the cracks, as ice has a lower bulk density than water (0.9 versus 1.0 kg dm^{-3}) and, over time, the rock will disintegrate into angular fragments. In soils, the annual freezing and thawing results in the process of 'cryoturbation,' giving such features as horizon involution (Figure 7), frost heave, cryogenic sorting, and patterned grounds (e.g., stone polygons).

The low rate of humification of organic matter and the wet conditions during the thawing periods result in the accumulation of peat at the surface. Peat soils are among the most widespread soils in the polar regions. Approximately half of the Earth's peat soils or Histosols (ST and WRB) occur in the circumpolar region of the northern hemisphere.

Other soils are dominantly Gelisols (ST) or Cryosols (WRB), soils that are characterized by the occurrence of permafrost within 100 cm depth, or within 200 cm depth in association with gelic materials above it, or having one or more cryic horizons within 100 cm of the soil's surface.

Soils as Climatological Archive

Some soil processes and resulting properties are clearly climate-related, such as the processes of ferralitization, rubefaction, and cryoturbation. Other processes such as brunification, decalcification, clay

illuviation, and cheluviation occur under a variety of climate conditions. Again others such as salinization and formation of secondary calcium carbonate are rapidly counteracted when climates become moister, and are not preserved in the soil. Therefore, the number of soil properties usable as climatological indicators is limited.

The process of ferralitization, which occurs in the moist tropical climates, leads to accumulation of iron (hydr)oxides, aluminum (hydr)oxides, and kaolinitic clay, all of which are very stable in soil environments. The occurrence of soils with large amounts of the above components almost invariably points to former moist tropical climates. A typical example of the inheritance of kaolinite from earlier weathering stages is the so-called 'arène,' i.e., the deeply weathered, sandy regolith over granites in France which has often been attributed to deep weathering in the Tertiary. Other evidence in soils for humid tropical conditions in Europe during the Tertiary era is found in the occurrence of laterites in the Mainz region of Germany.

Rubefaction – the process of red coloring as a result of the formation of hematite – is associated with climates that have a long, hot dry season. Rubified soils, now occurring in climates without a distinct dry period, are evidence of earlier climates that had a distinct dry period. Examples of these can be found in the cool climatic belt in western and central Europe, particularly in association with limestone and marls.

Frost action results, amongst other things, in the formation of frost wedges and polygonal structures. It is particularly these two phenomena that are frequently encountered in former periglacial areas. They are evidence of the formerly cold climatic conditions that occurred during the ice ages of the Quaternary Era (Figure 8).

Despite clay illuviation occurring in many different climates, it requires a net flow of water under gravity conditions from the surface downward. It is therefore unlikely that such conditions occur in the dry climates of the Earth. Yet the presence of clay illuviation features in present-day dry soils, as recognized in *Soil Taxonomy* by the suborder of Argids (Aridisols with an argic or nitric horizon), points to formerly moister climates with conditions more favorable to clay translocation.

Figure 8 Polygonal subsurface structures in a horizontal section in a loess soil in Belgium. (Reproduced with permission from ISRIC.)

Summary

It is evident that soils bear a strong stamp of past and present climatic conditions. Essentially, the ratio between precipitation and evaporation determines the flux of water in the soil (downward or upward). Also temperature is a key climatic component, as it not only regulates the speed of weathering (the higher the temperature, the faster chemical processes take place), but also the biological activity in the soil, which plays a crucial role in the breakdown of organic matter. Wind speed is important in the rate of evaporation; and wind action, of particular importance in dry climates with very little vegetation, contributes both to the removal of mainly finer particles and to the accumulation of mainly coarser particles.

See also: **Classification Systems:** FAO; **Clay Minerals**; **Flocculation and Dispersion**; **Grassland Soils**; **Metal Oxides**; **Salination Processes**

Further Reading

Doebl F (1973) Ein "Aquitan"-Profil von Mainz-Weissenau (Tertiär, Mainzer Becken). Mikrofaunische, sedimentpetrographische und geochemische Untersuchungen zu seiner Gliederung. (An Aquitanian age cross section at Mainz-Weissenau (Tertiary, Mainz Basin). Microfaunal, sedimentpetrographic and geochemical study of its subdivision.) *Geologisches Jahrbuch Series A5*.

Duchaufour P (1998) *Handbook of Pedology. Soils–Vegetation–Environment*. Rotterdam, the Netherlands: Balkema.

FAO-ISRIC-ISSS (1998) *World Reference Base for Soil Resources*. World Soil Resources Report 84. Rome, Italy: FAO.

Hsu PH (1977) Aluminium hydroxides and oxyhydroxides. In: Dixon JB and Weed SB (eds) *Minerals in Soil Environments*, pp. 99–143. Madison, WI: Soil Science Society of America.

Köppen W (1936) *Handbuch der Klimatologie,* vol. 1, part C. *Das geographische System der Klimate*. Berlin, Germany: Borntraeger-Verlag.

Marbut CF (1935) *Atlas of American Agriculture,* part III, *Soils of the United States*. US Department of Agriculture, Bureau of Chemistry and Soils. Washington, DC: US Government Printing Office.

Meyer A (1926) Über einige Zusammenhänge zwischen Klima und Boden in Europa. *Chemie der Erde* 2.
Rode AA (1962) Soil Science (Pochvovedenie, translated from Russian). Moscow, Russia: Goslesbumizdat.
Soil Survey Staff (1999) *Soil Taxonomy. A Basic System of Soil Classification for Making and Interpreting Soil Surveys*, 2nd edn. Agriculture Handbook 436. US Department of Agriculture, Natural Resources Conservation Service. Washington, DC: US Government Printing Office.
Trewartha GT (1968) *An Introduction to Climate*, 4th edn. New York: McGraw Hill.
VASKhNIL (1987) *Classification and Diagnostics of Soils in the USSR*. (Russian Translation Series **42**). Rotterdam, the Netherlands: Balkema.

Human Impacts

J Sandor, C L Burras, and M Thompson, Iowa State University, Ames, IA, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Humans, like many other organisms, are active agents of soil formation. Because soils comprise the dynamic, vibrant skin of the Earth's terrestrial surface, people have always interacted with, and therefore changed, soils and the course of their formation. While soils are subject to major change and even destruction by natural forces on the scale of geologic time, changes resulting from human activity usually occur on a much shorter time scale. People have impacted soil in a multitude of ways and extents, through farming, building, mining, and even war. In some cases, human activities enhance soils for particular uses. However, in a number of cases, the interplay between humans and soils has resulted in soil degradation, which is fundamentally a negative process of formation. Recognizing that soil use is literally and figuratively the base for most civilizations and that soil resources are essentially nonrenewable on the human time scale, understanding soil formation is imperative to developing agricultural and natural resource sustainability, and to protecting environmental quality.

Scales and Scope of Human Impacts on Soil Formation

To understand the effects of humans on soil, it is helpful first to consider soil-formation processes and their rates in natural environments to provide a frame of reference. From the perspective of pedology, the study of soil formation, soil is a complex assemblage of mineral and organic materials formed at the Earth's surface. In spite of the increasing impact of human activities on soil and the likelihood that all of Earth's ecosystems have been influenced to some degree by humans, many soils still retain their basic morphology imparted by natural processes and environmental factors. A hallmark of soil formation is the differentiation of horizons (layers). Soils are developed and organized into horizons by complex, interrelated physical, chemical, and biological processes that are determined by the factors of soil formation: climate, organisms, geology, topography, and time. The morphology of every soil is the expression of fundamental processes interacting with one another over multiple spatial scales (from micrometers to kilometers) and temporal scales (from days to millennia). Horizons at the surface (A horizons) are often enriched with organic matter, while deeper horizons (B horizons) may have accumulations of clay, calcium carbonate, metal–organic complexes, or other materials. Formation of A horizons is relatively rapid because organic matter accumulates in a few centuries to a millennium. Formation of B horizons commonly takes many thousands of years to become fully expressed. In this sense, soils can range from young to middle-aged to old. As soil formation progresses, soils generally tend to become increasingly anisotropic as they differentiate into greater kinds and numbers of horizons.

Landscape stability is a prerequisite for soil horizon development to proceed. Even so, disturbance and change are integral to the functioning of all natural ecosystems and their soils. Soils, landscapes, and associated biological communities are subject to disturbances ranging from minor perturbations such as low-intensity fires to major events such as volcanism, and to long-term climatic and environmental change. On the geologic time scale, soils are subject to major alteration, destruction, and renewal. In contrast to the relatively slow pace of natural soil development, soil changes resulting from human activity are often more rapid and far-reaching. Human-caused change may be so fast and irreversible that the impacted soil bears little resemblance to its original form. Human impacts usually reverse the anisotropic trend of soil formation, making soils simpler, less organized, and more homogeneous.

A wide array of land use and other human activities have altered paths of formation in many soils (**Table 1**). Corresponding soil changes also vary greatly in kind, intensity, time and spatial scale, and significance for soil and environmental quality (**Tables 2** and **3**). Human actions that change soil may act directly or indirectly by changing both soil morphology and the

Table 1 Human agents of change in soil properties and processes

	Examples
Agriculture and forestry	
• *Crop and animal agriculture*	Tillage, fertilization, cropping, flooding, irrigation, drainage, terracing, grazing
• *Forestry*	Altered vegetation type and cover, harvesting operations
• *Off-site effects*	Erosion and sedimentation, causing soil contamination and burial
Cities and industry	Excavation, urban cover, artificial fills, industrial pollution
Buildings, roads, and other structures	Excavation, artificial fills, pollution, paving, land leveling Dams and reservoirs, polders, dikes, mounds, other artificial landforms and fills
Mining and other earth and water resource extraction	Soil removal for pits and quarries, erosion and sedimentation associated with hydraulic mining, mixing and inversion of earth materials, reclamation, groundwater pumping
War and weapons testing	Bomb craters, military transport and engineering, contamination by radioactive elements
Human-linked environmental change	
• *Global climate change*	Atmospheric CO_2 increase, climate warming, increased climatic variability
• *Desertification*	Vegetation loss or change, accelerated wind and water erosion, salinization
• *Acid rain*	Soil acidification and other chemical changes

Table 2 Soil change resulting from human impact

Causes and characteristics of soil change[a]	*Examples*
Possible causes	
Direct	Compaction
Indirect	Slower or diverted water movement and reduced soil water storage after compaction
Deliberate	Fertilization, liming, irrigation
Unintended	Nutrient depletion or imbalance; salinization
Constructive	Plaggen soil; agricultural terracing and drainage
Destructive	Accelerated erosion, sulfide oxidation with drainage or exposure, salt/sodium buildup
Magnitude and extent	
Low impact	Organic matter change from light grazing
High impact	Urban expansion; land filling
Part of soil	Thinning of A horizon from cultivation-induced erosion
Whole soils	Removal of entire soil by intense erosion
Duration and rate	
Short-term/ephemeral	Alleviation of plow pans by freezing and thawing; liming
Long-term/permanent	Urban soil; mine soil
Slow rate	Agric horizon development
Fast rate	Oxidation following drainage
Response of soil to human impact	
Susceptible	Base-poor soil susceptible to acid rain
Resistant	Base-rich soil buffered against acid rain
Reversible (resilience)	Alleviation of compaction is faster in organic matter-rich A horizons
Irreversible	Laterite/plinthite hardening with exposure; construction of urban soils and soils on mined land
Outcomes of soil change for soil quality and use	
Beneficial	Organic matter addition
Neutral/benign	Fertilizing that balances crop removal of nutrients
Degradation (loss of productivity/ off-site environmental impacts)	A horizon erosion

[a]Bracketed terms illustrate range end members.

underlying soil-forming processes. They may be intentional changes based on land-management strategies to increase soil productivity, or rearrangements of landscapes and soil materials to construct buildings, roads, and other structures. Conversely, they may be unintended changes that can lead to degradation such as soil erosion and burial under sediment. Indirect soil change may result from off-site processes physically remote from the impacted soil, such as subsidence following groundwater pumping or downstream

Table 3 Spatial scales of human-induced soil change

Soil components	*Approximate spatial scale (m)*	*Examples of impacts*
Physical, chemical, and biological properties of soils (e.g., clay, microorganisms, organic matter)	10^{-10}–10^{-4}	Retention of pollutants such as heavy metals, pesticides, and industrial solvents
Morphological properties	10^{-3}–10^{-2}	Soil structure degradation; changes in texture, color, porosity, and pore distribution
Horizons	10^{-1}–10^{0}	A horizon erosion
Whole soils (pedon)	10^{0}–10^{1}	Plaggen soils; salinized soils; liming that changes Ultisols to Alfisols
Soil–watersheds–landscapes–ecosystems–biosphere	10^{2}–10^{7}	Broad changes in Mollisols, Histosols, Gelisols, and other soil orders

sedimentation. In land uses such as agriculture, surface horizons have usually been more altered than subsurface horizons, while engineering activities often change the entire sequence of soil horizons. Engineering activities lead to some of the most intensive impacts by removing soils entirely and replacing them with different earth materials, as well as by reshaping landforms or constructing new ones. Agriculture and engineering merge in management practices such as in paddy rice production and terracing of slopes, which also involve major geomorphic and soil change.

Some human activities such as agriculture have influenced soil processes and morphology for thousands of years, while others such as global climate change induced by human activity are becoming increasingly important. Some anthropogenic impacts on the environment, such as acid rain and other pollution, incur less visible chemical and biological changes, while processes such as desertification result in more sweeping changes through loss of vegetative cover, desiccation, wind and water erosion, and salt accumulation. While the atmospheric effects of global climate change have been detected relatively recently, their potential impacts on soil formation and distribution are large, because climate is a major determinant of soil formation over time.

Soils differ in their response to human actions in terms of their resistance to change, and their resilience, that is, their ability to rebound toward their original state. Processes of human-caused soil change may be reversible or irreversible, and the resulting changes may be ephemeral to permanent. Response depends on both external factors, such as the type of impact and environmental conditions, and internal soil properties. For example, soils with uniform textures or that are rich in organic matter tend to be resistant to compaction. After many thousands of years of natural formation, some subsurface soil horizons opened by deep plowing may return to their original condition in a few years (e.g., clay-rich argillic horizons), whereas more indurated horizons such as silica-cemented duripans remain fragmented longer.

Evaluating Human Impacts On Soil Formation

Human-caused soil change has been studied by pedologists in several ways. Hans Jenny and others have framed their views in the context of the factors of soil formation, placing humans within the biotic factor, while recognizing the unique role of human culture as distinct from other organisms. Jenny showed that human land use often alters soils through changes in the other soil-forming factors. For example, irrigation alters arid land soils by changing the climate factor, effectively increasing precipitation. Other researchers set humans apart from the natural factors of soil formation to emphasize the extraordinary scale and rate of anthropogenic effects on soils. In working with drastically altered soils such as mine soils and urban soils, pedologists have defined new soil taxonomic classes, because these soils bear little or no resemblance to their original state.

How is soil change detected and measured? While modern, large impacts on soil may be obvious, such as with wholesale change or destruction of soils and landscapes in urban development, longer-term soil changes are often more subtle and complex. Histories of soil change may be difficult to reconstruct, complicated by imprints of multiple land use activities and changing environmental conditions. An approach to evaluating long-term anthropogenic influence on soil has been to identify soils that are relatively undisturbed, or at least that have documented land-use histories, and to use these soils as reference points from which to measure soil change. Finding truly comparable sets of reference and altered soils is challenging and imprecise, and may not be possible in some situations. Although results of comparisons must be interpreted carefully, the paired-site approach represents one of the few methods available actually to measure long-term soil change. Soil changes from agriculture have been monitored using controlled experiments at a few locations such as Rothamsted Experimental Station (UK), Morrow Plots (Illinois,

USA), and Sanborn Plots (Missouri, USA) for more than a century. Evaluating ancient agricultural soils up to about 1000–2000 years in age has extended the time perspective on anthropogenic soil change. Much can be learned about long-term human effects on soils, and about successes and failures in soil use and management, from past cultures at archeological sites, as well as from contemporary traditional cultures who have lived on the same land for many generations. Studies of past and present traditional land use contribute information for evaluating soils on longer time scales inherent in the concept of sustainable land use, as well as for modeling and predicting the future condition of land resources.

Changes in Soil Properties and Processes Resulting From Human Activities

To facilitate understanding of impacts and appreciation of the many ways people depend on soil for sustenance and support, soil changes are presented here in the context of the type of land use or other human activity. Examples from agriculture, engineering for urban development, mining, waste disposal, war, and global environmental change are selected to represent the myriad of ways people change soil. It is important to recognize that human effects on soil are complex, because soils are dynamic systems with interrelated components; they are themselves part of ecosystems within Earth's biosphere. Soils vary in their sensitivity to disturbance and threshold for change. Single land-use practices can affect many soil properties in a cascading process, and different land uses may initiate similar processes of change in soils. Soils may be impacted by more than one land-use practice simultaneously or diachronically. Land uses and their variations often differ in the way they impact soils in terms of magnitude, spatial extent, rate, and duration.

Agriculture

Agriculture has profoundly and extensively impacted soils since its inception about 10 000 years ago. In this discussion, agriculture is framed broadly to include all plant and animal production for food, feed, fiber, and fuel, including crop and livestock farming, as well as forestry. All of these land uses rely on many forms of soils worldwide.

Some soils are so impacted by agriculture that their original horizons are wholly transformed or buried. Examples of truly anthropogenic soils are plaggen soils, common in western Europe (Figure 1). These constructed surface horizons, which can be a meter thick, are the product of centuries of cultivation, with additions of organic materials such as manure and sod and inorganic amendments such as sand or marl. Soils that have been similarly transformed through management practices such as terracing and long-term applications of fertilizing materials are found in other regions with long histories of intensive agriculture such as Southeast Asia and the Andes.

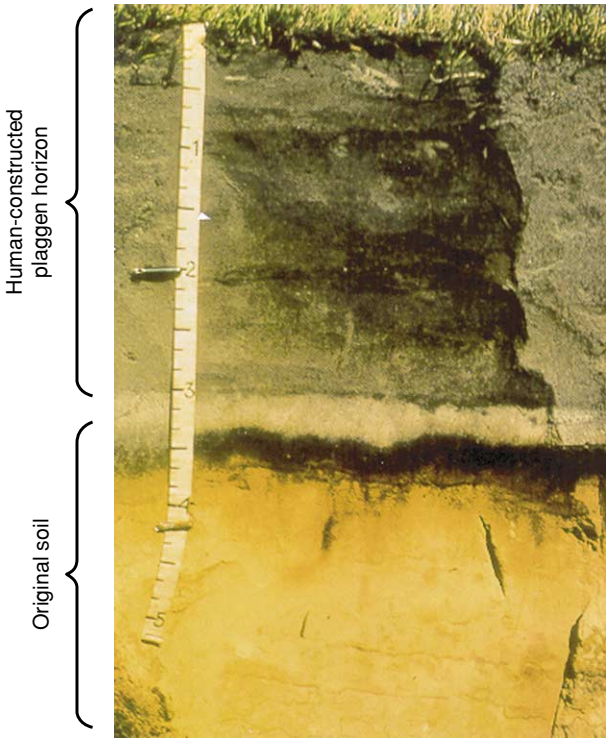


Figure 1 Plaggen horizon (approx. 1 m thick) from the Netherlands, created by long-term organic matter additions and cultivation, burying a natural Spodosol. (Reproduced from Soil Survey Staff (1999) *Soil Taxonomy,* 2nd edn. US Department of Agriculture. Washington, DC: US Government Printing Office.)

In contrast to both the intent and effect of constructed soils, many soils on slopes have been impacted by erosion accelerated by agriculture, some to the point of obliteration or deep burial under eroded sediments through mismanagement. Surface horizons, which generally contain the most plant nutrients and organic matter, are the most immediately vulnerable to erosion. Especially subject to major change are those soils with well-developed, organic matter-rich surface horizons such as Mollisols. The transformation of Mollisols by erosion to soils classified in other orders such as Alfisols, Inceptisols, and Entisols has been documented for past and present agriculture. A key reason for increased soil erosion is the loss of protective vegetation cover as natural ecosystems such as forests or prairies are converted to agricultural use. While accelerated erosion has been difficult to quantify on regional to global scales, estimated average erosion rates on cropland range from about 12 (USA) to 30–40 (Africa, Asia, South

America) metric tons per hectare annually, much greater than natural erosion rates. Human-induced water and wind erosion is estimated to have resulted in the degradation of approximately one-quarter of the world's cropland.

As surface horizons are eroded, subsurface horizons or bedrock effectively rise closer to the surface and may become exposed. Their incorporation into the topsoil under moderate to severe erosion alters soil properties such as color, structure, and texture, often lowering soil quality. If subsurface horizons are problematic for plant growth, serious degradation of productivity can ensue. In some tropical regions such as West Africa and Southeast Asia, exposure of horizons of cemented, iron-rich clay (petroplinthite) through erosion has resulted in abandonment of agricultural land. Erosion of Ultisols with highly weathered, acidic horizons of clay accumulation in warm, temperate to tropical environments has significantly reduced soil productive potential. Even in thick sediment parent materials that are considered 'forgiving' to erosion, productivity can diminish through exposure or shallow occurrence of dense or cemented horizons and layers that limit rooting, decrease available water capacity, or greatly slow drainage. For example, thick deposits of loess (wind-blown sediment rich in silt) such as in China and the Palouse and Midwest regions of the USA are favorable for crops but particularly susceptible to erosion. Erosion rates in loess can reach 100 metric tons per hectare per year, exceeding natural erosion rates by one to two orders of magnitude. This has resulted in the loss of A horizons, followed by the exhumation of buried soils (paleosols) or dense, brittle fragipans at shallower depths that reduce productivity.

Erosion also impacts soil formation off-site. As eroded sediments are transported downslope, soils in the lower terrain of watersheds may become buried by sediment that does not reach streams. Studies in several regions have measured burial of original soils by sediment that is centimeters to meters thick. Also, the freshly deposited sediment constitutes another parent material in which soil formation may begin anew. An opposite situation is where natural sedimentation processes and renewal of soil fertility are impaired by dam construction, as in the case of the Aswan dam on the Nile river. The extensive erosion and sedimentation that occur in many agricultural landscapes indicate that massive transformation of many soils is continuing. This situation underscores the need for documenting and responding to changes that undermine soil quality and productivity.

Because water deficiency or excess is of major concern in most agricultural systems, many soils have been altered by water-management practices. Some changes are direct, such as the effects of flooding soils for wet rice production, creating paddy soils with distinctive anthropogenic characteristics (Figure 2). 'Irragric' soils in arid to semiarid central Asia have been highly altered by long-term irrigation, including addition of suspended sediment in the irrigation water. Unintended chemical effects from irrigation in some regions have led to severe soil degradation through salt accumulation and transformation to saline or sodic soils. In contrast to paddy soil management, conversion of wetlands to row crop production by artificial drainage in management, regions such as the Midwest USA has changed soils from dominantly anaerobic to dominantly aerobic states, leading to changes such as organic matter oxidation. In some areas of the world, many meters of organic soil thickness have vanished through oxidation processes, resulting in significant land subsidence.

Figure 2 Sample (approx. 5 cm thick) of paddy soil managed for wet rice production in Southeast Asia. Anthropogenic features include variegated colors from flooding and alternating reduction/oxidation, and vesicles formed from gases trapped beneath the platy, puddled surface layer. (Reproduced with permission from Moorman F and van Breemen N (1978) *Rice: Soil, Water, Land*. Los Baños, Philippines: International Rice Research Institute.)

More subtle agricultural impacts may not completely alter the original soil, yet over time cause significant change. For example, since the mid nineteenth century, conventional cultivation of organic matter-rich, prairie-derived soils (Mollisols) throughout the Midwest USA has led to marked decreases in organic matter in upper soil horizons (Figure 3). Many Mollisols have lost approximately one-third to one-half of their original organic matter. In addition to erosion, a principal cause is the disruption of soil aggregates by cultivation, making previously protected humus accessible to microorganisms. Compared with native

prairie ecosystems that are characterized by abundant organic matter and conservative nutrient cycling, agricultural soils have lower inherent fertility and are more 'leaky' with respect to nutrients such as nitrogen. Soil organic matter loss has a cascading effect on soil properties such as structure, reducing aggregate stability, and making soils more prone to compaction. Organic matter contents can be at least partially restored to soils through management, by using more diverse, conservation-oriented cropping systems, or by a return to more natural vegetation.

Cities and Industry

The growth of cities and industry has profound impacts on soils and their continued formation. Some of the impacts are direct; others are indirect. The creation of constructed urban soils is perhaps the most easily recognized direct impact of city growth on soils (Figure 4). Constructed soils have distinct profiles that are typically the product of the cheapest engineering fill available at the moment of creation. Soil scientist Phillip J. Craul wrote that the history of urban development has been based upon the premise 'dirt is dirt and it's cheap.' These human-created soils may be highly stratified and composed of very different types of fill in terms of their texture, mineralogy, and chemistry. Total fill – and consequently soil – thicknesses can reach multiple meters, with a soil's maximum age corresponding with the age of the city. That is, a constructed urban soil in Rome, Athens, Mexico City, and Beijing can be several thousands of years old, whereas a similar soil in New York or Buenos Aires must be less than 500 years. Of course, many new hectares of these soils are created each year in every city as new projects are completed. Many ancient cities occur on tells, human-created mounds or hills formed through time as cities were rebuilt on top of previous ones. Other construction materials used in urban soils include composted sewage sludge, municipal solid wastes, heavy metals, glass, plastic, and metal. Still others consist of clean sand mined from nearby rivers or topsoil transported from local farms. In other words, urban soils have an exceptionally high degree of disorganized variability.

A second direct impact of city development on soil formation is at the landscape scale as houses, gardens, parks, and streets replace the natural terrain. The result is fragmented landscapes, where soil-forming processes become controlled by constructed topographic, hydrologic, and ecologic factors. Thus, while the morphology of many of these soils does not show the dramatic changes of their downtown counterparts,

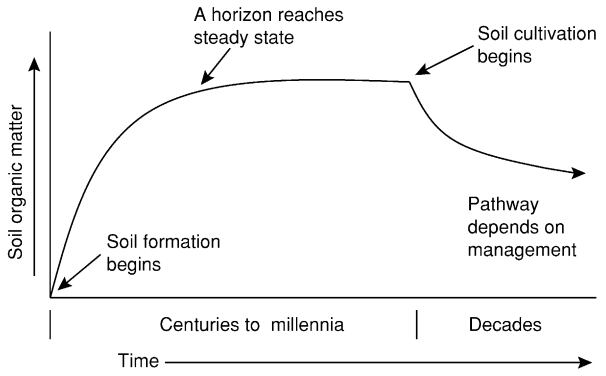


Figure 3 Accumulation of organic matter in topsoil (A horizon) during long-term natural soil formation, and subsequent rapid decrease following cultivation (based on the work of Hans Jenny).

Figure 4 Urban soil being created after the natural soil has been removed. Note natural soil profile in background. Photo by Julie McLaughlin.

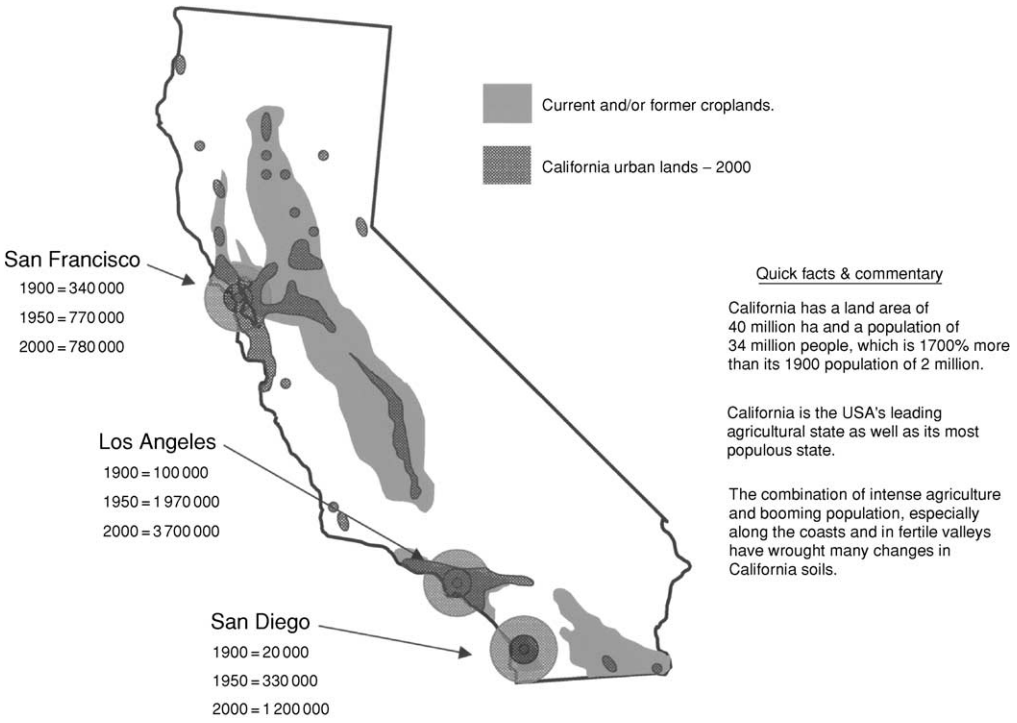


Figure 5 California: an example wherein soils are increasingly a product of human activities such as agriculture and urbanization. Compiled from a variety of sources including US Census Bureau and US Census of Agriculture data and maps for c. 1900, 1950, and 2000, several atlases as well as the Great Valley Center and US Geological Survey's ''Preliminary Assessment of Urban Growth in California's Central Valley.''

their soil-forming processes are considerably altered. This can be illustrated by considering two hypothetical, adjacent gardens: one growing turf that is heavily fertilized and irrigated, while the other is planted to conifer shrubs and trees but receives no fertilizers and extra water. Over time, the properties of the soils under these two gardens will diverge in response to their different microclimates (wet versus dry), biota, and chemical inputs. Like their constructed counterparts, they will be characterized by disorganized variability, although in this case the disorganization occurs at the landscape (i.e., garden-to-garden) scale.

The impact of cities and their continued growth on soil formation is increasing as large areas of prime farmland as well as wild, often fragile, lands get converted to urban uses (Figure 5). In the USA the area of urban land increased from 6 million to 26 million hectares between 1945 and 1992, with some parts of California and New England developing into '100-mile cities' as urban areas became interwoven through express highways and suburbs. Urbanization is even more of a critical factor in developing nations, because their populations are growing at a rate fourfold greater than the world population. The World Resources Institute estimates that nearly 500 000 ha of rural soils are converted to urban land annually. For example, the urban core of São Paulo, Brazil, expanded from 18 000 to 90 000 ha between 1930 and 1988, with the overall metropolitan area encompassing more than 800 000 ha.

Industry, like cities, directly and indirectly affects soil formation. Land excavation and drainage are two important direct agents of change. Dredging of wetlands as well as sediments from beneath shallow waters to improve harbors, build canals, and meet other commercial needs is an example of industrial excavation. The US Army Corps of Engineers estimates that billions of cubic meters of sediment are dredged annually around the globe with over 300 million cubic meters being dredged annually just in the USA. Dredging has three direct impacts with respect to soils and their formation. First, the dredging itself destroys whatever natural soil existed. Second, the dredged material – or 'spoil' – is piled somewhere and is later used in soil construction. Soil created from dredge spoil may be used for buildings, roads, parks, farms, or even wild areas. Over time it will develop a sequence of horizons in response to its environment. The third pedological impact of dredging is oxidation of mineral and organic matter. This occurs because spoil that was excavated from below the water table is now directly exposed to the oxygen-rich atmosphere. In many cases, the oxidation is fairly benign in terms of soil properties and

formation. A noteworthy problem occurs when the spoil contains sulfides. Exposure of sulfidic compounds to air results in the rapid formation of large quantities of sulfuric acid in these soils. Such acid sulfate soils are also created when naturally sulfidic soils are drained. It is estimated there are 24 million hectares of potential acid sulfate soils globally, with many located in prime settings for agricultural and urban development.

The indirect impacts of industry on soil formation are as manifold as the indirect impacts of urbanization. One of these is the addition to soil of contaminants such as lead, cadmium, and other heavy metals. In some locations these metals are deposited on soils as airborne contaminants that originated in factory or vehicle exhausts. In other cases they are added to soil directly and intentionally as constituents of fly ash and other wastes. 'Land farming' refers to disposing of waste by applying and incorporating it into soil a few tons per hectare at a time. These soils are subsequently cropped, although the crops are not normally used for human consumption. Compounds that contain sodium comprise another class of industrial contaminant that changes soil morphology – ultimately creating sodic soils. This process is increasingly noted worldwide because of the mind-boggling number of industrial uses for sodium and sodium compounds (e.g., coolant in nuclear processes; fumigants; solvents in manufacturing of brass, paper, ceramic glazes, textiles, and fertilizers; and in food processing) and the frequency at which sodium wastes are applied to land. Radionuclides are another class of industrial contaminants that alter soil processes. For example, 8.4 million hectares around Chernobyl, Ukraine, remain contaminated by ^{90}Sr and ^{137}Cs from the 1986 nuclear reactor disaster there.

The influence of most industrial processes on soil formation is proportional to the intensity of industrial activity and distance of the soil from the industry. This means that soils in areas having numerous factories, automobiles, and other industrial activities are more highly enriched in contaminants than soils where there is little industry. Likewise, the magnitude of impact radiates out with distance from the source of impact.

Mining

Humans have mined the Earth for metals, minerals, and fuel for thousands of years. Currently, 8 billion tons of minerals and oil are extracted from the Earth annually. Since the nineteenth century, mining activities have radically transformed landscapes and altered natural soil development processes. Mining requires that the overburden of soil and its parent materials be displaced so that the product (e.g., ore, rock, or coal) can be removed, concentrated, and transported off-site as economically as possible. Mining often results in large pits, high walls, and large accumulations of unconsolidated overburden, smelting residue, and finely ground residual ore. When abandoned at the Earth's surface, these materials often pose significant environmental risks due to subsidence, settlement, and slope instability. Spoil may also be contaminated with soluble metals. In addition, when some types of spoil are exposed to air and percolating water, oxidation reactions produce strongly acidic water with high metal concentrations, severely restricting subsequent biological activity. In the USA alone, it has been estimated that surface mining has disturbed more than 23 000 km^2.

Concern about mines and their spoils prompted many nations to enact environmental-quality laws beginning in the late twentieth century. For example, before the Resource Conservation and Recovery Act of 1976 and the Surface Mining Control and Reclamation Act of 1977, many mines in the USA were simply abandoned when the product had been exhausted or was too expensive to extract from the remaining ore. Since that time, however, planning for and completing reclamation of mined land has been part of every mining operation in the USA. Thus a number of practices are actively followed to allow the site to be used again for other purposes and to ensure that surface water and groundwater resources are protected. Construction of new soil at reclaimed mine sites is an example of the most radical of human impacts on soil formation.

Successful mine land reclamation must address a number of large-scale environmental issues while at the same time creating a near-surface environment that promotes vegetative growth and movement of water through the watershed. For example, a major problem in reclamation of coal and metallic ore mines is the oxidation of sulfide minerals and subsequent drainage and runoff of acidic water (**Figure 6**). Acidic spoil, surface water, and seeps must be treated with large amounts of calcite or hydrated lime to neutralize acidity before vegetation can be established. Organic matter and plant-available nutrients are commonly lacking in spoil, so application of sewage sludge or other organic amendments to the spoil surface is often helpful. Heavy applications of plant nutrients such as phosphorus and potassium may also be required.

Large-scale topographic reconstruction of mined land is expensive, but it is usually essential to ensure that reconstructed soils maintain ecologic and environmental quality. In some reconstructions, it is possible to approximate the general contours of the premining landscape (**Figure 7**). Throughout the process of consolidating, grading, and shaping the new landscape, control of settlement, slope instability, and surface erosion will minimize off-site damage of

Figure 6 Water that drains through spoil from mines is often highly acidic because of the oxidation of sulfur and iron. This water is toxic to plants and aquatic organisms. Acid mine drainage at the Minnesota Ridge Mine, South Dakota. This site was remediated in 2001. (Photograph by Tom Durkin. Reproduced with permission from the South Dakota Department of Environment and Natural Resources, Minerals and Mining Program.)

Figure 7 The Kennecott Flambeau Mine near Ladysmith, Wisconsin, was mined for copper, gold, and silver in the 1990s. At the end of mining, the open pit was completely backfilled and native vegitative communities were established. The site now includes high quality wetlands, native prairie, woodlands, and nature trails. Photographs before mining (1991), during mining (1996), and after restoration (2001) by T-B0 and courtesy of the Kennecott Flambeau Mining Company.

sediments and drainage as well as maintain optimum conditions for plant growth and development of natural soil horizons. Water management in reclaimed mine lands includes designing watersheds for directing both run-on and runoff water. The new landscape may be designed to retain or slow the movement of water by passing it through constructed wetlands.

Where the original materials were not systematically stockpiled during mining, reclaimed soils often display extreme variability in particle size and composition. Problems of inhomogeneous mixtures of coarse and fine particles include both subsidence and inadequate water-holding capacity for plants to grow. For this reason, modern reclamation approaches usually call for restoration of the natural sequence of geologic materials (e.g., bedrock-derived spoil placed below till-derived spoil placed below organic matter-rich topsoil). Rates of soil development in reconstructed mine soils are slow. Reclaimed mine soils tend to have 'AC' horizon sequences, in which incipient topsoil is underlain by unconsolidated layers lacking soil structure or other pedogenic features. Still, where reclamation has been well managed, the thickness of surface horizons in the reclaimed soil may exceed those of nearby unamended soils that have been degraded by farming or logging operations.

War

War, like mining, may have rapid and long-lasting impacts on soils. For example, nearly a century after an 11-month battle during World War I, extensive trench works and bomb craters remain near Verdun, France. Vietnam provides a documented example of the impacts of more recent wars. There, it has been estimated that bombs created some 21 million craters between 1965 and 1971, covering more than 50 000 ha (Figure 8). The explosion of a 2400-kg bomb typically created a hole approximately 10 m in diameter and 5 m deep; metal fragments from each bomb could be spread over an area of 0.5 ha. Some craters exposed soil horizons of iron oxide accumulation that subsequently irreversibly hardened to petroplinthite.

Similarly, land-clearing programs in Vietnam altered soil processes. Widespread and concentrated application of defoliating herbicides, as well as bulldozing of vegetation, resulted in extensive erosion, loss of organic matter-rich surface horizons, and increased flooding. Native mangrove forests have not returned to the drastically disturbed or herbicide-treated lands. More than 30 years after the application of herbicides ended, dioxin, a manufacturing contaminant in the herbicides, remains in the soil and sediments of the countryside. Thus soils are effectively beginning a new phase of soil genesis with topography, parent material, and vegetation altered from the previous state of dynamic equilibrium.

In addition to the persistence of war-related changes on soil properties and processes, there may be indirect effects that have a cascading environmental impact caused by displacement of civilian, agrarian populations. In rural Vietnam, bomb craters' soils contaminated with metal fragments, unexploded ordnance, and residual dioxin, and war-related destruction of water-diversion structures have made agricultural and forestry recovery very difficult for the inhabitants. In response, some farmers have moved to less-productive, marginal land that is more susceptible to erosion. Others have moved to urban slums, increasing the environmental pressures on urban areas. Thus the long-term impacts of war on soil properties and the processes of soil formation can extend beyond the zone of immediate impact.

Climate Change and Soil Formation

Soil formation is closely tied to climate, and, to the extent that human activities alter local or regional climatic variables such as temperature and effective precipitation, soil properties inevitably will be changed also. At the global scale, climatic change in the next century is likely to be driven by increasing atmospheric concentrations of 'greenhouse gases' such as carbon dioxide (CO_2) and methane (CH_4). As a result, it is expected that the mean temperature of the Earth's surface could rise as much as 1.5–4.5°C over the next 100 years. Although the local impacts of global warming are difficult to predict with certainty, significant shifts in the mean and seasonal variation of air temperature as well as in the total and seasonal distribution of rainfall are likely to perturb

Figure 8 A cratered mangrove forest in Gia Dinh province in August 1971. Bomb craters created during the war in Vietnam disrupted agriculture and most other land uses. (Photograph courtesy of Arthur Westing. Reproduced from Westing AH and Pfieffer W (1972) The cratering of Indo-China. *Scientific American* 226: 20–29.)

both natural and agricultural ecosystems. The impact of a global increase in temperature will be unevenly spread in both space and time. Some of the impacts include: (1) new precipitation patterns in which today's wet regions get wetter and dry regions get drier; (2) migration northward of boreal and hardwood forests; (3) melting of ice sheets, sea ice, and glaciers, with concomitant rise in sea level and flooding of coastal soils; and (4) melting of permafrost in northern latitudes and oxidation of organic matter now stored in cold-region soils. For example, the temperature increase has the potential to markedly shrink the extent of approximately 11 million square kilometers of Gelisols, the cold-region soils with permafrost, where an estimated 20% of the world's soil organic carbon is stored.

Human activities, mainly burning of fossil fuels, will thus lead to some global and regional climate change over the next century. Both regionally and locally, climate change will directly and indirectly impact soil processes by altering vegetation patterns, encouraging increased water and wind erosion, favoring mass movement, increasing the leaching of nutrients and organic matter through soils, and favoring microbial oxidation of organic matter in surface horizons. Although we are not now able to predict how fast soil properties will be altered as a result of global and regional climate change, ultimately some types of soil horizons will be lost (for example, surface horizons composed entirely of organic matter) and some types will be newly developed where they did not previously exist (for example, spodic horizons under new boreal forests). As human populations adapt agricultural practices to new climate conditions, further changes in soils may be accelerated. For example, salinization may increase due to more extensive irrigation, and landscape instability may become widespread as more forest land is cleared for production of crops and livestock.

Significance and Future of Human Impacts on Soil Formation

Soil is a critical, dynamic natural resource and vital component of ecosystems, but one that is often neglected. This is in part because soil lies beneath the surface and so is not as familiar as other resources such as water, plants, or animals. As a natural resource, soil is crucial for food-, fiber-, and fuel-production systems, for construction materials and foundations, for replenishing and maintaining the quality of surface water and groundwater, and for waste processing and containment. Natural variation in pathways of soil formation and soil properties make for a diverse range of suitabilities for different land uses and sensitivities to anthropogenic change. Conservation of soil resources and soil quality is a critical priority globally because soil is fundamentally nonrenewable on a human time scale.

Human activities dictate that soil change is inevitable and necessary. Anthropogenic soil change has a long history, but it has exponentially increased in intensity, spatial extent, and rate since the nineteenth century. Some changes are advantageous and improve soil functionality. However, many soil changes induced by past and present land use have resulted in environmental degradation. In developed and developing nations alike, accelerated erosion, compaction and disruption of structural aggregates, lowered fertility, and contamination by pollutants continue to be sources of serious concern. Because soil serves as a filter, substrate, and reservoir linked to other land, water, and biological resources, degradation is not just detrimental to the soil internally and locally, but it extends to all parts of the larger hydrologic, geologic, and biological system. Some soils, like plants and animals and the habitats and ecosystems with which they are closely connected, have become endangered and in some cases extinct. Human-altered soil is a significant factor in global environmental change, as both a cause and a recipient of change. Oxidation of soil organic matter is a source of increased atmospheric carbon dioxide; soil properties stand to be further altered by global climate change; and yet soil also has the potential to help reverse the trend by sequestering more carbon. Soils also play a key role in other forms of global change such as desertification, acid precipitation, and loss of tropical forest ecosystems.

What can be done to counter the negative impacts of human activities on soils and environment and to support conservation and more sustainable land use? Improved knowledge of soil formation and soil change is an important starting point. By more clearly recognizing the extent of soil change and understanding its causes, mechanisms, and consequences, the better our chances become to develop and implement management practices that can sustain soil resources and restore damaged land. Traditionally, soil maps have portrayed soils in their relatively original, undisturbed state. However, because of the unprecedented scale of soil change and massive transformation, it is imperative that anthropogenic soil change be documented, monitored, and acted upon to a much greater degree. Efforts to deal seriously with the extent and significance of human impact on soil formation and distribution have led to a proposal to include Anthrosols in the US Soil Taxonomy, and to recognition of Anthrosols at the highest level in the current international soil classification system (World Reference

Base for Soil Resources). Improved knowledge of soil formation processes in relation to natural and human-altered pathways is essential to the restoration of ecosystems and the development of sustainable land use. The future of human society and the Earth we inhabit depend on soil formation processes, and increasingly, on how we respond to the changes human actions cause in soils and the biosphere.

List of Technical Nomenclature

A horizon	Surface mineral horizon with some organic matter accumulation
argillic horizon	Subsurface diagnostic horizon with significant clay accumulation
B horizon	Subsurface mineral soil horizon (several kinds)
C horizon	Relatively unweathered, unconsolidated mineral horizon or layer (or soft bedrock), usually beneath the active zone of soil formation
fragipan	Dense, brittle subsurface horizon
irragric horizon or soil	Anthropogenic horizon formed by long-term irrigation and accompanying additions of sediment
paleosol	Soil formed in past geologic time
petroplinthite	Hardened or cemented horizon of highly weathered, iron-rich soil
plaggen horizon or soil	Anthropogenic horizon formed during long-term cultivation by additions of organic materials and/or inorganic amendments
spodic horizon	Refer to Spodosol below

The following terms are eight out of the 12 soil orders of US soil classification:

Alfisol	Mineral soil with subsurface clay accumulation, relatively low in organic matter, and relatively high in bases
Entisol	Undeveloped mineral soil
Gelisol	Mineral or organic soil of cold regions, with permafrost
Histosol	Organic soil
Inceptisol	Mineral soil with initial, but relatively weak development
Mollisol	Mineral soil with thick, dark, organic matter-rich surface horizon, high in bases
Spodosol	Mineral soil with subsurface accumulation of humus and amorphous aluminum, and commonly amorphous iron
Ultisol	Highly weathered mineral soil with subsurface clay accumulation, relatively low in organic matter, and relatively low in bases

See also: **Acid Rain and Soil Acidification**; **Archeology in Relation to Soils**; **Carbon Emissions and Sequestration**; **Civilization, Role of Soils**; **Classification of Soils**; **Classification Systems:** USA; **Degradation**; **Desertification**; **Erosion:** Water-Induced; **Heavy Metals**; **Morphology**; **Organic Matter:** Genesis and Formation; **Pedology:** Basic Principles; **Quality of Soil**; **Salination Processes**; **Structure**

Further Reading

Amundson R and Jenny H (1991) The place of humans in the state factor theory of ecosystems and their soils. *Journal of Soil Science* 151: 99–109.

Amundson R, Guo Y, and Gong P (2003) Soil diversity and land use in the United States. *Ecosystems* 6: 470–482.

Bidwell OW and Hole FD (1965) Man as a factor of soil formation. *Soil Science* 99: 65–72.

Bryant RB and Galbraith JM (2003) Incorporating anthropogenic processes in soil classification. In: Eswaran H *et al.* (eds) *Soil Classification: A Global Desk Reference*, pp. 57–66. Boca Raton, FL: CRC Press.

Courty MA, Goldberg P, and Macphail RI (1989) *Soils, Micromorphology, and Archaeology.* Cambridge, UK: Cambridge University Press.

Craul PJ (1999) *Urban Soils: Applications and Practices*. New York: John Wiley.

Dudal R, Nachtergaele FO, and Purnell MF (2002) *The Human Factor of Soil Formation.* Transactions of the 17th World Congress of Soil Science, Bangkok, Thailand.

Effland WR and Pouyat RV (1997) The genesis, classification, and mapping of soils in urban areas. *Urban Ecosystems* 1: 217–228.

Gong ZT (1983) Pedogenesis of paddy soil and its significance in soil classification. *Soil Science* 135: 5–10.

Goudie A (1994) *The Human Impact on the Natural Environment*, 4th edn. Cambridge, MA: MIT Press.

Jenny H (1984) The making and unmaking of a fertile soil. In: Jackson W *et al.* (eds) *Meeting the Expectations of the Land*, pp. 42–55. San Francisco, CA: North Point Press.

Johnson DL and Lewis LA (1995) *Land Degradation: Creation and Destruction.* Oxford, UK: Blackwell.

Lal R (ed.) (1999) *Soil Quality and Soil Erosion*. Boca Raton, FL: CRC Press.

Lal R, Sobecki TM, Livari T, and Kimble JM (2003) *Soil Degradation in the United States*. Boca Raton, FL: CRC Press.

Ruddiman WF (2001) *Earth's Climate: Past and Future*. New York: WH Freeman.

Sandor JA and Eash NS (1991) Significance of ancient agricultural soils for long-term agronomic studies and sustainable agriculture research. *Agronomy Journal* 83: 29–37.

Sencindiver JC and Ammons JT (2000) Minesoil genesis and classification. In: Barnhisel RI *et al.* (eds) *Reclamation of Drastically Disturbed Lands*. Madison, WI: Soil Science Society of America.

Westing AH and Pfieffer W (1972) The cratering of Indochina. *Scientific American* 226: 20–29.

Yaalon DH and Yaron B (1966) Framework for man-made soil changes – an outline of metapedogenesis. *Soil Science* 102: 272–277.

Parent Material

K R Olson, University of Illinois, Urbana, IL, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Parent material is the starting material from which soil develops and can include both consolidated rock and unconsolidated material. Mineral matter originating from rocks is referred to as 'soil parent material,' because it is the principal ingredient from which soils are formed. Unconsolidated material includes material deposited by gravity, wind, or water (Table 1) and consists of specific minerals of different sizes. However, the primary parent materials of organic soils are decomposing materials of various plant types.

Table 1 Bedrock type and parent materials by origin

Origin	*Parent material*	*Bedrock type*
Residual materials	Residium	
Sedimentary materials	Sedimentary	Sandstone
		Shale
		Limestone and dolomite
	Metamorphic	Marble
		Slate
		Quartzite
		Schist
		Gneiss
	Igneous	Basalt
		Granite
		Quartzite
Gravity deposits	Colluvium	
Ice deposits	Glacial till	
Water deposits		
Stream	Alluvium	
	Outwash	
Lake	Organic deposit	
	Beach deposit	
	Lacustrine	
Wind deposits	Loess	
	Eolian sand	
	Volcanic ash	

Defining Soil Parent Material and Role in Soil Genesis

Parent material is commonly defined as unconsolidated and chemically weathered mineral or organic matter from which the solum (set of horizons that are related through the same cycle of pedogenic or soil-forming processes) of soil is developed. 'Soil genesis' refers to the mode of origin of the soil, particularly the processes or soil-forming factors responsible for the development of the solum, or true soil, from unconsolidated parent material.

Soil-Forming Factors

The character, properties, and development of soils are controlled by external factors. Soil formation is difficult to view and often takes place over a long time. Parent material was initially identified as one of four significant soil-forming factors, as shown in the following equation:

$$S = f(\mathrm{cl}, \mathrm{o}, \mathrm{p})t^{\mathrm{o}} \qquad [1]$$

where S equals some attribute or measurable property of soil, $f()$ means 'as a function of,' cl is the climate of a region, o represents the organisms (both plants and animals), p is the geologic substratum, and t^{o} is the relative age (young, mature, or senile).

Later, a generalized equation, which related soil behavior to genetic factors and included relief, was developed:

$$S = f(\mathrm{cl}, \mathrm{o}, \mathrm{r}, \mathrm{p}, t, \ldots) \qquad [2]$$

where S equals some attribute or measurable property of soil, $f()$ means 'as a function of,' cl is the environmental climate, o is species of organisms, r is relief or topography (including hydrologic features such as the water table), p is parent material (the state of soil formation at time zero), and t is time (age of soil, absolute period of soil formation), with additional, unspecified factors.

The factors define the state of the soil system. The soil system has been defined as an arbitrary volume of the solum, a soil body, a pedon (three-dimensional body of soil with lateral dimensions large enough to permit the study of horizon shapes and relations in it), or an entire ecosystem of a component tessera (operation unit which is collected in the field, examined, and analyzed). The ecosystem approach avoids the impossible task of separating living material (nonsoil) from inanimate material (true soil) in a sample where they intermingle. The ecosystem is open to energy and matter influxes and outfluxes. Influx of energy includes solar radiation, heat transfer, and entropy

transfer; outflux of energy includes heat radiation and reflection of light. Influx of matter involves gases entering the ecosystem by mass flow (wind), water in liquid and solid forms, solids dispersed in air or dissolved or dispersed in water, and organisms that migrate into the ecosystem.

A lithosequence (Table 2) has been defined using functional factorial analyses as a set of soils with property differences due solely to differences in parent material with all other soil-forming factors constant as expressed in the following equation:

$$S = f(\mathrm{pm}), \mathrm{cl}, \mathrm{o}, \mathrm{r}, t, \ldots \quad [3]$$

where S equals some attribute or measurable property of soil, $f()$ means 'as a function of,' pm is parent material, cl is the environmental climate, o is species of organisms, r is relief or topography, t is time, and there are additional, unspecified factors.

Such an array or sequence of soils is difficult to recognize and establish because of the problem in establishing that all soils in the set have property differences due solely to parent material without some effects from environmental or local landscape position. Several sets of soils have been defined as approaching this condition, usually on young and relatively simple landscapes such as in glaciated regions. On these landscapes, attempts have been made to analyze the effects of differences in parent material composition.

Early approaches to soil survey and classification were based on the geology and composition of the soil-forming material. The geologic origin and composition of the initial material (Table 2) was identified by the terms 'granite soils,' 'glacial soils,' or 'till soils.' Soils that originate from a parent material inherit the mineral types found in them. Over time, these original minerals are weathered (dissolved) and new minerals form and accumulate in the soils.

Relief prior to and during soil formation is related to the nature of the starting soil material or parent material. In steeper topography, where the valleys below the mountain ranges are characterized by broad colluvial–alluvial fans, the starting or parent material near the mountain range contains more coarse and angular material than areas farther away from the mountain range. In broad river deltas, crests of natural levees near the stream channels contain coarser material than the areas beyond the levees, which are very nearly level and contain the finer-textured parent material.

Table 2 Lithosequence of soils with different composition and physical properties of till

Parent material	*Sequence of soil horizons*	*Soil order*
High-lime till	A, B, C	Mollisols Inceptisols
Medium-lime till	A, E, Bt, BC, C	Alfisols
Mixed sandstone and limestone till	O, E, Bs, BC, C	Spodosols
Sandstone and granite till	O, E, Bh, Bs, C	Spodosols

Weathering

'Weathering' describes the combined effects of all the physical and chemical processes that break down and transform preexisting parent materials at or near the surface of the Earth. These products are more stable under the physical and chemical conditions at the surface. Weathering of parent materials is the response to several forms of energy inputs as a function of time. The outcome of the weathering process is affected by physical and chemical factors. The products of weathering include liquids (including the solutions of salts present in rivers and the ocean) and solids (i.e., sediments and soils). The rock cycle, including erosion, transportation, and sedimentation, operates as a complex chemical sorting system which distributes weathering products of different composition to various sites.

Soil chemical composition is the product of parent material being altered over geologic time by the action of climate, topography, and biota. Physical disintegration and chemical weathering transform parent material into secondary minerals that often differ in type and composition with profile depth. Soils are transient bodies of the landscape. The initial soils will differ from the current and future soils. Parent material is often unrecognizable as a result of these chemical transformations and physical disintegration. The greatest change from the parent material occurs when soils in open systems gain new substances and lose their minerals through dissolution, leaching, or erosion. Minerals differ in their susceptibility to weathering. Physical disintegration and chemical weathering transform parent material into secondary minerals that often differ in kind and composition. Soils are the result of leaching, oxidation, and dissolution of surface materials by the percolation of groundwater and by humic acids derived from oxidized organic material. New minerals such as clays are formed by the chemical alteration of the parent material. The chemical weathering process develops a soil profile ranging from heavily weathered surface materials down to unaltered parent material.

Effects of Parent Material on Soil Properties

Soil formation starts after unconsolidated parent material is deposited on a stable landscape or after bedrock has been exposed at the Earth's surface and continues over time. Rate of soil formation depends on the climate, including temperature and rainfall. It also depends on vegetation and the activity of other organisms which live on or in the parent material. These organisms help convert parent materials to soil.

The properties of a modern soil are the result of the composition of the surficial layer which existed when the other factors began to impact and the alterations resulting from the effect of these factors over time. Properties of younger soils are significantly influenced by parent material. Weathering and pedogenic processes result in characteristics of the original parent material being eliminated. In old, weathered soils, extremely resistant starting material, such as quartz sand, may still exist. It can be difficult to separate the effects of the other soil-forming factors on the parent material of this soil, the nature (properties) of the initial material (Table 3) and its influence on soil, and the kind of 'preweathering' of the starting material before becoming parent material for the soil. Climatic and vegetational changes in the recent geologic past make it difficult to separate parent material influences on soil properties from the effects of other factors.

Modern soil properties are influenced by rock types (Table 1). The impacts of rock types on soil properties have been categorized into the following subdivisions: igneous, sedimentary, and combinations of mineralogically similar igneous and metamorphic rocks. Sedimentary parent materials (Table 1) include loessial deposits, unconsolidated glacial deposits, and coastal plain sediments, as well as consolidated rock such as sandstones, shales, limestone, and dolomites. Siliceous crystalline parent materials (Table 1) include both igneous and metamorphic rocks such as gneiss, granites, and schists, as well as more 'acidic' quartzose. Volcanic ash is a parent material composed of noncrystalline, glass fragments, bits of the easily weatherable ferromagnesian minerals, feldspars, and varying amounts of quartz. Dark-colored ferromagnesian rocks include andesites, diorites, basalt, and hornblende gneiss.

Mineral components of many soils are either developed *in situ* during the course of weathering and pedogenesis or primarily inherited from parent materials. Primary minerals such as igneous and metamorphic rocks are formed at high temperatures. Secondary minerals, including those in sedimentary rocks and soils, are those formed at lower temperatures. The elemental and mineralogical composition of the parent materials determine the soils' elemental and mineralogical composition. Approximately 96% by volume of the elemental composition of the Earth's crust is oxygen, silicon, aluminum, and iron.

Two of the most important properties of soil parent material are its mineral composition and its texture. Both of these characteristics are retained as properties of the soil formed from the parent material. They can be altered somewhat over time. Such alteration often reduces the particle size by weathering action. The fine particles are subject to removal from the surface layer by erosion, eluviation, and leaching.

Table 3 Parent material and bedrock type effects on soils and soil classification

Parent material or bedrock type	*Diagnostic epipedon or subsurface layer*	*Soil order*
Alluvium	None	Entisols
Beach deposits	None	Entisols
	Spodic horizon	Spodosols
Outwash	Cambic horizon	Inceptisols
	Spodic horizon	Spodosols
Lacustrine	Cambic horizon	Inceptisols
Organic deposits	Histic epipedon	Histosols
Volcanic deposits		Andisols
Loess	Mollic epipedon	Mollisols
Glacial till	Argillic horizon	Alfisols
		Ultisols
Sandstone	Cambic horizon	Inceptisols
Shale	Argillic horizon	Alfisols
Limestone	Mollic epipedon	Mollisols

Influence of Parent Materials in Soil Genesis

Lack of vertical uniformity in parent materials is a problem frequently faced by pedologists. Any of the parent materials (Table 1) could be located above or below another parent material and result in a lithologic discontinuity. Discontinuities can result from additions of loess, volcanic ash, sediments or colluvium over residual materials or older deposits. A lithosequence is defined as a set of soils with property differences due solely to differences in parent material with all other soil-forming factors constant. Parent materials affect soil property differences including diagnostic epipedons (surface layers) and subsurface layers which do affect the soils and soils classification (Table 3). In New York state, soil properties and soil horizons are attributed to differences in the composition and physical properties of the parent materials (Table 2). This set of four New York soils with different composition and physical properties of till (Table 2) is probably not a true lithosequence, since parent material differences

may not account for all soil property and soil horizon differences, but approaches one sufficiently to illustrate the point.

The basic model of soil implies that soils are dynamic and geographic. Dynamics is more appropriate than statistics in soil systems because of implied processes or driving forces. Morphological properties of soil develop as a result of processes acting on parent materials.

Time has the effect of modifying the influence of the parent material so that only in young or relatively immature soils will parent material be the dominant factor in the soil-forming process. The influence of the parent material is an inverse function of time. To illustrate this, soil scientists have suggested that soils derived from dissimilar parent materials such as granite and basalt will become indistinguishable given sufficient time to attain equilibrium.

Parent Material and Soil Classification

The central concept of an Entisol is a slightly developed soil with properties determined primarily by parent material (Table 3). The weak soil development is usually due to youthfulness, wetness or dryness extremes, or resistant parent materials. Mineral components of many soils are inherited from parent materials, while others are developed *in situ* during the weathering and pedogenesis. Soils derived from different parent materials often classify as distinct soil series based primarily on properties inherited from the parent material (Table 3).

Summary

The five soil-forming factors – parent material, climate, vegetation, topography, and time – determine the distribution and nature of soils over the Earth's surface. Mineral components of many soils are inherited from parent materials, while others develop *in situ* during weathering and pedogenesis. Morphological properties of soil develop as a result of processes acting on parent materials. Soil scientists have attempted to show the controlling effects of parent material on soil properties to the extent that parent material is proposed as an independent soil-forming factor and defined as the state of the soil system at time zero of soil formation. The physical body of soil and its associated mineralogical and chemical properties are the basis of other soil-forming factors. Previously weathered rock or even a previous soil could be considered parent material, which can be modified over time. Parent material exerts its strongest influence on the soil-forming process of immature soils.

List of Technical Nomenclature

cl	climate
$f()$	as a function of
o	organisms
p or pm	parent material
r	relief
S	a quantifiable attribute of soil
t or t^o	time

See also: **Factors of Soil Formation:** Biota; Climate; Human Impacts; Time

Further Reading

Birot P (1960) *The Cycle of Erosion in Different Climates*. (Translated from French, 1968, by Jackson CI and Clayton KM.) Berkeley, CA: University of California Press.

Buol SW, Hole FD, and Mc Cracken RJ (1980) *Soil Genesis and Classification*, 2nd edn. Ames, IA: Iowa State University Press.

Chesworth W (1973) The parent rock effect in the genesis of soil. *Geoderma* 10: 215–225.

Chesworth W (1976) Conceptual models in pedogenesis. A rejoinder. *Geoderma* 16: 257–260.

Cline MG (1953) Major kinds of profiles and their relationships in New York. *Soil Science Society of America Proceedings* 17: 23–27.

Dokuchaev VV (1883) Russian chernozem (Russkii Chernozen). In: *Collected Writings (Sochineniya)*, vol. 3. Moscow, USSR: Academy of Science.

Jackson C (1973) *Geology Today. Communications Research Machines*. Del Mar, CA: CRM Books.

Jackson ML (1964) Chemical composition of soils. In: Bear FE (ed.) *Chemistry of Soil*, 2nd edn, pp. 71–141. New York: Reinhold.

Jenny H (1941) *Factors of Soil Formation*. New York: McGraw-Hill.

Joffe JS (1949) *Pedology*, 2nd edn. New Brunswick, NJ: Pedology Publications, Rutgers University Press.

Russell RJ (1967) *River Plains and Sea Coasts*. Berkeley, CA: University of California Press.

Singer MJ and Munns DN (1996) *Soils: An Introduction*, 3rd edn. Upper Saddle River, NJ: Prentice-Hall.

Soil Survey Staff (1999) *Soil Taxonomy: A Basic System of Soil Classification for Making and Interpreting Soil Surveys*, 2nd edn. Agriculture Handbook No. 436, USDA–Natural Resource Conservation Service. Washington, DC: US Government Printing Office.

Troeh FR and Thompson LM (1993) *Soils and Soil Fertility*, 5th edn. New York: Oxford University Press.

Troeh FR, Hobbs JA, and Donahue RL (2004) *Soil and Water Conservation for Productivity and Environmental Protection*, 4th edn. Upper Saddle River, NJ: Prentice-Hall.

Time

E F Kelly and C M Yonker, Colorado State University, Fort Collins, CO, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

The many hundreds of thousands of soil types at the Earth's surface are each composed of unique biological, chemical, and physical properties. These properties are the direct result of a complex set of soil-forming processes that are conditioned by unique environmental factors. The environmental conditions under which soils and their properties develop are governed by unique combinations of the climate, biota, geology, and topography, which operate over time.

Climate affects soil development, as moisture and temperature control the rate of biological, chemical, and physical processes. The effect of biota is primarily related to controls on inputs and outputs of organic and inorganic materials; the effect of topography is the modification of temperature and precipitation relationships. The chemical and mineralogical composition of the parent material influences such key soil properties as nutrient status, alkalinity or acidity, and texture.

Whereas climate, biota, topography, and parent material have a direct effect on soil-forming processes, time has an indirect effect. Time affects soil-forming processes by controlling their duration and, therefore, the degree of soil development. Thus, we can differentiate between 'young' and 'old' soils in a given ecosystem on the basis of degrees of chemical and physical weathering and the resulting expression of soil properties.

Key Soil Properties that Vary with Time

The amount and distribution of soil organic carbon (SOC), carbonate, and various other elements, and the amount and distribution of clay and its mineralogy all change as a function of time in a generally consistent and predictable manner (Figure 1). The amounts and distribution of these soil constituents are often used to determine the relative age of a soil.

Once vegetation is established in an ecosystem, SOC is continually added through both above- and belowground inputs of living and dead plant tissues. The expectation is that older soils have a greater accumulation of SOC than younger soils. SOC accumulation eventually reaches a steady state wherein inputs equal outputs and there is no further net accumulation with increasing soil age. The length of time needed for SOC to reach steady state varies among environments, but is within a relatively short time frame of 10^2–10^3 years.

The accumulation of calcium carbonate ($CaCO_3$) in soils over time is generally limited to soils of semiarid and arid regions, where precipitation of carbonate occurs under conditions typical of dry soil environments, high pH, and high evapotranspiration. Under more humid conditions, the increased production of carbonic acid (H_2CO_3) results in the dissolution of $CaCO_3$ and a decrease in $CaCO_3$ content over time. Sources of calcium in arid and semiarid soils may be from *in situ* weathering of calcium-bearing minerals or from atmospheric deposition. Unlike SOC, $CaCO_3$ accumulation may not reach a steady state, but continues to increase through time.

The most abundant elements in soils, excluding oxygen and carbon, are silicon, aluminum, iron, magnesium, calcium, sodium, and potassium. Although these elements have varying solubilities, their relative abundance is a direct function of the degree of soil weathering, which increases through time. In general, the simple soluble cations, magnesium, calcium, sodium, and potassium, are readily released from the soil through weathering processes and are easily leached (Figure 2a). Barring an input of these elements to the soil system via the atmosphere or some other mechanism, their quantities steadily decrease as a function of time. Soil silicon, aluminum, and iron exist primarily as relatively insoluble hydroxides and are leached at a slower rate than the simple soluble cations (Figure 2b).

Phosphorus is an element which exists in soils in a variety of organic and inorganic forms, the distribution of which indicates the relative weathering of the soil profile and, therefore, soil age. As soil development begins and the soil matures, primary mineral sources are depleted, and organic and secondary, occluded mineral forms of phosphorus accumulate (Figure 2c).

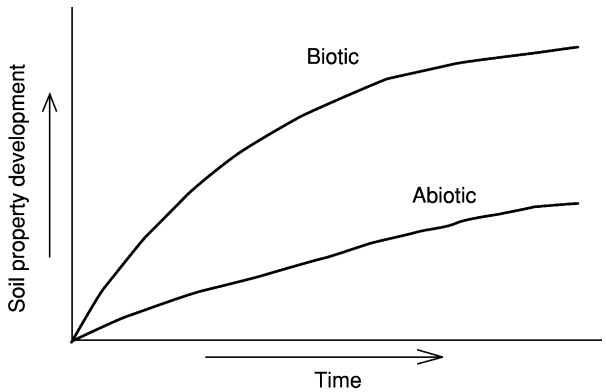


Figure 1 The effect of time on the development of biotic (e.g., soil organic carbon) and abiotic (e.g., calcium carbonate) soil properties.

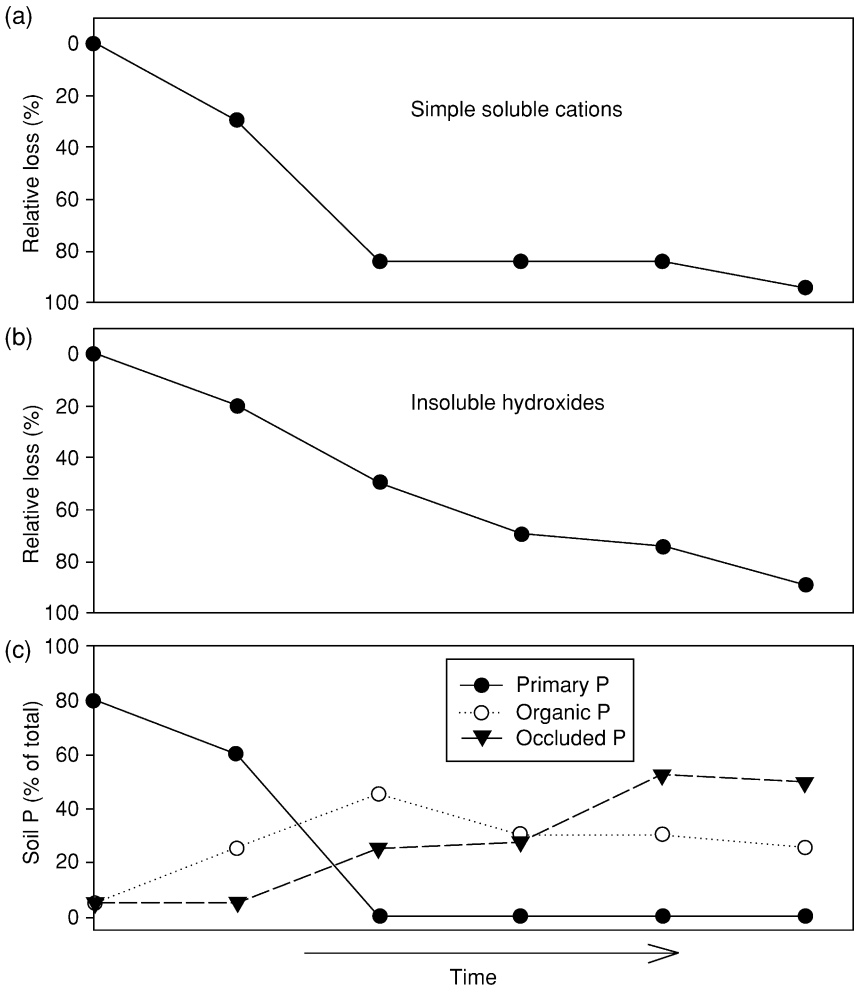


Figure 2 The effect of time on the loss of chemical soil constituents (a) and (b) and the transformation of phosphorus (c).

The accumulation and distribution of soil clay minerals are indicators of soil age. Clays accumulate through time as a function of eolian deposition of dust, translocation from superjacent soil horizons, and *in situ* formation of secondary clay minerals. Layers of accumulated clay can form in as little as a few hundred years if there are significant atmospheric inputs in the form of dust or mud rains. Where there is little atmospheric input, rates of accumulation are largely a function of translocation and weathering. Clay-enriched horizons can develop within 10^3–10^4 years, depending on whether the environment is humid or arid.

Clay mineralogy is partly a function of soil age because, over time, soils are increasingly more depleted in soluble framework cations and silicon, major constituents of clay minerals. Clay minerals, therefore, often form in a recognized sequence, with the earliest stages commonly consisting of illites and/or chlorites, intermediately weathering to smectites and/or vermiculites, typically resulting in kaolinite in the oldest soils that have undergone the lengthiest time of soil formation.

The preceding discussion theoretically describes the effect of time on soil development when climate is a constant variable. Given the same duration of soil formation, a soil forming in a warm, humid environment will have soil properties that are significantly more developed and appear to be older than a soil forming in a cool, dry environment. Increasing precipitation causes increased chemical and physical weathering, effectively mimicking the effect of a long period of soil formation in a relatively short time period. In humid areas, where precipitation exceeds potential evapotranspiration, biological and chemical weathering processes rapidly transform organic and mineral matter. Leaching transfers materials down, and eventually through, the soil profile. In drier areas, where precipitation is less than potential evapotranspiration, the same soil-forming processes operate; however, the availability of water limits the intensity and duration of these processes.

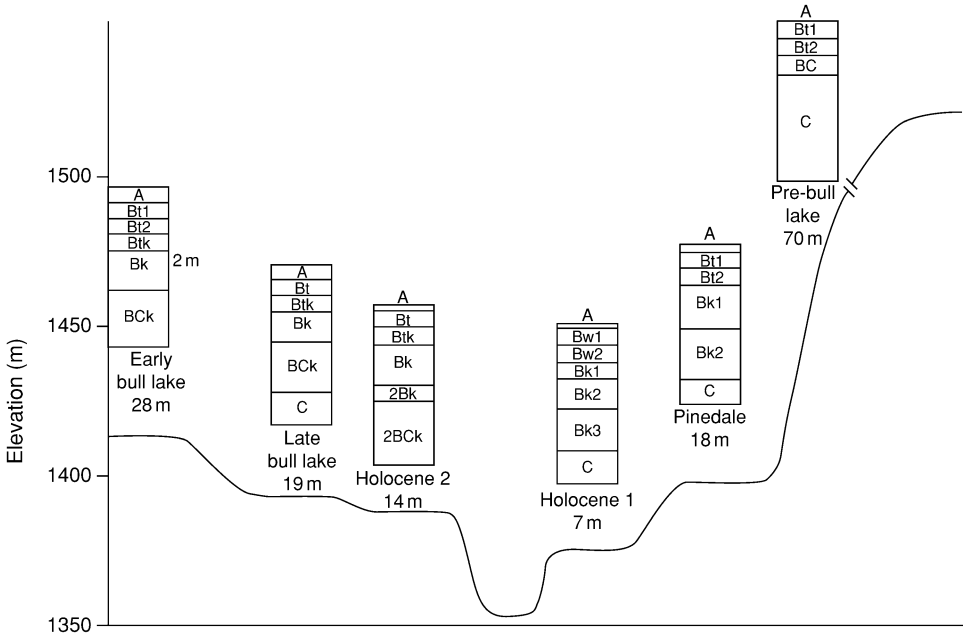


Figure 3 Soil chronosequence developing in a semiarid ecosystem. The youngest soil is located at the lowest elevation (7 m) above the floodplain; the oldest at the highest elevation (70 m). The soil ages are: Holocene (<10 000 years), Pinedale (13 000–20 000 years), Late Bull Lake (130 000–190 000 years), Early Bull Lake (190 000–300 000 years), Pre-Bull Lake (300 000–600 000 years).

Studies of Time and Soil Formation

The differences in properties between soils can be explained as a function of time if the soils being compared are forming in the same climate, maintain the same type of vegetation, have developed from the same or similar parent materials, and are located at the same topographic position. In other words, time is the only factor that varies between the soils being compared. Climate, biota, parent material, and relief can be held constant under a limited number of circumstances, including the comparison of soils developing on stream terraces sequentially incised by a single river system. In this case, the soils of the floodplain are developing in the most recently deposited material and are therefore the youngest; the soils on the highest stream terrace relative to the floodplain are therefore the oldest (Figure 3). Such a series of related terrace soils, which differ only with respect to their duration of formation, are called a chronosequence. Soils developing on chronosequences show broadly similar trends in morphology with time, including: (1) a thickening zone of organic matter accumulation; (2) increasing pedon thickness; (3) increasing horizon complexity; and (4) development of zones of secondary mineral accumulation. Such trends continue until such time that the soil becomes so weathered it begins to degrade.

Chronosequences are problematic in that, over millennial time scales, a constant climate cannot be assumed. Absolute soil age, or duration of soil development, can also be difficult to establish. In spite of these hindrances, chronosequences provide an excellent opportunity to study the influence of time on soil formation.

List of Technical Nomenclature

Calcium carbonate	Carbonate mineral that forms coatings on soil that react with an acid to form carbon dioxide gas
Evaporation	Water on the Earth's surface or in the soil absorbs heat from the sun to the point that it vaporizes or evaporates and becomes part of the atmosphere
Horizon	An individual layer within the soil which has its own unique characteristics (such as color, structure, texture, or other properties) that make it different from the other layers in the soil profile
In situ	Latin for 'the original position'
Soil horizons	An identifiable soil unit due to color, structure, or texture
Soil profile	The 'face' of a soil when it has been cut vertically that shows the individual horizons and soil properties with depth
Soil texture	The way soil 'feels' when it is squeezed between the fingers or in the hand. The texture depends on the amount of sand, silt, and clay in the sample (particle-size distribution), as well as other factors

Translocation	The movement or transfer of material from one part of the soil profile to another
Transpiration	Water in plants escapes or transpires into the atmosphere as the leaf stomates open to exchange carbon for oxygen

See also: **Factors of Soil Formation:** Biota; Climate; Human Impacts; Parent Material; **Pedology:** Basic Principles

Further Reading

Birkeland PW (1999) *Soils and Geomorphology*, 3rd edn. New York: Oxford University Press.

Buol SW, Hole FD, McCracken RJ, and Southard RJ (1997) *Soil Genesis and Classification*, 4th edn. Ames, IA: Iowa State University Press.

Dixon JB and Weed SB (eds) (1989) *Minerals in the Soil Environment*, 2nd edn. Book Series No. 1. Madison, WI: Soil Science Society of America.

Gerrard AJ (1981) *Soils and Landforms*. London, UK: George Allen & Unwin.

Harland WB, Armstrong RL, Craig LE, Smith AG, and Smith DG (1990) *A Geologic Time Scale*. Cambridge, UK: Press Syndicate of University of Cambridge.

Jackson JA (ed.) (1997) *Glossary of Geology*, 4th edn. Alexandria, VA: American Geological Institute.

Jenny H (1941) *Factors of Soil Formation*. New York: McGraw-Hill.

Jenny H (1980) *The Soil Resource*. New York: Springer-Verlag.

Merrits DJ, Chadwick OA, and Hendricks DM (1991) Rates and processes of soil evolution on uplifted marine terraces, northern California. *Geoderma* 51: 241–275.

Schumm SA (1991) *To Interpret the Earth: 10 Ways to be Wrong*. Cambridge, UK: Cambridge University Press.

Soil Science Society of America (1997) *Glossary of Soil Science Terms*. Madison, WI: Soil Science Society of America.

FAUNA

T Winsome, University of California–Davis, Davis, CA, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

The soil fauna is tremendously diverse and range in size from the smallest protists to fossorial rodents. The focus here is upon the soil macrofauna, their natural history and role in ecosystem function.

Characterization of Soil Macrofauna

The soil macrofauna generally includes animals that are greater than 1 cm in length or 2 mm diameter. Many of the invertebrates share physiological adaptations to life in the soil, including reduced tolerance for desiccation, high temperatures, and light, but increased tolerance to carbon dioxide. Along with soil-dwelling vertebrates, many have evolved characteristic morphological adaptations to a fossorial lifestyle, including reduced or no eyes, and forelimbs modified for digging (Figure 1). Others have evolved streamlined body plans with limbs reduced or absent altogether to assist in tunneling.

Soil macrofauna may be differentiated between the euedaphic forms, such as earthworms, which spend their entire life cycles within the litter and soil horizons, and those that are transients, such as some beetles, which carry out their larval stage in the soil but live aboveground as adults. The macrofauna may also be differentiated on the basis of their preferred habitat within the soil profile. Epigeic macrofauna dwell within the litter layer and seldom burrow into the mineral soil. Endogenic macrofauna reside within the mineral soil horizons and may be further differentiated on the basis of whether they live in surface or subsurface horizons.

Finally, soil macrofauna may be classified in functional terms as: (1) saprovores and microbivores, which feed on dead and decaying matter and the associated microorganisms, (2) herbivores and granivores, which feed on live plants and seeds, (3) predators, and (4) omnivores, which feed on a variety of living and dead substrates (Figure 2). The remainder of this section provides a brief description of common, important members of each functional class. The families, genera, and species names included in the descriptions are intended as examples and are by no means exhaustive.

Saprovores and Microbivores

Millipedes (Diplopoda), isopods (Isopoda), roaches (Blattaria), termites (Isoptera), earthworms (Annelida) and some species of ants (Hymenoptera: Formicidae), beetles (Coleoptera), and flies (Diptera) are included in the saprovores and microbivores group.

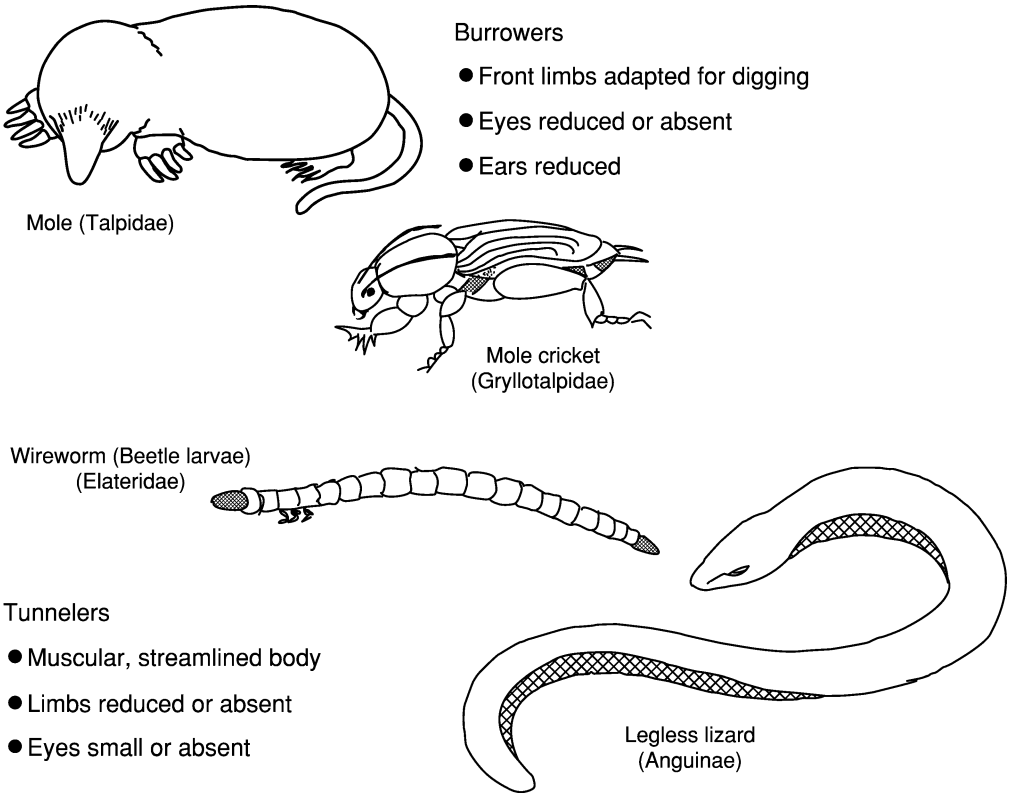


Figure 1 Morphological features for a life underground are shared by invertebrates and vertebrates alike. Animals that excavate soil, ''burrowers,'' have strong front limbs with large claws or claw-like appendages adapted for digging in soil or sand. Animals that push their way through the soil, ''tunnelers,'' have muscular, streamlined bodies with reduced or no appendages. Features common to both groups included reduced or no eyes, ears reduced in size, and sensory whiskers or appendages on or near the face.

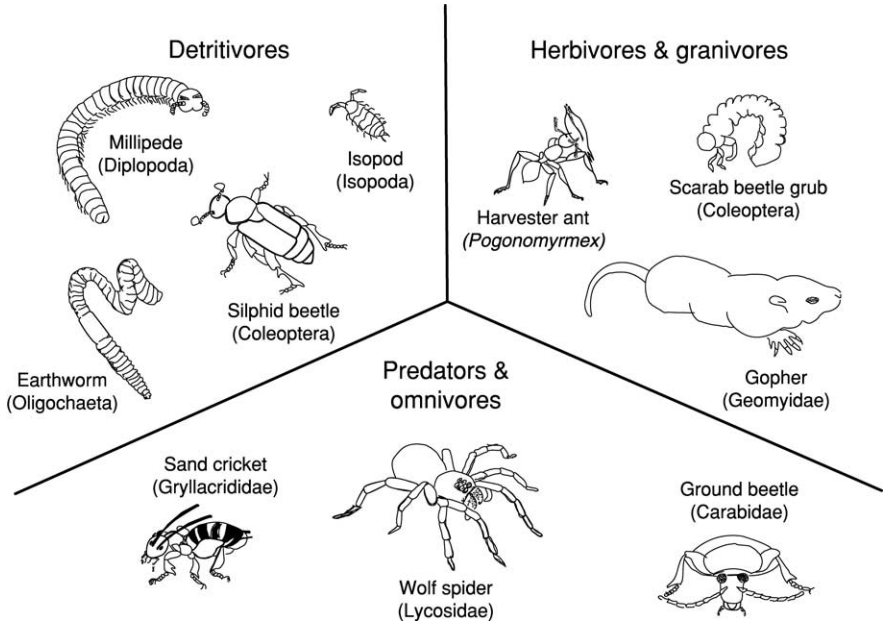


Figure 2 Commonly encountered soil macrofauna, grouped according to functional group. See text for details. (Not to scale.)

A few are generalist in their feeding habits and scavenge a wide variety of plant and animal residues, but most exhibit some degree of specialization with respect to diet. Many, perhaps most, of those that feed on plant residues derive their nutrition not from the residue itself but rather from the microorganisms that

have colonized the residue, and a large number feed exclusively on fungi and other microorganisms.

The millipedes, isopods, and roaches are widely distributed in many habitats, including temperate and tropical forests, grasslands, and deserts. Densities vary from fewer than 10 to hundreds per square meter, depending on species, season, and location. Most isopod species are relatively small, 1–2 cm in length, and are generally epigeic. Millipedes can be either epigeic or endogeic, depending on species, and are amongst the largest soil macroinvertebrates, ranging from 5 to 6 cm in length for temperate species to greater than 20 cm for some tropical species. Most roaches are epigeic and can also be quite large, greater than 5 cm for subtropical and tropical species. Both groups feed primarily on plant residues and the associated microflora, although they will also feed on fallen fruit and animal dung, and isopods will feed on seedlings. In the process of feeding, millipedes, roaches, and isopods physically fragment organic residues, in a process known as comminution, and haphazardly distribute the fragments within the litter and surface soil horizons.

Beetle species of the family Scarabaeidae (scarab beetles) are important scavengers of dung and carrion. Both larvae and adults of other scarabaeid species and beetle families occur within dead and dying trees and are important in the early stages of wood decomposition. Some species of fly are specialists on the mushrooms of various species of fungi (Drosophilidae). The larvae of many flies (Calliphoridae and Sarcophagidae) and beetles (Scarabaeidae, Silphidae, Dermestidae) are necrophages and occur exclusively on carrion. Many of these species are useful in criminal forensics, as the timing of their arrival on a corpse is highly dependent on time of death, degree of decomposition, and location.

Termites and ants are social insects, with caste systems consisting of workers, soldiers, winged male and female reproductives, and a queen. They occur in almost all habitats and are particularly important in arid and semiarid ecosystems where they are the most abundant soil-dwelling animals. The size of termite and ant colonies varies widely, from a few dozen individuals to hundreds of thousands, and colony densities range from a few to more than 5000 ha^{-1}. Once established, colonies persist for the life of the queen, which may be up to 2 decades for some species. Termites are dependent upon microbial symbionts for nutrition. Wood-feeding termite species in some families (e.g., Kalotermitidae) digest cellulose with the help of protozoan symbionts that reside in special compartments within the gut. Some of these species are important pests and cause tremendous economic damage to houses and other structures. Many soil-feeding termite species (Termitidae) lack protozoan symbionts but rely instead upon a mixture of bacteria and fungi, with which they derive nutrition from humified soil organic matter. Other groups (e.g., *Macrotermes*) lack intestinal symbionts altogether and feed on decayed leaves and other substrates that have been colonized by fungi (*Termitomyces*) that the termites culture in special chambers within their nests. Detritivorous ants include the carpenter ants, which dismantle snags, stumps, and downed logs, and various litter-collecting species. Most of these species use these substrates as culture material for microorganisms within nest chambers in a manner analogous to the termites. Many species are also scavengers of dead animals and fallen fruit. Like millipedes, roaches, and isopods, termites and ants fragment and redistribute organic materials, but differ in that most collect organic materials in an organized fashion, transport them over relatively long distances, and concentrate them in storage chambers within their nest structures.

Herbivores and Granivores

The herbivorous members of the soil macrofauna include species of beetles, flies, ants, and fossorial vertebrates. The larvae of many species of beetles and flies feed on plant roots and seedlings, and some feed on seeds. Many soil-dwelling herbivores are important agricultural pests, but in natural communities may play an important role in plant community structure and species diversity by regulating plant populations. The leaf-cutter ants common in tropical rainforests can defoliate entire trees, cutting the leaves into small pieces and transporting them back to the nest for use in their fungus gardens. Granivorous ant species serve a critical ecosystem function by transporting and burying seeds. They collect the seeds of many different species and store them in nest chambers, where they are used as food for larvae. Those that are dropped along the way or escape predation go on to germinate or remain protected within the soil seed bank. Fossorial mammalian herbivores include the pocket gopher (*Thomomys*) and related genera within the Rodentia. Some, such as the bathyergid mole-rats of South Africa, are social and form huge, subterranean colonies. They feed on roots, bulbs, and plant material that they pull down from the surface, and often store food in cache chambers within the soil.

Predators and Omnivores

Predatory soil macroinvertebrates include spiders (Aranae), scorpions (Scorpiones), centipedes (Chilopoda), and some species of beetles, ants, and flies. Many are ambush predators; the trap-door spiders (Ctenizidae),

for instance, construct a burrow with a 'trap door' of silk, soil, and litter residues behind which they hide in wait for prey to approach. Others, such as centipedes, ground beetles (Carabidae), and ants actively hunt for prey. These predators vary widely in size, from 1 to 2 cm to more than 15 cm in length for the largest centipedes and scorpions. Fire ants (*Solenopsis*) are capable of mobbing animals many times their own size, including small vertebrates. The victim is then dismembered and transported back to the nest. Predatory vertebrates include moles (Talpidae), shrews (Soricidae), and fossorial lizards and snakes (Reptilia). Many species of moles are specialized predators of earthworms, and often paralyze their prey with a bite to the head and then store them in cache chambers. Shrews are extremely small, epigeic mammals, most less than 10 cm in length. They have voracious appetites and feed on arthropods, earthworms, and small amphibians and reptiles, often attacking animals many times their own size. Some species have toxic compounds in their saliva that help to subdue prey. Fossorial reptiles feed on rodents and other soil fauna.

Common omnivores include mole crickets (Gryllotalpidae) and other members of the Orthoptera. These species feed on live plants, other soil fauna, and also scavenge dead plant residues. Omnivory is unusually common in soil food webs, and many of the species nominally assigned above as predators or detritivores in fact consume a wide variety of foods. Centipedes, for instance, long thought to be obligate predators, have recently been shown to include plant residues as part of their diet. The reason for the high incidence of omnivory in soil fauna remains poorly understood, but may include high mineral requirements or the need to supplement generally poor quality foods.

The Role of Soil Macrofauna in Ecosystem Function

Members of the soil macrofauna carry out three related functions that have ecosystem-level impacts: decomposition and nutrient cycling, modification of soil structure, and participation in both aboveground and belowground food webs.

Decomposition and Nutrient Cycling

Soil macrofauna have both direct and indirect effects on litter decomposition and nutrient cycling, but their net effects are best considered within the context of the entire soil fauna community (Figure 3). As a class, detritivores are characterized by relatively low assimilation efficiencies; estimates vary but range from 10 to 30%. (Some termite species are a notable

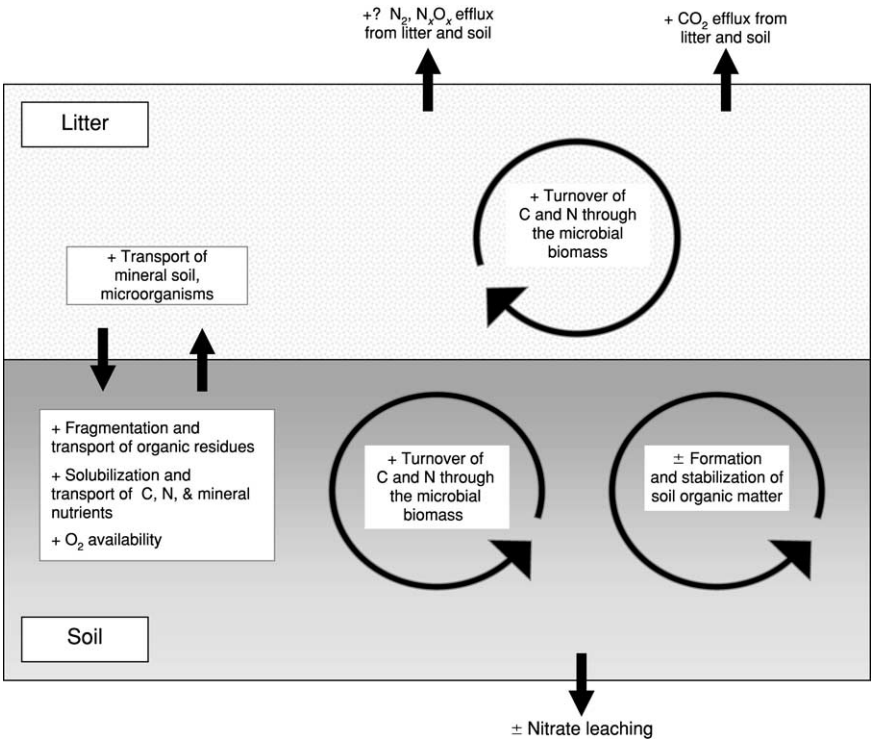


Figure 3 General effects of soil fauna on litter decomposition and nutrient cycles. A 'plus' symbol indicates enhancement; 'minus' indicates reduction. The 'plus-or-minus' symbol indicates that both effects can occur, depending on the organism(s) and ecosystem. The '+?' next to the efflux of gaseous forms of N (N_2 and N_xO_x) shows that evidence is inconclusive or contradictory.

exception and have assimilation efficiencies that exceed 50%). This requires a high consumption rate, and the soil fauna in total processes 20–40% of the annual litter input.

The process of litter decomposition is greatly facilitated by the physical comminution and redistribution of plant residues by soil macrofauna. These 'macroshredders' break down leaf and woody residues into smaller particles and, along with other soil animals, transport them from the surface deeper within the litter layer and soil mineral horizons. This physical fragmentation and transport of residues renders them more accessible as substrates for smaller fauna and soil microorganisms. Soluble nutrients are more readily leached into the soil from these small particles and enhance both microbial activity and plant growth. Decomposition rates are greatly reduced in the absence of soil fauna, especially in forests, where the volume of litter input is large and tends to be of relatively low quality, and in arid environments where climate limits microbial activity.

Soil macroinvertebrates also stimulate microbial activity through the production of large numbers of fecal pellets that are deposited throughout the soil profile and which serve as resource-rich microsites for microbial activity. Fecal pellets are generally associated with higher concentrations of soluble C, mineral forms of N and P, and available forms of mineral nutrients such as Ca, Mg, and K. The fecal pellets produced by earthworms and other soil-dwelling macroinvertebrates consist of a mixture of mineral soil particles and organic material, and contribute to the formation of stable soil aggregates. They are often characterized by numerous, small voids that provide habitable pore space for microorganisms (**Figure 4**).

The net influence of soil macrofauna on nutrient cycles at the ecosystem level is complex and remains poorly understood. Soil animal activity is generally associated with higher rates of C, N, and P mineralization. Termite mounds in arid regions of Australia, for instance, may account for up to 20% of the ecosystem flux of CO_2; in more mesic ecosystems, soil macrofauna contribute proportionally much less

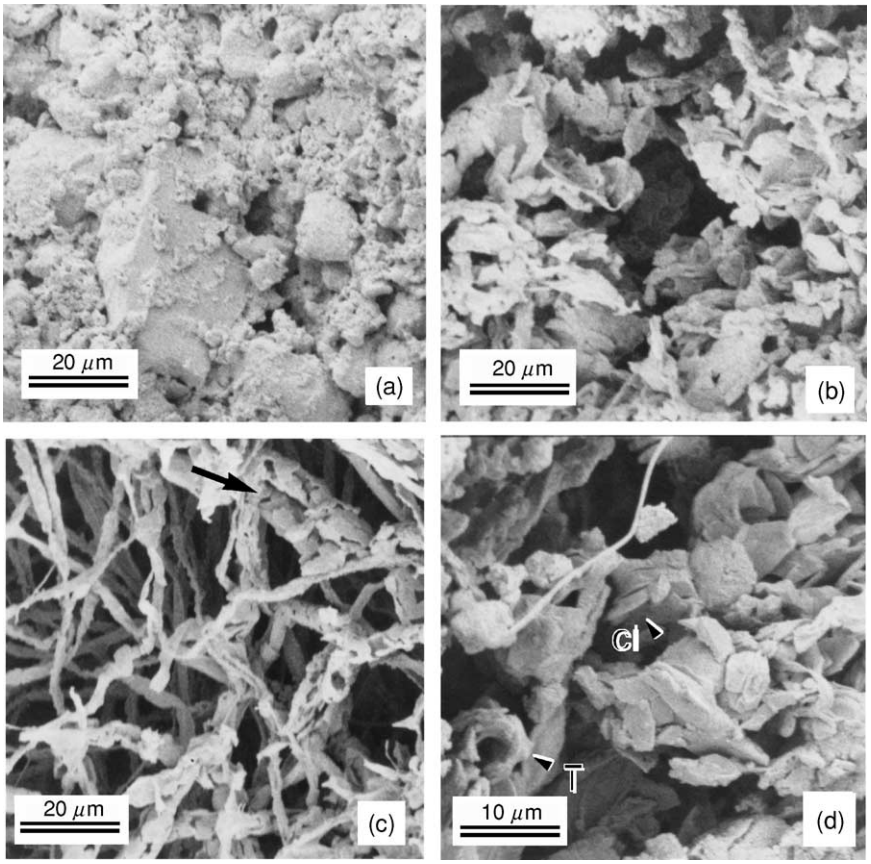


Figure 4 A comparison of micromorphological features of bulk soil and earthworm casts: (a) interior of bulk soil aggregate. Note dense, featureless structure; (b) interior of cast, with foliate structures surrounding voids; (c) interior of cast, showing fungal filaments bridging voids; (d) interior of cast, showing kaolinite domains (Cl) and clay particles forming a tube around fungal filaments (T). (Adapted from Winsome T and McColl JG (1998) Changes in chemistry and aggregation of a California forest soil worked by the earthworm *Argilophilus papillifer* Eisen (Megascdecidae). *Soil Biology & Biochemistry* 30: 1677–1687, with permission.)

to CO_2 flux, within the range of 2–10%. Experiments comparing N mineralization in soils with and without fauna (e.g., millipedes, earthworms) have shown that mineralization is enhanced by 10–30% in the presence of fauna. Through litter fragmentation and the production of fecal pellets, the fauna increase microbial production and activity, but through microbivory also regulate the degree to which nutrients are immobilized within the microbial biomass. While the fragmentation of detrital material enhances mineralization, the concomitant process of burial within the soil ensures that a proportion of it that would otherwise decompose on the surface is physically protected within soil aggregates and eventually contributes to the formation of stable soil organic matter. Feedback from all these processes may serve to maintain net ecosystem productivity.

Soil Modification

The physical disturbance and modification of soil horizons by animal activity is defined as 'biopedturbation.' This disturbance is extremely important in maintaining both spatial and temporal heterogeneity in ecosystems and has profound impacts on the structure and species diversity of both plant and animal communities. There are four general forms of biopedturbation, distinguished on the basis of spatial and temporal scale: (1) soil eject mounds, (2) burrow systems, (3) nest mounds, and (4) Mima and Mima-like terrain (Figure 5 and Table 1).

Soil eject mounds consist of the loose piles of soil thrown up on the surface as the animal burrows. These mounds are created by a wide variety of soil fauna, from ants and mole crickets to fossorial rodents. Accordingly, the size of these mounds varies from a few centimeters in diameter to greater than a meter. At any one point in time, the number of mounds formed by all species combined may total several hundred to thousands per hectare and occupy up to 10% of the surface. Over the course of time, the entire soil surface may thus be worked in this way. Burrow systems are created by various species of ants, termites, earthworms, and mammals, and consist of a network of burrows that run more or less parallel to

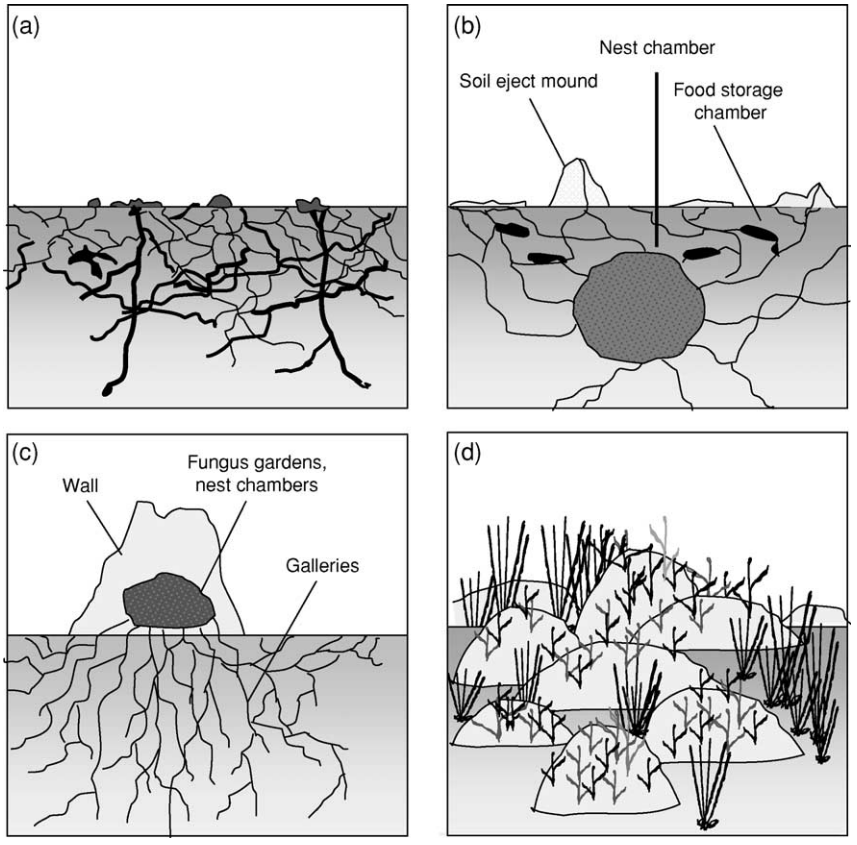


Figure 5 Forms of soil modification mediated by soil macrofauna: (a) burrow systems are dense, interconnected tunnels, created by a variety of soil animals, ranging from ants and termites to gophers; (b) subterranean nest structures are built by ants, termites, and rodents. The figure shows a central nest chamber, associated galleries and storage chambers, and soil eject mounds on the surface; (c) nest mounds are constructed by both ants and termites. Shown here is a mound constructed by fungus-cultivating termites; (d) Mima-type mound terrain is a landscape form, characterized by dense aggregations of low mound structures. Plant communities on the surface of the mounds are often very different from those that inhabit the intermound soils.

Table 1 Density, mass, or volume, and amount of soil displaced or transported to the surface by biopedturbation

Disturbance type, organism(s) and location	*Density ($n\,ha^{-1}$)*	*Mass ($Mg\,ha^{-1}$) or volume ($m^3\,ha^{-1}$)*	*Soil displaced or transported to surface ($Mg\,ha^{-1}\,y^{-1}$)*	*Source*[a]
Fossorial mound				
Tuco-tuco (*Ctenomys azarae*); Argentina	1050	14.7 $Mg\,ha^{-1}$	520	Roig *et al.*, 1988
Burrow system				
Earthworms, various species; USA, Europe, Africa, Asia	50 000–8 000 000[b]	—	2–91 (temperate) 3–507 (tropical)	Edwards and Bohlen, 1996
Nests and nest-mound systems				
Ants, various species; Australia	5100–15 400		0.1–0.31	Lobry de Bruyn and Conacher, 1990, 1994; Wiken *et al.*, 1976; Lee and Wood, 1971
Ants, (*Formica fusca*); British Columbia	1150		54 $m^3\,ha^{-1}$	
Termite, mixed species; tropics	1–200	0.1–300 $Mg\,ha^{-1}$	1–4	
Mima-mound terrain				
Pocket gopher (*Thomomys bottae*); USA	20–50	15.0–74.8 $m^3\,ha^{-1}$	101	Cox, 1984, 1990

[a]Sources:
Cox GW (1984) The distribution and origin of Mima mound grasslands in San Diego County, California. *Ecology* 65: 1397–1405.
Cox GW (1990) Soil mining by pocket gophers along topographic gradients in a Mima moundfield. *Ecology* 71: 837–843.
Edwards CA and Bohlen PJ (1996) *Biology and Ecology of Earthworms*. London, UK: Chapman and Hall.
Lee K and Wood TG (1971) *Termites and Soils*. New York: Academic Press.
Lobry de Bruyn LA and Conacher AJ (1990) The role of termites and ants in soil modification: a review. *Australian Journal of Soil Research* 28: 55–93.
Lobry de Bruyn LA and Conacher AJ (1994) The bioturbation activity of ants in agricultural and naturally vegetated habitats in semi-arid environments. *Australian Journal of Soil Research* 32: 555–570.
Roig VG, Loyarte Gonzalez MM, and Rosa ML (1988) Ecological analysis of mound formation of the Mima type in Rio Quinto, Province of Córdoba, Argentina. *Studies on Neotropical Fauna and Environment* 23: 103–116.
Wiken EB, Boersma LM, Lavkulich LM, and Farstead L (1976) Biosynthetic alteration in a British Columbia soil by ants (*Formica fusca* Linn). *Soil Science Society of America Proceedings* 40: 422–426.
[b]Measured on horizontal soil surfaces at various depths.

the soil surface. They can extend over many hectares and persist for decades, maintained and expanded by successive generations of animals.

Many soil macrofauna build large nests that are characteristic features of the landscape. Termites construct some of the most spectacular and complex nests. *Macrotermes* in central African savanna construct nests with 'chimneys' that can reach 9 m in height. In Australia, *Amitermes* constructs mounds several meters in height that have a flattened, fluted shape, while in tropical Africa *Cubitermes* constructs mushroom-shaped mounds. Although these shapes appear bizarre, they are in fact carefully engineered to regulate temperature and humidity within the nest or, in the case of the mushroom-shaped mounds, to prevent rainwater from collecting on the surface of the nest. In the process of building their nests, termites transport soil fines (clay-sized soil mineral particles) from the soil subsurface to the surface, altering the particle-size distribution of the soil in the nest area. Depending on species, they mix the soil with saliva, their feces, and/or organic matter and use this material to construct the galleries and chambers that form the interior of the nest. Termites that construct subterranean nests also transport soil fines and mix it with various materials to construct their nests. Their nests can be very large; for example, *Odontotermes* in Africa and India construct nests several meters in diameter, with foraging galleries that extend out 50 m or more from the central nest. For both mound-building and subterranean-nesting species, the area occupied by foraging galleries, storage chambers, and brood chambers can be quite extensive; for example, one nest system constructed by *Macrotermes michaelseni* in Kenya covered an area of 8000 m^2 and was associated with 6 km of galleries and 72 000 storage chambers. Some desert species construct vertical galleries that descend as much as 70 m to water sources.

Many species of ants also create large, elaborate nests from soil, with extensive subterranean galleries and chambers. Some species of carpenter ant build their nests within and beneath snags and stumps, cutting the wood away in small cubes. Over time,

their nests come to resemble piles of sawdust. There are also a large number of vertebrates that forage aboveground for plants and seeds, but nest in the soil and have important impacts on soil properties. The black-tail prairie dog (*Cynomys ludovicianus*) of North America forages at the surface, but spends much of its time belowground in extensive burrow systems known as 'towns.' Other mammals with similar life habits include the kangaroo rat (*Dipodomys*) of the American southwest, and the gerbil (*Tatera* and related genera) of Africa and the Middle East.

Mima and Mima-like terrain is characterized by low mounds, ranging from less than 1 m to more than 3 m in height and approximately 1–20 m in diameter, that occur in a regular formation across the landscape. The name 'Mima' derives from their type locality, Mima Prairie, in the state of Washington. They occur worldwide, however, and their various origins have long been the subject of some controversy. Competing hypotheses focus either on strictly geomorphologic processes or on the long-term effect of animal activity. Whatever the cause of their origin, recent research has confirmed that the maintenance of these landscapes is often due to animal activity. For example, the heuweltjies landscape (pronounced hear-vill-keys) near Cape Town in South Africa consists of low, nearly circular mounds some meters in diameter that are up to 4000 years old and have been maintained by successive generations of termites. Similar landscapes in wetlands in Argentina are associated with fire ants (*Solenopsis*). Rodent activity is also frequently associated with mima mound terrain; some of the best-studied examples include pocket gophers (*Thomomys*) in southern California and tuco-tucos (*Ctenomys*) in Argentina.

An important feature of systems characterized by biopedturbation is a marked difference between the animal-worked soil and surrounding soil in soil physical, chemical, and biological properties (Table 2). Over time, the net effects of biopedturbation on soil physical properties are generally to reduce bulk density and improve water infiltration. These effects vary, however, according to disturbance type, species, location, and whether or not the structure is actively maintained or has been abandoned. Eject mounds serve to mix subsurface and surface soils, but are loosely structured and thus easily eroded by wind or water unless quickly colonized by plants.

The burrow systems of the smaller soil macrofauna, such as ants, termites, and earthworms, create numerous small channels that increase water infiltration if they are open to the soil surface. The burrow systems of termites and ants are often kept sealed from the surface while the colony is active and do little to improve infiltration, but greatly enhance water infiltration once they are abandoned and open to the surface. The larger burrow systems created by mammals may improve preferential flow, but may also increase surface erosion and runoff if the animals are active in removing the overlying vegetation and compacting the soil. Nest mounds and mima mounds tend to consist of fine-textured soil material, often underlain by a layer of gravels and stones. The fine material used in termite nest construction may become compacted over the mound and form a seal over the soil surface as it erodes away from the nest, effectively reducing water infiltration and negatively impacting plant growth. In areas with a lot of termite activity, much of the soil surface horizon may consist of material worked by termites that has eroded away from nests constructed over many thousands of years.

The burrow systems created by earthworms and millipedes are generally enriched in C, N, and other nutrients, due to the incorporation of organic materials within the mineral soil and the deposition of fecal pellets. The nest mound soils of ants, termites, and rodents are also often enriched in C, N, and mineral nutrients, due to the concentration of urine, feces, and stored food within the mound. Termite mounds, in particular, are enriched in Ca relative to surrounding soil, so much so that in parts of central Africa traditional farmers collect the soil from termite mounds and apply it to their fields for lime and fertilizer.

The soil biological properties of burrow systems and nest mounds depend on the species, ecosystem, and season. In the nest mounds of kangaroo rats, the concentration of organic matter and nutrients within the mound stimulates soil microbial activity relative to the surrounding soil; conversely, in some termite nests the high polyphenol content of the nest material suppresses microbial activity. Ant and termite nests often support large inquiline ('animals that share a common space') communities. The numbers of protozoa, nematodes, microarthropods, and beetles may be much greater in and around the vicinity of ant nests than the surrounding soil, due to the accumulation of organic debris and availability of moisture. Many of these species are dependent upon their host nests and occur nowhere else. Mima-like mounds provide space for burrowing mammals such as ardvaarks in the South African heuweltjies and armadillos in Argentina.

These differences in soil properties in turn affect the structure and diversity of associated plant communities. The porous, nutrient-enriched burrow systems created by earthworm and millipede activity invariably improve plant growth. Large eject mounds in

Table 2 The impacts of various types of biopedturbation on soil properties, plant communities, and other soil biota

Disturbance type organism and location	*Effect on soil properties*	*Effects on plant community*	*Effects on other biota*	*Source*[a]
Fossorial soil eject mound				
Pocket gopher (*Thomomys bottae*); California serpentine grasslands; USA	Reduced bulk density, increased Ni, Mn, and other minerals associated with serpentine	Fewer exotic species not adapted to serpentine soils		Koide *et al.*, 1987
Burrow system				
Earthworm (Lumbricidae); temperate grasslands; USA, Europe	Reduced bulk density, increased water infiltration, increased C, N, mineral N	Increased pasture biomass		Reviewed by Lee, 1985
Bathyergid mole rats		Reduced biomass; a change in community structure		Reichman and Jarvis, 1989
Nest mound				
Kangaroo rat (*Dipodomys*); Chihuahuan desert; USA	Reduced bulk density, decreased moisture, increased C, N	Fewer perennials, increased biomass annuals, increased species diversity, number of annuals	Increased microbial biomass and respiration increased microarthropods, nematodes, insects, lizards, other rodents	Ayarbe and Kieft, 2000; reviewed by Whitford and Kay, 1999
Harvester ant (*Pogonomyrmex barbatus*); Chihuahuan desert; USA	Litter accumulation and burial, increased mineral N, P, K	Seed burial and dispersal	Increased protozoa, microarthropods	Wagner *et al.*, 1997
Mima-type mound system				
Fire ant, (*Solenopsis richteri*); Argentina	Decreased salinity	Salt-tolerant community on mounds	Habitat for armadillo	Cox, 1992
Termites, (*Odontotermes*); heuweltjies, South Africa	Increased N, K, Ca, Mg, increased disturbance	Disturbance-adapted plant community on mounds, increased productivity	Habitat for ardvaark and other mammals	Esler and Cowling, 1985

[a]Sources:

Ayarbe JP and Kieft TL (2000) Mammal mounds stimulate microbial activity in a semiarid shrubland. *Ecology* 81: 1150–1154.

Cox GW (1992) Fire ants (Hymenoptera: Formicidae) as major agents of landscape development. *Environmental Entomology* 21: 281–286.

Esler KJ and Cowling RM (1995) The comparison of selected life-history characteristics of Mesembryanthema species occurring on and off Mima-like mounds (heuweltjies) in semiarid southern Africa. *Vegetatio* 116: 41–50.

Koide RT, Huenneke LF, and Mooney HA (1987) Gopher mound soil reduces growth and affects ion uptake of two annual grassland species. *Oecologia* 72: 284–290.

Lee KE (1985) *Earthworms: Their Ecology and Relationships with Soils and Land Use*. Sydney, Australia: Academic Press.

Reichman OJ and Jarvis JUM (1989) The influence of three sympatric species of fossorial mole-rats (Bathyergidae) on vegetation. *Journal of Mammalogy* 70: 763–771.

Wagner D, Brown MJF, and Gordon DM (1997) Harvester ant nests, soil biota and soil chemistry. *Oecologia* 112: 232–236.

Whitford WG and Kay FR (1999) Biopedturbation by mammals in deserts: a review. *Journal of Arid Environments* 41: 203–230.

grasslands may bury and kill the vegetation on which they fall, and thus create vegetation gaps that favor colonization by fast-growing, pioneer species. In the South African heuweltjies, the mounds support a plant community adapted to chronic disturbance, while the intermound space supports an entirely different community. In the fire ant-maintained mound terrain in Argentina, a high, brackish water table restricts the vegetation in intermound areas to salt-tolerant species, while the mounds are markedly lower in salinity and support a diverse community of salt intolerant species. Thus, the spatial heterogeneity created by the mounds results in an overall increase in plant species diversity in these ecosystems. However, with the recent spread of invasive plant species throughout many parts of the world, these

mound systems often inadvertently promote the invasion of fast-growing, exotic plant species at the expense of native species. For example, gopher eject mounds in California grasslands are quickly colonized by fast-growing, exotic annual grasses at the expense of native perennials. Mima-like mounds created by rodents in Argentina are the preferred habitat for exotic trees and shrubs.

Participation in Belowground and Aboveground Food Webs

The soil macrofauna have been described as 'facilitators' or 'regulators' of interactions between other groups of organisms, and this may well be one of their most important roles in ecosystem function. Species that burrow extensively within the rhizosphere or transport organic material from the surface into soil horizons transport microorganisms to new habitats. For instance, a number of soil animals have been shown to disseminate the spores of mycorrhizal fungi throughout the rhizosphere, increasing the inoculation of plant roots. Soil animals have also been shown to serve as important vectors and reservoirs for microorganisms that cause both plant and animal diseases, including those that affect humans. Recent research has shown that the transfer of genetic material between different strains of bacteria is facilitated by the soil-mixing activity of earthworms, which may have important implications for the spread of disease-causing organisms.

Soil macrofauna serve as important food sources for a wide variety of animals, many of which are specialist feeders on certain species. Termites and ants are high in protein and fat and are consumed by humans in some cultures, as well as by a large number of other mammals, birds, and reptiles. Earthworms are also rich in protein and are consumed by many mammals and birds. In polluted environments, soil macrofauna form a critical link in the transfer of toxic compounds from the soil to the aboveground biota in a process known as bioaccumulation. In the course of feeding on soil and litter residues, detritivores accumulate in their bodies any heavy metal or chemical pollutants that may be present in low levels in the environment. The spiders, centipedes, and small mammals that feed on these animals further concentrate these compounds within their tissues. The birds and mammals that feed upon the soil animals ingest these compounds in concentrations that greatly exceed those in the soil and are frequently poisoned, or simply concentrate the compounds yet further and pass them along to their predators and scavengers.

Soil macrofauna play important and sometimes critical roles in ecosystem nutrient cycles, the shape and structure of landscapes, and the flow of energy and matter between belowground and aboveground components of ecosystems. They are also fascinating in their own right. Few groups of organisms are more diverse and complex with respect to their evolution, behavior, and ecology. From termites to trapdoor spiders, the capacity of the soil macrofauna to modify their physical and biological environment stands unparalleled in the animal kingdom.

See also: **Forensic Applications**; **Vadose Zone:** Microbial Ecology

Further Reading

Abe T, Bignell DE, and Higashi M (eds) (2000) *Termites: Evolution, Sociality, Symbioses, Ecology*. Dordrecht, The Netherlands: Kluwer Academic Publishers.

Benckiser G (ed.) (1997) *Fauna in Soil Ecosystems: Recycling Processes, Nutrient Fluxes, and Agricultural Production*. New York: Marcel Dekker.

Coleman DC and Crossley DA (eds) (1996) *Fundamentals of Soil Ecology*. London, UK: Academic Press.

Cox GW, Contreras LC, and Milewski AV (1995) Role of fossorial mammals in community structure and energetics of Pacific Mediterranean ecosystems. In: Kalin Arroyo MT, Zedler PH, and Fox MD (eds) *Ecology and Biogeography of Mediterranean Ecosystems in Chile, California, and Australia*, pp. 383–398. New York: Springer-Verlag.

Hopkins SP and Read HJ (1992) *The Biology of Millipedes*. Oxford, UK: Oxford University Press.

Jones CG, Lawton JH, and Shachak M (1997) Positive and negative effects of organisms as physical ecosystem engineers. *Ecology* 78: 1946–1957.

Lal R (1987) *Tropical Ecology and Physical Edaphology*. Chichester, UK: John Wiley & Sons, Ltd.

Lee KE (1985) *Earthworms: Their Ecology and Relationships with Soils and Land Use*. Sydney, Australia: Academic Press.

Lee KE and Wood TG (1971) *Termites and Soils*. New York: Academic Press.

Lobry de Bruyn LA (1999) Ants as bioindicators of soil function in rural environments. *Agriculture, Ecosystems and Environment* 74: 425–441.

Lobry de Bruyn LA and Conacher LA (1990) The role of termites and ants in soil modification: a review. *Australian Journal of Soil Research* 28: 55–93.

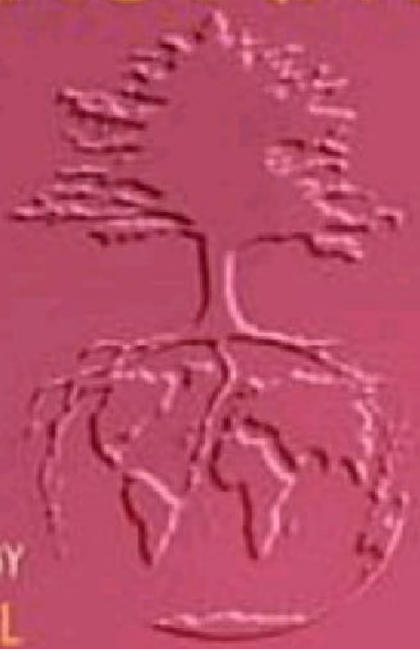
Meadows PS and Meadows A (eds) (1991) *The Environmental Impact of Burrowing Animals and Animal Burrows*. Oxford, UK: Clarendon Press.

Wallwork JA (1976) *The Distribution and Diversity of Soil Fauna*. London, UK: Academic Press.

Wardle DA (2002) *Communities and Ecosystems: Linking the Aboveground and Belowground Components*. Princeton, NJ: Princeton University Press.

Whitford WG and Kay FR (1999) Biopedturbation by mammals in deserts: a review. *Journal of Arid Environments* 41: 203–230.

Encyclopedia of Soils in the Environment



Edited by
Daniel Hillel
Cynthia Rosenzweig
David Powlson
Kate Scow
Michail Singer
Donald Sparks

Volume Two

ENCYCLOPEDIA OF SOILS IN THE ENVIRONMENT

FOUR-VOLUME SET

by Daniel Hillel (Editor-in-Chief)

Hardcover: 2200 pages
Publisher: Academic Press; 1 edition (November 8, 2004)
Language: English
ISBN-10: 0123485304
ISBN-13: 978-0123485304

Book Description

More than ever before, a compelling need exists for an encyclopedic resource about soil the rich mix of mineral particles, organic matter, gases, and soluble compounds that foster both plant and animal growth. Civilization depends more on the soil as human populations continue to grow and increasing demands are placed upon available resources.
The Encyclopedia of Soils in the Environmentis a comprehensive and integrated consideration of a topic of vital importance to human societies in the past, present, and future.

This important work encompasses the present knowledge of the world's variegated soils, their origins, properties, classification, and roles in the biosphere. A team of outstanding, international contributors has written over 250 entries that cover a broad range of issues facing today's soil scientists, ecologists, and environmental scientists.
This four-volume set features thorough articles that survey specific aspects of soil biology, ecology, chemistry and physics. Rounding out the encyclopedia's excellent coverage, contributions cover cross-disciplinary subjects, such as the history of soil utilization for agricultural and engineering purposes and soils in relation to the remediation of pollution and the mitigation of global climate change.

This comprehensive, yet accessible source is a valuable addition to the library of scientists, researchers, students, and policy makers involved in soil science, ecology, and environmental science.

Also available online via ScienceDirect featuring extensive browsing, searching, and internal cross-referencing between articles in the work, plus dynamic linking to journal articles and abstract databases, making navigation flexible and easy. For more information, pricing options and availability visit www.info.sciencedirect.com.

* A distinguished international group of editors and contributors
* Well-organized encyclopedic format providing concise, readable entries, easy searches, and thorough cross-references
* Abundant visual resources — photographs, figures, tables, and graphs — in every entry
* Complete up-to-date coverage of many important topics — essential information for scientists, students and professionals alike

EDITOR-IN-CHIEF

Daniel Hillel
Columbia University
New York, NY
USA

EDITORS

Jerry L Hatfield
National Soil Tilth Laboratory
Ames, IA
USA

Kate M Scow
University of California
Davis, CA
USA

David S Powlson
Rothamsted Research
Harpenden
UK

Michael J Singer
University of California
Davis, CA
USA

Cynthia Rosenzweig
NASA Goddard Institute for Space Studies
New York, NY
USA

Donald L Sparks
University of Delaware
Newark, DE
USA

EDITORIAL ADVISORY BOARD

R Bardgett
Lancaster University
Lancaster
UK

J L Boettinger
Utah State University
Logan, UT
USA

G Gee
Pacific Northwest National Laboratory
Richland, WA
USA

R Keren
The Volcani Center
Bet Dagan
Israel

J Kimble
USDA Natural Resources Conservation Service
Lincoln, NE
USA

M B Kirkham
Kansas State University
Manhattan, KS
USA

M Kutilek
Prague
Czech Republic

D Martens
Southwest Watershed Research Center
Tucson, AZ
USA

K Mengel
Justus Leibig University
Giessen
Germany

K Reichardt
Center for Nuclear Energy in Agriculture
Piracicaba
Brazil

K Ritz
Cranfield University
Silsoe
UK

R Schulin
Swiss Federal Institute of Technology Zurich
Schlieren
Switzerland

N Senesi
Università di Bari
Bari
Italy

J T Sims
University of Delaware
Newark, DE
USA

K Smith
University of Edinburgh
Edinburgh
UK

R L Tate
Rutgers University
New Brunswick, NJ
USA

N van Breemen
Wageningen Agricultural University
Wageningen
The Netherlands

W H van Riemsdijk
Department of Soil Quality
Wageningen
The Netherlands

FOREWORD

The *Encyclopedia of Soils in the Environment* is a vitally important scientific publication and an equally important contribution to global public policy. The *Encyclopedia* brings together a remarkable range of cutting-edge scientific knowledge on all aspects of soil science, as well as the links of soils and soil science to environmental management, food production, biodiversity, climate change, and many other areas of significant concern. Even more than that, the *Encyclopedia* will immediately become an indispensable resource for policy makers, analysts, and students who are focusing on one of the greatest challenges of the 21st century. With 6.3 billion people, our planet is already straining to feed the world's population, and is failing to do so reliably in many parts of the world. The numbers of chronically poor in the world have been stuck at some 800 million in recent years, despite long-standing international goals and commitments to reduce that number by several hundred million. Yet the challenge of food production will intensify in coming decades, as the human population is projected to rise to around 9 billion by mid-century, with the increased population concentrated in parts of the world already suffering from widespread chronic under-nourishment.

Unless the best science is brought to these problems, the situation is likely to deteriorate sharply. Food production systems are already under stress, for reasons often related directly to soils management. In Africa, crop yields are disastrously low and falling in many places due to the rampant depletion of soil nutrients. This situation needs urgent reversal, through increasing use of agro-forestry techniques (e.g. inter-cropping cereals with leguminous nitrogen-fixing trees) and increasing the efficient applications of chemical fertilizers. In other impoverished, as well as rich, parts of the planet, decades of intensive agriculture under irrigation have led to salinization, water-logging, eutrophication of major water bodies, dangerous declines of biodiversity and other forms of environmental degradation. These enormous strains are coupled with the continuing pressures of tropical deforestation and the lack of new promising regions for expanding crop cultivation to meet the needs of growing populations. Finally, there looms the prospect of anthropogenic climate change. Global warming and associated complex and poorly understood shifts in precipitation extremes and other climate variables all threaten the world's natural ecosystems and food production systems in profound yet still imperfectly understood ways. The risks of gradual or abrupt climate change are coupled with the risks of drastic perturbations to regional and global food supplies.

The *Encyclopedia* offers state-of-the-art contributions on each of these challenges, as well as links to entries on the fundamental biophysical processes that underpin the relevant phenomena. The world-scale and world-class collaboration that stands behind this unique project signifies its importance for the world community. It is an honor and privilege for me to introduce this path-breaking endeavor.

Jeffrey D Sachs
Director
The Earth Institute at Columbia University
Quetelet Professor of Sustainable Development
Columbia University, New York, USA

PREFACE

The term 'soil' refers to the weathered and fragmented outer layer of our planet's land surfaces. Formed initially through the physical disintegration and chemical alteration of rocks and minerals by physical and biogeochemical processes, soil is influenced by the activity and accumulated residues of a myriad of diverse forms of life. As it occurs in different geologic and climatic domains, soil is an exceedingly variegated body with a wide range of attributes.

Considering the height of the atmosphere, the thickness of the earth's rock mantle, and the depth of the ocean, one observes that soil is an amazingly thin body – typically not much more than one meter thick and often less than that. Yet it is the crucible of terrestrial life, within which biological productivity is generated and sustained. It acts like a composite living entity, a home to a community of innumerable microscopic and macroscopic plants and animals. A mere fistful of soil typically contains billions of microorganisms, which perform vital interactive biochemical functions. Another intrinsic attribute of the soil is its sponge-like porosity and its enormous internal surface area. That same fistful of soil may actually consist of several hectares of active surface, upon which physicochemical processes take place continuously.

Realizing humanity's utter dependence on the soil, ancient peoples, who lived in greater intimacy with nature than many of us today, actually revered the soil. It was not only their source of livelihood, but also the material from which they built their homes and that they learned to shape, heat, and fuse into household vessels and writing tablets (ceramic, made of clayey soil, being the first synthetic material in the history of technology). In the Bible, the name assigned to the first human was Adam, derived from 'adama,' meaning soil. The name given to that first earthling's mate was Hava (Eve, in transliteration), meaning 'living' or 'life-giving.' Together, therefore, Adam and Eve signified quite literally 'Soil and Life.'

The same powerful metaphor is echoed in the Latin name for the human species – Homo, derived from humus, the material of the soil. Hence, the adjective 'human' also implies 'of the soil.' Other ancient cultures evoked equally powerful associations. To the Greeks, the earth was a manifestation of Gaea, the maternal goddess who, impregnated by Uranus (god of the sky), gave birth to all the gods of the Greek pantheon.

Our civilization depends on the soil more crucially than ever, because our numbers have grown while available soil resources have diminished and deteriorated. Paradoxically, however, even as our dependence on the soil has increased, most of us have become physically and emotionally detached from it. Many of the people in the so-called 'developed' countries spend their lives in the artificial environment of a city, insulated from direct exposure to nature, and some children may now assume as a matter of course that food originates in supermarkets.

Detachment has bred ignorance, and out of ignorance has come the delusion that our civilization has risen above nature and has set itself free of its constraints. Agriculture and food security, erosion and salination, degradation of natural ecosystems, depletion and pollution of surface waters and aquifers, and decimation of biodiversity – all of these processes, which involve the soil directly or indirectly – have become abstractions to many people. The very language we use betrays disdain for that common material underfoot, often referred to as 'dirt.' Some fastidious parents prohibit their children from playing in the mud and hurry to wash their 'soiled' hands when the children nonetheless obey an innate instinct to do so. Thus soil is devalued and treated

as unclean though it is the terrestrial realm's principal medium of purification, wherein wastes are decomposed and nature's productivity is continually rejuvenated.

Scientists who observe soil closely see it in effect as a seething foundry in which matter and energy are in constant flux. Radiant energy from the sun streams onto the field and cascades through the soil and the plants growing in it. Heat is exchanged, water percolates through the soil's intricate passages, plant roots extract water and transmit it to their leaves, which transpire it back to the atmosphere. Leaves absorb carbon dioxide from the air and synthesize it with soil-derived water to form the primary compounds of life. Oxygen emitted by the leaves makes the air breathable for animals, which consume and in turn fertilize plants.

Soil is thus a self-regulating bio-physio-chemical factory, processing its own materials, water, and solar energy. It also determines the fate of rainfall and snowfall reaching the ground surface – whether the water thus received will flow over the land as runoff, or seep downward to the subterranean reservoir called groundwater, which in turn maintains the steady flow of springs and streams. With its finite capacity to absorb and store moisture, and to release it gradually, the soil regulates all of these phenomena. Without the soil as a buffer, rain falling over the continents would run off entirely, producing violent floods rather than sustained river flow.

Soil naturally acts as a living filter, in which pathogens and toxins that might otherwise accumulate to foul the terrestrial environment are rendered harmless. Since time immemorial, humans and other animals have been dying of all manner of disease and have then been buried in the soil, yet no major disease is transmitted by it. The term *antibiotic* was coined by soil microbiologists who, as a consequence of their studies of soil bacteria and actinomycetes, discovered streptomycin (an important cure for tuberculosis and other infections). Ion exchange, a useful process of water purification, also was discovered by soil scientists studying the passage of solutes through beds of clay.

However unique in form and function, soil is not an isolated body. It is, rather, a central link in the larger chain of interconnected domains and processes comprising the terrestrial environment. The soil interacts both with the overlying atmosphere and the underlying strata, as well as with surface and underground bodies of water. Especially important is the interrelation between the soil and the climate. In addition to its function of regulating the cycle of water, it also regulates energy exchange and surface temperature.

When virgin land is cleared of vegetation and turned into a cultivated field, the native biomass above the ground is often burned and the organic matter within the soil tends to decompose. These processes release carbon dioxide into the atmosphere, thus contributing to the earth's greenhouse effect and to global warming. On the other hand, the opposite act of reforestation and soil enrichment with organic matter, such as can be achieved by means of conservation management, may serve to absorb carbon dioxide from the atmosphere. To an extent, the soil's capacity to store carbon can thus help to mitigate the greenhouse effect.

Thousands of years are required for nature to create life-giving soil out of sterile bedrock. In only a few decades, however, unknowing or uncaring humans can destroy that wondrous work of nature. In various circumstances, mismanaged soils may be subject to erosion (the sediments of which tend to clog streambeds, estuaries, lakes, and coastal waters), to leaching of nutrients with attendant loss of fertility and eutrophication of water bodies, to waterlogging and impaired aeration, or to an excessive accumulation of salts that may cause a once-productive soil to become entirely sterile. Such processes of soil degradation, sometimes called 'desertification,' already affect large areas of land.

We cannot manage effectively and sustainably that which we do not know and thoroughly understand. That is why the tasks of developing and disseminating sound knowledge of the soil and its complex processes have assumed growing urgency and importance. The global environmental crisis has created a compelling need for a concentrated, concise, and definitive source of information – accessible to students, scientists, practitioners, and the general public – about the soil in all its manifestations – in nature and in relation to the life of humans.

Daniel Hillel
Editor-in-Chief
May 2004

INTRODUCTION

The *Encyclopedia of Soils in the Environment* contains nearly 300 articles, written by the world's leading authorities. Pedologists, biologists, ecologists, earth scientists, hydrologists, climatologists, geographers, and representatives from many other disciplines have contributed to this work. Each of the articles separately, and all of them in sequence and combination, serve to summarize and encapsulate our present knowledge of the world's variegated soils, their natural functions, and their importance to humans.

Concise articles surveying specific aspects of soils (soil genesis, soil chemistry and mineralogy, soil physics and hydrology, and soil biology) are complemented by articles covering transdisciplinary aspects, such as the role of soils in ecology, the history of soil utilization for agricultural and engineering purposes, the development of soil science as a discipline, and the potential or actual contributions of soils to the generation, as well as to the mitigation, of pollution and of global climate change.

This comprehensive reference encompasses both the fundamental and the applied aspects of soil science, interfacing in general with the physical sciences and life sciences and more specifically with the earth sciences and environmental sciences.

The *Encyclopedia of Soils in the Environment* manifests the expanding scope of modern soil science, from its early sectarian focus on the utilitarian attributes of soils in agriculture and engineering, to a wider and much more inclusive view of the soil as a central link in the continuous chain of processes constituting the dynamic environment as a whole. Thus it both details and integrates a set of topics that have always been of vital importance to human societies and that are certain to be even more so in the future.

Daniel Hillel
Editor-in-Chief
May 2004

CONTENTS

Contents are given as follows: CHAPTER NAME Author(s) Page number

VOLUME 1

A

ACID RAIN AND SOIL ACIDIFICATION L Blake 1

ACIDITY N Bolan, D Curtin and D Adriano 11

AERATION D E Rolston 17

AGGREGATION

Microbial Aspects S D Frey 22

Physical Aspects J R Nimmo 28

AGROFORESTRY P K R Nair 35

AIR PHASE see AERATION; DIFFUSION

ALBEDO see ENERGY BALANCE; RADIATION BALANCE

ALLOPHANE AND IMOGOLITE see AMORPHOUS MATERIALS

ALLUVIUM AND ALLUVIAL SOILS J L Boettinger 45

ALUMINUM SPECIATION D R Parker 50

AMMONIA D E Kissel and M L Cabrera 56

AMORPHOUS MATERIALS J Harsh 64

ANAEROBIC SOILS P W Inglett, K R Reddy and R Corstanje 72

ANION EXCHANGE see CATION EXCHANGE

APPLICATIONS OF SOILS DATA P J Lawrence 78

ARCHAEA J E T McLain 88

ARCHEOLOGY IN RELATION TO SOILS J A Homburg 95

B

BACTERIA

Plant Growth-Promoting Y Bashan and L E de-Bashan 103

Soil L J Halverson 115

BACTERIOPHAGE M Radosevich, K E Williamson and K E Wommack 122

BIOCONTROL OF SOIL-BORNE PLANT DISEASES C E Pankhurst and J M Lynch 129

BIODIVERSITY D H Wall 136

BUFFERING CAPACITY B R James 142

BULK DENSITY see POROSITY AND PORE-SIZE DISTRIBUTION

C

CALCIUM AND MAGNESIUM IN SOILS N Bolan, P Loganathan and S Saggar 149

CAPILLARITY D Or and M Tuller 155

CARBON CYCLE IN SOILS

Dynamics and Management C W Rice 164

Formation and Decomposition C A Cambardella 170

CARBON EMISSIONS AND SEQUESTRATION K Paustian 175

CATION EXCHANGE L M McDonald, V P Evangelou and M A Chappell 180

CHEMICAL EQUILIBRIA A P Schwab 189

CHEMICAL SPECIATION MODELS see SURFACE COMPLEXATION MODELING

CHERNOZEMS see GRASSLAND SOILS

CHILDS, ERNEST CARR E G Youngs 195

CIVILIZATION, ROLE OF SOILS D Hillel 199

CLASSIFICATION OF LAND USE see LAND-USE CLASSIFICATION

CLASSIFICATION OF SOILS R W Arnold 204

CLASSIFICATION SYSTEMS

Australian R W Fitzpatrick 211

FAO F O Nachtergaele 216

Russian, Background and Principles M Gerasimova 223

Russian, Evolution and Examples D Konyushkov 227

USA D J Brown 235

CLAY MINERALS D G Schulze 246

CLIMATE CHANGE IMPACTS P Bullock 254

CLIMATE MODELS, ROLE OF SOIL P Smith 262

COLD-REGION SOILS C-L Ping 268

COLLOID-FACILITATED SORPTION AND TRANSPORT R Kretzschmar 276

COMPACTION J J H van den Akker and B Soane 285

COMPOST T L Richard 294

CONDITIONERS R E Sojka, J A Entry and W J Orts 301

CONSERVATION see EROSION: Water-Induced; Wind-Induced; SUSTAINABLE SOIL AND LAND MANAGEMENT; TERRACES AND TERRACING

CONSERVATION TILLAGE M R Carter 306

COVER CROPS L Edwards and J Burney 311

CROP ROTATIONS C A Francis 318

CROP WATER REQUIREMENTS L S Pereira and I Alves 322

CROP-RESIDUE MANAGEMENT D C Reicosky and A R Wilts 334

CRUSTS

Biological J Belnap 339

Structural R L Baumhardt and R C Schwartz 347

CULTIVATION AND TILLAGE M R Carter and E McKyes 356

D

DARCY'S LAW D Swartzendruber 363

DEGRADATION C J Ritsema, G W J van Lynden, V G Jetten and S M de Jong 370

DENITRIFICATION D A Martens 378

DESERTIFICATION D Hillel and C Rosenzweig 382

DIFFUSION T Addiscott and P Leeds-Harrison 389

DISINFESTATION A Gamliel and J Katan 394

DISPERSION see FLOCCULATION AND DISPERSION

DISSOLUTION PROCESSES, KINETICS K G Scheckel and C A Impellitteri 400

DRAINAGE, SURFACE AND SUBSURFACE N R Fausey 409

DRYLAND FARMING G A Peterson 414

E

EARTHWORMS see FAUNA

EDAPHOLOGY A L Ulery 419

ELECTRON PARAMAGNETIC RESONANCE see ELECTRON-SPIN RESONANCE SPECTROSCOPY

ELECTRON-SPIN RESONANCE SPECTROSCOPY N Senesi and G S Senesi 426

ELECTROSTATIC DOUBLE-LAYER see CATION EXCHANGE

ENERGY BALANCE M Fuchs 438

ENVIRONMENTAL MONITORING P J Loveland and P H Bellamy 441

ENZYMES IN SOILS R P Dick and E Kandeler 448

EROSION

Irrigation-Induced G A Lehrsch, D L Bjorneberg and R E Sojka 456

Water-Induced J E Gilley 463

Wind-Induced T M Zobeck and R S Van Pelt 470

ESSENTIAL ELEMENTS E A Kirkby 478

EUTROPHICATION A J Gold and J T Sims 486

EVAPORATION OF WATER FROM BARE SOIL C W Boast and F W Simmons 494

EVAPOTRANSPIRATION G Stanhill 502

F

FACTORS OF SOIL FORMATION

Biota A H Jahren 507

Climate O C Spaargaren and J A Deckers 512

Human Impacts J Sandor, C L Burras and M Thompson 520

Parent Material K R Olson 532

Time E F Kelly and C M Yonker 536

FAUNA T Winsome 539

VOLUME 2

FERTIGATION U Kafkafi and S Kant 1

FERTILITY J L Havlin 10

FERTILIZERS AND FERTILIZATION H W Scherer 20

FIELD CAPACITY see WATER CYCLE

FLOCCULATION AND DISPERSION I Shainberg and G J Levy 27

FLUORESCENCE SPECTROSCOPY N Senesi and V D'Orazio 35

FOLIAR APPLICATIONS OF NUTRIENTS M Tagliavini and M Toselli 53

FOOD-WEB INTERACTIONS P C de Ruiter and J C Moore 59

FORENSIC APPLICATIONS W F Rowe 67

FOREST SOILS J R Boyle 73

FOURIER TRANSFORM INFRARED SPECTROSCOPY D Peak 80

FRACTAL ANALYSIS Y Pachepsky and J W Crawford 85

FREEZING AND THAWING

Cycles B Sharratt 98

Processes G N Flerchinger, G A Lehrsch and D K McCool 104

FUNGI K Ritz 110

G

GEOGRAPHICAL INFORMATION SYSTEMS J Bo¨hner, T Selige and R Ko¨the 121

GERMINATION AND SEEDLING ESTABLISHMENT A Hadas 130

GLOBAL WARMING see CARBON EMISSIONS AND SEQUESTRATION; CLIMATE CHANGE IMPACTS; GREENHOUSE GAS EMISSIONS

GRASSLAND SOILS J A Mason and C W Zanner 138

GREEN MANURING see COVER CROPS

GREENHOUSE GAS EMISSIONS K A Smith 145

GROUNDWATER AND AQUIFERS Y Bachmat 153

GROUNDWATER POLLUTION see POLLUTION: Groundwater

H

HEAT AND MOISTURE TRANSPORT R Horton and A Globus 169

HEAT CAPACITY see THERMAL PROPERTIES AND PROCESSES

HEAT FLOW see THERMAL PROPERTIES AND PROCESSES

HEAVY METALS D C Adriano, N S Bolan, J Vangronsveld and W W Wenzel 175

HILGARD, EUGENE WOLDEMAR R Amundson 182

HOOGHOUDT, SYMEN BAREND P A C Raats and R R van der Ploeg 188

HUMIFICATION T C Balser 195

HYDRAULIC PROPERTIES, TEMPERATURE EFFECTS S A Grant 207

HYDRIC SOILS G W Hurt 212

HYDROCARBONS P Kostecki, R Morrison and J Dragun 217

HYDRODYNAMIC DISPERSION see SOLUTE TRANSPORT

HYDRODYNAMICS IN SOILS T P A Ferre´ and A W Warrick 227

HYSTERESIS J H Dane and R J Lenhard 231

I

IMMISCIBLE FLUIDS R J Lenhard, J H Dane and M Oostrom 239

INCEPTISOLS A Palmer 248

INDUSTRIAL POLLUTION see POLLUTION: Industrial

INFILTRATION T P A Ferre´ and A W Warrick 254

INFRARED SPECTROSCOPY see FOURIER TRANSFORM INFRARED SPECTROSCOPY

IRON NUTRITION K Mengel and H Kosegarten 260

IRRIGATION

Environmental Effects S Topcu and C Kirda 267

Methods D L Bjorneberg and R E Sojka 273

ISOTOPES IN SOIL AND PLANT INVESTIGATIONS K Reichardt and O O S Bacchi 280

ISOTROPY AND ANISOTROPY T-C J Yeh, P Wierenga, R Khaleel and R J Glass 285

J

JENNY, HANS R Amundson 293

K

KELLOGG, CHARLES J D Helms 301

KINETIC MODELS P M Jardine 307

KIRKHAM, DON D R Nielsen and R R van der Ploeg 315

L

LAMINAR AND TURBULENT FLOW see HYDRODYNAMICS IN SOILS

LANDFILLS see WASTE DISPOSAL ON LAND: Municipal

LAND-USE CLASSIFICATION J A LaGro Jr 321

LAWES, JOHN BENNET AND GILBERT, JOSEPH HENRY A E Johnston 328

LEACHING PROCESSES B E Clothier and S Green 336

LIEBIG, JUSTUS VON R R van der Ploeg, W Bo¨hm and M B Kirkham 343

LIMING E J Kamprath and T J Smyth 350

LIPMAN, JACOB G. J C F Tedrow 358

LOESS A J Busacca and M R Sweeney 364

LOWDERMILK, WALTER CLAY J D Helms 373

LYSIMETRY T A Howell 379

M

MACRONUTRIENTS C W Wood, J F Adams and B H Wood 387

MACROPORES AND MACROPORE FLOW, KINEMATIC WAVE APPROACH P F Germann 393

MAGNESIUM IN SOILS see CALCIUM AND MAGNESIUM IN SOILS

MANURE MANAGEMENT J T Sims and R O Maguire 402

MARBUT, CURTIS FLETCHER J P Tandarich 410

MATRIC POTENTIAL see HYDRODYNAMICS IN SOILS; WATER POTENTIAL; WATER RETENTION AND CHARACTERISTIC CURVE

MEDITERRANEAN SOILS J Torrent 418

METAL OXIDES A C Scheinost 428

METALS AND METALLOIDS, TRANSFORMATION BY MICROORGANISMS S M Glasauer, T J Beveridge, E P Burford, F A Harper and G M Gadd 438

METALS, HEAVY see HEAVY METALS

MICROBIAL PROCESSES

Environmental Factors P G Hartel 448

CommunityAnaly sis C H Nakatsu 455

Kinetics N S Panikov 463

MICRONUTRIENTS L M Shuman 479

MINERAL–ORGANIC–MICROBIAL INTERACTIONS P M Huang, M C Wang and M K Wang 486

MINERALS, PRIMARY P M Huang and M K Wang 500

MINERALS, SECONDARY see CLAY MINERALS

MINIMUM TILLAGE see CONSERVATION TILLAGE

MISCIBLE DISPLACEMENT see SOLUTE TRANSPORT

MORPHOLOGY P R Owens and E M Rutledge 511

MULCHES C L Acharya, K M Hati and K K Bandyopadhyay 521

MYCORRHIZAL FUNGI L M Egerton-Warburton, J I Querejeta, M F Allen and S L Finkelman 533

VOLUME 3

N

NEMATODES D A Neher and T O Powers 1

NEUTRON SCATTERING M J Fayer and G W Gee 6

NITROGEN IN SOILS

- Cycle M S Coyne and W W Frye 13
- Nitrates D S Powlson and T M Addiscott 21
- Nitrification J I Prosser 31
- Plant Uptake A Hodge 39
- Symbiotic Fixation J I Sprent 46

NITROGEN FERTILIZERS see FERTILIZERS AND FERTILIZATION

NUCLEAR WASTE DISPOSAL G W Gee, P D Meyer and A L Ward 56

NUTRIENT AVAILABILITY N K Fageria and V C Baligar 63

NUTRIENT MANAGEMENT G D Binford 71

O

ORGANIC FARMING C A Francis 77

ORGANIC MATTER

- Principles and Processes M Schnitzer 85
- Genesis and Formation K M Haider and G Guggenberger 93
- Interactions with Metals N Senesi and E Loffredo 101

ORGANIC RESIDUES, DECOMPOSITION A J Franzluebbers 112

ORGANIC SOILS D L Mokma 118

OVERLAND FLOW T S Steenhuis, L Agnew, P Ge´rard-Marchant and M T Walter 130

OXIDATION–REDUCTION OF CONTAMINANTS C J Matocha 133

P

PADDY SOILS C Witt and S M Haefele 141

PARENT MATERIAL see PEDOLOGY: Basic Principles; FACTORS OF SOIL FORMATION: Parent Material

PEDOLOGY

- Basic Principles M J Singer 151
- Dynamic F C Ugolini 156

PEDOMETRICS I O A Odeh and A B McBratney 166

PENMAN, HOWARD LATIMER J L Monteith 176

PENMAN–MONTEITH EQUATION R Allen 180

PERCOLATION see HYDRODYNAMICS IN SOILS

PERMAFROST see POLAR SOILS

PERMEABILITY see HYDRODYNAMICS IN SOILS

PERSISTENT ORGANIC POLLUTANTS (POPS) see POLLUTANTS: Persistent Organic (POPs)

PESTICIDES R H Bromilow 188

PETROLEUM see HYDROCARBONS

pH N Bolan and K Kandaswamy 196

PHOSPHORUS IN SOILS

Overview J T Sims and P A Vadas 202

Biological Interactions M D Mullen 210

PHYTOTOXIC SUBSTANCES IN SOILS M Qadir, S Schubert and D Steffens 216

PLANT–SOIL–WATER RELATIONS R A Feddes and J C van Dam 222

PLANT–WATER RELATIONS C Gimenez, M Gallardo and R B Thompson 231

POISEUILLE'S LAW see HYDRODYNAMICS IN SOILS

POLAR SOILS J C F Tedrow 239

POLLUTANTS

Biodegradation P B Hatzinger and J W Kelsey 250

Effects on Microorganisms M E Fuller 258

Persistent Organic (POPs) D Johnson 264

POLLUTION

Groundwater H Rubin 271

Industrial S P McGrath 282

POLYMERS AND MICROORGANISMS M C Rillig 287

POORLY CRYSTALLINE ALLUMINOSILICATES see AMORPHOUS MATERIALS

POROSITY AND PORE-SIZE DISTRIBUTION J R Nimmo 295

POTASSIUM IN SOILS P M Huang, J M Zhou, J C Xie and M K Wang 303

PRECIPITATION, WATERSHED ANALYSIS J V Bonta 314

PRECIPITATION–DISSOLUTION PROCESSES W P Robarge 322

PRECISION AGRICULTURE see SITE-SPECIFIC SOIL MANAGEMENT

PREFERENTIAL FLOW see UNSTABLE FLOW; MACROPORES AND MACROPORE FLOW, KINEMATIC WAVE APPROACH

PRODUCTIVITY D L Karlen 330

PROFILE see MORPHOLOGY

PROTOZOA W Foissner 336

Q

QUALITY OF SOIL B J Wienhold, G E Varvel and J W Doran 349

R

RADIATION BALANCE J L Hatfield, T J Sauer and J H Prueger 355

RADIONUCLIDES see ISOTOPES IN SOIL AND PLANT INVESTIGATIONS

RAINFED FARMING see DRYLAND FARMING

RANGE MANAGEMENT G L Anderson 360

RECYCLING OF ORGANIC WASTES see POLLUTANTS: Biodegradation

REDISTRIBUTION see WATER CYCLE

REDOX POTENTIAL R D DeLaune and K R Reddy 366

REDOX REACTIONS, KINETICS P S Nico and S Fendorf 372

REMEDIATION OF POLLUTED SOILS E Lombi and R E Hamon 379

REMOTE SENSING

Organic Matter D K Morris, C J Johannsen, S M Brouder and G C Steinhardt 385

Soil Moisture T J Jackson 392

RHIZOSPHERE A C Kennedy and L Z de Luna 399

RICHARDS, LORENZO A. W R Gardner 407

ROOT ARCHITECTURE AND GROWTH L E Jackson 411

ROOT EXUDATES AND MICROORGANISMS B-J Koo, D C Adriano, N S Bolan and C D Barton 421

S

SALINATION PROCESSES I Shainberg and G J Levy 429

SALINITY

Management D Hillel 435

Physical Effects D Russo 442

SALT BALANCE OF SOILS see SALINATION PROCESSES

SALT-AFFECTED SOILS, RECLAMATION R Keren 454

SAND DUNES H Tsoar 462

SATURATED AND UNSATURATED FLOW see HYDRODYNAMICS IN SOILS; VADOSE ZONE: Hydrologic Processes

SCALING

Physical Properties and Processes G Sposito 472

Transport Processes R P Ewing 477

SEPTIC SYSTEMS R L Lavigne 485

SHIFTING CULTIVATION R Lal 488

SITE-SPECIFIC SOIL MANAGEMENT C J Johannsen and P G Carter 497

SLASH AND BURN AGRICULTURE see SHIFTING CULTIVATION

SLUDGE see WASTE DISPOSAL ON LAND: Liquid; Municipal

SODIC SOILS G J Levy and I Shainberg 504

SOIL–PLANT–ATMOSPHERE CONTINUUM J M Norman and M C Anderson 513

SOLUTE TRANSPORT M C Sukop and E Perfect 521

SORPTION

Metals D L Sparks 532

Organic Chemicals B Xing and J J Pignatello 537

Oxyanions C P Schulthess, H Wijnja and W Yang 548

SORPTION–DESORPTION, KINETICS D L Sparks 556

SPATIAL PATTERNS J H Go¨rres and J A Amador 562

VOLUME 4

SPATIAL VARIATION, SOIL PROPERTIES R Webster 1

SPECIFIC SURFACE AREA K D Pennell 13

STATISTICS IN SOIL SCIENCE R Webster 19

STERILIZATION see DISINFESTATION

STOCHASTIC ANALYSIS OF SOIL PROCESSES D Russo 29

STRESS–STRAIN AND SOIL STRENGTH S K Upadhyaya 38

STRUCTURE V A Snyder and M A Va´zquez 54

SUBSOILING R L Raper 69

SULFUR IN SOILS

- Overview M A Tabatabai 76
- Biological Transformations S D Siciliano and J J Germida 85
- Nutrition M A Tabatabai 91

SURFACE COMPLEXATION MODELING S Goldberg 97

SUSTAINABLE SOIL AND LAND MANAGEMENT J L Berc 108

SWELLING AND SHRINKING D Smiles and P A C Raats 115

T

TEMPERATE REGION SOILS E A Nater 125

TEMPERATURE REGIME see THERMAL PROPERTIES AND PROCESSES

TENSIOMETRY T K Tokunaga 131

TERMITES see FAUNA

TERRA ROSSA see MEDITERRANEAN SOILS

TERRACES AND TERRACING G R Foster 135

TESTING OF SOILS A P Mallarino 143

TEXTURE G W Gee 149

THERMAL PROPERTIES AND PROCESSES D Hillel 156

THERMODYNAMICS OF SOIL WATER P H Groenevelt 163

TILLAGE see CONSERVATION TILLAGE; CULTIVATION AND TILLAGE; ZONE TILLAGE

TILTH D L Karlen 168

TIME-DOMAIN REFLECTOMETRY G C Topp and T P A Ferre´ 174

TROPICAL SOILS

Arid and Semiarid H C Monger, J J Martinez-Rios and S A Khresat 182

Humid Tropical S W Buol 187

U

UNSTABLE FLOW T S Steenhuis, J-Y Parlange, Y-J Kim, D A DiCarlo, J S Selker, P A Nektarios, D A Barry and F Stagnitti 197

URBAN SOILS J L Morel, C Schwartz, L Florentin and C de Kimpe 202

V

VADOSE ZONE

Hydrologic Processes J W Hopmans and M Th van Genuchten 209

Microbial Ecology P A Holden and N Fierer 216

VIRUSES see BACTERIOPHAGE

VOLCANIC SOILS G Uehara 225

W

WAKSMAN, SELMAN A. H B Woodruff 233

WASTE DISPOSAL ON LAND

Liquid C P Gerba 238

Municipal D A C Manning 247

WATER AVAILABILITY see PLANT–SOIL–WATER RELATIONS

WATER CONTENT AND POTENTIAL, MEASUREMENT G S Campbell and C S Campbell 253

WATER CYCLE D K Cassel and B B Thapa 258

WATER EROSION see EROSION: Water-Induced

WATER HARVESTING D Hillel 264

WATER MANAGEMENT see CROP WATER REQUIREMENTS

WATER POTENTIAL D Or, M Tuller and J M Wraith 270

WATER REQUIREMENTS see CROP WATER REQUIREMENTS

WATER RETENTION AND CHARACTERISTIC CURVE M Tuller and D Or 278

WATER TABLE see GROUNDWATER AND AQUIFERS

WATER, PROPERTIES D Hillel 290

WATER-REPELLENT SOILS J Letey 301

WATERSHED MANAGEMENT M D Tomer 306

WATER-USE EFFICIENCY M B Kirkham 315

WEED MANAGEMENT D D Buhler 323

WETLANDS, NATURALLY OCCURRING E K Hartig 328

WIDTSOE, JOHN A. AND GARDNER, WILLARD G S Campbell and W H Gardner 335

WIND EROSION see EROSION: Wind-Induced

WINDBREAKS AND SHELTERBELTS E S Takle 340

WOMEN IN SOIL SCIENCE (USA) M J Levin 345

WORLD SOIL MAP H Eswaran and P F Reich 352

Z

ZERO-CHARGE POINTS J Chorover 367

ZONE TILLAGE J L Hatfield and A T Jeffries 373

Table of Contents, Volume 2

F (cont.) **1**

Fertigation **1**

Introduction 1
Advantages of Fertigation 1
Drip-Fertigation System 1
Containers for Fertilizers 1
Filtration 2
Distribution of Water and Nutrients in Soil 2
Water Flow 2
Nutrient Movement 2
Selection of Fertilizers 3
Solubility of Fertilizers 4
Compatibility of Fertilizers 5
Precipitation in the Irrigation Lines 5
Scheduling Fertigation 5
Behavior of Nutrients in Soil 6
Root Growth 7
Subsurface Fertigation 9
Further Reading 9

Fertility **10**

Introduction 10
The Importance of Soil Fertility 10
Essential Plant Nutrients 10
Basic Soil Properties and Processes 11
Cation and Anion Exchange 11
Base Saturation 12
Buffer Capacity 12
Soil Organic Matter 12
Soil Acidity and Alkalinity 13
Specific Nutrients 14
Nitrogen 14
N mineralization/immobilization 14
Nitrification 15
NH4+ fixation 15
Denitrification 15
Volatilization 15
Phosphorus 15
Organic P 16
Inorganic P 16
Potassium 16
Sulfur 17
Calcium and Magnesium 18
Micronutrients 19
Iron, zinc, manganese, copper 19
Boron, chloride, molybdenum, nickel 19
Further Reading 19

Fertilizers and Fertilization **20**

Introduction 20
Nitrogen 20
Soil Nitrogen 20
N Uptake of Plants 20
Soil Test 20
Time of N Fertilizer Application 21
N Fertilizers 21

Slow-Release and Bioinhibitor-Amended Nitrogen Fertilizers 22
Coated fertilizers 22
Uncoated fertilizers 22
Nitrification Inhibitors 22
Methods of N Application 22
Phosphorus 23
Soil Phosphorus 23
Soil Test 23
P Application Rates 23
P Fertilizers 23
Potassium 24
Soil Potassium 24
K Fertilizer Recommendation 24
K Fertilizers 25
Complex Fertilizers 25
Bulk Blends 25
Micronutrient Fertilizers 25
Iron 25
Manganese 26
Zinc 26
Copper 26
Boron 26
Molybdenum 26
Further Reading 26
Flocculation and Dispersion 27
Introduction 27
Shape and Charge of Soil Clays 27
Forces Between Clay Platelets 28
Repulsion Forces 28
The electrical double layer 28
Hydration repulsion forces 28
Flocculation and Attraction Forces 28
Van der Waals force of attraction 29
Electrostatic EF attraction forces 29
Clay charge density and electrostatic attraction forces 29
Spontaneous and Mechanical Dispersion 30
Flocculation and Dispersion in Na/Ca Clays 30
The flocculation value of sodium and calcium montmorillonites 30
Effect of pH on the FV of sodium montmorillonite 31
Dispersion and flocculation in Na/Ca clay systems 31
The effect of peptizing agents 32
Soil clay systems 32
Dispersion and the Hydraulic Properties of Soils 32
Summary 34
Further Reading 34
Fluorescence Spectroscopy 35
Introduction 35
Basic Principles, Methodology, and Instrumentation 35
The Fluorescence Phenomenon 35
Instrumentation and Experimental Procedures 36
Fluorescence Spectra 36
Fluorescence Efficiency, Lifetime, and Intensity 37
Limitations and Problems 37
Molecular and Environmental Factors Affecting Fluorescence 38
Fluorescence Quenching 38
Fluorescence Polarization 39

Fluorescence Correlation Spectroscopy 39
Soil Organic Components 39
Fluorescence Spectra 39
Nature of Fluorophores 40
Molecular and Environmental Factors Affecting Fluorescence 41
Humification Indexes and Differentiation of Humic Substances by Fluorescence Properties 41
Fluorescence Efficiency and Lifetime 43
Conformational Studies by Fluorescence Polarization 43
Diffusion Coefficients Determined by Fluorescence Correlation Spectroscopy 44
Fluorescence Quenching of Soil Organic Components by Metal Ions 44
Molecular and Mechanistic Aspects 44
Quantitative Aspects 46
Sorption/Binding of Fluorescent Organic Pollutants to Soil Organic Components 46
Fluorescence Polarization 47
Fluorescence Quenching 48
Soil Mineral Components 48
Structural Quenching 49
Adsorption of Organic Chemicals to Mineral Surfaces 49
Surface Acidity 49
Polarity of Mineral Surfaces 50
Molecular Distribution, Conformation, Aggregation, and Mobility 50
Biochemical Transformations on Clay Surfaces 51
Ligand-Promoted Mineral Dissolution 51
List of Technical Nomenclature 51
Further Reading 52
Foliar Applications of Nutrients 53
Introduction 53
Efficiency of Foliar-Applied Fertilizers 53
Application of Foliar Nutrients 54
Mechanisms of Penetration of Foliar-Applied Nutrients 55
Factors Affecting Cuticular Penetration 57
Plant Use of Foliar-Applied Nutrients 58
List of Technical Nomenclature 59
Further Reading 59
Food-Web Interactions 59
Introduction 59
Trophic Structures in Soil Food Webs 60
Trophic Interactions and Soil Ecosystem Processes 61
Food Web Models 62
Trophic Interactions and the Dynamics and Stability of Soil Populations and Food Webs 64
Further Reading 66
Forensic Applications 67
Introduction 67
Collection of Evidentiary Soil Samples 67
Laboratory Analysis of Forensic Soil Samples 68
Soil Organic Content 68
Soil Color 68
Percentage Organic Content 69
Soil Particle-Size Distributions 69
Soil Mineralogical Analysis 70
Soil as a Matrix 70
Buried Remains 70

Trace Evidence in Soil ... 71
Further Reading ... 72
Forest Soils ... 73
Introduction ... 73
Roots and Mycorrhizae ... 74
Litterfall and Litter: Organic Matter Recycling in Forest Soils ... 75
Biogeochemistry and Hydrology ... 76
Disturbance Effects on Forest Soils ... 77
Human Effects on Forest Soils ... 77
'Ecosystem Services' of Forest Soils ... 78
Criteria and Indicators for Soils: Sustainable Forest Management ... 78
Summary: Soil Conditions for Forest Productivity ... 78
Further Reading ... 79
Fourier Transform Infrared Spectroscopy ... 80
Introduction ... 80
The Nature of Molecular Vibrations ... 80
Infrared Peaks of Interest ... 80
Infrared Spectral Analysis ... 80
FTIR as a Quantitative Tool ... 82
Instrumentation ... 82
Common Experimental Techniques ... 82
Transmission-mode Experiments ... 83
Diffuse Reflectance Infrared Fourier Transform Experiments ... 83
Attenuated Total Reflectance Experiments ... 84
Further Reading ... 85
Fractal Analysis ... 85
Introduction ... 85
Fractal Objects ... 86
Physical Models of Emerging Fractal Scaling ... 89
Measuring Fractal Parameters of Soils ... 90
Direct Geometric Measurements ... 90
Proxy Measurements ... 93
Transport Processes ... 95
Uncertainty in Measuring Fractal Parameters ... 96
The Use of Soil Fractal Parameters ... 96
List of Technical Nomenclature ... 97
Further Reading ... 97
Freezing and Thawing ... 98
Cycles ... 98
Introduction ... 98
Global Significance ... 98
Guiding Principles ... 99
Tillage Practices ... 100
Crop Residue Management ... 101
Soil Amendments ... 102
List of Technical Nomenclature ... 103
Further Reading ... 103
Processes ... 104
Introduction ... 104
Heat and Water Relations During Freezing and Thawing ... 104
Water Migration and Frost Heave ... 105
Freezing Dynamics ... 105
Freeze-Thaw Impacts on Infiltration ... 106
Freeze-Thaw Impacts on Soil Erodibility ... 107
Freeze-Thaw Impacts on Aggregate Stability ... 108
Mode of Action ... 108

Factors Affecting the Impact 108
Soil Freezing and Thawing Effects on Solute Migration 109
Further Reading 110
Fungi 110
Introduction 110
Classification 110
Biology 111
Morphology and Form 111
Physiology and Reproduction 111
Biotic Interactions 112
Fungal Associations with Other Microbes 113
Fungal Associations with Plants 113
Mutualism 113
Parasitism 113
Fungal Associations with Fauna 114
Fungal Contributions to Soil-System Function 115
Nutrient Cycling 115
Soil Architecture 115
Methods of Study 116
Isolation Methods 116
Direct Observation 117
Molecular Methods 117
Activity 118
Exploitation of Soil Fungi 118
Further Reading 119
G 121
Geographical Information Systems 121
Introduction 121
Elementary Features and Basic Principles of GIS Applications to Soils 121
Data Inventory and Management Applications 123
Data Analyses and Mapping Applications 125
Modeling and Decision-Support applications 126
Summary 129
Further Reading 129
Germination and Seedling Establishment 130
Introduction 130
Seed Germination and Environmental Conditions 130
Water Requirements 130
Temperature Requirements 130
Aeration Requirements 131
Light Requirements 131
Soil Mechanical Constraints 131
Water Uptake by Seeds and Seedlings 131
The Imbibition Phase 132
The Transition Phase 132
The Growth Phase 132
Differentiation Between the Seed Germination Phases 132
Physical Principles of Water Uptake by Seeds and Seedlings 133
Specific Effects of Matric and Osmotic Soil-Water Potential Components 134
Seed-Water Potential 134
Seed and Soil Conductivities and Diffusivities to Water 134
Seed-Coat Diffusivity to Water 134
Seed-Coat and Seed-Soil Interface Geometrical Configuration 135
Impedance to Water Flow Across the Seed-Soil Interface 135
Seedling Emergence and Establishment 136

Future Research 136
List of Technical Nomenclature 136
Further Reading 137
Grassland Soils 138
Introduction 138
Global Distribution of Grasslands and Grassland Soils 138
Environment of Pedogenesis in Grasslands 139
Organic Matter Accumulation, Bioturbation, and Structure Development 140
Mineral Weathering, Clay Translocation, and Secondary Carbonate Mineral Formation 142
Grassland Soil Hydrology and Its Influence on Pedogenesis 144
Classification of Grassland Soils 144
Further Reading 145
Greenhouse Gas Emissions 145
Introduction 145
Methane Sources and Sinks 146
Soil Sources 146
Soil Sink for Methane 147
Efforts to Reduce Emissions 148
Nitrous Oxide 149
Agricultural Emissions 151
Potential for Reducing N2O Emissions 152
Further Reading 153
Groundwater and Aquifers 153
Basic Concepts 153
Types of Aquifer 154
Groundwater Basin 154
Prominent Features of an Aquifer 154
Large Areal Extent Compared with Saturated Thickness 154
Permanently yet Slowly Renewable Reservoir 154
Large Storage Capacity Compared with Annual Recharge/Discharge 154
Hidden Medium 155
Functions of an Aquifer 156
Easily Accessible and Durable Source of Water 156
Subsurface Storage Reservoir 156
Facility for Improving Water Quality 156
Aquifer as a Conduit 157
Aquifer Characteristics 157
Groundwater State and Behavior 157
Levels of Groundwater Study 158
Tracing the Movement of Groundwater 158
Darcy's Law 158
Aquifers and Groundwater as Part of a Water Resources System 159
Groundwater Contamination Sources and Events 161
Groundwater and Aquifers Need to be Managed 162
Groundwater Management Levels and Models 163
Use and Usefulness of Formal Models in Groundwater Management 165
Anthropogenic Constraints on Groundwater Management 165
Nonphysical Means of Groundwater Management 166
Acknowledgment 167
Further Reading 167
H 169
Heat and Moisture Transport 169
Examples of Coupled Heat and Water Flow in Soils 171
Limitations in Our Understanding of Coupled Heat and Water Flow 173
List of Technical Nomenclature 174

Further Reading....174
Heavy Metals....175
Introduction....175
Sources of Metals in the Environment....176
Environmental Contamination....176
Retention and Dynamics of Metals in Soils....177
Adsorption....177
Complexation....178
Precipitation....178
Bioavailability-Natural Attenuation Interactions....178
Biological Response to Metals....179
Soil Remediation....180
Acknowledgments....181
Further Reading....181
Hilgard, Eugene Woldemar....182
Further Reading....187
Hooghoudt, Symen Barend....188
Introduction....188
From Physical Chemistry to First Contacts with Drainage Theory and Practice....188
Physical Characterization of Soils in the Laboratory....190
Auger-Hole Method for Determining Hydraulic Conductivity in the Field....190
Flow of Water in Shallow Soils Without or With Structure in Their Natural Condition....191
The Hooghoudt Drainage Equation....192
Applications of Drainage Theory....192
Water Management of the Plant Root Zone....193
Mentor of Young Talents: The Theorists Van Deemter and Ernst....194
Hooghoudt's Legacy....194
Further Reading....194
Humification....195
Introduction....195
Historical Perspective....196
Chemical Perspective on HS Genesis....197
Traditional (Degradative) Methods of HS Characterization....198
Nondegradative and Newer Degradative Techniques....198
Chemical Perspective on Microbial Role in Humification....199
Microbiological Perspective....200
Phase 1: Initial Decomposition of Primary Resources....201
Phase 2: Slow Decomposition, Energy Limited....201
Phase 3: Direct Genesis and Degradation....201
Regulation of Humic Substances Genesis and Turnover....203
Climate and Soil Type....203
Amount and Chemical Composition of Starting Resources....204
Role of Microbial Community Composition....204
Summary....205
List of Technical Nomenclature....206
Further Reading....206
Hydraulic Properties, Temperature Effects....207
Introduction....207
Soil-Water Characteristics....207
Relative Permeability....207
Effect of Temperature on Soil-Water Characteristics....208
Effect of Temperature on Hydraulic Conductivity....208
Water and Energy Transport in a Nonisothermal Soil....209
Summary....210

List of Technical Nomenclature 210
Further Reading....211
Hydric Soils 212
Introduction....212
Hydrology of Hydric Soils....212
Important Chemical Reactions....213
Reduction of Oxygen....213
Reduction of Nitrogen 213
Iron and Manganese Reduction 213
Sulfur Reduction and Methane Production....214
Redox Potential 214
Factors that Control Redox Potential 214
Hydric Soil Indicators....214
Carbon Features 215
Iron and Manganese Features 215
Sulfur Features....216
Marl Features 216
Summary 216
Further Reading....216
Hydrocarbons....217
Introduction....217
Chemistry of Petroleum Hydrocarbons....217
Alkanes 217
Alkenes 218
Aromatics 218
Homologous Series....218
Isomers 219
Transport of Petroleum Hydrocarbons in Soil....219
Residual Saturation....220
Estimation of the Distribution of Hydrocarbons in Soil 220
Hydrocarbon Interactions at the Capillary Fringe 221
Solubilization of Hydrocarbons into Groundwater 221
Petroleum Hydrocarbon Transport in Groundwater....222
Free-Phase Groundwater Plume....222
Dissolved-Phase Groundwater Plume 222
Length of Petroleum Hydrocarbon Plumes 223
Transport in Fractures....223
Remediation of Petroleum Hydrocarbon-Impacted Soil and Groundwater....224
Physical Separation 224
Soil vapor extraction 224
Soil-washing (solvent extraction) 224
Thermal desorption....224
Biochemical Reactions....225
Natural attenuation and biopiles 225
Land farming....225
Landfills....225
Physicochemical Reactions....225
Solidification....225
Asphalt incorporation 225
Incineration 225
Vitrification 226
Summary 226
Further Reading....226
Hydrodynamics in Soils 227
Introduction....227
Steady-State Flow Through a Saturated Medium....227

Transient Flow Through a Saturated Medium 228
Steady-State Flow Through an Unsaturated Medium....229
Transient Flow Through an Unsaturated Medium 230
Further Reading....230
Hysteresis 231
Introduction....231
The Capillary Pressure Concept....232
Water-Retention Curve....232
Causes of Hysteresis....233
Modeling 234
Summary 236
Further Reading....237
I....239
Immiscible Fluids 239
Introduction....239
Fluid Distribution in the Pore Spaces....239
Movement of Immiscible Fluids in Soils....243
Partitioning of NAPL Components Among the Fluid and Solid Phases....245
Summary 246
Acknowledgment 246
Further Reading....246
Inceptisols 248
Introduction....248
Occurrence 248
Inceptisols with Permanent or Intermittent High Water Tables....248
Anthrepts 249
Cryepts 250
Ustepts....250
Xerepts 250
Udepts 251
Further Reading....253
Infiltration....254
Introduction....254
Plug Flow....255
Constant Infiltration at the Ground Surface....255
Constant Pressure Head at the Ground Surface....255
Richards Equation - The Shape of the Wetting Front....256
The Advance of a Wetting Front....256
The Effects of Infiltration Rate 257
The Effects of the Antecedent Water Content 258
The Effects of the Saturated Hydraulic Conductivity 258
The Effects of Soil Texture 258
The Effects of Gravity 258
Two-Dimensional Infiltration 259
Further Considerations 259
Further Reading....260
Iron Nutrition 260
Introduction....260
Iron in the Soil and its Acquisition by Plants....260
Iron Uptake, Long-Distance Transport in the Plant, and Storage in the Cell..262
Biochemical Functions of Fe in Plants....263
Lime-Induced Chlorosis....264
Iron Deficiency and Toxicity....265
Iron Application....266
List of Technical Nomenclature 266
Further Reading....266

Irrigation **267**
Environmental Effects 267
Introduction 267
Environmental Impact of Irrigated Agriculture 268
Water Shortage 268
Waterlogging and Salinity 269
Pollutants Attributable to Irrigation 270
Plant nutrients 270
Pesticides 271
Heavy metals 271
Public health concerns 272
Sustainability 272
Further Reading 273
Methods 273
Introduction 273
Surface Irrigation 274
Furrow Irrigation 274
Basin and Border Irrigation 275
Sprinkler Irrigation 276
Solid-Set Sprinkler Systems 276
Set-Move Sprinkler Systems 276
Moving Sprinkler Systems 277
Microirrigation 278
Subirrigation 278
Salinity Hazards 279
Choosing an Irrigation System 279
List of Technical Nomenclature 279
Further Reading 280
Isotopes in Soil and Plant Investigations **280**
Introduction 280
Radioisotopes as Tracers 280
Stable Isotopes as Tracers 281
Use of Radiation in Soil-Plant Investigations 283
Gamma Attenuation 283
Tomography 283
Neutron Moderation 283
Sterilization and Mutation Breeding 284
Other Nuclear Techniques 284
Activation Analysis 284
X-Ray Fluorescence 284
Autoradiography 284
C-14 dating 284
Further Reading 284
Isotropy and Anisotropy **285**
Introduction 285
Darcy Law 285
Control Volume, Representative Control Volume, Homogeneity, and Heterogeneity 285
Isotropy and Anisotropy 286
Scales and Anisotropy 286
Pressure Head- or Moisture-Dependent Anisotropy 287
Influences of Anisotropy on Movement of Moisture Plumes in Soils 289
List of Technical Nomenclature 291
Further Reading 291
J **293**
Jenny, Hans **293**

Further Reading....300
K....301
Kellogg, Charles....301
Introduction....301
Education and Educator....301
Soil Survey Division, US Department of Agriculture....302
Kellogg as Scientist-Administrator....302
Soil Taxonomy....303
Soil-Geomorphology Studies....303
Training Soil Scientists....303
International Influence....304
Tropical Soils....304
Interactions in Agricultural Productivity....305
Science and Policy....305
Further Reading....306
Kinetic Models....307
Introduction....307
Single-Component, Multiprocess Models....307
Physical Nonequilibrium....307
Two-region approach....307
Multiregion approach....308
Chemical and Microbiological Nonequilibrium....308
Adsorption....309
Transformation....309
Oxidation-Reduction....311
Multicomponent, Multiprocess Models....312
Nonequilibrium Processes Involving Dissolved Aqueous Phase Constituents....312
Nonequilibrium Processes Involving Dissolved Nonaqueous-Phase Constituents....313
Further Reading....314
Kirkham, Don....315
Kirkham's Scientific Contributions....317
Further Reading....319
L....321
Land-Use Classification....321
Introduction....321
Land and Society....321
Spatial Information and Mapping....321
Land-Use Classification....322
Land-Use Suitability....322
Subsurface Conditions....324
Surface Conditions....324
Land-Use Planning....325
Site Selection....325
Programming....325
Site Inventory....325
Site Analysis....326
Plan Development and Implementation....326
Conclusion....327
Further Reading....327
Lawes, John Bennet and Gilbert, Joseph Henry....328
John Bennet Lawes....328
Joseph Henry Gilbert....329
The Agricultural Background in the Second Quarter of the Nineteenth Century....329

The Years at Rothamsted Immediately Before 1843 330
The Earliest Experiments 330
Early Observations on the Need for Phosphorus and the Production of Superphosphate 330
The Start of the Experimental Program at Rothamsted 330
The Place of Long-Term Experiments in Agricultural Research and the Role of Lawes and Gilbert 334
Further Reading 336
Leaching Processes 336
Introduction 336
Leachate Quantity 337
Leachate Quality 338
Preferential Leaching Processes 338
Modeling and Risk Assessment 340
Summary 342
Further Reading 343
Liebig, Justus von 343
Introduction 343
Life History 344
The Work of Liebig 345
Liebig and Sprengel 347
Summary 348
Further Reading 349
Liming 350
Introduction 350
Development of Acid Soils 350
Nature of Soil Acidity 350
Soil Solution Acidity 350
Exchangeable Acidity 350
Nonexchangeable Acidity 351
Acid Soil Constraints to Plant Growth 351
Hydrogen Toxicity 351
Aluminum Toxicity 351
Manganese Toxicity 352
Calcium and Mg Deficiencies 352
Response to Liming 353
Cation Exchange Properties Affecting Liming of Acid Soils 355
Lime Requirement Determinations 355
Soil pH 355
Titration of Soil with a Base 355
Buffer pH Method 356
Exchangeable Al 356
Factors to Consider in Lime Application 357
Liming Materials 357
Neutralizing Value 357
Fineness of Liming Materials 357
Incorporation of Lime in the Soil 357
Conclusion 357
Further Reading 358
Lipman, Jacob G. 358
The Early Years 358
Lipman the Scientist 359
Lipman the Administrator 361
The End of an Era 363
Further Reading 364
Loess 364

Introduction 364
Distribution 365
Origin of Silt and Eolian Transport of Loess 365
Sedimentology 366
Stratigraphy 366
Geomorphology 367
Paleosols in Loess and Studies of Paleoclimate 368
Surface Soils in Loess 370
Environmental Concerns 370
Dust Additions to Other Soils 371
List of Technical Nomenclature 371
Further Reading 373
Lowdermilk, Walter Clay 373
Introduction 373
Further Reading 378
Lysimetry 379
Introduction 379
History 379
Lysimeter Types 379
Lysimeter Water Balance 380
Lysimeter Design Considerations 380
Lysimeter Shape and Area 381
Lysimeter Depth 382
Soil Profile Characteristics 382
Soil monolith collection 382
Lysimeter tanks 383
Lysimeter drainage systems 383
Weighing-Lysimeter Scale Systems 383
Measuring ET or Latent Heat Fluxes with Weighing Lysimeters 384
Summary 385
Further Reading 385
M 387
Macronutrients 387
Introduction 387
Nitrogen 387
N Fixation 387
Fertilizers and Manures 388
Phosphorus 389
Potassium 390
Calcium 390
Magnesium 391
Sulfur 391
List of Technical Nomenclature 392
Further Reading 392
Macropores and Macropore Flow, Kinematic Wave Approach 393
Introduction 393
Early Considerations on Flow and Transport in Soils 393
Deviations from Richards Equation and Convection-Dispersion Approaches 394
Alternative Approaches to Flow and Transport 395
General Requirements 395
Kinematic Wave Theory 396
Application of the Kinematic Wave Model 397
Further Developments of Kinematic Wave Approaches to Preferential Flow 400
Applications 400

Summary 400
List of Technical Nomenclature 401
Further Reading....401
Manure Management 402
Introduction....402
Manure Generation, Use, and Properties....402
Summary of Global Trends in Animal Production 402
Trends in Manure Production....404
Manure Characteristics by Species....405
Major Uses of Animal Manures 407
Best Management Practices (BMPs) for Animal Manures as Soil Amendments....407
Overview 407
Manure management for grassland agriculture....407
Managing manures for arable crop production....408
Contemporary Approaches to Manure Management....408
Summary 410
Further Reading....410
Marbut, Curtis Fletcher....410
Missouri Ozarks Native and Academic....411
Scientist in Soil Survey 413
Internationalist-Explorer....414
In Service to Humanity....417
Epilog....417
Further Reading....417
Mediterranean Soils 418
Soil-Forming Factors 418
Climate 418
Vegetation 420
Lithology and Relief....421
Time 421
Soil-Forming Processes....421
Mass Additions and Losses....421
Leaching and Redistribution of Calcite and Silica 421
Clay Translocation....422
Rubefaction 422
Development of Vertic Features....423
Major Soil Types 424
Physical, Chemical, and Mineralogical Properties....425
Physical Properties....426
Mineralogical Properties....426
Chemical Properties....427
Further Reading....427
Metal Oxides 428
Introduction....428
Fe Oxides 428
Goethite (alpha-FeOOH) 429
Hematite (alpha-Fe2O3) 431
Ferrihydrite (Fe5HO8 · 4H2O) 431
Lepidocrocite (gamma-FeOOH) 432
Magnetite (Fe3O4) and Maghemite (gamma-Fe2O3) 432
Rare Fe Oxide Minerals 432
Mn Oxides....433
Phyllomanganates (Layer Structures) 433
Tectomanganates (Tunnel Structures) 434
Chain Structures 434

Al Oxides 434
Gibbsite 435
Rare Al Oxide Minerals 435
Al Hydrolysis and Poorly Crystalline Al Hydroxides 436
Si Oxides 436
Quartz and Cristobalite....437
Opal....437
Ti Oxides....437
Further Reading....438
Metals and Metalloids, Transformation by Microorganisms....438
Introduction....438
Microbially Induced Metal Transformations 439
Mobilization....439
Acidification....439
Complexation by Ligands....440
Siderophores....440
Biomethylation....440
Redox Reactions....440
Immobilization....442
Biosorption 442
Precipitation of Metals and Minerals 443
Partitioning in the cellular environment....443
Oxalates....444
Carbonates 444
Redox Reactions....445
Bacterial and fungal oxidation....445
Dissimilatory metal-reduction by bacteria 445
Biotechnological Applications of Microbial Metal Transformations 445
Ore Leaching....445
Bioremediation 446
Summary 447
Acknowledgments....447
Further Reading....447
Microbial Processes 448
Environmental Factors....448
Introduction 448
Abiotic Factors 448
Nutrients 448
Water 450
Oxygen 451
Temperature 452
pH 452
Light....452
Biotic Factors 453
Symbiosis or Mutualism....453
Commensalism and Amensalism 453
Competition....453
Predation and Parasitism....453
Integrating Abiotic and Biotic Factors....455
Further Reading 455
Community Analysis 455
Introduction 455
Background 456
Nucleic Acids 456
Denaturation and Renaturation....456
Soil Nucleic Acid Extraction 456

Ribosomal RNA 457
Direct Community Analysis 457
G+C Content....457
DNA Reassociation....457
Nucleic Acid Hybridization 458
Microscopic Examination and Fluorescent In Situ Hybridization 459
Fatty Acid Analysis....460
Indirect Community Analysis....460
Polymerase Chain Reaction 460
Community Richness....460
PCR-Based Community Structure Analysis....461
Terminal restriction fragment length polymorphism....462
Gradient gel electrophoresis....462
Data Analysis 462
Further Reading 463
Kinetics 463
Introduction 463
The Main State Variables and Growth Parameters....464
Amount of Soil Microorganisms....464
Substrate Concentration....465
Products....466
The Models of Microbial Growth....466
Exponential Growth 466
Monod Model....467
Structured Models....467
Synthetic Chemostat Model....468
Growth Kinetics In Situ and Ex Situ....468
Diversity of Growth Characteristics: Life Strategy Concept....471
Decomposition Kinetics....474
Decomposition as Self-Decay....474
Decomposition of Simple Individual Compounds Added to Soil....476
Decomposition of Cellulose 476
Decomposition of Plant Litter....479
Further Reading 479
Micronutrients 479
Introduction....479
Zinc....479
Copper....480
Manganese....480
Iron....481
Boron 482
Molybdenum 483
Micronutrient Soil Tests 483
Micronutrient Fertilizers 484
List of Technical Nomenclature 485
Further Reading....486
Mineral-Organic-Microbial Interactions 486
Introduction....486
Mechanisms of Binding of Organic Matter by Soil Mineral Colloids 486
Nature of Mineral Colloid Surfaces....486
Binding of Organic Matter....487
Ion exchange 487
Protonation 487
Coordination and ion dipole....487
Hydrogen bonding 487
Hydrophobic bonding on clay-organic complexes 487

Anion adsorption....488
Adsorption on external mineral surfaces....488
Adsorption in clay interlayers....488
Influence of Organic Matter on the Formation and Surface Chemistry of Metal Oxides....488
Aluminum Oxides....488
Iron Oxides....489
Role of Soil Minerals in Abiotic Catalysis of the Formation of Humic Substances....490
Polyphenol Pathway....490
Maillard Reaction....491
Interactions of Soil Mineral Colloids with Microorganisms and Enzymes....491
Surface Interactions of Soil Minerals with Microorganisms....491
Formation of Mineral Colloid- and Humic-Enzyme Complexes....493
Soil Mineral Weathering Transformations as Influenced by Microbial Activity....493
Mineral Weathering....493
Fine-Grained Mineral Formation....494
Impacts of Mineral-Organic-Matter-Microbe-Interactions on Soils in the Environment....494
Soil Structural Stability....494
Biochemical Cycling of C, N, P, and S in Soil....496
Transformations of Organic Pollutants....496
Transformations of Metals....498
Further Reading....499
Minerals, Primary....500
Introduction....500
Quartz....500
Feldspars....501
Micas....505
Common Accessory Primary Minerals....506
Olivines....508
Pyroxenes and Amphiboles....509
Further Reading....510
Morphology....511
Introduction....511
Site Description....511
Horizons....511
Soil Texture....512
Soil Color....513
Redoximorphic Features and Mottles....514
Redoximorphic Features....514
Mottles....516
Soil Structure....516
Shape....516
Size....517
Grade....517
Weak....518
Moderate....518
Strong....518
Compound Structure....518
Soil Consistence....518
Roots and Pores....518
Chemical Response....518
pH....518
Effervescence....518

Reaction to Alpha,Alpha'-Dipyridyl 519
Special Features 519
Pedon Features 519
Horizon or Layer Features 519
Ped Surface Features 519
Location 519
Amount 519
Boundary 520
Soil Micromorphological Description 520
Further Reading 520
Mulches 521
Introduction 521
How Mulching Influences the Soil Environment 521
Soil Moisture Regime 521
Controling evaporation from the soil surface 521
Improving infiltration rate 522
Soil moisture retention 523
Water condensation at night 523
Controlling Runoff and Soil Erosion 523
Modifying Soil Thermal Regime 524
Soil Aeration 524
Improving Soil Structure 524
Improving the Soil Chemical Environment 525
Improving Biological Activity in Soil 526
Types of Mulches 526
Organic Mulches 526
Inorganic Mulches 527
Plastic mulch 527
Black plastic mulch 528
Transparent/clear plastic mulch 528
White plastic mulch 528
Colored plastic mulch 528
Infrared transmitting mulches (IRT) 529
Gravel mulch 529
Crushed rock 529
Other materials used as mulch 529
Methods of Mulch Application 529
Impact of Mulching on Crop Growth and Input-Use Efficiency 530
Modification of Crop Growth 530
Weed competition 530
Seedling emergence 530
Root growth 530
Improvement of Input-Use Efficiencies 530
Limitations of Mulching 531
Further Reading 532
Mycorrhizal Fungi 533
Introduction 533
Complexity 534
Defining the Mycorrhizal System and Participants 535
Governance of the System 536
Diversity and the Delivery of Ecosystem Services 536
Multiple Genes Coordinate the Development of Mycorrhizae 538
Genomic Hegemony over Nutrient and Carbon Transport 538
Functional Genomics 538
Games and Models for Mapping Causality 539
Hierarchy and Scaling 540

Autonomous Networks.....541
Summary 542
Further Reading.....542

FERTIGATION

U Kafkafi and S Kant, The Hebrew University of Jerusalem, Rehovot, Israel

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

'Fertigation' is a fusion of two words: 'fertilizer' and 'irrigation.' Fertigation is the process of applying mineral fertilizers to crops along with the irrigation water. This word was first used in the USA to describe the bubbling of anhydrous ammonia into irrigation water; later, other fertilizers were applied through a sprinkler system. The widest use of fertigation is in drip irrigation. The concept of irrigation with dissolved nutrients dates back to Roman times, when city sewage was used to irrigate crops. The use of jute bags containing ammonium sulfate ($(NH_4)_2SO_4$) fertilizer at the entrance of canals and flood irrigation for banana crops has been used by growers in the Jordan Valley since the early 1930s. Mixing of micronutrients such as zinc sulfate ($ZnSO_4$) with a flood or furrow irrigation system is still in use in many developing countries, where farmers place bags of the soluble nutrients at the inlet of the irrigation water.

Fertigation in various trickle-irrigation technologies involves the injection of fertilizer solutions into irrigation systems via calibrated injection pumps. The corrosive nature of fertilizers prevents the use of fertigation when aluminum or zinc-plated metal pipes are used for irrigation. The development of plastic systems for irrigation coupled with the development of trickle-irrigation systems has enabled the refinement of fertilizer application through the irrigation lines. The fertilizer industry has adapted to field demands by introducing pure and soluble fertilizers. The time-consuming process of the dissolution of various fertilizers has led to the development of liquid fertilizer blends, according to specific recipes as ordered by growers to meet the demands of specific plants at particular growing stages and under certain climatic conditions. The use of conventional, broadcast application of fertilizers becomes ineffective with a drip-irrigation system. An adequate supply to satisfy plant demands for nutrients from a limited soil root volume can only be achieved by synchronizing the supply of water and nutrients during the various growing stages of the plants. The units used to plan fertigation are: milligrams of a nutrient consumed per day per plant, rather than kilograms of a nutrient per hectare. The daily application rate of fertigation changes during the growing season and is planned to meet the plant's daily demands according to its nutrient-uptake strategy.

Advantages of Fertigation

Fertigation has some specific advantages over broadcast and band fertilization: (1) a frequent supply of nutrients reduces fluctuation of nutrient concentration in soil; (2) there is efficient utilization and precise application of nutrients according to the nutritional requirements of the crop; (3) fertilizers are applied throughout the irrigated soil volume; (4) nutrients can be applied to the soil when soil or crop conditions would otherwise prohibit entry into the field with conventional equipment. Drip fertigation has added advantages over other methods of fertigation: (1) application of nutrients only to the wetted soil volume where roots are active reduces loss of nutrients by leaching or soil fixation and increases fertilizer-use efficiency; (2) the crop foliage remains dry, thus retarding incidence of pests or diseases and escaping foliage burn; (3) the wind has no effect and runoff is avoided.

Drip-Fertigation System

A drip-fertigation system is designed specifically for particular crop, climatic, and soil conditions.

Containers for Fertilizers

There are two main methods of delivering fertilizer solution to the irrigation lines:

1. Fertilizer dilution tanks: these are connected at the head of the irrigation line. The difference in water head between inlet and outlet ports drives some of the irrigation water through the fertilizer container and transfers it to the irrigation line;
2. External pump: the fertilizer solution is injected by positive pressure directly into the irrigation line. There are several types of pump: some use the energy of the irrigation water (the pressure and flow in the main water line is used to move a piston); others use water flow through a Venturi-type pump to suck the fertilizer solution into the irrigation line. For small areas and home gardening, accurate, electric-driven pumps are available.

The dilution tank delivers a known amount of fertilizer during the irrigation cycle until all the fertilizer is washed from the tank. In this method the concentration of the nutrients in the trickling water is varied with time, but a known total amount of

nutrient per unit area is delivered. With the injection pump, it is possible to control the concentration of the nutrients in the irrigation solution.

Filtration

Filtration is a prerequisite in drip irrigation to avoid clogging of drip lines and emitters, and to maintain a predicted level of uniformity of water and fertilizer application. The main filtration system depends on the quality and source of water supply. It is used to remove gravel or sand from deep well water sources. Open surface water (ponds, rivers, or lakes) supplies may contain organic matter and algae that must be removed before entry to the lines. In a fertigation system, the second filtration step (screen filter) after the fertilizer container is necessary to remove any particulate or precipitate that comes from fertilizer mixtures. Deep water sources may contain divalent soluble iron (Fe). This iron, when it comes in contact with phosphate solution, produces a gel-like precipitate that can block the tricklers and filters. The water quality and composition must be taken into consideration at the planning stage of fertigation systems.

Distribution of Water and Nutrients in Soil

Water Flow

Two main forces, gravity and capillary, govern the movement of water in the soil during the irrigation process. During evaporation the net movement of water is in an upward vertical direction. The classic infiltration one-dimensional flow equation is:

$$\frac{\partial w}{\partial t} = \frac{\partial}{\partial z}\left[K(w)\frac{\partial H}{\partial z}\right] \qquad [1]$$

Change in water content (∂w) with time (∂t) at a position in the soil must be equal to change in specific flux in the vertical direction (z). According to the Darcy law, specific flux $= -K(w)\,\partial H/\partial z$, where K is the hydraulic conductivity, which is a function of water content (w), and $\partial H/\partial z$ is the hydraulic head gradient. Hydraulic head (H) is the sum of gravitational head (z) and matric head (h), i.e., $H = z + h$.

When water is applied from a point source by a drip system, water movement is not one-dimensional but rather is spread in vertical and radial (r) dimensions. The equation of water flow from a point source takes the form:

$$\frac{\partial w}{\partial t} = \frac{\partial}{\partial r}\left[K(w)\frac{\partial h}{\partial r}\right] + \frac{K(w)}{r}\times\frac{\partial h}{\partial r} + \frac{\partial}{\partial z}\left[K(w)\frac{\partial H}{\partial z}\right] \qquad [2]$$

In drip irrigation, water spreads from a trickler in two dimensions and creates a wetted front of different shapes (Figure 1) depending on soil type and water-discharge rate. When emitter discharge rate is higher than the soil-infiltration rate, lateral water movement dominates and a shallower penetration of water is expected when a constant amount of water is applied.

Nutrient Movement

Mass flow and diffusion govern the movement of nutrients in soil solution. Mass flow is the movement of nutrient ions with the moving water. Diffusion occurs because of concentration gradient. Two-dimensional (x and z) movement of reactive nutrient ions in soil is given by the following equation:

$$[b(c)+w]\frac{\partial c}{\partial t} = \frac{\partial}{\partial x}\left[D_{h_{xx}}\frac{\partial c}{\partial x} + D_{h_{xz}}\frac{\partial c}{\partial z}\right] + \frac{\partial}{\partial z}\left[D_{h_{xz}}\frac{\partial c}{\partial x} + D_{h_{zz}}\frac{\partial c}{\partial z}\right] - \left[\frac{\partial (q_x c)}{\partial x} + \frac{\partial (q_z c)}{\partial z}\right] \qquad [3]$$

where c is nutrient concentration in the soil solution, b is buffering capacity of the soil, q is flux of the soil solution according to the Darcy law, and D_h is the coefficient of diffusion. For nutrient application with a drip system, radial (r) movement of nutrients with water is added, so the nutrient flow will follow:

$$[b(c)+w]\frac{\partial c}{\partial t} = \frac{\partial}{\partial r}\left[D_{h_{rr}}\frac{\partial c}{\partial r} + D_{h_{rz}}\frac{\partial c}{\partial z}\right] + \frac{1}{r}\left[D_{h_{rr}}\frac{\partial c}{\partial r} + D_{h_{rz}}\frac{\partial c}{\partial z}\right] + \frac{\partial}{\partial z}\left[D_{h_{zz}}\frac{\partial c}{\partial z} + D_{h_{rz}}\frac{\partial c}{\partial r}\right] - \left[\frac{\partial (q_r c)}{\partial r} + \frac{\partial (q_z c)}{\partial z} + \frac{q_r c}{r}\right] \qquad [4]$$

Equation [4] suggests that strongly adsorbed ions (such as phosphate; $b(c) > 1$) have lower mobility in soil than nonsorbed ions (such as nitrate or chloride). For a given nutrient concentration, $b(c)$ of fine-textured soils exceeds that of coarse-textured soils, therefore mobility of sorbed ions in clay soils is less than in sandy soils.

Salt mass movement in two different soils and two discharge rates by drip irrigation is shown in Figure 2. Both soils have the same initial salt concentration and the same salt concentration at the soil inlet. The movement of nonsorbed solute in the vertical component of the wetted zone is deeper and in the radial direction it is narrower as the soil become coarser and the emitter discharge rate becomes lower.

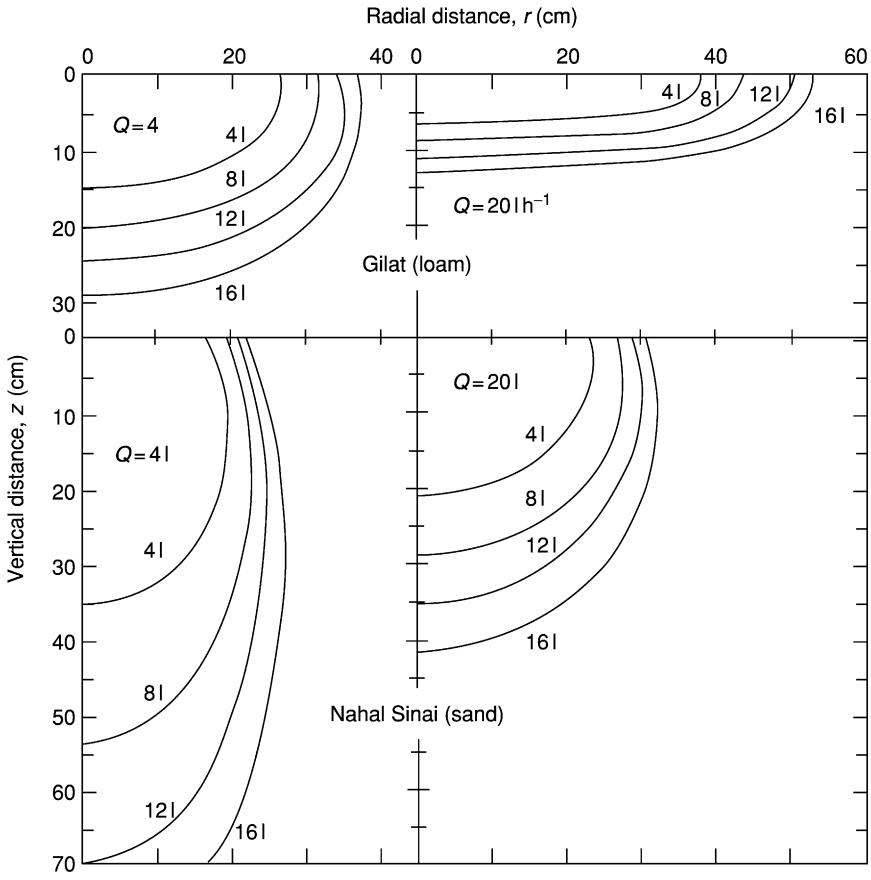


Figure 1 Water content distribution as a function of infiltration time and cumulative infiltration water (in liters) for two soils and two discharge rates (Q). (Adapted from Bresler E (1977) Trickle-drip irrigation: principles and application to soil-water management. *Advances in Agronomy* 29: 343–393, with permission.)

The concentration and movement of nonsorbed nitrate ions, added with water, follow the distribution pattern shown in Figure 2.

In repeated cycles of fertigation, there is a balance between the distance of lateral water flow and evaporation. As a result, at the border between the dry and the wet zone, a circle of soluble salts tends to accumulate, especially in hot dry areas with no intermittent rains. The salt accumulated in the wet zone periphery can reach very high levels and, in the case of a single flush of rain, this high-salt soil can be washed into the root zone and cause considerable damage. To avoid this soluble-salt accumulation on the soil surface due to evaporation, irrigation under plastic is used, especially for a permanent utilization of saline water for irrigation of vegetables and orchards. In arid climates where the evaporation rate is high, mobile nutrient ions (NO_3^-, Cl^-, Ca^{2+}, and K^+) tend to accumulate at the wet zone periphery in the soil surface. This higher concentration of soluble salts is detrimental for young seedlings, because their root systems might face high salt concentrations even with good-quality water.

Selection of Fertilizers

Most of the water-soluble and liquid fertilizers are suitable for fertigation. When choosing a fertilizer, three main considerations have to be taken into account: (1) plant type and stage of growth; (2) type of soil and irrigation system; and (3) quality of available water. Consideration has to be given to the sensitivity of a plant to the forms of N during various physiological stages. Some plants such as tomato are very sensitive to the concentration of ammonium near the root, so a nitrate-rich nutrient solution is selected for this type of plant.

At high root temperatures, ammonium might damage the roots by placing a burden on their root-soluble sugar content. In aerated soils at high temperatures, nitrate sources outweigh ammonium sources. On heavy clay soils, a zone of ponded water under the trickle point might develop where, at high soil temperatures, local anaerobic conditions can cause severe nitrate-N (NO_3^--N) losses to the atmosphere as N_2 or N_2O gases. In such conditions the plants can show N deficiency despite sufficient nitrate supply

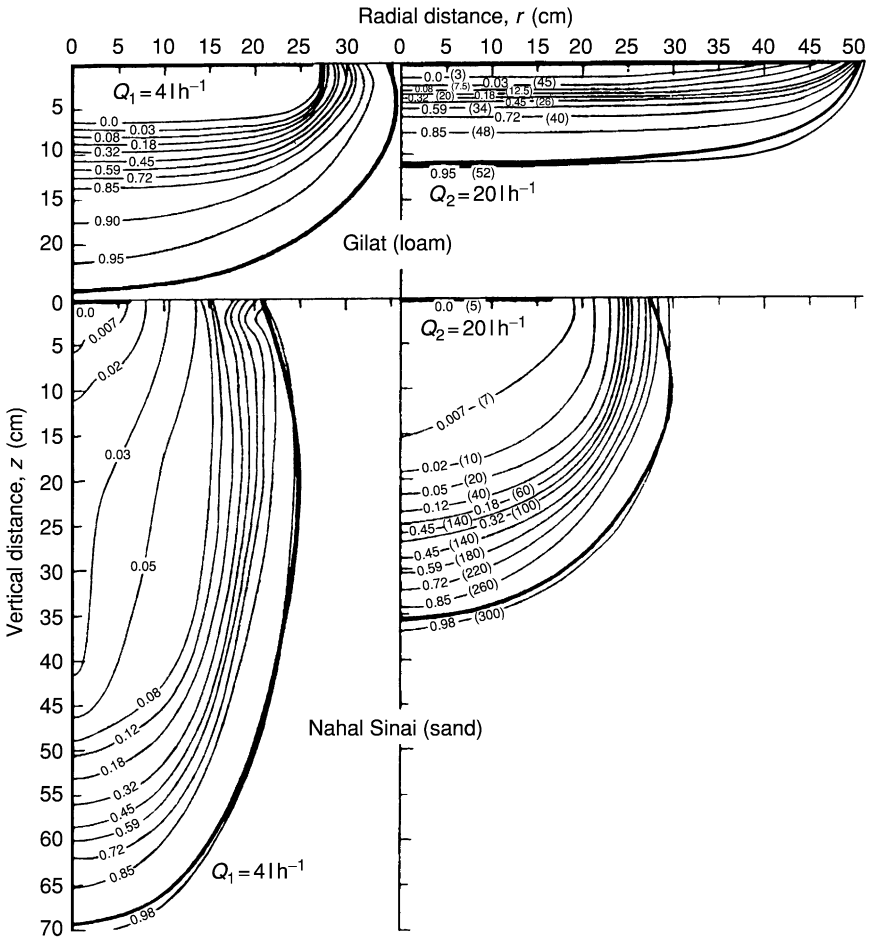


Figure 2 Distribution pattern of two-dimensional salt concentration during infiltration for two trickler discharges (Q_1 and Q_2) in loam and sandy soils. The numbers labeling the curves indicate relative salt concentration in terms of $(C - C_o)/C_n$. The numbers in parentheses are salt concentration (C) in soil solution (millimoles per liter); C_o is inlet salt concentration and C_n is initial salt concentration in soil. (Adapted from Bresler E (1977) Trickle-drip irrigation: principles and application to soil-water management. *Advances in Agronomy* 29: 343–393, with permission.)

through the irrigation line. Low concentrations of N as urea or ammonium in the irrigation solution might prevent loss and deficiency of N in plants.

Lowering the pH of the irrigation water to approximately 5.0 is often needed to keep phosphorus (P) in solution during the fertilizer injection. P application as phosphoric acid is preferable during the cold season. If micronutrients are needed, their soluble forms as chelates are less susceptible to precipitation in the irrigation lines and in the soil.

Solubility of Fertilizers

Solid fertilizers vary in their dissolution rate and in the amount that can be dissolved in a given volume of water. Solubility of fertilizers is generally reduced when two or more fertilizers are mixed together. The solubility of some commonly used fertilizers is presented in Table 1. This characteristic is crucial for the suitability of a fertilizer to be used in fertigation. Solubility of fertilizers generally increases with temperature. When present as a single salt, the main factor affecting solubility of potassium (K) and nitrate fertilizers is temperature. Whether the fertilizer dissolution is endothermic (the solution temperature is reduced during dissolution) or exothermic affects the dissolution rates and energy needed to prepare fertilizer solution. Most solid N fertilizers absorb heat when mixed with water. This results in cooling the solution and slows down the preparation of the liquid solution in the mixing tank.

The solubility of calcium nitrate ($Ca(NO_3)_2$) is higher at 30°C (Table 1); however, in practice its endothermic reaction is strong and only 50% of the amount expected to be dissolved at 30 °C is dissolved in 5 min, which results in a temperature drop to 2 °C. When the mixing tank is placed in an open field, this characteristic could reduce the expected concentration in the irrigation line, because some of the fertilizer

Table 1 Solubility, nutrient content, and soil reaction of some common fertilizers used in fertigation

Fertilizer	*Chemical formulae*	*Grade*[a] *N–P–K*	*Other nutrients*	*Soil reaction*	*Solubility (g per 100 ml water)*			
					0°C	*10°C*	*20°C*	*30°C*
Urea	$CO(NH_2)_2$	45-0-0	–	Acidic	–	78	–	–
Ammonium nitrate	NH_4NO_3	34-0-0	–	Acidic	118	152	192	242
Ammonium sulfate	$(NH_4)_2SO_4$	20-0-0	24% S	Acidic	71	73	75	78
Calcium nitrate	$Ca(NO_3)_2 \cdot 4H_2O$	15-0-0	17% Ca	Basic	102	112	121	150
Monoammonium phosphate	$NH_4H_2PO_4$	11-48-0	1.4% Ca 2.6% S	Acidic	23	29	37	46
Diammonium phosphate	$(NH_4)_2HPO_4$	21-53-0	–	Acidic	43	–	–	–
Phosphoric acid	H_3PO_4	0-52-0	–	Very acidic	548	–	–	–
Muriate of potash	KCl	0-0-62	–	Neutral	28	31	34	37
Potassium nitrate	KNO_3	13-0-44	–	Basic	13	21	31	45
Potassium sulfate	K_2SO_4	0-0-53	8% S	Neutral	7	9	11	13

[a]Concentrations of N, P_2O_5, and K_2O, expressed as a percentage of weight of the fertilizer.

will remain as a solid in the tank. The diverse solubility characteristics of fertilizers and the problems they cause to the end user were the trigger to the establishment of dilution plants that supply to the farmers a nutrient cocktail to order, to meet specific plant needs. Another approach to the solubility problem of a mixed nutrient solution has been adopted in the modern glasshouse industry, where different tanks of separate nutrient sources and acid and bases are used. With the help of a computer, it is possible to program the amounts of each nutrient element to be pumped into the irrigation line in small doses. This technology is too expensive to operate in large, open-field and plantation crops, where the Venturi or proportional pumps are used to inject the dose of fertilizer solution from a storage tank into the irrigation line.

Urea, ammonium nitrate, calcium nitrate, potassium nitrate, and ammonium phosphate are easily soluble in water and are used extensively in the preparation of single-nutrient or multinutrient fertilizer solutions. Monoammonium phosphate, phosphoric acid, and urea phosphate are water-soluble, but they may have precipitation problems when injected at high rates into hard water (containing large amounts of calcium and magnesium carbonates). All K fertilizers are water-soluble, but they vary in their rates of dissolution and sensitivity to the solution temperature. A common field problem in tanks of mixed fertilizer solution is precipitation during cold nights.

Compatibility of Fertilizers

Interaction between fertilizers can affect their availability and uptake by plants. Mixing of two fertilizers can sometimes result in precipitation. Injection of Ca salts with phosphate or sulfate may increase the chances of precipitate formation as calcium phosphate or calcium sulfate ($CaSO_4$). Fertilizer compatibility is presented in Figure 3. The pH of the irrigation solution should be within the range 5.5–7.0. A pH that is too high will reduce the availability of P, zinc, and iron and may result in precipitation of Ca and Mg phosphates or carbonates in the irrigation lines. A pH value that is too low is detrimental to roots and may increase Al and Mn concentrations in the soil solution. Nitric (HNO_3) or phosphoric (H_3PO_4) acids are used to reduce the pH level in fertigation. The advantage of using these acids is that, beside dissolution of basic precipitates in the line, they also supply the plants with the essential nutrients nitrate and phosphate, replacing other fertilizers containing N and P. In saline water and calcareous clay soils, nitric acid minimizes the increased salinity in the root zone and supplies only nitrate, which counterbalances excess chloride.

Precipitation in the Irrigation Lines

Precipitation of insoluble dicalcium phosphate and dimagnesium phosphate and calcium carbonate can develop with high-pH water. Iron phosphate precipitate can form in drip lines with low-pH water. Water containing high levels of Mg ions may cause an ammonium magnesium phosphate precipitation in the fertilizer tank. Avoiding the use of ammonium fertilizer in such conditions can prevent the risk of blocking the emitters. Using potassium sulfate (K_2SO_4) or ammonium sulfate ($(NH_4)_2SO_4$) with water containing high concentrations of Ca can result in precipitation as $CaSO_4$ (gypsum precipitates), which can clog the drip lines.

Scheduling Fertigation

The nutrient elements are taken up by plants depending upon the plants' requirements at a specific developmental stage. Injecting fertilizer into a

	Urea	AN	AS	MAP	CAN	CN	MKP	DAP	PA	PN	MOP
AN	C										
AS	LC	C									
MAP	C	LC	LC								
CAN	LC	LC	NC	NC							
CN	LC	C	NC	NC	LC						
MKP	C	C	C	C	NC	NC					
DAP	C	NC	LC	C	NC	NC	C				
PA	LC	LC	C	LC	NC	NC	C	LC			
PN	C	C	LC	C	C	C	C	C	LC		
MOP	C	LC	C	C	NC	NC	LC	C	NC	LC	
PS	C	C	C	C	LC	LC	C	C	LC	C	C

Figure 3 Compatibility of common fertilizers. C, compatible; LC, limited compatibility; NC, not compatible; AN, ammonium nitrate; AS, ammonium sulfate; MAP, monoammonium phosphate; CAN, calcium ammonium nitrate; CN, calcium nitrate; MKP, monopotassium phosphate; DAP, diammonium phosphate; PA, phosphoric acid; PN, potassium nitrate; MOP, muriate of potash (KCl).

drip-irrigation system offers the benefits of nutrient application to the crop in optimum amounts and at times when it is most likely to be taken up by the plant. Fertigation provides optimum concentrations and quantities of nutrients directly into the root zone. Hence, accurate planning of when and how much fertilizer to apply is critical for fertigation management. Nutrient requirements by plants change with their physiological growth stages. The climatic factors, timing of fertilizer application, and irrigation timings affect crop growth and consequently affect the quantity of nutrients required. A fertigation schedule depends upon soil characteristics and fertilizer requirements of the growing plants. Nutrient (N, P, and K) uptake by some field and vegetable crops with respect to growth is given in Tables 2–4. Soil type, system design, and length of growing season govern the frequency of fertilizer application. The frequency of injecting fertilizers varies with soil type and specific plant requirements. In plants grown on sand dunes, several irrigations per week might be needed, while on heavy clay soils only one or two irrigations per week are needed. The smaller the root volume, the higher is the fertigation frequency.

Behavior of Nutrients in Soil

Soil chemical properties are important in planning fertigation. pH has a great effect on the availability of residual nutrients in soil as well as on those added via fertigation. The balance between cation and anion uptake by the plant affects the pH in the rhizosphere. When a plant takes up more nutrient cations, such as when NH_4^+ is the main N source, protons are exuded by the roots and acidify the soil surrounding the roots. If the anion uptake is predominant, as when NO_3^- is the main source of N, the roots exude OH^- or HCO_3^-, resulting in a pH rise in the rhizosphere. Nitrate and ammonium are the main forms of N available for plant uptake. Ammonium nutrition leads to excretion of 0.9 mol of H^+ by roots with 1 mol of ammonium assimilation; while, where there is nitrate reduction in the leaves, 0.1–0.3 mol of OH^- is released from the root per mole of NO_3^- taken up. The rhizosphere pH can change with the form and concentration of the N fertilizer, but the extent of the pH change around the root depends on the buffer capacity of the soil.

Cation exchange capacity (CEC) of the soil is an important consideration in determining the amount and frequency of cations that need to be added during fertigation. In most agricultural soils and irrigation water, Ca and Mg are present in larger quantities than those taken up by any crop, and their supply is usually satisfied with water mass flow. K^+ is the main cation that must be supplied with the irrigation water. In order to maintain an acceptable concentration of K in the soil solution, a soil with low CEC must

Table 2 Nitrogen uptake by various field and vegetable crops with respect to relative time of growth

Crop	Relative time of growth (%)					Total uptake ($g\,plant^{-1}$)	Plants ha^{-1}	Projected yield ($t\,ha^{-1}$)
	0–20	20–40	40–60	60–80	80–100			
Cotton								
g $plant^{-1}$	0.20	1.80	3.80	2.20	1.60	9.60	25 000	1.3[b]
mg N $plant^{-1}day^{-1}$[a]	6	58	123	71	52	62		
Maize								
g $plant^{-1}$	0.25	1.58	1.00	0.83	0.50	4.17	60 000	8
mg N $plant^{-1}day^{-1}$[a]	11	70	44	37	22	37		
Tomato								
g $plant^{-1}$	0.50	0.75	2.50	4.25	3.25	11.25	20 000	100
mg N $plant^{-1}day^{-1}$[a]	19	28	91	156	119	83		
Sweet pepper								
g $plant^{-1}$	0.40	1.80	1.10	0.70	0.60	4.60	50 000	55
mg N $plant^{-1}day^{-1}$[a]	20	90	55	35	30	46		
Potato								
g $plant^{-1}$	0.08	1.00	1.08	0.50	0.17	2.83	60 000	50
mg N $plant^{-1}day^{-1}$[a]	4	50	54	25	9	28		
Muskmelon								
g $plant^{-1}$	0.20	0.60	1.60	2.80	0.80	6.00	25 000	50
mg N $plant^{-1}day^{-1}$[a]	10	30	80	140	40	60		
Watermelon								
g $plant^{-1}$	0.83	1.67	3.33	6.67	2.50	15.00	12 000	75
mg N $plant^{-1}day^{-1}$[a]	41	84	166	333	125	150		
Cabbage								
g $plant^{-1}$	0.10	0.20	0.80	1.90	0.60	3.60	50 000	29
mg N $plant^{-1}day^{-1}$[a]	8	16	63	150	47	56		
Cauliflower								
g $plant^{-1}$	0.10	0.20	0.50	2.00	1.40	4.20	50 000	9
mg N $plant^{-1}day^{-1}$[a]	8	16	40	157	110	66		
Eggplant								
g $plant^{-1}$	0.50	3.25	2.00	2.50	1.50	9.75	20 000	40
mg N $plant^{-1}day^{-1}$[a]	14	89	55	69	41	54		

[a]Daily amounts of nitrogen (mg N $plant^{-1}$ day^{-1}) added with drip fertigation; this amount includes 10% additional nitrogen for root consumption.
[b]Seed cotton (lint) yield.

get a fresh supply of K more frequently than in soil with high CEC, which can hold higher quantities of K. Fertigation is most practical in sandy soils and soils of dry and arid regions that have a low CEC, because these soils need frequent irrigation and quick nutrient replenishment. Old farming practices considered sand dunes as nonagricultural soils, but the introduction of fertigation has turned desert sand dunes into productive agricultural soils.

Nitrate-N is highly mobile and is more likely to be lost through surface runoff and leaching. In trickle irrigation, ponding under the trickler creates an oxygen-devoid space in which denitrification of applied nitrate is observed during the irrigation cycle. The rate of water discharge from an emitter is chosen to meet the rate of water entry to the soil from a point source. Volatilization of ammonia during urea application might also occur if the irrigation solution is basic. Acidifying the irrigation water prevents ammonia losses.

Added phosphate is rapidly adsorbed and precipitated in the soil, leading to a rapid decline in water-soluble phosphate concentration in the soil solution. Its movement is retarded owing to retention by soil oxides and clay minerals. P application with drip irrigation is more efficient than with sprinkler irrigation or broadcasting. P in fertigation is supplied directly into the zone where active roots are present, allowing its immediate uptake by active roots before drying and irreversible fixation occur in the soil.

Root Growth

To achieve optimum plant growth, the root zone must be well supplied with water, nutrients, and oxygen. Water potential kept by frequent irrigation at a consistently low water tension, especially in clay soil, can result in a suboptimal supply of oxygen in the root zone. Roots respond within minutes to a reduction in oxygen supply by cessation of root extension. The elongation zone of a cotton root dies after only 30 min without oxygen. In drip irrigation the soil is saturated only in the vicinity of the trickler (**Figure 1**), while most of the soil volume is wetted by an

Table 3 Phosphorus uptake by various field and vegetable crops with respect to relative time of growth

Crop	Relative time of growth (%)					Total uptake ($g\,plant^{-1}$)	Plants ha^{-1}	Projected yield ($t\,ha^{-1}$)
	0–20	20–40	40–60	60–80	80–100			
Cotton								
$g\,plant^{-1}$	0.17	0.24	0.80	0.44	0.17	1.80	25 000	1.3[b]
mg $P_2O_5\,plant^{-1}\,day^{-1}$[a]	5.2	7.7	25.8	14.2	5.2	11.6		
Maize								
$g\,plant^{-1}$	0.07	0.30	0.28	0.25	0.10	1.00	60 000	8
mg $P_2O_5\,plant^{-1}\,day^{-1}$[a]	2.9	13.2	12.1	11.0	4.4	8.8		
Tomato								
$g\,plant^{-1}$	0.03	0.05	0.17	0.45	0.25	0.95	20 000	100
mg $P_2O_5\,plant^{-1}\,day^{-1}$[a]	1.1	1.8	6.2	16.5	9.0	7.0		
Sweet pepper								
$g\,plant^{-1}$	0.03	0.10	0.20	0.08	0.04	0.45	50 000	55
mg $P_2O_5\,plant^{-1}\,day^{-1}$[a]	1.5	5.0	10.0	4.0	2.0	4.5		
Potato								
$g\,plant^{-1}$	0.01	0.05	0.10	0.14	0.09	0.39	60 000	50
mg $P_2O_5\,plant^{-1}\,day^{-1}$[a]	0.5	2.5	5.0	7.0	4.5	3.9		
Muskmelon								
$g\,plant^{-1}$	0.02	0.08	0.20	0.32	0.20	0.82	25 000	50
mg $P_2O_5\,plant^{-1}\,day^{-1}$[a]	1.1	4.0	10.0	16.0	10.0	8.2		
Eggplant								
$g\,plant^{-1}$	0.03	0.12	0.18	0.42	0.35	1.10	20 000	40
mg $P_2O_5\,plant^{-1}\,day^{-1}$[a]	0.8	3.3	5.0	11.5	9.6	6.0		

[a]Daily amounts of phosphorus (mg $P_2O_5\,plant^{-1}\,day^{-1}$) to be added with drip fertigation; this amount includes 10% additional phosphorus for root consumption.
[b]Seed cotton (lint) yield.

Table 4 Potassium uptake by various field and vegetable crops with respect to relative time of growth

Crop	Relative time of growth (%)					Total uptake ($g\,plant^{-1}$)	Plants ha^{-1}	Projected yield ($t\,ha^{-1}$)
	0–20	20–40	40–60	60–80	80–100			
Cotton								
$g\,plant^{-1}$	0.60	2.00	3.60	0.60	0.20	7.00	25 000	1.3[b]
mg $K_2O\,plant^{-1}\,day^{-1}$[a]	20	65	117	20	7	45		
Maize								
$g\,plant^{-1}$	0.25	1.83	1.00	0.33	0.08	3.50	60 000	8
mg $K_2O\,plant^{-1}\,day^{-1}$[a]	11	80	44	14	4	31		
Sugarcane								
$g\,plant^{-1}$	0.50	0.60	0.70	1.80	0.60	4.20	50 000	140
mg $K_2O\,plant^{-1}\,day^{-1}$[a]	11	13	15	40	13	19		
Tomato								
$g\,plant^{-1}$	0.70	0.80	3.50	7.00	4.50	16.50	20 000	100
mg $K_2O\,plant^{-1}\,day^{-1}$[a]	25	30	128	256	165	121		
Sweet pepper								
$g\,plant^{-1}$	0.50	2.00	1.40	1.40	0.40	5.70	50 000	55
mg $K_2O\,plant^{-1}\,day^{-1}$[a]	25	100	70	70	20	57		
Potato								
$g\,plant^{-1}$	0.20	0.80	1.80	1.50	0.40	4.70	60 000	50
mg $K_2O\,plant^{-1}\,day^{-1}$[a]	10	40	90	75	20	47		
Muskmelon								
$g\,plant^{-1}$	0.40	1.20	4.00	4.40	2.00	12.00	25 000	50
mg $K_2O\,plant^{-1}\,day^{-1}$[a]	20	60	190	220	100	120		
Eggplant								
$g\,plant^{-1}$	0.75	5.00	3.00	1.75	1.00	11.50	20 000	40
mg $K_2O\,plant^{-1}\,day^{-1}$[a]	21	138	82	48	28	64		

[a]Daily amounts of potassium (mg $K_2O\,plant^{-1}\,day^{-1}$) to be added with drip fertigation; this amount includes 10% additional potassium for root consumption.
[b]Seed cotton (lint) yield.

unsaturated flow. This situation arises when tricklers of high discharge rates (more than $1\,l\,h^{-1}$) are used in clay soils. Oxygen is excluded from the saturation zone when there is a continuous supply of water at high flux. A slow flow rate can maintain the moisture and oxygen regime in the wet soil volume at optimum conditions.

The nitrate-to-ammonium ratio has a significant influence on the development of the root system. A higher concentration of ammonium has deleterious effects on root growth specifically when root temperature is high. At a high root temperature, additional sugar is required simultaneously in large amounts for respiration and ammonium metabolism in the root. The supply of sugar from the leaves lags behind the respiration and metabolism consumption, resulting in momentary excess of ammonia in the root that causes root death. Under the same high-temperature conditions, nitrate reduction occurs mainly in leaves, where the sugar supply is near the site of N metabolism. At high soil temperatures, especially in soilless cultures, nitrate-N is a safe form of N during fertigation. At low root temperatures, a high level of NO_3^- accumulates in the roots and low amounts are transported to the shoot, resulting in N deficiency. Therefore, the proper concentration and form of N applied in fertigation should be chosen differently for winter and summer seasons for intensively grown crops.

Subsurface Fertigation

Installation and disassembling of laterals in drip fertigation every season require extra labor and energy. Damage to drip lines by movement of heavy machinery and animals restricts the use of surface drip-irrigation, although placement of trickle lines at appropriate soil depths can solve such problems. Recent improvements in liquid, soluble, and compatible fertilizers, filtration devices, and reduced clogging problems have opened the way for development of subsurface fertigation. Special slow-release chemicals imbedded in the plastic of the filters prevent root entry into the trickler. In addition to cost-effectiveness and energy saving, subsurface drip-fertigation has added agronomic advantages over surface drip-fertigation: (1) placement of nutrients in the region where root activity is maximal and the daily and seasonal temperature fluctuations are minimal. Distribution of nutrients in subsurface applications is more uniform and spherical around emitters than surface fertigation, where movement is hemispheric below the point source and evaporation causes accumulation of soluble salts at the wet zone periphery; (2) roots can explore a greater soil volume for water and nutrients, since roots grow deeper in a subsurface system; (3) the top 4–5 cm of soil remains dry in subsurface fertigation, thereby reducing evaporation losses and weed germination.

Fertigation gives the plant grower a tool to control the daily supply of water and nutrients in tandem with plant growth rate and nutrient demand. When properly employed, plant potential growth can be achieved that is only limited by radiation and temperature. To reach maximum plant performance, the accurate design of fertigation system, efficient utilization of equipment, combined with the timely application and selection of compatible fertilizers, are essential criteria. More detailed research on surface and subsurface fertigation is needed for different commercial crops. Fertigation in protected agriculture and potted plants must take into consideration the root zone temperature when choosing N fertilizer. Further study on the availability and movement of dissolved nutrients in different soil types to meet the nutrient requirements of important crops as a function of time is needed for open-field-grown crops.

See also: **Fertilizers and Fertilization**; **Irrigation: Methods**; **Nutrient Availability**; **Nutrient Management**

Further Reading

Bar Yosef B (1999) Advances in fertigation. *Advances in Agronomy* 65: 1–77.

Bresler E (1977) Trickle-drip irrigation: principles and application to soil-water management. *Advances in Agronomy* 29: 343–393.

Burt C, O'Connor K, and Ruehr T (1998) *Fertigation*. CA: Irrigation Training and Research Center.

Dasberg S and Bresler E (1985) *Drip Irrigation Manual*. Bet Dagan, Israel: International Irrigation Information Center.

Hanan JJ (1998) *Greenhouses*. New York: CRC Press.

Haynes RJ (1985) Principles of fertilizer use for trickle irrigated crops. *Fertilizer Research* 6: 235–255.

Hochmuth GJ (1995) *Fertigation of Vegetable Crops in Florida, USA*, pp. 199–214 Dahalia Greidinger International Symposium on Fertigation, 26 Mar–1 Apr, 1995. Haifa, Israel: Technion-Israel Institute of Technology.

Kafkafi U (1994) Combined irrigation and fertilization (fertigation) in arid zones and protected agriculture. *Israel Journal of Plant Science* 42: 301–320.

Mmolawa K and Or D (2000) Root zone solute dynamics under drip irrigation: a review. *Plant and Soil* 222: 163–190.

Scaife A and Bar-Yosef B (1995) *Fertilizing for High-Yield and -Quality Vegetables*. IPI Bulletin 13. Basel, Switzerland: International Potash Institute.

Sneh M (1995) *The History of Fertigation in Israel*, pp. 1–10. Dahalia Greidinger International Symposium on Fertigation, 26 Mar–1 Apr, 1995. Haifa, Israel: Technion-Israel Institute of Technology.

FERTILITY

J L Havlin, North Carolina State University, Raleigh, NC, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Fertility of the soil is characterized by physical, chemical, and biological processes that control plant nutrient availability. Knowledge of these processes and interactions is critical to optimizing nutrient supply and plant productivity. Plants absorb nutrients dissolved in the soil solution. Nutrients in plant residue left in the field after harvest are ultimately returned to the soil and are subject to the same reactions as fertilizer nutrients. Although this cycle varies among nutrients, understanding nutrient dynamics in the soil–plant–atmosphere system is essential to successful nutrient management.

The Importance of Soil Fertility

World population will reach 9.3 billion people by 2050. Since additional arable land is only 15–20% more than current cultivated land, continued increases in productivity are essential to meet future food and fiber needs. While crop productivity has increased greatly since 1950, further advances in agricultural technologies must occur to enhance crop yield. Principal factors contributing to higher yields include development of improved cultivars, nutrient (soil fertility) and pest management, soil and water conservation, and cultural practices. Development and use of fertilizer and pesticides are directly related to increased crop productivity. Without effective and efficient management of soil fertility, 30–40% more land area would be needed today to produce the same yields. Therefore, growers must take advantage of technologies that increase productivity.

Maximum crop production depends on the growing-season environment and the ability of the producer to minimize all yield-limiting factors through appropriate management technologies. A well-established principle called the Law of the Minimum requires that the most yield-limiting factor be corrected first, the second most limiting factor next, and so on. For example, the correct variety planted at the optimum time and population may not result in the highest yield if plant-available water or nutrient availability were the most limiting factors. Thus, minimizing water as a yield-limiting factor must be accomplished before the crop can fully respond to management of any other factor.

Essential Plant Nutrients

An essential plant nutrient is involved in metabolic functions such that the plant cannot complete its life cycle without the nutrient. Seventeen nutrients are essential to plant growth (Table 1). Photosynthesis converts CO_2 and H_2O into carbohydrates that form proteins and other compounds. Macronutrients

Table 1 Essential plant nutrients and their relative and average plant concentrations

	Plant concentration[a]		
Element or nutrient	*Relative*	*Average*	*Classification*
Hydrogen H	60 000 000	6.0%	
Oxygen O	30 000 000	45.0%	
Carbon C	30 000 000	45.0%	
Nitrogen N	1 000 000	1.5%	
Potassium K	400 000	1.0%	Primary Macronutrients
Phosphorus P	30 000	0.2%	
Calcium Ca	200 000	0.5%	
Magnesium Mg	100 000	0.2%	
Sulfur S	30 000	0.1%	
Chloride Cl	3000	100 ppm (0.01%)	
Iron Fe	2000	100 ppm	
Boron B	2000	20 ppm	
Manganese Mn	1000	50 ppm	Micronutrients
Zinc Zn	300	20 ppm	
Copper Cu	100	6 ppm	
Molybdenum Mo	1	0.1 ppm	
Nickel Ni	1	0.1 ppm	

[a]Concentration expressed on a dry-matter-weight basis.

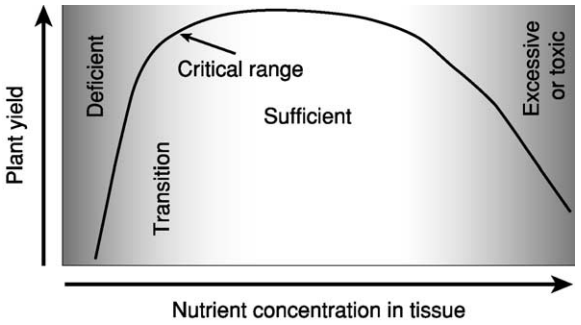


Figure 1 Relationship between plant nutrient concentration and yield. Nutrient-deficient plants exhibit visual deficiency symptoms (abnormal leaf color or other growth characteristic) and respond in yield to addition of the deficient nutrient. Plant nutrient concentration below which plant growth or yield is increased with nutrient addition is the critical level or range. Sufficiency or luxury consumption is the concentration range where added nutrient does not increase yield but increases nutrient concentration. Excessive or toxic nutrient concentrations reduce plant growth and yield.

and micronutrients are classified by their relative abundance in plants (Table 1). Sodium (Na), cobalt (Co), vanadium (Va), and silicon (Si) are beneficial in some specific plants.

When plant nutrient concentration is deficient enough to reduce plant growth severely, distinct visual deficiency symptoms appear (Figure 1). Extreme deficiencies can result in plant death. Visual symptoms may not appear when a nutrient is marginally deficient, but plant yield may still be reduced. Plant nutrient concentration above which plant growth or yield is not increased is the critical level or range. Critical nutrient ranges vary among plants but always occur in the transition between nutrient deficiency and sufficiency (Figure 1). Sufficiency or luxury consumption is the concentration range where added nutrient does not increase yield but increases nutrient concentration. Nutrient concentration is considered excessive or toxic when plant growth and yield are reduced. Plants also absorb many nonessential elements in the soil solution. For example, high Al^{3+} in plants occurs in acid soils with relatively large amounts of soluble Al. In severely acid soils Al toxicity can reduce plant yield.

Basic Soil Properties and Processes

Plant nutrient availability is a dynamic interaction of many soil processes (Figure 2). As plants absorb nutrients from soil solution, reactions occur to resupply the solution. Specific nutrients are influenced more by some reactions than by others. For example, biological processes are more important to N and S availability, and surface exchange reactions are important for K, Ca, and Mg.

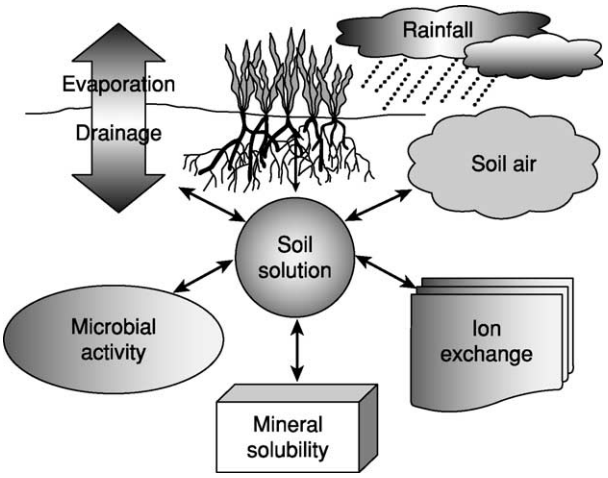


Figure 2 General diagram of nutrient cycling in soil. See text for description.

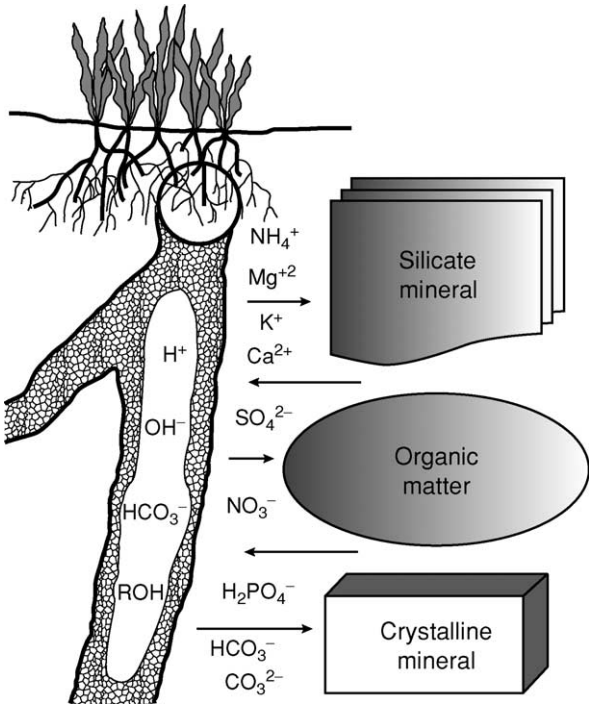


Figure 3 Ion exchange between nutrient sources (minerals and OM) and plant roots. As roots absorb nutrients from soil solution, nutrients adsorbed to the cation or anion exchange capacity are desorbed, nutrient-bearing minerals dissolve, and/or OM mineralization resupplies nutrients to soil solution.

Cation and Anion Exchange

Exchange of cations and anions on surfaces of clay minerals, inorganic compounds, organic matter (OM), and roots is one of the most important soil chemical properties influencing nutrient availability (Figure 3). Adsorbed ions are reversibly exchanged with other ions in solution. Cation exchange capacity (CEC) represents the quantity of negative charge

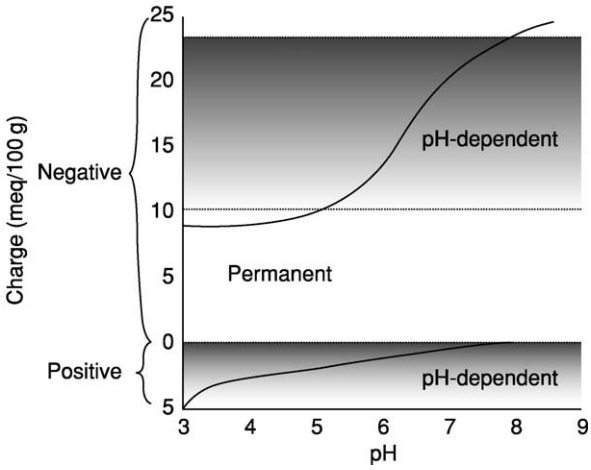


Figure 4 Influence of soil solution pH on CEC and AEC. Permanent charge is due to isomorphic substitution and pH-dependent charge is due to broken edges.

Table 2 Typical range in cation exchange capacities (CEC) for different-textured soils

Soil texture	*CEC (cmol kg^{-1})*
Sands (low OM)	3–5
Sands (high OM)	10–20
Loams	10–15
Silt loams	15–25
Clay and clay loams	20–50
Organic soils	50–100

available to attract cations. Anion exchange capacity (AEC) represents the positive charge available to attract anions in solution. In most soils CEC > AEC.

CEC generated through isomorphic substitution is permanent charge unaffected by solution pH (Figure 4). Both CEC and AEC also occur on mineral edges and OM surfaces, and vary with pH (Figure 4). Soils with greater clay and OM contents have a higher CEC (and AEC) than sandy, low-OM soils (Table 2).

CEC greatly influences nutrient availability, as most adsorbed cations are plant nutrients, except for Al^{3+}. In acid soils the primary cations are Al^{3+}, H^+, Ca^{2+}, Mg^{2+}, and K^+, whereas when soil pH $\geq$6.5 major exchangeable cations are Ca^{2+}, Mg^{2+}, K^+, and Na^+. Cations are adsorbed with different strengths, which influences the relative ease of desorption. The lyotropic series, or relative strength of adsorption, is $Al^{3+} > H^+ > Ca^{2+} > Mg^{2+} > K^+ = NH_4^+ > Na^+$. Cations with greater charge are more strongly adsorbed. For cations with similar charge, adsorption strength is determined by the size of the hydrated cation; greater hydrated cation radii reduce adsorption because larger hydrated cations cannot get as close to the exchange site as smaller cations. For anions, the relative strength of adsorption is $H_2PO_4^{2-} > SO_4^{2-} \gg NO_3^- = Cl^-$. In most soils

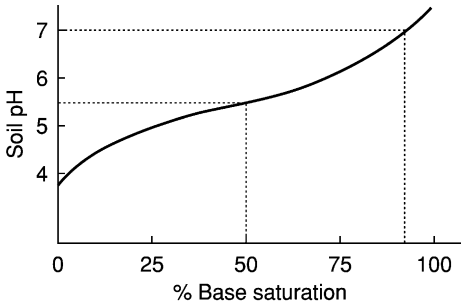


Figure 5 General relationship between soil pH and base saturation.

$H_2PO_4^{2-}$ is the primary anion adsorbed, although in severely acid soils SO_4^{2-} adsorption represents a major source of plant available S.

Base Saturation

Base saturation (BS) represents the percentage of CEC occupied by bases (Ca^{2+}, Mg^{2+}, K^+, and Na^+). The %BS increases with increasing soil pH (Figure 5). The availability of Ca^{2+}, Mg^{2+}, and K^+ increases with increasing %BS. For example, an 80% BS soil provides cations to plants more easily than 40% BS soil. At pH 5.5 most soils have 45–55% BS, whereas at pH 7 BS is $\geq$90%.

Buffer Capacity

Plant nutrient availability depends on nutrient concentration in solution and on the buffer capacity (BC). For example, as roots absorb K^+, K^+ on the CEC is desorbed to resupply solution K^+. With P, adsorbed $H_2PO_4^{2-}$ will desorb, while P minerals may also dissolve to resupply solution $H_2PO_4^{2-}$. The ratio of adsorbed nutrient quantity (Q) to solution nutrient intensity (I) describes soil BC:

$$BC = \frac{\Delta Q}{\Delta I}$$

In Figure 6, $BC_{soil\ A} > BC_{soil\ B}$. Thus, increasing adsorbed nutrients increases solution concentration in soil B much more than soil A. As plant uptake decreases nutrients in solution (ΔI), soil A more readily buffers solution concentration than soil B. Since $BC_{soil\ A} > BC_{soil\ B}$, soil A likely contains greater clay and/or OM, resulting in greater CEC. In low BC soils where exchangeable K^+ might be too low to provide sufficient plant-available K, K must be added to increase exchangeable K to prevent K deficiency. Thus, BC strongly influences nutrient availability.

Soil Organic Matter

Soil OM represents organic materials in various stages of decay. The most important fraction to soil fertility

is humus that is relatively resistant to microbial degradation, but is essential for maintaining optimum soil physical conditions important for plant growth (Figure 7). Fresh crop residues undergo rapid microbial decomposition. The heterotrophic biomass represents soil microorganisms and fauna responsible for mineralization and immobilization processes that influence availability of N, P, S, and other nutrients. Residue decomposition and humus formation processes depend on climate, soil type, and soil and crop management practices. Soil OM can have substantial effects on soil processes and properties that influence nutrient availability (Table 3). For example, increasing soil OM increases soil aggregation and decreases bulk density, which improves water infiltration, air exchange, root proliferation, and crop productivity. Soil OM levels depend on soil and crop management. If management is changed, a new OM level is attained that may be lower or higher than the previous level. Management systems should be adopted that sustain or increase soil OM and productivity. Proper nutrient management will produce high yields, which will increase the quantity of residues and soil OM.

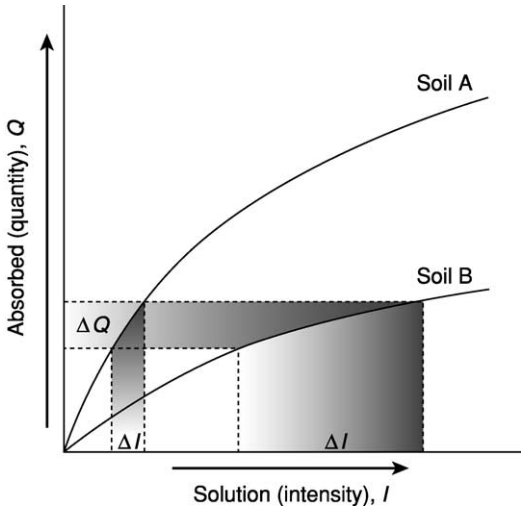


Figure 6 The relationship of adsorbed and solution nutrients as influenced by soil type. As nutrient concentration in solution (ΔI) decreases, soil A more readily buffers the solution concentration than soil B. Since $BC_{soil\ A} > BC_{soil\ B}$, soil A likely contains greater clay and/or OM content, resulting in greater CEC.

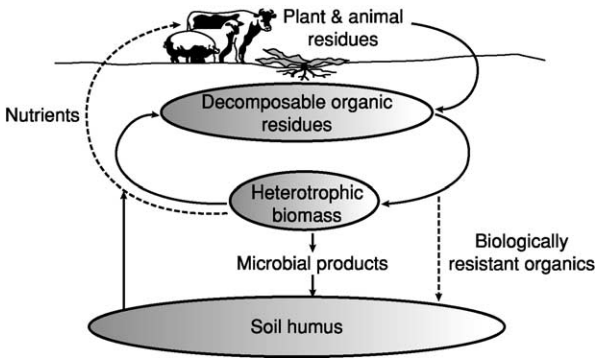


Figure 7 General degradation process of crop residue or other organic material added to soil.

Soil Acidity and Alkalinity

Acid soils usually occur in regions where annual precipitation >600–800 mm. Optimum soil pH varies between 4.5 and 7.5 depending on the crop. Neutralizing soil acidity by liming to ≥pH 5.6 will reduce exchangeable Al^{3+} to <10% of the CEC and will usually prevent Al toxicity. Decreasing soil pH under cropping systems is attributed to: (1) use of acid-forming fertilizers and manures; (2) crop removal and/or leaching of cations that decreases %BS; and (3) decomposition of organic residues. Soil acidification is enhanced with increasing rainfall since rain has a pH of 5.7 or less, depending on pollutants such as SO_2, NO_2, and others. Soil pH can strongly influence nutrient availability depending on the specific nutrient and the magnitude of the pH change. These influences will be discussed below under each specific nutrient.

Calcareous soils contain solid $CaCO_3$, exhibit soil pH ≥7.2, and occur in regions of <500 mm annual precipitation. Increasing rainfall increases depth to $CaCO_3$ where no $CaCO_3$ exists in the rooting zone at >800–1000 mm annual rainfall. If $CaCO_3$ exists, all of it must be dissolved or neutralized before soil pH could decrease. In most situations, reducing soil pH by neutralizing $CaCO_3$ is not practical. High soil

Table 3 Soil organic matter (OM) influences on soil properties and processes

Property	*Remarks*	*Effect on soil*
Soil color	Increasing OM darkens color	Darker color causes soils to warm up faster
H_2O-holding capacity	OM adsorbs ~20 times weight in H_2O	Enhances plant-available water
Soil aggregation	Binds clay particles into aggregates	Aggregation enhances permeability to H_2O/air
Chelation	Forms complexes with cations	Enhances micronutrient availability
Ion exchange/buffering	OM exhibits both CEC and AEC	Higher buffer capacity increases nutrient supply
Mineralization	OM degradation releases nutrients	Important nutrient source
Water solubility	OM is relatively insoluble in H_2O	OM does not extensively leach
Organic complexation	OM complexes organic compounds	Influences degradation/persistence of pesticides

Table 4 Classification of saline, sodic, and saline/sodic soils

Classification	*Conductivity (mmhos cm^{-1})*	*Soil pH*	*Exchangeable sodium percentage*	*Soil physical condition*
Saline	>4.0	<8.5	<15	Normal
Sodic	<4.0	>8.5	>15	Poor
Saline/sodic	>4.0	<8.5	>15	Normal

pH caused by the presence of $CaCO_3$ can reduce availability of P and several micronutrients.

Saline and sodic soils are characterized by their electrical conductivity (EC_{se}), soil pH, and exchangeable Na% (Table 4). When exchangeable Na > 15%, soil aggregates disperse, reducing permeability to water. In saline soils soluble salt concentration interferes with plant growth, although salt tolerance varies with plant species. In sodic soils, excess Na disperses clays and can create nutritional disorders. Saline-sodic soils exhibit both high salt and Na content.

Specific Nutrients

Nitrogen

Most soils do not supply sufficient N to optimize productivity of nonlegume crops. Thus, supplemental N must be provided. Understanding N cycling is necessary to optimize crop yield and minimize the environmental impact of N inputs. Most soil N originated from atmospheric N_2; however, plants cannot directly metabolize N_2 into protein. Atmospheric N_2 is transformed to plant-available N by: (1) symbiotic microorganisms associated with legume roots; (2) nonsymbiotic soil microorganisms on the soil surface; (3) inorganic N in rainfall caused by lightning; and (4) manufacture of N fertilizers (from N_2).

Rainfall contains NH_4^+, NO_3^-, NO_2^-, and organic N. Atmospheric electrical discharges and industrial waste gases annually account for 1–50 kg N ha^{-1}. Symbiotic bacteria (rhizobia) specific to each host species convert N_2 to NH_4^+ in legume root nodules. N fixation by legumes ranges from 10 to 200 kg ha^{-1} year^{-1}, depending on the legume, and many soil, crop, and environmental factors. Low soil pH reduces rhizobia function through Al^{3+} toxicity and/or low Ca^{2+} supply. High soil N availability also reduces N_2 fixation. Management practices that reduce legume yield will reduce N_2 fixation. For example, enhanced water, nutrient, weed, insect, and diseases stress will reduce N_2 fixation. Poor harvest management of forage legumes that include excessive harvest frequency and premature or delayed harvest will reduce legume stand and N_2 fixation. Some nonsymbiotic bacteria and blue-green algae also fix small amounts of N_2 (1–5 kg N ha^{-1} year^{-1}) but contribute little to N availability. Industrial N fixation is the major N source. The Haber-Bosch reaction converts N_2 and H_2 to NH_3 gas used directly (anhydrous NH_3 fertilizer) or manufactured into fertilizer N products.

Table 5 Typical C:N ratios for selected of plant residues, soil microbes, and soil OM

Organic substances	*C:N ratio*
Soil microorganisms	8:1
Soil organic matter	10:1
Sweet clover (young)	12:1
Barnyard manure (rotted)	20:1
Clover residues	23:1
Green rye	36:1
Corn/sorghum stover	60:1
Small grain straw	80:1
Oak	200:1
Pine	286:1
Spruce, fir	>1000:1

N mineralization/immobilization Soil comprises about 5% inorganic N and 95% inorganic N. Ammonium (NH_4^+), nitrate (NO_3^-), and to a minor extent nitrite (NO_2^-) are in soil solution and absorbed by plants. Plant-available NH_4^+ and NO_3^- depend on N applied as fertilizers or waste materials, and N mineralized from soil OM. Mineralization is a two-step process (aminization and ammonification) that converts organic N to NH_4^+. N mineralization increases with soil OM content, temperature, moisture, and O_2. During the growing season 1–4% organic N is mineralized. N immobilization is the microbial conversion of NH_4^+ and NO_3^-, to organic N. As soil microorganisms decompose plant residues, N immobilization can substantially lower NH_4^+ and NO_3^-, resulting in potential crop N deficiency. N input rates are established to satisfy N demand by the crop and soil microbes.

Relative N mineralization or immobilization depends on the plant residue C:N ratio. Soil microbes have a low (8:1) C:N ratio relative to plant residues (Table 5). Mineralization occurs with plant residues ≤20:1 C:N because residue N is sufficient to meet the microbial demand during decomposition. Immobilization occurs with plant residues ≥30:1 C:N. For ratios between 20 and 30, no net N release or immobilization occurs. Residue N content also

indicates potential N immobilization/mineralization. With >2% N content, mineralization usually occurs. The residue decomposition rate depends on the quantity of residue added, soil inorganic N content, temperature, moisture, and residue characteristics (particle size and lignin/waxes content). Uncultivated soils have a relatively constant amount of plant residue returned to the soil and usually low N mineralization. When tilled, N mineralization increases, reducing soil OM content and ultimately N availability. Any change in OM content dramatically reduces the quantity of N mineralized and thus soil N availability to crops.

Nitrification The NH_4^+ produced from mineralization or other N inputs is converted to NO_3^- through a two-step nitrification reaction: $NH_4^+ \rightarrow NO_2^- \rightarrow NO_3^-$. The NO_3^- produced is mobile and subject to potential leaching losses. Knowledge of the factors affecting nitrification and careful management of N inputs can reduce N leaching losses. Generally the environmental conditions that favor plant growth are optimum for nitrification (Table 6).

NH_4^+ fixation Vermiculite and mica clay minerals can fix NH_4^+ in interlayer spaces. Some fixed NH_4^+ can be replaced with Ca and Mg. The availability of fixed NH_4^+ ranges from negligible to relatively high. NH_4^+ fixation is important in certain soils, fixing nearly 30% of applied N.

Denitrification Under low O_2 conditions, anaerobic organisms obtain O_2 by converting NO_2^- and NO_3^- to N_2, N_2O, NO, and NO_2 gases. Denitrifying organisms also exist in well-oxygenated soils mostly near roots (rhizosphere C source) and in anaerobic microsites or soil aggregates where O_2 diffusion is reduced. Denitrification is enhanced by plants because of readily available C in root exudates and sloughed-off root tissues. Denitrification potential is high in most soils; however, the quantity of gaseous N loss varies due to fluctuations in environmental conditions (H_2O content, temperature, etc.) and other factors (Table 6). Denitrification occurs primarily during periods of high precipitation (heavy rains, irrigation, snow melt), and continuously as small losses from anaerobic microsites in soil aggregates. Under denitrifying conditions, NO_2^- can accumulate, which is toxic to plants. High rates of band-applied N fertilizers elevate NH_4^+ levels and pH, enhancing NO_2^- accumulation. Denitrification is a useful remediation tool for reducing excessive NO_3^- in soil water. The use of riparian buffers to reduce NO_3^- in groundwater is well established.

Volatilization Compared with NH_3 volatilization associated with urea or NH_4^+ fertilizers, NH_3 loss from N mineralized from OM is small. Volatilization of NH_3 depends on high NH_4^+ in soil solution and high soil pH given by $NH_4^+ \rightarrow NH_3 + H^+$. The quantity of NH_3 loss depends on the interaction of many soil, environmental, and N fertilizer factors, but is highest in coarse-textured, calcareous soils, and with surface crop residues (Table 6).

Phosphorus

Availability of P depends on solution P concentration and the ability of soil P fractions to resupply or buffer solution P. As solution P decreases with plant uptake,

Table 6 Factors affecting nitrification, denitrification, and volatilization of soil N

Factor	*Comment*
Nitrification	
Supply of NH_4^+	Conditions favoring NH_4^+ production (mineralization) increase nitrification
Soil pH	Wide pH range, optimum pH is 8.5. Nitrifying bacteria need adequate Ca and P
Soil aeration	Nitrifying bacteria require O_2; good soil structure is essential to gas exchange/aeration
Soil moisture	Nitrification is highest at field capacity soil water content
Soil temperature	Optimum temperature 25–35°C. To minimize N leaching losses, soil temperature <10°C to fall apply N
Denitrification	
Decomposable organic matter	Fresh crop residues (C source) stimulate microbial growth and respiration
Soil moisture	High H_2O content impedes O_2 diffusion, increasing anaerobic microbe activity
Soil pH	Denitrifying bacteria prefer pH > 5.0. N gas released is also influenced by pH
Soil temperature	Denitrification increases rapidly from 2 to 25°C, decreasing above 60°C
NO_3^- supply	Increasing NO_3^- increases denitrification potential
Volatilization	
Soil pH	NH_3 loss increases with increasing pH
Buffer capacity	High BC decreases NH_3 loss by increased adsorption and pH buffering
N placement	Volatilization losses are greater with broadcast versus banded N fertilizers
Soil temperature	Volatilization increases with increasing temperature up to about 45°C
Soil moisture	Maximum NH_3 loss occurs at field capacity and during slow drying conditions
Crop residue	Residues increase NH_3 through wet, humid conditions at soil surface

P minerals dissolve, P adsorbed on clay mineral surfaces desorb, and organic P mineralizes to resupply solution P. Maintaining adequate plant-available P in solution (intensity) depends on P desorption and dissolution reactions (quantity) that resupply solution P taken up by plants (Figure 6). The larger the BC, the greater the ability to buffer solution P. Plants absorb $H_2PO_4^-$ and HPO_4^{2-}. Average solution P concentration is 0.05 ppm and varies widely among soils. Solution P required by plants varies from 0.003 to 0.3 ppm that is replenished nearly 100 times during the growing season. Rapid replenishment of solution P is essential near actively absorbing roots.

Organic P About 50% of total soil P is organic and occurs primarily as inositol phosphates, phospholipids, and nucleic acids. P mineralization represents microbial degradation of organic P in plant/animal residues and soil OM to release $H_2PO_4^-/HPO_4^{2-}$. P mineralization increases with increasing OM. The C/P ratio or P content of decomposing residues regulates whether P mineralization or immobilization occurs (Table 7). P and N mineralization/immobilization processes are similar. Inorganic mineralized P can be removed from solution through adsorption to mineral surfaces and precipitation as Al-, Fe-, or Ca-P minerals. The quantity of P mineralized varies widely but contributes about 3–10 $kg\,ha^{-1}$ in temperate region soils.

Inorganic P Inorganic P in solution can be: (1) absorbed by plants; (2) immobilized by microorganisms; (3) adsorbed to mineral surfaces; and/or (4) precipitated as secondary P compounds. Adsorption and precipitation reactions are referred to as P fixation and are a continuous sequence of precipitation and adsorption. In acid soils, inorganic P precipitates as Fe-P or Al-P minerals and/or is adsorbed to Fe/Al oxide surfaces. In neutral or high-pH soils, Ca-P minerals precipitate and/or P is adsorbed to clay and $CaCO_3$. Adsorption dominates under low solution P, while precipitation occurs when solution P levels exceed a mineral's solubility product (K_{SP}). Ultimately, solution P concentration is controlled by the solubility of inorganic P minerals. For example, decreasing solution $H_2PO_4^-$ by plant uptake causes variscite ($AlPO_4 \cdot 2H_2O$) to dissolve and resupply solution P according to:

$$AlPO_4 \cdot 2H_2O + H_2O \leftrightarrow H_2PO_4^- + H^+ + Al(OH)_3$$

Many factors affect P fixation in soils that ultimately influence P availability to plants (Table 8). P fixation is strongly influenced by soil pH, where P availability is maximum near pH 6.5. At low pH values, fixation occurs largely through adsorption to Fe/Al oxides and precipitation as $AlPO_4$ and $FePO_4$. As pH increases, Fe and Al concentration decreases, reducing P adsorption/precipitation. At pH >7.0, Ca-P minerals precipitate and P is adsorbed to $CaCO_3$, causing solution P to decrease. Thus, minimum P adsorption occurs at pH 6.0–6.5.

In soils with high P fixation capacity (high clay content or BC), solution P will decrease quickly after P fertilization. With low P fixation capacity solution P can be elevated for several months or years depending on P rate and application method. With high P fixation, fertilizer P should be applied in smaller, more frequent applications. When fertilizer P is broadcast-applied, P is in contact with more soil, increasing P fixation, compared to band-applied P. Band placement decreases contact between soil and fertilizer, reducing P fixation and increasing plant-available P.

Table 7 Influence of residue C:P on P mineralization and/or immobilization

C:P ratio	*%P in residue*	*Mineralization/immobilization*
<200	>0.3	Net mineralization of organic P
200–300	0.2–0.3	No gain or loss of inorganic P
>300	<0.2	Net immobilization of inorganic P

Potassium

Soil K content ranges between 0.5 and 2.5% and is substantially greater than that absorbed by plants. Depending on the soil, <10% soil K is plant-available. Exchangeable and solution K provides plant-available K during the season and nonexchangeable K resupplies solution and exchangeable K. Nonexchangeable

Table 8 Factors affecting P fixation in soils

Factor	*Comment*
Soil minerals	Fe/Al oxides can adsorb large amounts of solution P. P adsorption is greater with 1:1 clays than 2:1 clays. P adsorption increases with increasing clay content
Soil texture	Increasing clay content increases P adsorption
Soil pH	P adsorption to Fe/Al oxides decreases with increasing pH
Soil organic matter	Organic matter can 'coat' minerals, reducing P adsorption
Soil temperature	P adsorption increases with increasing temperature

K release is slow and essentially unavailable to crops during a season. Unlike N, P, and other nutrients, K availability is not influenced by soil OM.

Solution K averages about 5 ppm and is in equilibrium with exchangeable K. Equilibrium between solution and exchangeable K is represented by the *Q/I* ratio described for P. *Q/I* is directly proportional to CEC and measures relative ability to resupply solution K^+. High *Q/I* represents a high K buffer. Adding fertilizer K will increase *Q/I* and K availability. Low *Q/I* represents low CEC, sandy soils exhibiting low K buffer. The ability of soil to buffer solution K^+ as plants remove solution K^+ is partially controlled by the release rate of interlayer K. About 90% total soil K is nonexchangeable or mineral K. Nonexchangeable K resupplies exchangeable and solution K^+ as they are reduced during plant uptake. The K release rate is controlled through weathering of K minerals (micas > feldspars). Low solution K encourages interlayer K release. Mica and vermiculite minerals can 'fix' K, which represents solution K^+ entering the interlayer region. NH_4^+ ions can also be fixed. Fixation of K^+ and NH_4^+ occurs more in clay soils. Fixation of K reduces solution K, but can be released again to available K. Many factors influence K availability (Table 9).

Sulfur

Like N, sulfur (S) occurs in organic (≥90%) and inorganic (SO_4^{2-}) forms. Although large temporal variation occurs, solution S ranges from 1 to 20 ppm SO_4^{2-}, where sandy soils often contain ≤4 ppm. Near industrial areas, solution SO_4^{2-} can be increased through deposition caused by SO_2 pollution. Like NO_3^-, SO_4^{2-} can be leached, especially in sandy soils with a high quantity of percolating water. Adsorbed SO_4^{2-} can contribute to plant-available S in acid soils with relatively high AEC and/or Al/Fe oxide content. Many soil factors affect SO_4^{2-} adsorption (Table 10). The amount and type of clay and soil pH have the greatest influence on the quantity of SO_4^{2-} adsorbed. The availability of S depends mostly on mineralization of organic S to SO_4^{2-}. S mineralization ranges from 5 to 15 kg SO_4-S ha^{-1} $year^{-1}$. Immobilization converts SO_4^{2-} to organic S. Factors that affect microbial activity influence S mineralization/immobilization

Table 9 Factors affecting K availability to crops

Factor	*Comment*
Clay minerals	Increasing 2:1 clay content increases K availability
CEC	Increasing CEC increases K availability
Exchangeable K	Increasing exchangeable K increases K availability
K fixation	Soils with greater K fixation potential require greater K inputs to increase K availability
Soil moisture	Increasing solution K or H_2O content increases K diffusion and availability
Soil temperature	Increasing soil temperature increases K availability (> diffusion and K release rate)
Soil pH	At low pH K uptake can be reduced by Al toxicity. Under higher pH, > exchangeable Ca/Mg competes with K for uptake sites, reducing K uptake

CEC, cation exchange capacity.

Table 10 Factors affecting SO_4^{2-} adsorption and S mineralization/immobilization in soils

Factor	*Comment*
Adsorption	
Clay content	Adsorption increases with clay content. SO_4^{2-} adsorption follows kaolinite > illite > montmorillonite
Hydrous oxides	Fe/Al oxides contribute most of the SO_4^{2-} adsorption in severely acid soils
Soil depth	SO_4^{2-} adsorption is greater in subsoils due to greater clay and Fe/Al oxides
Soil pH	AEC increases with decreasing pH. Little SO_4^{2-} adsorption at pH >6.5
SO_4^{2-} concentration	Adsorbed SO_4^{2-} increases with solution SO_4^{2-}
Other anions	Adsorption strength decreases according to $OH^- > H_2PO_4^- > SO_4^{2-} > NO_3^- > Cl^-$
Soil OM	AEC site on OM can contribute to SO_4^{2-} adsorption.
Mineralization and immobilization	
S content of OM	S mineralized increases with increasing S content in residue. C/S ratio ≤ 200 S mineralization occurs; C/S = 200–400 no net mineralization occurs; C/S ≥ 400 S immobilization occurs
Soil temperature	S mineralization is optimum between 20 and 40°C
Soil moisture	S mineralization is optimum at 60% of field capacity

OM, organic matter; AEC, anion exchange capacity.

(Table 10). Soils with low OM (<1.2–1.5%) and/or course texture are often S-deficient.

Calcium and Magnesium

In acid and neutral pH soils, Ca and Mg occur on exchange sites and in minerals. In calcareous soils Ca/Mg also exist as $CaCO_3$, $MgCO_3$, and/or $CaMg(CO_3)_2$. As exchangeable cations, they behave similarly to K. As Ca/Mg in solution decreases with plant uptake or leaching, Ca/Mg desorbs from CEC to resupply the solution (Figure 2). Plant uptake of Ca/Mg depends mostly on the amount of solution Ca/Mg, soil pH, and %Ca/Mg on CEC. Solution Ca/Mg decreases by: (1) plant uptake; (2) leaching; (3) adsorbed on CEC; and/or (4) precipitated as minerals (in arid climates).

Solution Ca and Mg ranges between 5 and 300 ppm Ca and 5 and 50 ppm Mg (15 ppm Ca and 10 ppm Mg is adequate for most crops), with the higher levels in arid regions. Sandy, humid-region soils are generally low in Ca and/or Mg, whereas clay soils are usually higher in exchangeable Ca/Mg. Low-Ca soils require supplemental Ca, usually through liming ($CaCO_3$), to increase pH. Many crops respond to Ca and Mg when Ca <25% and Mg <10% of CEC. High exchangeable K can interfere with Mg uptake. High NH_4^+ levels from fertilization can induce Mg deficiency in low-Mg soils.

Table 11 Factors affecting micronutrient availability to crops

Factor	*Comment*
Iron	
Soil pH	Deficiencies most often observed on high-pH (>7.2), calcareous, arid-region soils
Poor aeration	Increase in CO_2 in wet, poorly drained soils increases HCO_3^-, enhancing Fe stress
Soil OM	Low soil OM can enhance Fe stress due to low chelation, especially on high-pH soils
Nutrient interactions	High Cu, Mn, Zn, Mo, and P can reduce Fe uptake
Plant species	Plant species and varieties vary in susceptibility to Fe stress
Zinc	
Soil pH	Deficiencies most often observed on high-pH soils
Soil OM	Low soil OM can enhance Zn stress due to low chelation, especially on high-pH soils
Nutrient interactions	High Fe, Cu, Mn, and P can reduce Zn uptake
Climate	Zn deficiencies are more pronounced during cool, wet seasons
Plant species	Plant species and varieties vary in susceptibility to Zn stress
Copper	
Soil texture	Solution Cu is lower in leached and calcareous sands
Soil pH	Deficiencies most often observed on high-pH soils
Soil OM	Cu strongly complexes with OM. High OM can reduce Cu availability
Nutrient interactions	High Fe, Zn, and P can reduce Cu uptake
Climate	Cu deficiencies are more pronounced during cool, wet seasons
Plant species	Plant species and varieties vary in susceptibility to Cu stress
Manganese	
Soil pH	Deficiencies most often observed on high-pH, calcareous, coarse-textured soils
Poor aeration	Mn availability can be increased by wet, poorly aerated soil
Soil OM	Mn availability increases with OM. Very high OM can reduce Mn availability
Nutrient interactions	High Cu, Fe, or Zn can reduce Cu uptake
Climate	Mn deficiencies are more pronounced during cool, wet seasons
Plant species	Plant species and varieties vary in susceptibility to Mn stress
Boron	
Soil pH	B availability decreases with increasing pH, especially at pH >6.5
Soil moisture	B deficiency is enhanced with low soil moisture conditions
Soil OM	Increasing OM increases B availability
Soil texture	Coarse-textured, well-drained, sandy soils are often low in B
Nutrient interactions	Low Ca increases B uptake and potential toxicity
Plant species	Plant species and varieties vary in susceptibility to B deficiency and toxicity
Molybdenum	
Soil pH	Mo availability increases with increasing pH
Fe/Al oxides	High Fe/Al oxides increase Mo adsorption (decrease available Mo)
Nutrient interactions	P enhances Mo absorption by plants. S, Cu, Mn can decrease Mo uptake
Climate	Mo deficiency is more severe under dry soil conditions
Plant species	Plant species and varieties vary in susceptibility to Mo deficiency

OM, organic matter.

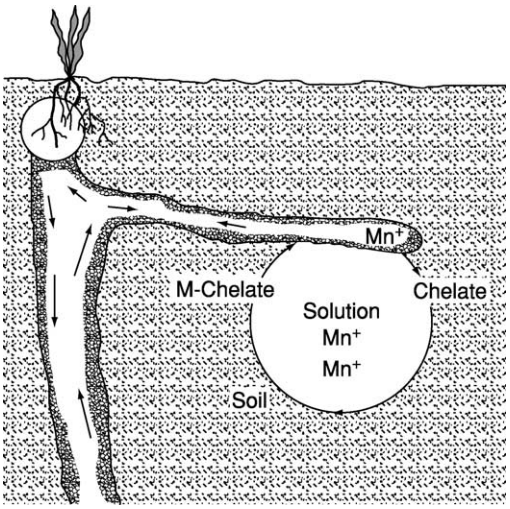


Figure 8 Cycling of chelate-bound micronutrients in soil solution to provide plant-available micronutrients. Chelates in solution bind micronutrients to metal cations and transport them to the root surface where micronutrient is released to plant uptake.

Micronutrients

Micronutrients occur in soil as: (1) primary and secondary minerals; (2) adsorbed to mineral and OM surfaces; (3) constituents of OM; and (4) solution (Figure 2). The proportion of each depends on the micronutrient, but all impact plant availability.

Iron, zinc, manganese, copper Plant-available Fe, Zn, Mn, and Cu are controlled by mineral solubility and soil OM. The solubility of Fe, Zn, Mn, and Cu minerals is generally very low and decreases with increasing pH. In most soils, solution Fe is not sufficient to meet plant need, thus another process increases Fe availability since most soils (except high-pH, calcareous soils) provide adequate Fe for plant growth. Soluble organic compounds (chelates) in soil solution complex micronutrients and increase solution Fe, Zn, Mn, and Cu concentration and diffusion to plant roots (Figure 8). Increasing soil OM can improve micronutrient availability. Factors influencing micronutrient availability are illustrated in Table 11.

Boron, chloride, molybdenum, nickel Boron (B) occurs in solution ($H_3BO_3^0$), minerals, adsorbed on clays and Fe/Al oxides, and in OM. Adsorbed and mineral B buffer solution B. Increasing pH, clay content, and OM favor B adsorption. Soil OM complexes with B and increases B availability. Most of the chloride (Cl^-) exists in soil solution, with very little occurring in mineral, organic, or adsorbed forms. In arid-region soils, Cl salts can precipitate and accumulate in surface soil, causing potential salinity and sodicity problems. Low-Cl soils can enhance disease incidence on some crops. Excess Cl in irrigation water can reduce yield of salt-sensitive crops. Molybdenum (Mo) occurs in minerals, adsorbed on Fe/Al oxides, complexed in OM, and in soil solution (MoO_4^{2-}). Solution Mo increases with increasing soil pH. Plant-available Ni is provided by adsorbed and mineral dissolution. Factors affecting B, Cl, Mo, Ni availability are shown in Table 11.

See also: **Fertilizers and Fertilization**; **Nutrient Availability**; **Nutrient Management**; **Quality of Soil**

Further Reading

Adams F (ed.) (1984) *Soil Acidity and Liming*. Madison, WI: Soil Science Society of America.

Barber SA (1984) *Soil Nutrient Bioavailability: A Mechanistic Approach*. New York: John Wiley.

Bresler E, McNeal BL, and Carter DL (1982) *Saline and Sodic Soils*. New York: Springer-Verlag.

Follett RF (ed.) (1989) Nitrogen management and groundwater protection. In: *Developments in Agricultural Managed-Forest Ecology #21*. New York: Elsevier.

Khasawneh FE (ed.) (1980) *The Role of Phosphorus in Agriculture*. Madison, WI: American Society of Agronomy, Crop Science Society of America, Soil Science Society of America.

Lindsay WL (1979) *Chemical Equilibria in Soils*. New York: John Wiley.

Mengel K and Kirkby EA (1987) *Principles of Plant Nutrition*. Bern, Switzerland: International Potash Institute.

Mortvedt JJ, Giordano PM, and Lindsay WL (eds) (1991) *Micronutrients in Agriculture, No. 4*. Madison, WI: Soil Science Society of America.

Mortvedt JJ, Murphy LS, and Follett RH (1999) *Fertilizer Technology and Application*. Willoughby, OH: Meister.

Munson RD (ed.) (1985) *Potassium in Agriculture*. Madison, WI: Soil Science Society of America.

Rendig VV and Taylor HM (1989) *Principles of Soil–Plant Relationships*. New York: McGraw-Hill.

Stevenson FJ (ed.) (1982) *Nitrogen in Agricultural Soils*. Madison, WI: American Society of Agronomy.

Stevenson FJ (1986) *Cycles of Soil*. New York: John Wiley.

Tabatabai MA (ed.) (1986) *Sulfur in Agriculture. No. 27*. Madison, WI: ASA, CSSA, Soil Science Society of America.

FERTILIZERS AND FERTILIZATION

H W Scherer, University of Bonn, Bonn, Germany

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

The main nutrient supplies for plants are released from soil reserves, decomposing biological residues (leaves, roots, etc.), aerial deposition, biological nitrogen fixation and, in the case of crop plants, fertilizers. When considering the availability of elements for plants, soil factors such as cation exchange capacity, acidity, redox potential, temperature, synergistic and antagonistic effects of other elements must be taken into account. The most important sources of elements in soils are the parent materials from which they are derived and soil provides the principal natural source of nutrients, while fertilizers provide the additional input to increase crop yields.

Total world consumption of fertilizer nitrogen (N), phosphorus (P), and potassium (K), in 1999/2000 was 85, 15, and 19 million t year^{-1}, respectively. In the future the demand for fertilizers is expected to increase with world population growth and the need for food. The projected increase in fertilizer consumption will mainly occur in developing countries, where food production urgently needs to be increased.

Nitrogen

Soil Nitrogen

N is the mineral nutrient most in demand by plants. Inorganic soil N mainly occurs as ammonium (NH_4^+) and nitrate (NO_3^-). In the surface layer of most soils the share of inorganic N of total soil N is usually less than 5%, while over 90% of N is organically bound. Most of the N in soil organic matter is in the form of stable humic compounds, while a minor part, mainly amino acids and amino sugars, is labile with a rapid turnover. Such compounds represent prime organic substrates in deamination reactions, where the removal of NH_4^+ is mostly carried out by enzymes. Once NH_4^+ is formed by mineralization it can be taken up by plants. However, mineralization depends on different factors that cannot be controlled and the release of NH_4^+ from organic N compounds is not necessarily timed to fit the plant's need. In most soils NH_4^+ may be fixed by clay minerals or can be held on negatively charged soil complexes or humic compounds. This reduces its mobility and therefore the availability for plants as compared to the more mobile NO_3^- ion, which is formed by nitrification of NH_4^+. Nitrification is important from the standpoint of soil fertility.

N Uptake of Plants

Plants take up N as NO_3^-, which is the most abundant form of inorganic N in the soil solution, and as NH_4^+. The uptake of both N forms depends mainly on their availability. In many crops the combination of NO_3^- and NH_4^+ results in better growth than when either form is applied alone: optimal proportions depend strongly on the total concentrations supplied. Whether NO_3^- or NH_4^+ as sole nitrogen source is better for plant growth depends on different factors. One important factor is the plant species. Plants that are adapted to acid soils or to low soil redox potential have a preference for NH_4^+, while plants with a preference for high-pH soils prefer NO_3^-.

Soil Test

In intensive agriculture deciding the optimal N fertilizer rate is a major challenge from an economic as well as ecological point of view. The optimal N fertilizer rate is the difference between the N demand of the crop and the N supply from different sources. Therefore, in predicting N fertilizer needs, in addition to the N demand of the crop, both the residual mineral N as well as the mineralizable N must be taken into consideration. The importance of determining residual mineral N as a basis for N fertilizer recommendation is already well established. Especially in the deeper soils of cooler, drier climates the amount of mineral N must be taken into account. Renewed interest in developing improved methods, taking the N mineralization capacity into account, has mainly been stimulated by concern over the environmental consequences of applying more fertilizer N than is demanded by the crop. However, the real crop need for N is not known before harvest. Therefore, yield is forecast using recent average yield data from the local area. Because an exact yield estimate is not possible, the crop has to be observed during the growing period and the N fertilizer plan adjusted. Depending on the N fertilizer form applied, N is more or less rapidly available to plants. However, calculating the optimum N application rate, it must be considered that even under optimal conditions N fertilizers are never fully utilized by the crop.

Time of N Fertilizer Application

Depending on soil and especially on climatic conditions, splitting the total N application rate into two or even more increments is recommended for maximum N fertilizer efficiency, whereby the first N fertilizer application is recommended shortly before the period of highest demand and of most rapid plant growth. The time of sidedressing depends on the crop. For cereals, maximum N fertilizer efficiency is reached by the latest possible application which is compatible with the development stage that permits ready uptake. However, for crops like sugar beet with a high N demand at the early growing stage for developing an extensive leaf surface, a late N fertilizer application is not desirable. To assure high sugar concentrations in the roots, sugar beets should be under conditions of mild N deficiency in the end of the growing period.

N Fertilizers

Most N fertilizers contain NO_3^- and NH_4^+ as the N carrier (Table 1). Caused by the sorption of NH_4^+ on soil colloids, the uptake rate of NO_3^- is normally higher under field conditions and therefore plants respond faster to NO_3^- application. However, N losses by leaching or denitrification may be higher with NO_3^- application compared with NH_4^+ application. Especially on paddy soils, denitrification losses may be high. For this reason fertilizers containing NH_4^+ or urea are used on these soils. On the other hand, NH_3 losses may occur in alkaline soils, when NH_4^+-containing fertilizers are applied, caused by the conversion of NH_4^+ to NH_3.

Table 1 Nitrogen (N) fertilizers

Fertilizer	*Chemical composition*	*N concentration (%)*
Ammonium nitrate	NH_4NO_3	35
Nitrochalk	$NH_4NO_3 + CaCO_3$	27
Ammonium nitrate sulfate	$(NH_4)_2SO_4 \cdot NH_4NO_3$	26
Ammonium sulfate	$(NH_4)_2SO_4$	21
Aqueous ammonia	NH_3, NH_4OH	>10
Anhydrous ammonia	NH_3	82
Urea	$CO(NH_2)_2$	46
Calcium cyanamide	$CaCN_2$	22

Adapted from Mengel K and Kirkby EA (2001) *Principles of Plant Nutrition*. Dordrecht: Kluwer Academic, with permission.

Ammonium nitrate, which is an explosive salt, is prohibited in some countries for direct use as an N fertilizer, but is often used to prepare liquid fertilizers. To reduce the danger of fire, limestone is added to ammonium nitrate, resulting in so-called nitrochalk. Neutralization of nitric and sulfuric acid with NH_3 results in the formation of the double salt of ammonium sulfate and ammonium nitrate (Figure 1).

Ammonium nitrate sulfate, a double salt of NH_4NO_3 and $(NH_4)_2SO_4$ in equal proportions, is formerly known as leunasalpeter. It is widely used on forage and grass seed crops and on autumn-seeded small grains. (Potassium nitrate is often used for foliar application.) Some years ago ammonium sulfate, which results in acidification of the soil, was an important fertilizer, but its consumption has declined in recent years and it was substituted by urea, which has a higher N content. Urea is a major N fertilizer. It is converted to NH_4^+ by urease in the soil. This rapid hydroloysis may result in significant N losses through NH_3 volatilization. Volatile losses of NH_3 originating from urea may be reduced by applying urease inhibitors that delay the formation of NH_4^+.

Calcium cyanamide is a slow-reacting N fertilizer, which contains N in the amide and cyanide form. This N fertilizer must be converted to urea, which is split into NH_3 and CO_2 before plant uptake. During conversion in the soil intermediate toxic products are formed, which are nitrification inhibitors as well as herbicides. For this reason the application of calcium cyanamide is recommended before sowing or transplanting.

Anhydrous ammonia is a liquid N fertilizer, which contains NH_3 in a liquid under pressure. It is a high-grade N fertilizer and therefore of considerable

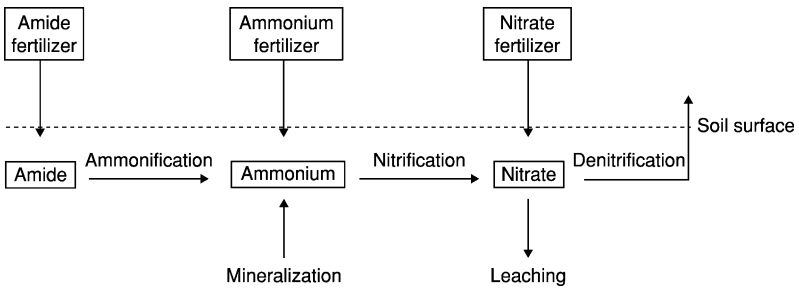


Figure 1 Turnover of nitrogen fertilizers in the soil.

advantage in terms of transport costs. However, its use is often restricted, because the liquid is under pressure, and therefore it needs special safety precautions and equipment for transportation and application. To minimize NH_3 losses by volatilization an injection assembly is used to apply it to the soil.

Aqueous ammonia contains about 25% NH_3 and the solution is only under low pressure. Therefore the handling of this solution is less complicated compared with anhydrous ammonia. NH_3 losses can usually be avoided by applying it at sufficient depth and under appropriate soil physical and moisture conditions.

Slow-Release and Bioinhibitor-Amended Nitrogen Fertilizers

Only 50–60% of fertilizer N added to the soil is usually taken up by the crop during the season of application. An approach to increase the efficiency of N fertilizers is to control the rate of N fertilizer dissolution. Controlled-release fertilizers, which are expensive to produce and therefore mainly used in the production of high-value crops, can be divided into two groups: coated fertilizers and uncoated fertilizers with a low solubility. The aim of using controlled-release fertilizers is to provide nutrients according to crop demand.

Coated fertilizers Different types of coating are used: impermeable coatings with tiny holes through which solubilized materials diffuse, and semipermeable coatings through which water diffuses until the internal osmotic pressure ruptures the coating or impermeable coatings that must be broken, for example, by biological actions. Further coatings may function only as physical barriers or be a source of plant nutrient. The most important coating materials, which are usually applied in intimate contact with the surface of the fertilizer particle are waxes, polymers, and sulfur.

Osmocotes are surrounded by a plastic shell, which allows water to diffuse into the shell, which tears as a result of the uptake of water and the nutrients diffuse into the soil. With sulfur-coated urea (SCU), water vapor transfers through the sulfur coating, solubilizes the urea within the shell, builds up sufficient osmotic pressure to disrupt the coating, and urea solution is released. The decomposition of the sulfur coat also depends on the oxidation of the sulfur, which is mainly brought about by soil microorganisms. Hence the availability of the sulfur-coated fertilizers depends on microbial activity.

Because the release of nutrients from the fertilizer depends on soil water content, and plant growth depends on soil moisture, nutrient release may also be adapted to plant growth.

Uncoated fertilizers While urea is readily soluble in water and quickly decomposed to release NH_4^+, it forms several chemical reaction products that are useful as slow-release N fertilizers. Urea reacts with several aldehydes to form compounds that are sparingly soluble in water. Commercial products are ureaform (UF), crotonylidendiurea (CDU), and isobutylidenurea (IBDU). After application to the soil, these fertilizers are prone to hydrolysis and release urea. The rate of microbiological decomposition is controlled by the property of the aldehyde as well as by soil temperature.

Nitrification Inhibitors

A controlled supply of N may, presumably, also be achieved by the use of any ammonium source in combination with a nitrification inhibitor (DCD = dicyandiamide; N-serve = nitrapyrin), which blocks the oxidation of NH_4^+ to NO_2^- (Figure 2). The loss of N by leaching or denitrification is therefore reduced. The amount of nitrification inhibitor needed to reduce the rate of nitrification can be reduced when it is granulated with the ammonium fertilizer.

Methods of N Application

Most the N fertilizer is applied by surface broadcasting, followed by injection of liquids and gases, incorporation by attachment to plow, planter, or cultivator, and distribution in irrigation water.

Especially in orchard or sod conditions where immediate fertilizer incorporation is not possible, surface broadcasting has some shortcomings. Under dry conditions on soils with a high pH, significant amounts of N can be lost by NH_3 volatilization when N is applied as urea or as NH_4^+ salts.

The injection of liquids and gases requires careful adaptation to the conditions of soil water. If the soil is too wet or too dry, volatilization losses may occur. Even under optimum soil water conditions injection must be at sufficient depth.

By application with the irrigation water (fertigation), N is applied during the whole growing season according to crop needs. A disadvantage associated with sprinkler irrigation is the chance of nonuniform N distribution caused by excessive wind.

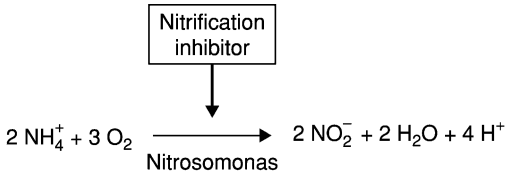


Figure 2 Inhibition of nitrification by a nitrification inhibitor.

Phosphorus

Soil Phosphorus

P in soils is present in the soil solution (plant-available), as labile phosphate precipitates and adsorbed to soil particles, mainly clay minerals (potentially available to plants), as nonlabile phosphate in the form of calcium, iron, and aluminum phosphate (not plant-available), in organic form, including P in soil organic matter (released after mineralization), and in living soil biomass (Figure 3). The quantity of P in soil solution in the upper soil layer is in the range of 0.3–3 kg P ha^{-1} and must be replenished by mobilization of phosphate from the labile pool within a time span of hours or days. There is a critical limit of readily soluble soil phosphate below which yields become severely constrained. For a high production level as well as for sustainable agriculture the maintenance of the readily soluble soil phosphate pool is a prerequisite.

Soil Test

In order to recommend an appropriate P fertilizer application, a soil test is indispensable for both soils with insufficient as well as for those with high amounts of available P. However, although reliable soil tests for available phosphate are at our disposition, regular P soil testing only covers a limited percentage of agricultural land. Soil tests avoid wasting the limited raw-material phosphate and reduce the risk of P leaching.

Different extraction methods are used as an index of plant-available P in soils. However, these methods never measure the quantity of P available to a crop, but determine more or less a pool of soil P that is somehow related to that portion of soil P which is plant-available.

To determine the P concentration of the soil solution the soil is extracted with water or a diluted salt solution (i.e., 0.02 mol l^{-1} $CaCl_2$). The objective of this method is to determine the P concentration level that directly limits plant growth. The most common methods used worldwide to extract plant-available P are probably the acid fluoride extract and the alkaline bicarbonate extract. In Europe an extraction using lactate is common.

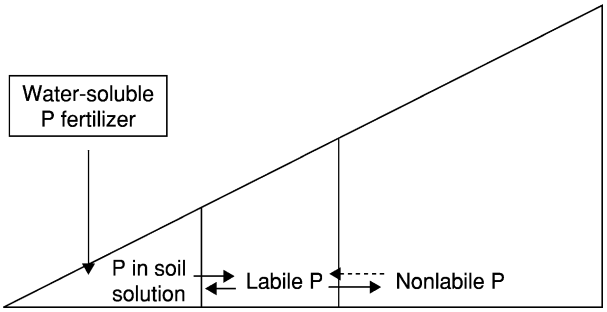


Figure 3 Phosphorus (P) dynamics in the soil. (Adapted from Mengel K and Kirkby EA (2001) *Principles of Plant Nutrition*. Dordrecht: Kluwer Academic, with permission.)

Saving P is also a question of an efficient use of soil and fertilizer P. Agronomic measures for improving phosphate efficiency are fertilizing according to the results of soil tests, providing optimum pH conditions for phosphate availability, using crop species in the crop rotation, which are able to mobilize less soluble soil phosphates, and applying appropriate P fertilizers.

P Application Rates

According to available soil P and crop species, P application rates generally range from 20 to 80 kg P ha^{-1}. As compared with slow-growing crops, crops with a high growth rate, for example corn, alfalfa, or sugar beet, require higher amounts of available P. On soils with a high P fixation capacity, up to 200 kg P ha^{-1} is recommended.

To avoid a decline in soil fertility, phosphate that has been removed by crops or is lost needs to be replenished by P fertilizer application. Less than 15% of P fertilizer applied to a soil is normally taken up by the crop grown immediately after application, while the rest is converted to soil phosphate and the recovery becomes less in the following years.

At neutral soil pH, fertilizer phosphate added to soils will at first be adsorbed or precipitated in more-soluble calcium phosphate forms. Later the phosphate of these forms will be shifted to apatite. In more acid soils rich in sesquioxides, adsorption is the strongest phosphate sink. Besides adsorption, phosphate occlusions occur in soils with a low pH.

P Fertilizers

The raw material for the production of P fertilizers is rock phosphate (apatite). Depending on their provenance, rock phosphates differ in fertilizer value. The solubility of apatites depends on the degree of isomorphic substitution of PO_4^{3-} by CO_3^{3-} and is higher the more phosphate is substituted by carbonate. While hard crystalline apatites are very insoluble and useless as a P fertilizer, soft rock phosphates may be used after grinding under acidic soil conditions (pH <5.5), for example in tropical regions, where soluble P fertilizers are prone to leaching. Further, organic farmers, who do not apply chemically treated rock phosphate, are allowed to use ground soft rock phosphate. Over the past decade the direct application of rock phosphate declined and at present it constitutes 0.8% of total global mineral P consumption.

Table 2 Phosphorus (P) fertilizers

Fertilizer	*Chemical composition*	*P concentration (%)*
Superphosphate	$Ca(H_2PO_4)_2 + CaSO_4$	8–10
Triple superphosphate	$Ca(H_2PO_4)_2$	20
Monoammonium phosphate	$NH_4H_2PO_4$	21–22
Diammonium phosphate	$(NH_4)_2H_2PO_4$	24
Basic slag	$Ca_3P_2O_8 \cdot CaO + CaO \cdot SiO_2$	4–10
Sinterphosphate	$CaNaPO_4 \cdot Ca_2SiO_4$	11–13
Partially acidulated rock phosphate	$Ca(H_2PO_4)_2$, $CaHPO_4$, apatite	10
Ground rock phosphate	Apatite	13

Adapted from Mengel K and Kirkby EA (2001) *Principles of Plant Nutrition*. Dordrecht: Kluwer Academic.

Treating ground rock phosphate with sulfuric acid results in the formation of superphosphate, a mixture of $Ca(H_2PO_4)_2$ and $CaSO_4$, and treating with phosphoric acid results in the formation of triple superphosphate (Table 2). Both are water-soluble P fertilizers and suitable for most soil types. However, after the application of water-soluble P fertilizers, reactions between the phosphate and soil particles take place which reduce P availability. Under tropical conditions with a high risk of P leaching, water-soluble P fertilizers are not recommended. Partially acidulated P fertilizers are derived from rock phosphates which have been treated with amounts of sulfuric or phosphoric acids: only part of the apatite is converted to water-soluble phosphate and the rest remains as apatite. Therefore this P fertilizer contains only a part of plant-available P.

Mono- and diammonium phosphates are produced by adding NH_3 to phosphoric acid. They are often used for band application, which increases P efficiency, because less P is adsorbed by soil particles and P uptake is increased by the root attraction by NH_4^+.

Sinterphosphate is produced by disintegration of rock phosphate with Na_2CO_3 and silica at a temperature of about 1250°C and contains $CaNaPO_4$ and Ca_2SiO_4.

Basic slags $(Ca_3(PO_4) \cdot CaO + CaO \cdot SiO_2)$ are a by-product of the steel industry.

Potassium

Soil Potassium

Potassium is one of the three major crop nutrients. Its main source for growing plants under natural conditions is the weathering of K^+-bearing minerals, whereby the rate of K^+ release not only depends on the K^+ content of the minerals but also on structural differences between K^+-bearing minerals. The K^+ status of soils mainly depends on the parent material and its degree of weathering. Whether the rate of K^+ release is adequate to meet the demand of the crop depends mainly on the kind of soil and the intensity of cropping.

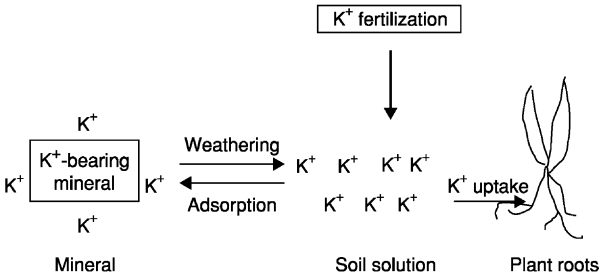


Figure 4 Potassium dynamics in the soil. (Adapted from Mengel K and Kirkby EA (2001) *Principles of Plant Nutrition*. Dordrecht: Kluwer Academic, with permission.)

Considering the mobility of soil K^+, different K^+ fractions can be distinguished: K^+ present in soil solution, K^+ adsorbed in exchangeable form to soil colloids, interlayer K^+ between the tetrahedral layers of minerals such as micas, and K^+ as a structural element of soil minerals. K^+ in soil solution and adsorbed K^+ can be easily adsorbed by plant roots (Figure 4).

K Fertilizer Recommendation

In agricultural practice it is important to know for a given location whether K^+ fertilizer application can increase crop yield. The main factors which have an impact on the crop response to K^+ fertilization are plant-available soil K^+, soil moisture, and the crop species or even sometimes different cultivars of the same species. In this respect differences in root metabolism and rooting pattern are of particular interest. Because of the longer root system of monocots like grass they profit better from soil K^+ than dicots.

High K^+ responses are likely to occur if the top soil is rather dry during the early growing stage of the crop. Because it is difficult to provide general K^+ fertilizer recommendations that are applicable to different crops, soils, and climates, the basis for recommendations should be made using soil analysis as well as K^+ fertilizer experiments for representative crops and soils. If the level of plant-available soil K^+ declines below the optimum, K^+ fertilizer must be applied. However, when K^+ fertilizers are added to soils, a portion remains in soil solution, while another portion will be held in exchangeable form and yet another portion may become unavailable, especially in soils with a high K^+ fixation capacity. Under these conditions K^+ application rates may be extremely

Table 3 Potassium (K) fertilizers

Fertilizer	*Chemical composition*	*K concentration (%)*
Muriate of potash	KCl	50, 41, 33
Potassium sulfate	K_2SO_4	43
Potassium nitrate	KNO_3	37
Sulfate potash magnesia	$K_2SO_4 \cdot MgSO_4$	18
Kainite	$KCl + NaCl + MgSO_4$	10

Adapted from Mengel K and Kirkby EA (2001) *Principles of Plant Nutrition*. Dordrecht: Kluwer Academic, with permission.

high. If the K^+ level of the soil is higher than the optimum, lower K^+ fertilizer rates are recommended or K^+ application may even be omitted. In the case of optimum K^+ conditions the amount of K^+ applied should maintain the K^+ status. This quantity can be calculated from the K^+ removal of the crop.

K Fertilizers

The usual potassium fertilizers are two main types in which K^+ is combined with either chloride (muriate of potash) or sulfate (sulfate of potash) (Table 3). Potassium chloride (KCl) is available in three different grades: 50% K, 41% K, and 33% K. The two latter forms contain substantial amounts of sodium chloride (NaCl) and are recommended as K^+ fertilizers for natrophilic crops.

Potassium sulfate (K_2SO_4) is usually manufactured by reacting the chloride with sulfuric acid. Normally K_2SO_4 contains 43% K. The fact that the use of modern highly concentrated fertilizers leads to unbalanced plant nutrition suggests that sulfate-based fertilizers, which have up to now only been used for certain special crops, may have a wider application.

Sulfate potash magnesia ($K_2SO_4 \cdot MgSO_4$) is essentially a mixture of potassium sulfate and kieserite with 18% K and 11% Mg. This is a useful fertilizer to apply to nonchloride crops when there is a need for Mg.

Potassium nitrate (KNO_3) is mainly used in intensive horticulture under glass. Low-analysis chloride-based fertilizers (kainite) are crude or partly processed mineral salts. Chemically, they are mixtures of KCl and NaCl. Some also contain kieserite, which may be part of the raw material or may have been mixed in.

The needs of the plant, soil conditions, and climatic factors will determine which form of K^+ fertilizer is best suited to obtain high yields and good quality in any particular case.

Complex Fertilizers

Complex fertilizers are stable granules with all the nutrients of the guaranteed formula. They contain two or three of the main plant nutrients N, P, and K. Sometimes secondary nutrients (calcium, magnesium, and sulfur) as well as micronutrients in soluble form are incorporated, depending on the respective formula or special grade.

NPK fertilizers commonly differ in their NPK ratios. An NPK compound 13-13-21, for example, means that the ratio of $N : P_2O_5 : K_2O$ of these plant nutrients in the compound is 13% N, 13% P_2O_5, and 21% K_2O. Further NPK combinations are 10-15-20, 15-15-15, 12-12-17, 12-12-20. Combinations of N and P, N and K, as well as of N and K are also available. The complex fertilizer granules are free-flowing, resistant to physical damage and moisture, and easy to handle. The hardness of the granules as well as their resistance to abrasion keep the fertilizer from disintegration and dust formation even with frequent handling. Complex fertilizers take less time and labor to apply than individual straight fertilizers.

Bulk Blends

Bulk blends are an alternative to complex fertilizers. One of the major differences between bulk blends and complex fertilizers is the way in which the nutrients are combined. While bulk blends are a physical mixing or blending of separate dry ingredients with no chemical reaction between the different components, in complex fertilizers the chemical reaction during the production within the slurry guarantees that each granule contains the correct proportion of all nutrients.

Bulk blends are usually sold at a lower price than complex fertilizes. Their disadvantage is that bulk blends will segregate during transport, handling, and application into the individual components, due to differences in granule size and density.

Micronutrient Fertilizers

An inadequate supply of one of the micronutrients may impair plant growth and reduce yield and quality of the crop. The main issues with micronutrients are to ensure adequate but not excessive supply to plants. Micronutrients are often applied in combination with NPK fertilizers, as salts or chelates to the soil or dissolved in water sprayed on to the crop foliage (Table 4).

Iron

The most common sources of inorganic iron (Fe) sources are ferrous sulfate ($FeSO_4$) and ferrous carbonate ($FeCO_3$). Organic chelates include Fe salts of ethylenediaminetetraacetic acid (EDTA),

Table 4 Micronutrient fertilizers

Fertilizer	*Chemical composition*	*Element concentration (%)*
Ferrous sulfate	$FeSO_4 \cdot 7\ H_2O$	20
Ferrous carbonate	$FeCO_3$	48
Iron chelate	Fe-EDTA	5
Iron chelate	Fe-EDDHA	6
Manganese sulfate	$MnSO_4 \cdot 4\ H_2O$	30
Manganese oxide	Mn_2O_3, MnO_2	20
Manganese chelate	Mn-EDTA	12
Zinc sulfate	$ZnSO_4 \cdot 7\ H_2O$	23
Zinc carbonate	$ZnCO_3$	52
Zinc oxide	ZnO	80
Zinc chelate	Zn-EDTA	15
Copper sulfate	$CuSO_4 \cdot 5\ H_2O$	14
Copper oxychloride	$Cu_2Cl(OH)_3$	48
Copper chelate	Cu-EDTA	14
Boric acid	H_3BO_3	28
Borax	$Na_2B_4O_7 \cdot 10\ H_2O$	11
Solubor	$Na_2B_8O_{13} \cdot 5\ H_2O$	21
Sodium molybdate	$Na_2MoO_4 \cdot 2\ H_2O$	40
Ammonium molybdate	$3\ (NH_4)_2MoO_4 \cdot 4\ MoO_3 \cdot 4\ H_2O$	54

ethylenediamine di(*o*-hydroxyphenyl) acetic acid (EDDHA), and diethylenetriaminepentaacetic acid (DTPA).

Soils in which iron chlorosis is likely to occur are referred to as high-lime, alkaline, or calcareous. Under these conditions iron is generally not available for a continuous supply to the plant. As iron is fixed under high pH values, chelates have been shown to be superior sources of Fe for plants because lower rates are required than inorganic sources. Because of the high costs of chelates and the generally ineffective nature of inorganic Fe sources, foliar applications are the most effective method of correcting Fe chlorosis.

Manganese

Manganese sulfate ($MnSO_4$) is the most famous manganese (Mn) fertilizer. It is recommended for soil as well as for foliar application. However, when soils are treated with $MnSO_4$, banded placement rather than broadcasting should be carried out. Of the organic Mn fertilizers, Mn-EDTA, which is mainly used for foliar application, appears to give the best response. Water-insoluble Mn fertilizers are different Mn oxides (Mn_2O_3, MnO_2).

Zinc

There are many zinc (Zn) fertilizers available, but the most common inorganic compounds are zinc sulfate ($ZnSO_4$), which is the most used Zn fertilizer because of its high solubility, and zinc oxide (ZnO). Other inorganic Zn compounds are Zn carbonate ($ZnCO_3$) and Zn sulfide (ZnS). The main organic Zn fertilizer is Zn-EDTA. Both inorganic and organic Zn fertilizers are used for soil and foliar application.

Copper

Inorganic as well as organic copper (Cu) fertilizers are used to control Cu deficiency. Copper sulfate ($CuSO_4$) is the oldest Cu fertilizer, which may be used for soil and foliar application. However, caused by the scorching of $CuSO_4$ of the leaves, Cu oxychloride ($Cu_2Cl(OH)_3$) is recommended for foliar application. Cu-EDTA is recommended for soil and foliar application.

Boron

The most important boron (B) fertilizers are boric acid (H_3BO_3), borax ($Na_2B_4O_7 \cdot 10\ H_2O$), sodium tetraborate ($Na_2B_4O_7 \cdot 5\ H_2O$), sodium pentaborate ($Na_2B_{10}O_{16} \cdot 10\ H_2O$), and solubor ($Na_2B_8O_{13} \cdot 5\ H_2O$). Borax is the most used B fertilizer, and, like the other Na borates except solubur, which is applied both to soils and leaves, it is used for soil application. On high-B-fixing soils, H_3BO_3 is also used as a foliar spray. As compared with the other micronutrients the limited range between B deficiency and toxicity is extremely narrow.

Molybdenum

Molybdenum (Mo) is the micronutrient with the lowest demand. Mo deficiency is corrected by the application of sodium molybdate ($Na_2MoO_4 \cdot 2\ H_2O$) or ammonium molybdate ($3\ (NH_4)_2MoO_4 \cdot 4\ MoO_3 \cdot 4\ H_2O$). Besides soil application, foliar applications or treating seeds are common.

See also: **Fertigation**; **Fertility**; **Nutrient Availability**

Further Reading

Engelstad OP (1985) *Fertilizer Technology and Use*, 3rd edn. Madison, WI: Soil Science Society of America.

Finck A (1992) *Fertilizers and Fertilization* (in German). Weinheim, Germany: VCH.

IFA (1992) *World Fertilizers Use Manual*. Paris: International Fertilizer Industry Association.

Laegreid M, Bockman OC, and Kaarstad O (1999) *Agriculture, Fertilizers and the Environment*. New York: CAB.

Mengel K (1985) Dynamics and availability of major nutrients in soils. *Advances in Soil Science* 2: 65–131.

Mengel K (1997) Agronomic measures for better utilization of soil and fertilizer phosphates. *European Journal of Agronomy* 7: 221–233.

Mengel K and Kirkby EA (2001) *Principles of Plant Nutrition*. Dordrecht: Kluwer Academic.

Stevenson FJ (1982) *Nitrogen in Agricultural Soils*. Agronomy No. 22. Madison, WI: American Society of Agronomy, Crop Science of America, Soil Science Society of America.

Field Capacity *See* **Water Cycle**

FLOCCULATION AND DISPERSION

I Shainberg and G J Levy, The Volcani Center, Bet Dagan, Israel

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

The clay particles of soils are of colloid size (less than 2 μm) and usually carry a net negative charge. This charge is neutralized by exchangeable cations. The interaction between soil solution and soil clay and the phenomena of clay dispersion and flocculation depends on the type and amount of the soil clays, the nature of the exchangeable cations, and the electrolyte concentration in the soil solution. Clay dispersion (deflocculation) and flocculation are very important in determining the physical properties of soils such as hydraulic conductivity, aggregate stability, crusting, infiltration rate, and erosion.

In a stable clay suspension, dispersed particles collide frequently because of their Brownian motion, but separate again because of diffuse double-layer repulsion forces. When salt is added to the clay suspension, the particles may stick together upon collision, forming flocs that settle. The stability of the colloidal suspension is a balance between repulsive and attractive forces acting simultaneously among the suspended particles.

Shape and Charge of Soil Clays

The clay minerals of most soils are dominated by various layer silicates. Layer silicate clay minerals are classified as 1:1 where each layer consists of one tetrahedral silica sheet and one octahedral alumina sheet (e.g., kaolinite); 2:1 where each layer consists of one octahedral sheet sandwiched between two tetrahedral sheets (e.g., montmorillonite and vermiculite); or 2:1:1 where a metal hydroxide sheet is sandwiched between the 2:1 layers (chlorite). (*See* **Clay Minerals**.)

The clay minerals are characterized by isomorphic substitution of lower-valence cations in either the tetrahedral (Al replacing Si) or octahedral (Mg replacing Al) sheets or both. This excess of permanent negative charge inside the crystal is balanced by exchangeable cations on the external surface. Layer silicate clay minerals also possess variable charge located at the broken edges of clay minerals. At the edges of the 2:1 clay minerals, hydroxyl Al or hydroxyl Si is present. Negative charge may develop on the edges of the octahedral (Al—OH) layer at high pH and positive charge at low pH.

The colloidal activities of the clay minerals depend on their specific surface area (meters squared per gram) and the charge density expressed by their cation exchange capacity (CEC, in centimoles of cation charge per kilogram) divided by their specific surface area. Kaolinite is a widespread, 1:1 clay mineral with low CEC (1–10 $cmol_c\,kg^{-1}$), low surface area (20–50 $m^2\,g^{-1}$), and low colloidal activity. Conversely, montmorillonite, the dominant clay in soils from semiarid regions, belongs to the 2:1 clay mineral category, has a high CEC (80–120 $cmol_c\,kg^{-1}$), high surface area (800 $m^2\,g^{-1}$) and high colloidal activity. Vermiculite with high CEC (120–150 $cmol_c\,kg^{-1}$) and intermediate surface area and illite with intermediate CEC (20–40 $cmol_c\,kg^{-1}$) and intermediate surface area are clay minerals with low-to-intermediate colloidal activity. The mineralogy, charge characteristics, and CEC of several clay minerals are presented in Table 1.

Table 1 Charge characteristics and cation exchange capacities of clay minerals

Solid	Charge per unit half-cell		Cation exchange capacity ($cmol_c\,kg^{-1}$)
	Tetrahedral	Octahedral	
Kaolinite	0	0	1–10
Smectite			80–120
Montmorillonite	0	−0.33	
Vermiculite	−0.85	0.23	120–150
Mica			20–40
Muscovite	−0.89	−0.05	
Chlorite			10–40

Reproduced with permission from Goldberg S, Lebron I, and Suarez DL (1999) Soil colloidal behavior. In: Sumner ME (ed.) *Handbook of Soil Science*, pp. B195–B240. Boca Raton, FL: CRC Press.

Forces Between Clay Platelets

Repulsion Forces

The electrical double layer According to the Gouy–Chapman theory, the diffuse double layer consists of the lattice charge and the compensating counterions. The counterions are subject to two opposing tendencies: the cations are attracted electrostatically to the negatively charged clay surface; and the cations tend to diffuse from the surface of the particle, where their concentration is high, into the bulk of the solution, where their concentration is lower. The two opposing tendencies result in decreasing counterion concentration from the clay surface to the bulk solution (Figure 1). Divalent cations are attracted to the surface with a force twice as large as that of monovalent cations. Thus, the diffuse double layer in the divalent ion system is more compressed toward the surface. With an increase in the electrolyte concentration in the bulk solution, the tendency of the counterions to diffuse away from the surface is diminished and the diffuse double layer compresses toward the surface. The distribution of mono- and divalent cations in the diffuse double layer at high and low electrolyte concentration is also presented in Figure 1.

In Gouy–Chapman's diffuse double-layer theory, the counterions are considered as point charges without dimensions. As long as the ions are far from the clay surface, their concentration is low and their size has little influence on the statistical calculations of the diffuse double layer. However, in the vicinity of the clay surface where the concentration of the cations is high, the distance of closest approach to the surface is important, and the size of the ions and the energy of hydration of the cations must be considered to give an accurate description of the ions' distribution near the clay surface. Consequently, the double layer is divided into the Stern layer, i.e., the first layer of the counterions immediately next to the clay surface, and the Gouy layer, which is the diffuse layer. Adsorbed cations are classified as being in the diffuse or Stern layer depending on their dimension and hydration energy. Calculations using the amended theory predict differences in the adsorption characteristics and flocculation values of Li^+, Na^+, and K^+ montmorillonite in general agreement with experimental observations.

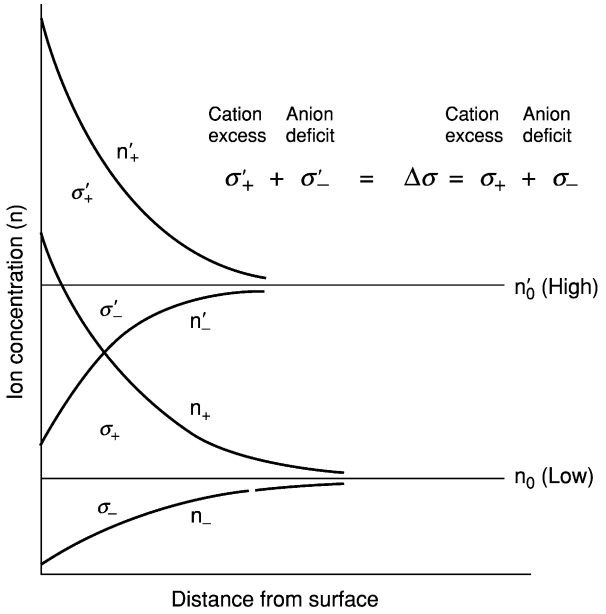


Figure 1 Distribution of counterions (n'_+ and n_+) and co-ions (n'_- and n_-) in the electrical double layer for a permanent charge surface (σ) at two electrolyte concentrations. (After Van Olphen H (1977) *An Introduction to Clay Colloid Chemistry*, 2nd edn. New York, NY: John Wiley.)

When two clay platelets approach each other, their diffuse counterion atmospheres overlap. Work must be performed to overcome the electrical repulsion forces between the two positively charged ionic atmospheres. The electrical double-layer repulsion force, also called the 'swelling pressure,' can be calculated by means of the diffuse double-layer theory. The greater the compression of the ionic atmosphere toward the clay surface, the smaller the overlap of atmospheres for a given distance between particles. Consequently, the repulsion forces between the particles decrease with increases in the salt concentration and in the valence of the adsorbed ions. Thus, high repulsion and high swelling persist in dilute solutions of sodium clay.

Hydration repulsion forces The adsorbed cations and the clay surfaces must lose some of their hydration water, because the clay particles approach each other very closely in the process of flocculation. The work that is needed to remove these water molecules manifests itself as the repulsion energy. Relative to the double-layer repulsion forces, the range of these forces is short. These forces partially negate van der Waals forces of attraction, resulting in particles not being trapped in a true primary minimum. This makes it easier for particles to disperse with little input of energy as the electrolyte concentration is reduced.

Flocculation and Attraction Forces

A stable clay suspension flocculates upon the addition of salt. Flocculation takes place only when a threshold electrolyte concentration, the 'flocculation value' (FV), is applied. A suspension of clay platelets flocculates in three possible modes of particle association (Figure 2): (1) association between flat oxygen planes

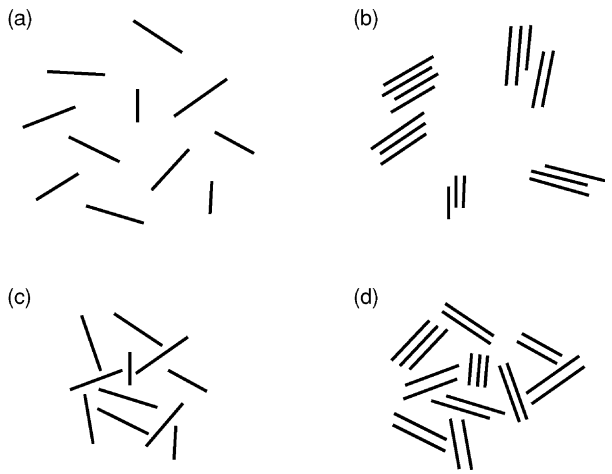


Figure 2 Schematic representation of modes of particle flocculation: (a) dispersed; (b) face-to-face flocculation; (c) edge-to-face flocculated; (d) face-to-face and edge-to-face flocculated. (After Van Olphen H (1977) *An Introduction to Clay Colloid Chemistry*, 2nd edn. New York, NY: John Wiley.)

of two parallel platelets (face to face, FF); (2) association between edge surface of neighboring particles (edge to edge, EE); and (3) association between edge and oxygen planar surface (edge to surface, EF). In the presence of low concentrations of NaCl, the diffuse double-layer repulsion prevents particle association and a stable suspension is obtained. As the concentration of NaCl increases, double layers at soil surfaces become compressed and, at the FV, attraction forces may prevail and EF, EE, and FF association may occur. Similarly, when polyvalent cations are adsorbed (e.g., Ca and Mg) on the clay surface, the diffuse double layer is compressed and attraction forces may prevail. Although the three types of association are really three modes of flocculation, only the EE and the EF types of association were discussed in the past and were labeled clay flocculation with high gel volume. The thicker particles, which result from simple FF association, were not called 'flocs,' because they were condensed with low gel volume. Special cases of FF association, which lead to high-volume flocs, have now been described (Figure 3) and it has been concluded that FF association may also lead to high-volume flocculation.

Van der Waals force of attraction The van der Waals attractive force between colloidal particles arises from the fluctuating dipole of one atom that polarizes another, and the two atoms attract each other. This attraction between atom pairs is additive and therefore the energy of interaction between particles decreases much more slowly with distance than that between individual atoms. The van der Waals attraction force leads to FF clay association.

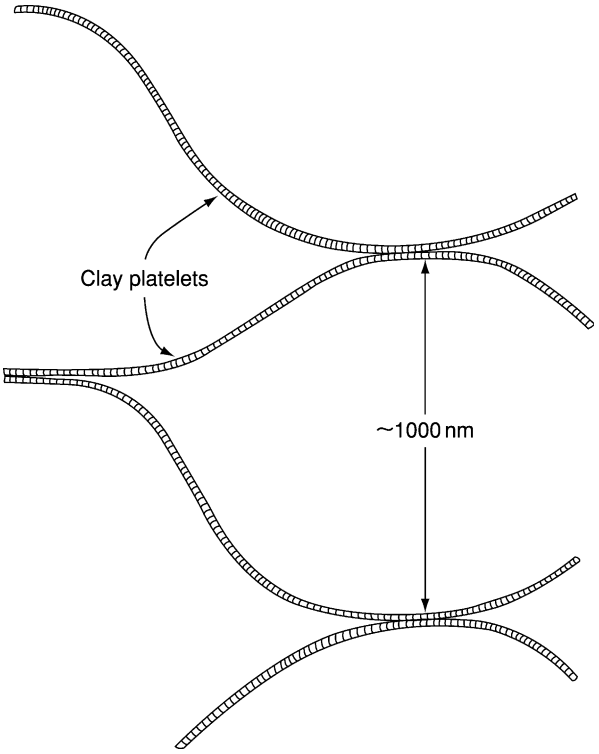


Figure 3 Mode of platelet association (face-to-face) in sodium montmorillonite suspension at moderate electrolyte concentration (e.g., 44 mmol dm^{-3} NaCl concentration). (Reproduced with permission from Keren R, Shainberg L, and Klein E (1988) Settling and flocculation value of Na-montmorillonite particles in aqueous media. *Soil Science Society of America Journal* 52: 76–80.)

Electrostatic EF attraction forces The flat surfaces are not the only surfaces of the plate-like clay particles; they also expose an edge surface area. At the edges, the octahedral alumina or iron sheets are disrupted and primary bonds are broken. This situation is analogous to that on the surface of alumina or iron particles in alumina and iron sols. On such surfaces an electric double layer is created by the adsorption of potential-determining ions. The hydroxyl-aluminum edge carries a positive charge in acid solution and a negative charge in high-pH solution. The zero point of charge does not necessarily occur at neutral pH. The charge becomes more positive with decreasing pH, and its sign may be reversed with increasing pH. When the charge on the edges is positive and the repulsive diffuse double-layer forces are diminished by a high concentration of electrolytes, a positive-edge to negative-face surface predominates and EF flocculation occurs.

Clay charge density and electrostatic attraction forces The colloid properties of clays depend to a large extent on the density of charge on the clay surface. This parameter can be derived from the

exchange capacity per gram and the surface area per gram (Table 1). Sodium montmorillonite in dilute suspensions exists as single platelets. Calcium montmorillonite, even in distilled water, exists in packets (tactoids or quasicrystals), each consisting of several (4–9) clay platelets with a film of water 0.45 nm on each internal surface and c-spacing of less than 1.9 nm. The swelling pressure of calcium montmorillonite can be predicted from the diffuse double-layer theory if one assumes that the number of platelets in a tactoid is 5 and that only the Ca ions on the external surfaces of the tactoids form a diffuse double layer. Sodium, magnesium, and calcium vermiculite do not swell beyond 1.5 nm (two molecular layers of H_2O between the platelets), even in distilled water. The illites exist in quasicrystals with c-spacing of less than 1 nm regardless of the adsorbed cations. It has been suggested that electrostatic attraction forces, which increase with the increase in surface charge density and with increase in cation valence, prevent clay swelling in clays with high surface charge density.

The charge density at clay surfaces is an average value, and heterogeneity of the charge density due to isomorphous substitution has been determined for montmorillonites ranging between 0.7×10^{-3} and 1.5×10^{-3} $mmol_c\,m^{-2}$. Wyoming bentonite is the most heterogeneous among six various clay types. It is possible, therefore, that at specific sites on the planar surfaces the charge density is very low or even zero (the surface area of these particular sites may not be detected). Thus, in the extreme situation where the charge density is very low (or even zero, such as in pyrophyllite), there are no repulsion forces (both double layer and hydration), and the van der Waals attraction forces exceed the repulsion forces and a FF association may occur (Figure 3). Similarly, in sites where the charge density is high (such as in vermiculite), the electrostatic attraction forces between the cations and the clay platelets is very high, swelling decreases, and a FF association may occur (Figure 3). Since the montmorillonite platelets are long and flexible and have the capability to bend, the platelets may form cohesive junctions at several locations of low or high surface charge densities to form a network structure, as illustrated in Figure 3. This results in a high gel volume, as demonstrated for Na-montmorrilonite at pH 9.8 (Figure 4).

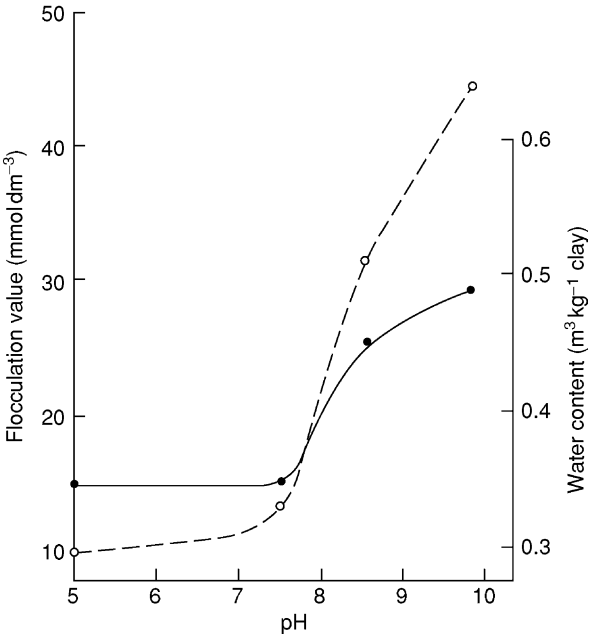


Figure 4 The flocculation value (dashed line) and gel volume (solid line) of sodium montmorillonite in NaCl solution as affected by suspension pH (the clay concentration was 0.1% and the apparent size of the platelets was less than 2 μm). (Reproduced with permission from Keren R, Shainberg L, and Klein E (1988) Settling and flocculation value of Na-montmorillonite particles in aqueous media. *Soil Science Society of America Journal* 52: 76–80.)

Spontaneous and Mechanical Dispersion

Spontaneous dispersion often takes place without energy inputs when sodic clay is placed in water of very low electrolyte concentration. The proportion of clay particles separated in this way increases with an increase in the sodium adsorption ratio (SAR) of the soil solution and with a decrease in the electrolyte concentration.

Hydrated clay particles in equilibrium with low-SAR solution, which have undergone limited separation, can be pushed further apart by applying external mechanical pressure such as stirring or raindrop impact. When the SAR of the clay solution is low to moderate and the electrolyte concentration is high enough to prevent spontaneous dispersion, the introduction of a mechanical force such as mixing the suspension may cause clay dispersion. The clay at the soil surface is especially susceptible to dispersion because of the mechanical energy applied by raindrops. This is one of the reasons why the soil surface is more susceptible to clay dispersion and aggregate disintegration by low sodic conditions, compared with the soil's profile.

Flocculation and Dispersion in Na/Ca Clays

The flocculation value of sodium and calcium montmorillonites Adsorbed Na ions form a diffuse double layer at the clay surface, and, because of the high swelling between the platelets, single platelets tend to persist in dilute solution. When salt is added to the clay suspension, the particles stick together upon collision, forming flocs that settle. The suspension is then separated into the bottom sediment and a

Table 2 Selected critical flocculation concentrations for some clay minerals

Mineral	CFC (mol m^{-3})	Background electrolyte	pH	Suspension strength (g kg^{-1})
Georgia kaolinite	5	NaCl	7	0.6
	245	$NaHCO_3$	8.3	0.6
	0.4	$CaCl_2$	7	0.6
Montmorillonite	12	NaCl	?	1
	0.13	$CaCl_2$	?	1
Vermiculite	38	NaCl	7	0.6
	58	$NaHCO_3$	8.3	0.6
	0.8	$CaCl_2$	7	0.6
Grundy illite	7.24	NaCl	6	1
	0.2	$CaCl_2$	6	1
Illite 36	9	NaCl	7	0.6
	185	$NaHCO_3$	8.3	0.6
	0.13	$CaCl_2$	7	0.6

CFC, critical flocculation concentration.
Based on data from Goldberg S, Lebron I, and Suarez DL (1999) Soil colloidal behavior. In: Sumner ME (ed.) Handbook of Soil Science, pp. B195–B240. Boca Raton., FL: CRC Press; Oster JD, Shainberg I, and Wood JD (1980) Flocculation value and gel structure of Na/Ca montmorillonite and illite suspensions. *Soil Science Society of America Journal* 44: 955–959.

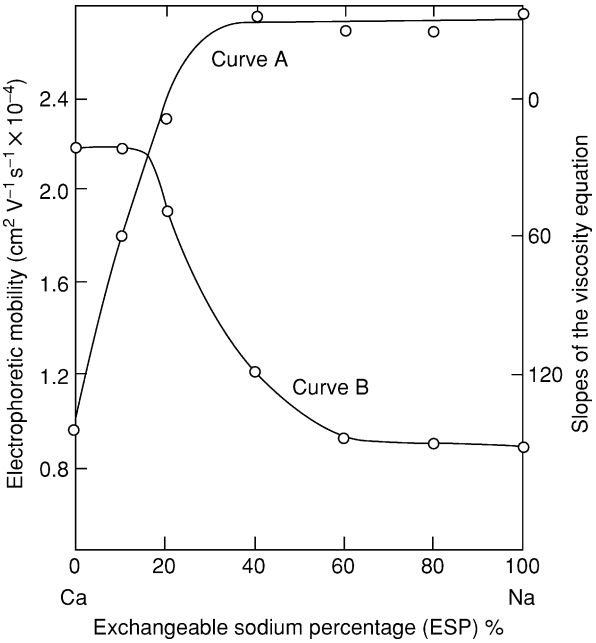


Figure 5 Dependence of electrophoretic mobility (curve A) and the relative size (curve B) of montmorillonite particles on the exchangeable sodium percentage (ESP). The relative size is expressed in units of the slope in Einstein's equation for the viscosity of suspension. (Reproduced with permission from Shainberg I and Letey J (1984) Response of soils to sodic and saline conditions. *Hilgardia* 52: 1–57.)

particle-free supernatant liquid. The FVs of clays depend on counterion valency. The FVs of sodium and calcium montmorillonite are 12 $mmol_c\,l^{-1}$ NaCl and 0.25 $mmol_c\,l^{-1}$ $CaCl_2$, respectively (Table 2). The sodium montmorillonite gel has approximately 330 g of water per gram of clay, which corresponds to a film thickness of 440 nm per surface (Figure 4). This thickness exceeds the range of diffuse double-layer forces and/or the van der Waals forces. The open structure typical of sodium montmorillonite gels suggests that EF attraction forces or the flexible platelets mechanisms (Figure 3) are operating.

Effect of pH on the FV of sodium montmorillonite

The FV of sodium montmorillonite by NaCl is pH-dependent and its value is 12.6 $mmol_c\,l^{-1}$ at pH 7, and 10, 13, 31, and 44 $mmol\,l^{-1}$ of NaCl at pH 5, 7.5, 8.5, and 9.8, respectively (Figure 4). The final gel volume of sodium montmorillonite samples at the above pH values is high (350–490 g of water per gram of clay) (Figure 4). The high gel volume indicates that a three-dimensional 'card house' structure is formed, even at pH 9.8, in spite of the fact that at pH 9.8 a negative charge predominates at the broken edges. The high gel volume at pH 9.8 suggests that an open structure with FF association between the platelets (at sites of low and high charge density) predominates rather than the EF card house structure.

Dispersion and flocculation in Na/Ca clay systems

Whereas single platelets tend to persist in a dilute solution of sodium montmorillonite, calcium montmorillonite platelets aggregate into tactoids or quasi-crystals. Each tactoid consists of several (4–9) clay platelets in parallel array, with interplatelet distances of 0.9 nm. A diffuse double layer exists only on the external surfaces of the tactoids. In a suspension containing a mixture of Na and Ca ions, 'demixing' of the cations occurs so that some interlayer spaces contain mainly Na ions and others mainly Ca ions. An indication of the location of the adsorbed ions in montmorillonite clay saturated with a mixture of mono- and divalent cations is obtained by measurements of the electrophoretic mobility of clay particles in suspension (Figure 5). A slight addition of exchangeable Na (e.g., 10%) to Ca-saturated clay has a considerable effect on the electrophoretic mobility of the clay (Figure 5). When the exchangeable sodium percentage (ESP) in a clay suspension is less than 40%, the Na ions concentrate on the external surfaces of the tactoids and the Ca ions concentrate on the internal surfaces.

Results of studies of the FVs of montmorillonite and illite suspensions saturated with mixtures of Na and Ca ions in the exchange phase (Figure 6) indicate that a small addition of exchangeable Na to Ca-saturated montmorillonite can have a considerable effect on the FV. The demixing model, which suggests that Na ions concentrate on the external surfaces of the tactoids

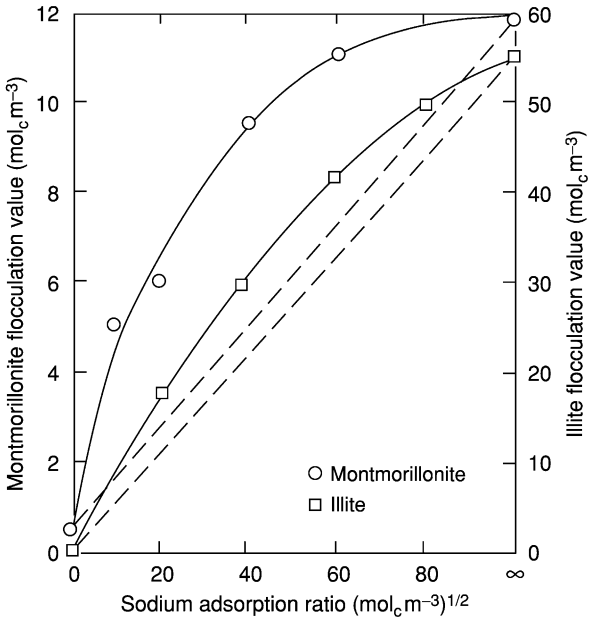


Figure 6 Flocculation value for Wyoming montmorillonite (API 25) and Fithian illite (API 35) as a function of the sodium adsorption ratio. (Reproduced with permission from Shainberg I and Letey J (1984) Response of soils to sodic and saline conditions. *Hilgardia* 52: 1–57.)

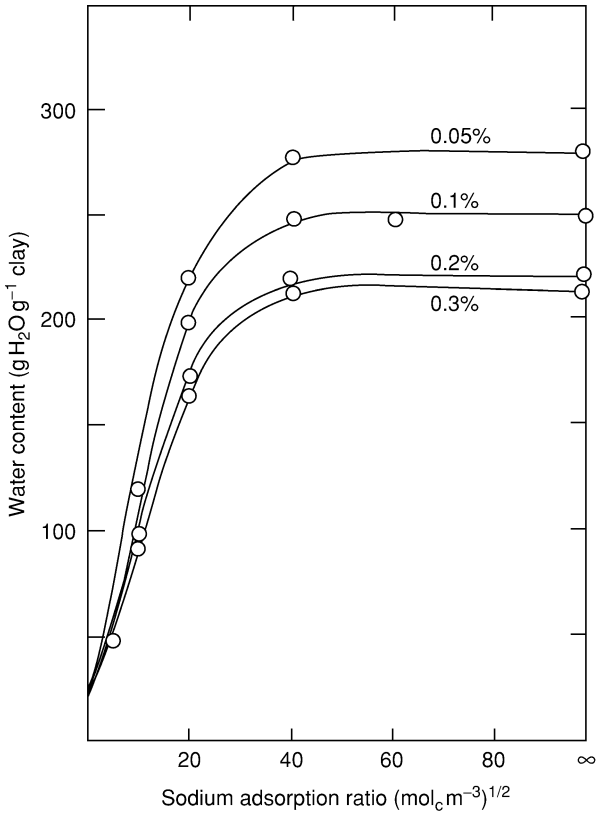


Figure 7 Water content of Wyoming montmorillonite gel as a function of Na adsorption ratio and suspension clay content. (Reproduced with permission from Oster JD, Shainberg I, and Wood JD (1980) Flocculation value and gel structure of Na/Ca montmorillonite and illite suspensions. *Soil Science Society of America Journal* 44: 955–959.)

and increase the flocculation values of the clay considerably, was applied. When the ESP of the clay suspension reaches 50, the FV of the Na/Ca mixture reaches the FV of pure sodium montmorillonite (Figure 5). Also the gel volume increases with an increase in ESP until it reaches the gel volume of pure sodium montmorillonite at ESP 50 (Figure 7). In clay suspensions with ESP of more than 50, disintegration of the tactoids is complete and single platelets predominate.

The effect of peptizing agents Sodium polymetaphosphate ($(NaPO_3)_{13}$) is a very effective peptizing agent. Organic anions may also act as peptizing agents. The FV of the peptizer-treated clay suspensions increases rapidly with increasing peptizer concentration. The anions are probably chemisorbed at the edges of the clay particles by reacting with the exposed Al. Thus, the charge of the edges becomes negative, which prevents the EF bond from taking place.

If the EF bond is eliminated by the adsorption of the polymetaphosphate, then one might expect that the open, card-house structure of the gel is also eliminated. However, an open structure with high gel volume has been obtained for the sodium gel (Table 3). This also indicates that a FF bond may lead to open gel structure.

Soil clay systems Flocculation and dispersion behavior of soil clays differs significantly from that of pure clay systems, possibly because soil clays usually occur as mixtures and because of their association with other minerals and organic matter present in the soil. A few studies demonstrate that the FV values of soil clays is two- to tenfold higher than for pure clay systems (Figure 8). Conversely, in the presence of hydrous oxides or sparingly soluble minerals such as $CaCO_3$, clay dispersivity is reported to be less severe. Hence, extrapolation from pure clay systems is of limited practical value.

Dispersion and the Hydraulic Properties of Soils

Figure 9 shows the importance of clay dispersion and movement in determining the hydraulic conductivity of soils. In this experiment a noncalcareous loam with 20% clay (montmorillonite) and a calcareous clay soil (50%, montmorillonitic) were equilibrated with a solution of 20 mmol$_c$ l^{-1} of SAR 10 to give soils with ESP 10. Following equilibration, the soils were leached with distilled water. The loam was completely sealed and the clay soil maintained a relative hydraulic conductivity of about 20% (Figure 9). This observation

contradicts the generalization that soils having a higher content of expansible clays are more susceptible to sodic conditions because of more swelling. In the calcareous clay soil, salt concentration in the soil solution was maintained at $4\,mmol_c\,l^{-1}$ by dissolution of carbonates. At ESP 10, this concentration prevents clay dispersion, and the hydraulic conductivity decrease is due only to swelling. Conversely, in the noncalcareous loam, soil solution concentration was below $1\,mmol_c\,l^{-1}$, thus clay dispersion has occurred and the dispersed clay has clogged the conducting pores.

The infiltration rate of the soil surface, exposed to rain, is more susceptible to dispersion and sodic conditions than the hydraulic conductivity of the soils

Table 3 The effect of sodium polymetaphosphate ($(NaPO_3)_{13} = 102 \times 13$) on the flocculation value and the water content of the montmorillonite gel prepared from suspensions that contained 0.1% and 0.2% clays

Peptizer concentration		*NaCl flocculation value ($mol_c\,m^{-3}$)*	*Water content of gel ($g\,H_2O\,g^{-1}$ clay)*			
			With peptizer		*Without peptizer*	
ppm	*$mol_c\,m^{-3}$*		*0.1%*	*0.2%*	*0.1%*	*0.2%*
0	0	12	240	220	240	220
1	0.01	20	260	235	200	190
10	0.1	80	170	180	150	140
100	1.0	120	160	165	130	130

Reproduced with permission from Oster JD, Shainberg I, and Wood JD (1980) Flocculation value and gel structure of Na/Ca montmorillonite and illite suspensions. *Soil Science Society of America Journal* 44: 955–959.

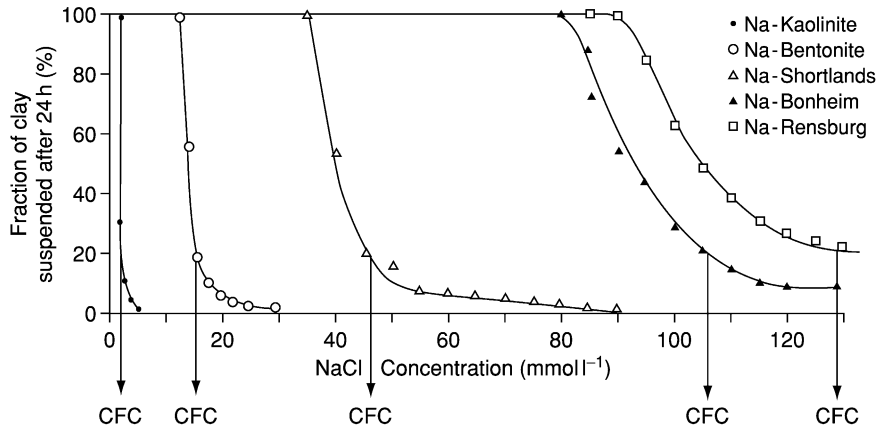


Figure 8 Flocculation behavior of Na clays suspended in NaCl solution. (Reproduced with permission from Frenkel H, Fey MV, and Levy GJ (1992) Organic and inorganic anion effects on reference and soil clay critical flocculation concentration (CFC). *Soil Science Society of America Journal* 56: 1762–1765.)

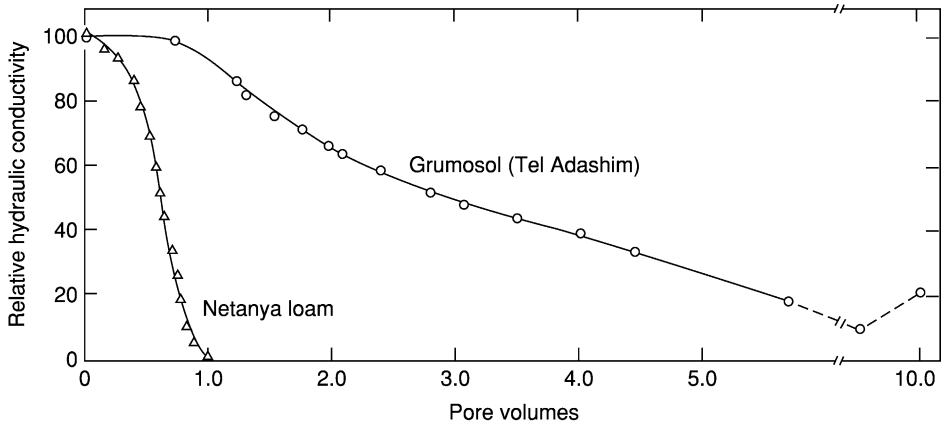


Figure 9 Hydraulic conductivity of two montmorillonite soils having an ESP of 10 and leached with distilled water. (Shainberg and Oster, 1978.) (Reproduced with permission from Shainberg I and Letey J (1984) Response of soils to sodic and saline conditions. *Hilgardia* 52: 1–57.)

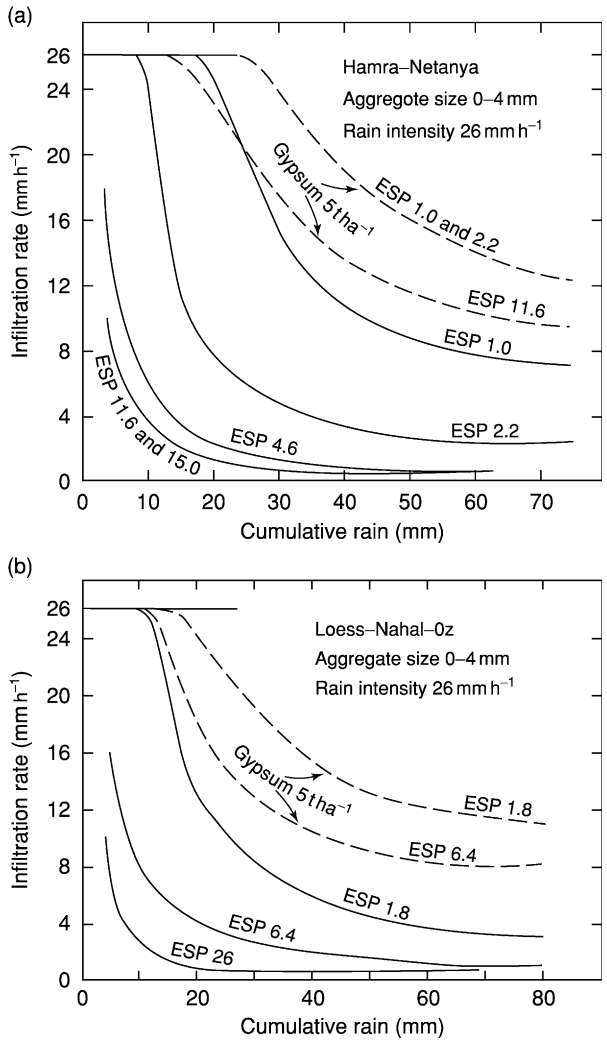


Figure 10 Effect of soil ESP (and phosphogypsum) on the infiltration rate of the (a) Hamra (Netanya) and (b) loess soil. (Adapted from Shainberg I and Lety J (1984) Response of soils to sodic and saline conditions. *Hilgardia* 52: 1–57, with permission.)

(Figure 10). Furthermore, the infiltration rates of calcareous and noncalcareous soils are equally sensitive to low levels of exchangeable sodium, whereas the hydraulic conductivity of the calcareous soils is less sensitive to low ESP than that of the noncalcareous soils. Both phenomena are explained by the clay-dispersion mechanism. At the soil surface, the concentration of the soil solution is determined solely by rain water and the low concentration of the salt enhances clay dispersion. Also, the impact of raindrops enhances clay dispersion. Both mechanisms enhance clay dispersion, aggregate disintegration, low infiltration rate, and high runoff.

Summary

The dispersion and flocculation of clays depend on the mineralogy of the clay and are affected by the composition of exchangeable cations, by the concentration of salt, by the pH, by the presence of peptizing agents, and by the application of a stirring mechanism. In a mixed Na/Ca suspension, the flocculation value of the clay increases sharply with the introduction of a small amount of sodium and with an increase in pH. Dispersion, movement, and deposition of soil colloids alter the geometry of soil pores and affect the hydraulic properties of soils. Thus, introducing a small amount of Na to soils exposed to leaching with rain water may cause clay dispersion and a deterioration of soil physical properties.

See also: **Clay Minerals**

Further Reading

Frenkel H, Fey MV, and Levy GJ (1992) Organic and inorganic anion effects on reference and soil clay critical flocculation concentration. *Soil Science Society of America Journal* 56: 1762–1766.

Goldberg S, Lebron I, and Suarez DL (1999) Soil colloidal behavior. In: Sumner ME (ed.) *Handbook of Soil Science*, pp. B195–B240. Boca Raton, FL: CRC Press.

Keren R, Shainberg I, and Klein E (1988) Settling and flocculation value of Na-montmorillonite particles in aqueous media. *Soil Science Society of America Journal* 52: 76–80.

Olson CG, Thompson ML, and Wilson MA (1999) Phyllosilicates. In: Sumner ME (ed.) *Handbook of Soil Science*, pp. F77–F123.

Oster JD, Shainberg I, and Wood JD (1980) Flocculation value and gel structure of Na/Ca montmorillonite and illite suspensions. *Soil Science Society of America Journal* 44: 955–959.

Quirk JP (1994) Interparticle forces: a basis for the interpretation of soil physical behavior. *Advances in Agronomy* 53: 121–183.

Rengasamy P and Sumner ME (1998) Processes involved in sodic behavior. In: Sumner ME and Naidu R (eds) *Sodic Soils*. New York: Oxford University Press.

Shainberg I and Kemper WD (1996) Hydration status of absorbed ions. *Soil Science of Society of America Proceedings* 30: 707–713.

Shainberg I and Letey J (1984) Response of soils to sodic and saline conditions. *Hilgardia* 52: 1–57.

Stul MS and Mortier WJ (1974) The heterogeneity of the charge density in montmorillonite. *Clays and Clay Minerals* 22: 391–396.

Van Olphen H (1977) *An Introduction to Clay Colloid Chemistry*, 2nd edn. New York, NY: John Wiley.

FLUORESCENCE SPECTROSCOPY

N Senesi and V D'Orazio, Università di Bari, Bari, Italy

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Classical monodimensional fluorescence spectroscopy in the emission, excitation, and synchronous-scan modes and total luminescence or tridimensional fluorescence are noninvasive, nondestructive, nonseparative, and sensitive techniques requiring minimal sample pretreatment. They have proven to be very promising and useful in the study of both organic and inorganic soil constituents. The variety of fluorescent structures indigenous in soil organic components, and especially in its humified fractions, i.e., humic substances (HS), mainly humic acids (HAs) and fulvic acids (FAs), has provided invaluable information on their structural and functional chemistry and reactivity as a function of relevant environmental factors such as pH and ionic strength. The capacity of metal ions to quench soil HS fluorescence has provided new insights into various aspects of metal–HS interaction/complexation processes. Fluorescence quenching and fluorescence polarization techniques have been applied successfully for mechanistic and quantitative studies of binding/adsorption of several organic pollutants to soil HS. Further, fluorescent organic probe molecules have proven to be very useful in studies of soil mineral–solution interfaces in various conditions. Recent advances in commercially available fluorescence instrumentation and in fluorescent derivatizing agents have opened up further possibilities which were previously unattainable.

Basic Principles, Methodology, and Instrumentation

The Fluorescence Phenomenon

Every molecule possesses a series of closely spaced energy levels (Figure 1). When a quantum of light impinges on a molecule and is absorbed, the electron is promoted from the groundstate, S_0, to an excited singlet state, S_1, S_2, etc. Typical transitions of this kind are very probable in complex molecular systems containing atoms with lone pairs of electrons, such as O and N, and aromatic and/or aliphatic conjugated

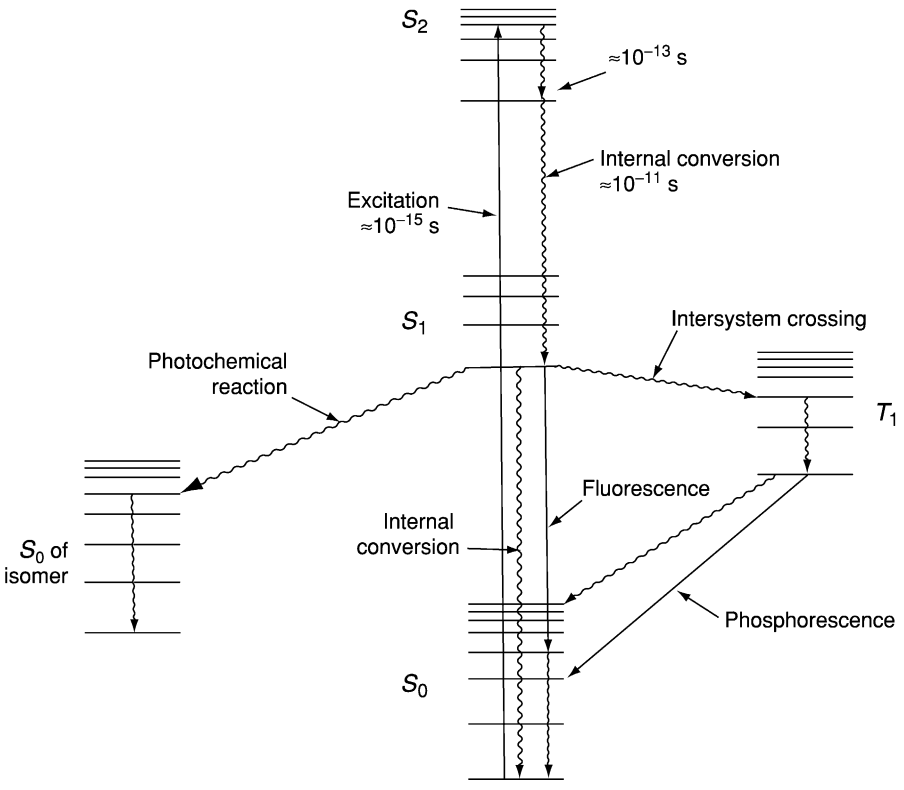


Figure 1 Schematic energy-level diagram for a hypothetical molecule.

unsaturated systems capable of a high degree of electron delocalization. Any excited state then relaxes down to the lowest vibrational level of that state, and, subsequently, energy is released by 'internal conversion,' dropping the electron to the lowest vibrational level of the lowest excited state, S_1. Relaxation from the S_1 state, which has a lifetime of the order of 10^{-9} s, to the groundstate S_0 can then occur by different mechanisms (Figure 1). These include: (1) the radiative photoprocess of fluorescence, which consists in the emission of photons less energetic than those absorbed as a result of the energy loss incurred during internal conversion processes; and (2) various nonradiative processes, such as: (a) internal conversion processes occurring by relaxation through internal vibrations or through collisions with solvent and/or solute molecules; (b) transfer of energy to another chemical compound resulting in a photochemical reaction or formation of an excited-state dimer (excimer) or excited-state complex (exciplex) which then emits photons at longer wavelengths than fluorescence; (c) intersystem crossing to a triplet state, T_1, that has a lifetime much longer (often $>10^{-5}$ s) than its S_1 precursor, which is followed by either a delayed release of energy known as the phenomenon of phosphorescence, or by relaxation through internal conversion processes.

Instrumentation and Experimental Procedures

Generally, minimal sample preparation is required for fluorescence measurements. Solutions and dilute colloidal suspensions (<1 g solid l^{-1}) can be placed directly into quartz cells, whereas the use of dense colloidal suspensions should be avoided because these result in increases in light scattering and light absorption, and a subsequent decrease in signal-to-noise ratio.

The principal components of a fluorescence spectrophotometer used for steady-state measurements are shown schematically in Figure 2. Excitation energy is provided by a high-energy continuum light source, commonly a xenon discharge lamp. The desired excitation wavelength is selected by passing incident radiation through a monochromator before entering the sample cell, preferentially a quartz cell. Radiation emitted by the sample is generally monitored at an angle of 90° to the excitation radiation to minimize signal distortion due to light scattering, or using front-surface detection. Wavelength selection of the emitted radiation is obtained with another monochromator. The emitted light is then measured by the detector, generally a photomultiplier tube. The instrument is completed by a readout device, generally a recorder output or, preferentially, a microprocessor for data analysis.

Thermostated cell holders to control sample temperature and cell stirrers can be added to most commercial

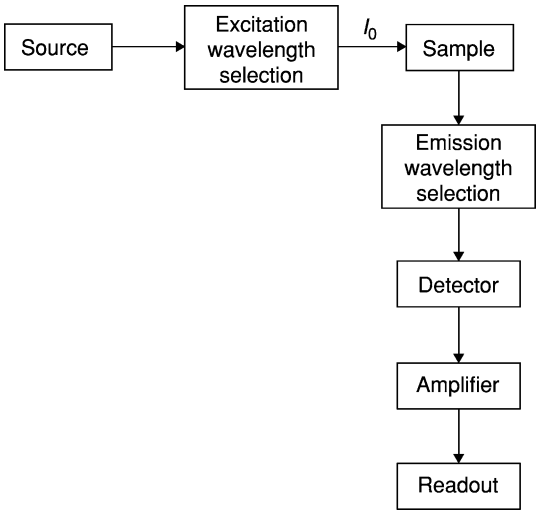


Figure 2 Schematic diagram of typical fluorescence instrumentation.

spectrofluorimeters. New instrumental developments include the use of lasers as excitation sources and the use of array or image detectors. The entire instrument can be put under computer control with several benefits in precision, calculation of quantum yields, and corrections for factors affecting fluorescence, such as the inner-filter effect and photodecomposition. Due to instrumental factors such as fluctuations in source intensity, common spectrofluorimeters provide uncorrected spectra that vary from instrument to instrument; thus a correction must be made to obtain a 'true' spectrum in terms of quanta versus wavelength. Spectrofluorimeters that include provisions for obtaining corrected spectra have been designed.

Time-resolved fluorescence spectra or fluorescence lifetime measurements, which allow the measurement of the time a molecule resides in the excited state, involve the use of pulsed or modulated light sources and a rapid-detecting photomultiplier, a photodiode array, or a time-correlated single-photon counter. Polarized fluorescence measurements can be achieved by adapting a conventional instrument by placing appropriated oriented polarizers between the excitation monochromator and the sample, and between the sample and the emission monochromator.

Fluorescence Spectra

Conventional monodimensional fluorescence spectra can be obtained in the three modes of emission, excitation, and synchronous-scan excitation on samples dissolved or suspended in aqueous media. The emission spectrum is recorded by measuring the relative intensity of radiation emitted as a function of the wavelength at a constant excitation wavelength. The excitation spectrum is recorded by measuring the emission

intensity at a fixed wavelength while varying the excitation wavelength. Synchronous-scan excitation spectra are obtained by measuring the fluorescence intensity while simultaneously scanning over both the excitation and emission wavelengths, and keeping a constant, optimized wavelength difference, $\Delta\lambda = \lambda_{em} - \lambda_{exc}$, between them. The synchronous-scan technique can selectively increase the intensity of specific peaks, thus improving the spectral structure resolution by decreasing the intrinsic broadness and band overlapping occurring in emission and excitation spectra of molecularly complex and heterogeneous multicomponent samples.

Total luminescence or tridimensional (3D) fluorescence spectra provide a complete representation of the fluorescence spectral features of a sample in the form of an excitation–emission matrix (EEM), in which fluorescence intensity is measured as a function of the excitation wavelength on one axis and emission wavelength on the other. To generate EEMs a series of emission spectra over a preselected wavelength range (generally chosen within 350–600 nm) are collected over a range of excitation wavelengths (generally chosen within 200–500 nm) by selecting an appropriate wavelength step size for both the excitation (e.g., from 4 to 10 nm) and the emission (generally, 0.5 or 1 nm). Fluorescence intensity data sets at each excitation wavelength are then converted into 3D spectra and EEMs using adequate macroprograms, and analyzed using dedicated computer graphic softwares.

The EEM spectra can be compared using the uncorrected matrix correlation (UMC) method. The UMC value equals 1 where the two matrices are identical and decreases to approaching zero with decreasing similarity between them. The recently developed fluorescence zone integration (FZI) method, which utilizes quantitatively all the wavelength-dependent fluorescence intensity data from an EEM spectrum, is capable of quantifying differences in EEM among various samples, and to measure the continuum of binding sites in metal and organic complexes of a sample.

Fluorescence Efficiency, Lifetime, and Intensity

Fluorescence efficiency, or quantum yield, Φ_f, measures the effectiveness with which the adsorbed energy is reemitted, and is defined as the ratio of the total energy emitted as fluorescence per total energy adsorbed. Absolute efficiency measurements are difficult because they require calibrated sources and detectors. However, the relative fluorescence efficiency of a sample, Φ_s, can be easily obtained by measuring the fluorescence intensity of a dilute solution (I_k) of a standard of known efficiency, Φ_k, such as quinine sulfate, and that of the sample (I_s):

$$\Phi_s = \Phi_k(I_s q_k A_k / I_k q_s A_s) \quad [1]$$

where I values are determined by integrating the area beneath the corrected fluorescence spectrum, q is the relative photon output of the source at the excitation wavelength taken directly from the curve, and A is the absorbance.

Fluorescence lifetime, τ, refers to the probability of finding a given molecule in the excited state at time t after the excitation source is turned off, and represents the mean lifetime of the excited state. The value of τ can be calculated by measuring the fluorescence intensity I at time t, $I(t)$, according to:

$$I(t) = I_M \exp(-t/\tau) \quad [2]$$

where I_M is the maximum intensity with the excitation source on, i.e., during excitation, and t is the time elapsed after the excitation source is turned off.

Fluorescence intensity, I, depends on the molar concentration of adsorbing species in solution, C, and the fluorescence efficiency, Φ_f, according to:

$$I = \Phi_f I_0[1 - \exp(-\epsilon b C)] \quad [3]$$

where I_0 is the intensity of incident radiation, ϵ is the molar absorptivity at the excitation wavelength, and b is the path length of the cell. For very dilute solutions, where ϵbC is sufficiently small, eqn [3] reduces to a linear relationship:

$$I = \Phi_f I_0 \epsilon b C \quad [4]$$

and the fluorescence intensity is essentially homogeneous throughout the sample. When ϵbC increases, the fluorescence intensity is no longer homogeneous within the cell but is increasingly localized at the front surface of the cell, thus the fraction of fluorescence measured by the common detection system is smaller (inner-cell effect). To avoid this effect, instruments are designed to look at an area in the front part of the cell rather than at the center.

Limitations and Problems

Fluctuations in intensity of the exciting light source, variations in instrumental sensitivity of the detection system, and other instrumental factors may cause distortions and variations of instrumental response. The shape of the excitation spectrum should be identical to that of the absorption spectrum of the molecule, but differences may be observed due to instrumental artifacts arising from variations in source intensity with wavelength and other causes. The shapes of emission and excitation spectra should be independent of the wavelength of the exciting

radiation and the wavelength at which fluorescence is measured, but spectra are often wavelength-dependent. In practice, however, the variations expected are likely to be small when the fixed excitation or emission wavelengths used are varied within a limited range of values, i.e., between 320 and 370 nm and 520 and 570 nm, respectively. Light-scattering effects such as Rayleigh, Tyndall, and Raman effects may also influence to a varying extent the fluorescence spectra recorded.

Corrections accounting for instrumental artifacts and for scattering effects should be applied to fluorescence spectra, particularly when quantitative comparisons are made between spectra measured on a variety of spectrofluorimeters by different operators. However, comparisons of uncorrected fluorescence spectra may be allowed on a qualitative basis when they are recorded on the same instrument using the same experimental conditions.

Molecular and Environmental Factors Affecting Fluorescence

Several molecular parameters can affect fluorescence intensity and wavelength. In particular, atoms and ions of large atomic number, such as halogens and paramagnetic metals, decrease the fluorescence efficiency of a fluorescent organic compound. Increasing the extension of the π-electron system increases fluorescence and shifts the emission wavelengths toward longer values. Steric hindrance generally reduces fluorescence, whereas increasing molecular structural rigidity has the effect of increasing fluorescence. Electron-withdrawing groups, such as carbonyls and carboxyls, largely reduce fluorescence intensity, whereas the opposite is true for electron-donating substituents such as hydroxide, alkoxide, and amino groups. Further, all these groups shift fluorescence to longer wavelengths.

A number of environmental factors can also affect the fluorescence response of a sample. For example, fluorescence efficiency is decreased by temperature increase, or by the presence of a heavy atom in the solvent molecule, whereas it is increased by increasing solvent viscosity. Fluorescence wavelength is shifted to shorter or longer values by an increase in solvent polarity. Fluorescence intensity is reduced by increasing solute–solvent interaction, e.g., by hydrogen bonding. Therefore, heavy-atom solvents and hydrogen-bonding solvents should be avoided whenever possible.

Light absorption by the solvent, large concentrations of a fluorescent solute, the presence in solution of a solute that absorbs radiation together with the fluorescent molecule, or a solute that absorbs the fluorescence of the sample molecule, give rise to the so-called 'inner-filter' effect that causes quenching of fluorescence of the sample. This effect can be reduced by selecting a suitable excitation wavelength, diluting the sample, increasing solvent viscosity, viewing fluorescence closer to the front surface of the cell, and/or using the method of standard additions or a correction factor.

The pH of the medium affects fluorescence spectra of acidic or basic aromatic compounds by rapid proton transfer reactions occurring to various extents during the lifetime of the excited state. Dissolved molecular oxygen can quench fluorescence of some organic molecules by several mechanisms, especially in polar solvents. This effect can be easily eliminated by bubbling nitrogen through the solution. Metal ions, especially paramagnetic ions, are also fluorescence quenchers, even where they do not form complexes with the fluorescent molecule. Metal quenching can be reduced markedly, e.g., by diluting the sample.

Fluorescence Quenching

The emission intensity of a fluorescent species can be quenched by its association and/or adsorption to organic and inorganic substrates. The 'simple' binding ($P-S$) between a fluorescent organic pollutant, P, and a single type of site of a sorbent, S, can be represented by:

$$P + S = P - S \qquad [5]$$

with a binding constant, K_b:

$$K_b = [P-S]/[P]\,[S] \qquad [6]$$

The mass balance on P is described by:

$$[P_{tot}] = [P] + [P-S] \qquad [7]$$

where $[P_{tot}]$ is the formal or total concentration of P in the system.

Combining eqns [6] and [7] and rearranging yields:

$$[P_{tot}]/[P] = 1 + K_b[S] \qquad [8]$$

Assuming that the fluorescence intensity is proportional to the concentration of free P in solution, $[P]$, then:

$$I_0/I = 1 + K_b[S] \qquad [9]$$

where I_0/I is the ratio of fluorescence intensities of the pollutant, P, in the absence (I_0) and in the presence (I) of a quenching sorbent S. The value of K_b can thus be readily calculated from fluorescence measurements. However, in the case where S and $P-S$ include a large number of adsorbing/complexing sites and sorbates/

complexes, a rigorous interpretation of fluorescence data is much more difficult, and only the estimation of an average binding quotient is possible.

Fluorescence Polarization

The fluorescence polarization method can be applied to study the conformation of a fluorescent organic molecule and the sorption/binding of a fluorescent organic pollutant to a variety of environmental sorbents. Any process that causes a decrease in the rate of rotation of a fluorophore, i.e., a variation in its size, during the lifetime of the excited state, including change of conformation, adsorption on suspended particulate matter, or binding to dissolved materials, produces a decrease in polarization. The polarization, p, can be calculated by the fluorescence intensities measured in the planes parallel and perpendicular to the plane of polarization of the excitation radiation, $I_{\parallel}$ and $I_{\perp}$, respectively, according to:

$$p = (I_{\parallel} - I_{\perp})/(I_{\parallel} + I_{\perp}) \qquad [10]$$

By measuring the polarization values for the free/dissolved fluorophore (p_f), the bound/adsorbed fluorophore (p_b) and their mixture (p), it is possible to obtain the ratio of concentrations of bound (F_b) and free (F_f) fluorophore, according to:

$$F_b/F_f = (\Phi_f/\Phi_b)[(p - p_f)/(p_b - p)] \qquad [11]$$

where Φ_f/Φ_b is the ratio of fluorescence efficiencies for free and bound fluorophore. The values of Φ_f and p_f can be easily determined from the fluorophore solution in the absence of sorbent material, whereas the values of Φ_b and p_b can be measured in the presence of a large excess of sorbent material so that practically all the fluorophore can be considered as adsorbed. Measuring the polarization, p, as a function of binding/sorbent titrated into a fluorophore solution at constant total concentration, or vice versa, can then provide data to construct an adsorption isotherm, and possibly calculate an equilibrium constant for binding, if stoichiometry is known or can be assumed.

Fluorescence Correlation Spectroscopy

In the fluorescence correlation spectroscopy (FCS) technique a laser light (excitation at 488 or 514 nm) is focused into the sample using confocal optics, thus realizing an open, illuminated volume element that is called 'confocal volume.' In the absence of chemical processes, temporal variations of measured fluorescence intensity in the confocal volume can be attributed uniquely to the translational diffusion of the fluorophore. These variations can be analyzed using an autocorrelation function that includes a parameter, τ_D, which is the diffusion time characteristic of the particle under study in the confocal volume. The system can be calibrated with a rhodamine 6G (R6G) standard that has a known diffusion coefficient. The values of diffusion times for R6G and the sample in each condition are then determined from the best fit of the autocorrelation function, and finally used to calculate the diffusion coefficient of the fluorescent sample.

Soil Organic Components

Fluorescence Spectra

Although fluorescent structures constitute only a minor portion of HA and FA macromolecules, fluorescence analysis has provided unique information on their structural and functional chemistry. Fluorescence emission spectra of soil HAs and FAs generally consist of a unique broad band featuring a flat maximum at a wavelength ranging from about 500 to 520 nm for HA and from about 445 to 465 nm for FA, as a function of their origin and nature (**Figure 3**). The overall fluorescence intensity of FA is generally larger than that of HA.

Fluorescence excitation spectra of soil HAs and FAs generally feature a number of distinct peaks and/or shoulders whose exact position and relative intensity depend on their origin and nature. In particular, HAs generally feature two closely spaced major peaks in the long-wavelength region (around 465 and 450 nm), often accompained by a minor peak or shoulder in the intermediate-wavelength range (at 400–390 nm) (**Figure 3**). In contrast, FAs generally exhibit a main excitation peak in the intermediate region of the spectrum (400–390 nm), with additional minor peaks and shoulders at longer and/or shorter wavelengths (**Figure 3**).

Synchronous-scan spectra of soil HAs generally feature only one peak in the long-wavelength region, often accompanied by faint shoulder(s) at longer and shorter wavelengths (**Figure 3**). Soil FAs generally exhibit two main synchronous-scan peaks at long (450–460 nm) and intermediate (390–400 nm) wavelengths, often with some less intense peaks and shoulders at both sides (**Figure 3**).

Total luminescence or EEM spectra of soil HAs and FAs generally show excitation/emission fluorescence pairs in the wavelength range of 300–400/400–500 nm, which is in agreement with monodimensional spectra. Soil HAs feature fluorescence maxima at longer excitation/emission wavelengths and are less intensely fluorescent than soil FAs. Typical EEM spectra of three reference HA and FA

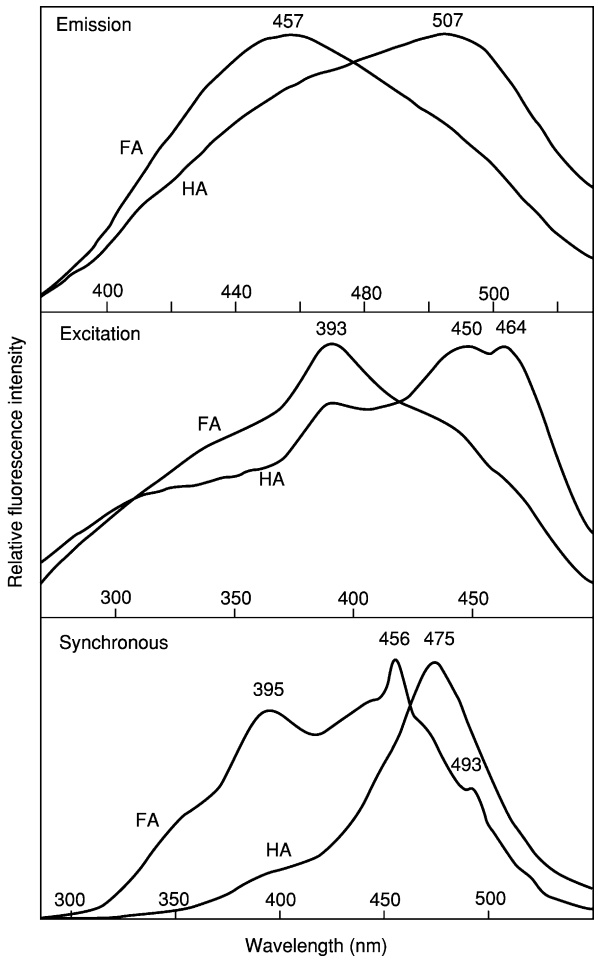


Figure 3 Fluorescence emission, excitation, and synchronous-scan spectra of the International Humic Substances Society (IHSS) reference humic acid (HA) and fulvic acid (FA) isolated from a Mollic Epipedon soil. Reproduced with permission from Senesi N, Miano TM, Provenzano MR, and Brunetti G (1991) Characterization, differentiation, and classification of humic substances by fluorescence spectroscopy. *Soil Science* 152: 259–271

samples of the International Humic Substances Society (IHSS) collection are shown in Figure 4.

The FZI method has been applied successfully to analyze EEMs of fractionated and bulk soil dissolved organic matter (DOM). Integration beneath EEM spectral regions selected on the basis of chemical criteria makes it possible to quantify DOM with similar fluorescent properties, and to interpret structural differences and the possible origin of DOM samples.

Nature of Fluorophores

Due to their molecular complexity and chemical heterogeneity, fluorescence spectra of HS probably represent the sum of the spectra of a number of different fluorophores. Although the identification of specific fluorophores in HAs and FAs is difficult, if not impossible, some hypotheses have been suggested on their possible chemical nature. In particular, the small fluorescence intensities and long wavelengths of the main fluorescence peaks of soil HAs can be attributed to large-molecular-weight (MW) components possessing linearly condensed aromatic ring systems bearing electron-withdrawing substituents, such as carbonyl and carboxyl groups, and/or to other unsaturated bond systems capable of a great degree of conjugation. Conversely, the large fluorescence intensities and short wavelengths typical of soil FAs can be ascribed to simple structural components of small MW bearing electron-donating substituents, such as hydroxyl, methoxyl, and amino groups, small degree of aromatic polycondensation, and small level of conjugated chromophores.

A number of structural components, which are known to be fluorescent as simple compounds and have been identified in HS macromolecules by

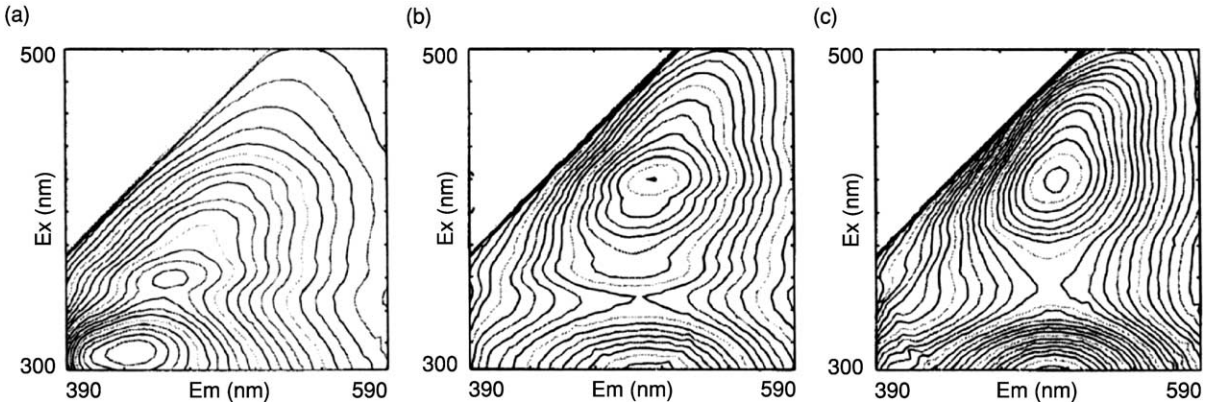


Figure 4 Absorbance-corrected excitation–emission matrix spectra of (a) a soil fulvic acid and (b and c) two soil humic acids obtained from the International Humic Substances Society (IHSS) reference collection (50 mg l^{-1}, 0.010 mol l^{-1} KCl, pH 6). Reproduced with permission from Mobed JJ, Hemmingsen SL, Autry JL, and McGown LB (1996) Fluorescence characterization of IHSS humic substances: total luminescence spectra with absorbance correction. *Environmental Science and Technology* 30: 3061–3065.

degradative methods and nuclear magnetic resonance (NMR) and infrared (IR) spectroscopies, have been suggested as potential contributors to the fluorescence of HS (Table 1). These include: (1) benzene rings bearing a hydroxyl group conjugated to a carbonyl, methylsalicylate units, and dihydroxybenzoic acid structures, such as protocatechuic, caffeic, and ferulic acids, which fluoresce at short wavelengths; (2) hydroxy- and methoxy-coumarin-like structures possibly originated from lignin, chromone, and xanthone derivatives of plant origin, and Schiff-base structures derived from polycondensation of carbonyls and amino groups, which fluoresce at intermediate wavelengths; and (3) flavonoid, naphtol and hydroxyquinoline structures that fluoresce at long wavelengths.

Molecular and Environmental Factors Affecting Fluorescence

The fluorescence intensity and spectral shape of HS can be affected by a number of molecular and environmental parameters. For example, a decrease in fluorescence intensity, shift to long wavelengths, and broadening of the emission maximum are observed with increasing MW of HS macromolecules. These effects can be related to the more extensively substituted aromatic network and greater degree of polymerization of the heavier HS fractions in comparison to the lighter fractions.

A concentration in the range 5–10 mg l^{-1} does not significantly affect the EEM spectra of HS. However, as the concentration increases from 10 to 120 mg l^{-1} at pH = 7.0 for soil FAs, fluorescence emission shifts to longer wavelengths, the overall fluorescence intensity increases and the relative intensities of synchronous-scan peaks change (Figure 5). In contrast, soil HAs fluoresce more intensely at intermediate concentrations (16–70 mg l^{-1}). These effects can be ascribed to modifications of the degree of association and/or to configuration rearrangements affecting HA and FA fluorophores.

The spectral shapes and relative fluorescence intensities can be variously affected by pH changes on dependence of the nature and origin of HS. For example, the synchronous-scan peaks of a soil FA reduce from four to two as the pH increases from 4.5 to 10 (Figure 6). Increasing the pH from 2 to 10, the fluorescence intensity increases and a red shift occurs in both the long- and short-wavelength regions of EEM spectra of some soil FAs (Figure 7). These effects can be attributed to variations in ionization of acidic groups, especially phenolic hydroxyl groups, changes in free radical concentration and/or particle association, and conformational rearrangements such as decoiling of macromolecular structures, e.g., by disruption of hydrogen bonds and consequent major exposure of functional groups to the solvent.

Ionic strength due to KCl in the range of 0–1 mol l^{-1} appears not to affect significantly the EEM spectra of HS. However, a gradual decrease in fluorescence intensity is measured for some soil HAs and FAs upon increasing ionic strength due to NaCl concentration from 0.001 m to 0.01 mol l^{-1}. This effect can be ascribed to both the gradual coiling-up of the HS macromolecular structure and the salt-depressing ionization of HS functional groups.

Increasing the temperature from 18 to 48°C causes a decrease in intensities and a shift to longer wavelengths of the major emission and excitation peaks of FAs. This effect can be ascribed to the increased deactivation of excited molecules by processes other than fluorescence, e.g., intermolecular collisions. The intensities of emission and excitation peaks of FAs tend to increase with increasing the redox potential from −1.17 to +2.5 V in solutions at small concentrations of mild reductants or oxidants.

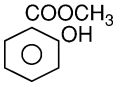
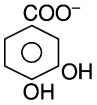
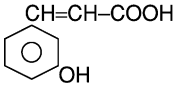
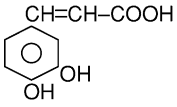
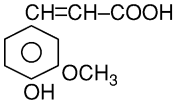
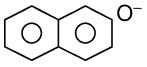
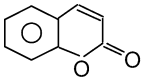
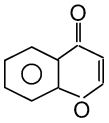
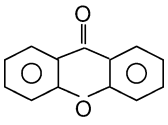
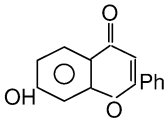
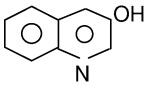
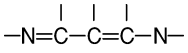
Humification Indexes and Differentiation of Humic Substances by Fluorescence Properties

Both conventional monodimensional and EEM fluorescence spectra can be used as diagnostic criteria for differentiating HA and FA from the same source and HAs and FAs of different origin, and for evaluating their humification degree. A humification index of soil DOM has been developed based on the ratio of the area under the 'upper quarter' ('H,' 435–480 nm) and the area under the 'lower quarter' ('L,' 300–445 nm) of the fluorescence emission spectrum of the sample measured at an excitation wavelength of 254 nm. Another humification index proposed for soil DOM is based on the ratio of synchronous-scan fluorescence intensity measured either at 400 and 360 nm or at 470 and 360 nm.

The humification degree of FAs of various origin has been related to their synchronous-scan spectra obtained at various pHs. Finally, a significant positive correlation has been shown to exist between the humification degree of several soil HAs, evaluated from the concentration of organic free radicals measured by electron paramagnetic resonance (EPR) spectroscopy, and total fluorescence, calculated as the area under the emission spectrum obtained at an excitation wavelength of 400 nm.

However, the classification and humification indexes developed on the basis of fluorescence properties of HS still need extensive validation on a much larger and differentiated number of HS samples before they can be considered fully applicable.

Table 1 Wavelengths of fluorescence emission and excitation maxima of potential molecular component contributors to the fluorescence of humic substances

Fluorescing molecule	*Structural formula*	λ_{em} *max (nm)*	λ_{exc} *max (nm)*
Methyl salicylate		448	302,366
Salicylic acid		410	314
3-Hydroxybenzoic acid		423	314
Protocatechuic acid (ionized)		455	340–370
3-Hydroxycinnamic acid		407	310
Caffeic acid		450	365
Ferulic acid		440	350
β-Naphtol (ionized)		460	350
Coumarins			
Unsubstituted		454	376
Hydroxy- and methoxy-		400–475	320–343
Esculetin		475	390
Scopoletin		460	390
Others disubstituted		430–462	350–419
Chromone derivatives		409–490	320–346
Xanthone and hydroxyxanthones		456 465	410 343, 365
Flavones and isoflavones		415–475	313–365
Hydroxyquinolines		450	350
Schiff-base derivatives		470	360–390

Adapted from Senesi N, Miano TM, Provenzano MR, and Brunetti G (1991) Characterization, differentiation and classification of humic substances by fluorescence spectroscopy. *Soil Science* 152: 259–271.

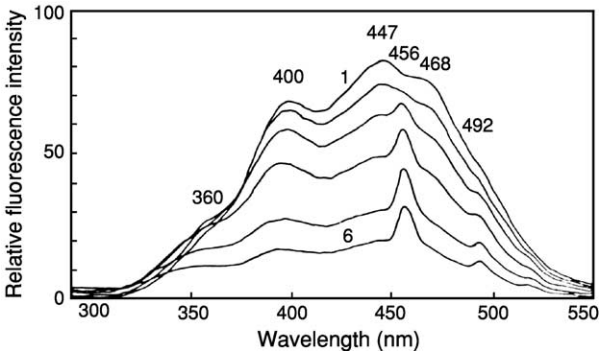


Figure 5 Effect of concentration (line 1, 120; 2, 100; 3, 75; 4, 50; 5, 25; 6, 10 mg l^{-1}) on the fluorescence synchronous-scan spectrum ($\Delta\lambda = \lambda_{em} - \lambda_{exc} = 18$ nm) of a soil fulvic acid. Reproduced with permission from Senesi N (1990) Molecular and quantitative aspects of the chemistry of fulvic acid and its interactions with metal ions and organic chemicals. Part II. The fluorescence spectroscopy approach. *Analisis Chimica Acta* 232: 77–106.

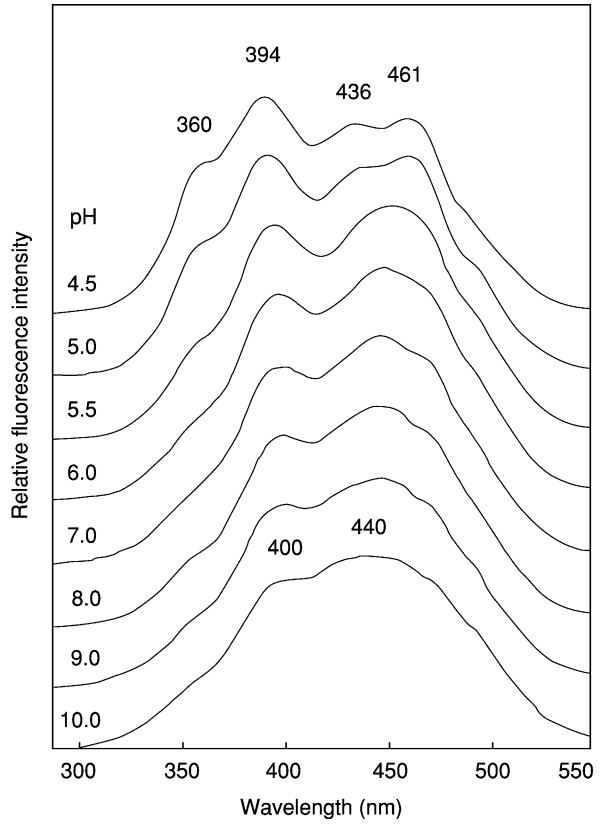


Figure 6 Effect of pH on the fluorescence synchronous-scan spectrum ($\Delta\lambda = \lambda_{em} - \lambda_{exc} = 18$ nm) of a soil fulvic acid. Reproduced with permission from Senesi N (1990) Molecular and quantitative aspects of the chemistry of fulvic acid and its interactions with metal ions and organic chemicals. Part II. The fluorescence spectroscopy approach. *Analisis Chimica Acta* 232: 77–106.

Fluorescence Efficiency and Lifetime

A fluorescence decay curve not fitting a single exponential function has been obtained for soil FAs; this indicates a range of fluorescence lifetimes as it would be expected for a mixture of fluorophores. Fluorescence efficiencies of about 0.1 are estimated for FAs from experimental lifetime data of the order of 15–20 ns. However, much smaller fluorescence efficiencies, ranging from 0.0017 to 0.0006, are measured by comparison of emission intensities of FA with quinine sulfate, which suggests the presence of only about 1% of efficient fluorophores in FAs.

The small fluorescence efficiency of FAs can also be attributed to quenching effects caused by the presence of paramagnetic metal ions and/or to a significant presence of free radicals and to other factors such as molecular size and degree of oxidation. Furthermore, fluorescence efficiency of FAs decreases with increasing MW.

Conformational Studies by Fluorescence Polarization

Fluorescence polarization is an attractive method for studying the conformation of HS because it does not require any chemical modification of the molecule, and allows measurements at concentrations close to those naturally occurring in aqueous media. The rotational relaxation time (RRT) of soil FAs obtained by this method does not change over the pH range 5–8, or at FA concentrations from 3.3×10^{-5} to 3.3×10^{-4} mol l^{-1}, or on varying the ionic strength of the medium, thus suggesting no aggregation or configuration change for FA in such conditions. The net RRT value of 2.0 ns calculated for FA based on a fluorescence lifetime of 2.1 ns is in favor of a flat extended conformation for FA. If the conformation of FA were cylindrical or rod-like, a much shorter RRT would be expected.

Values of RRT measured by time-dependent fluorescence depolarization for a soil FA in glycerol–water (40/60, w/w) solutions indicate that, at small ionic strength and FA concentration, disaggregation and molecular expansion occur for FA with increasing pH from 2.0 to 8.0, which is possibly due to the accumulation of negative charge from increasing carboxylate ionization. In contrast, RRT data obtained at low pH (2.0–4.0) and/or large FA concentration at large ionic strength are indicative of the existence of FA in an elongated conformation and of its tendency to aggregate, probably through the hydrogen bonding formation favored by the large presence of protonated acidic groups. These results confirm that FA behaves like a flexible polyelectrolyte in solutions approximating naturally occurring

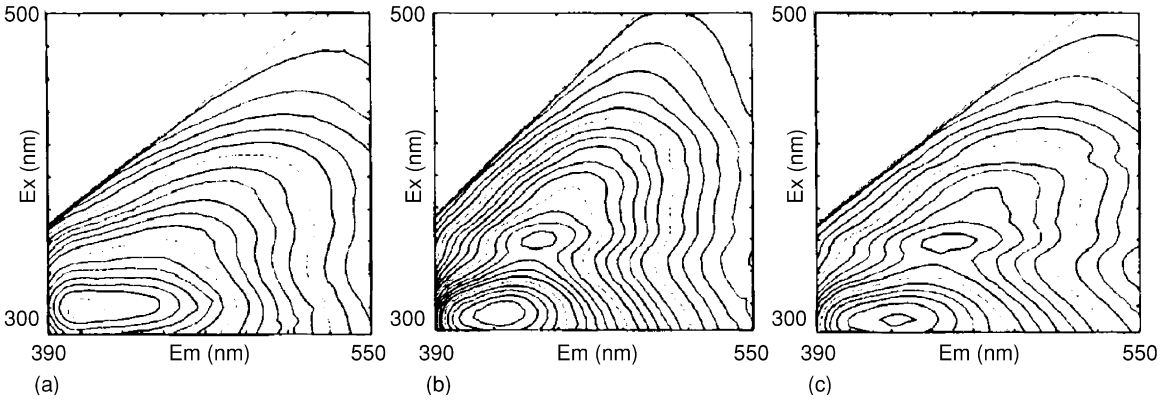


Figure 7 Absorbance-corrected excitation–emission matrix spectra of a soil fulvic acid (50 mg l^{-1}, 0.010 mol l^{-1} KCl) at pH 2.0 (a), 6.0 (b), and 10.0 (c). Reproduced with permission from Mobed JJ, Hemmingsen SL, Autry JL, and McGown LB (1996) Fluorescence characterization of IHSS humic substances: total luminescence spectra with absorbance correction. *Environmental Science and Technology* 30: 3061–3065.

concentrations. This information is very important in the solution chemistry of FA, as its structure and reactivity are markedly influenced by the conformation, i.e., molecular size and shape.

Diffusion Coefficients Determined by Fluorescence Correlation Spectroscopy

Average diffusion coefficients in the range of 2–3 $\times$ 10^{-10} m^2 s^{-1} are determined for FAs and HAs by the FCS method in a range of concentration, hydration time, pH, and ionic strength, which corresponds to hydrodynamic diameters of approximately 1.5–2.1 nm. The slight decrease in diffusion coefficients with decreasing pH from 10 to 3 suggests that some aggregation occurs with formation of dimers and trimers. The small or insignificant reduction of the diffusion coefficient measured with increasing ionic strength up to 100 mmol l^{-1} suggests the relative rigidity of HS molecules. No effect is observed as a result of changes in the concentration (1–50 mg l^{-1}) and hydration time (1–14 days) of HA and FA and of using a Ca^{2+} salt or an Na^{+} salt at an ionic strength of 5 mmol l^{-1}. For most conditions studied, FAs show a diffusion coefficient larger than that of HAs, which is in agreement with their respective molar masses.

Fluorescence Quenching of Soil Organic Components by Metal Ions

Molecular and Mechanistic Aspects

Metal ions, especially paramagnetic ions, are able to quench the fluorescence of HS by enhancing the rate of some nonradiative processes that compete with fluorescence, such as intersystem crossing. Large evidence is obtained that the 'close' association of metal ion and organic ligand, i.e., the formation of strong metal–HS complexes, is necessary for efficient quenching. Paramagnetic metal ions, such as Cu^{2+}, Fe^{3+}, Fe^{2+}, Ni^{2+}, Cr^{3+}, and VO^{2+}, can effectively quench HS fluorescence via intramolecular energy transfer by forming strong complexes. The paramagnetic ions Co^{2+} and Mn^{2+}, and diamagnetic ions such as Pb^{2+} and Al^{3+}, which form weaker complexes with HS, show a much less pronounced quenching effect, whereas cations like Cd^{2+}, K^{+}, Na^{+}, Ca^{2+}, and Ba^{2+} show no quenching effect.

Wavelength shifts of fluorescence emission maxima and/or excitation peaks are often observed upon interaction of HS with paramagnetic metal ions. For example, complexation of Cu^{2+} and Fe^{3+} to a soil FA at pH 4 and 6 causes a shift to longer wavelength (390 nm) of the 360-nm excitation peak of unreacted FA. In contrast, the excitation peaks of FA are shifted toward shorter wavelengths by FA interaction with Sn^{2+} more than with Ni^{2+}.

The fluorescence quenching of various soil FAs occurring upon complexation with Mn^{2+} ions decreases as the MW of FA increases. Synchronous-scan fluorescence analysis of three different MW fractions of a soil FA shows that the affinity for Al^{3+} and UO_2^{2+} ions of the large-MW fraction is larger than that of the small-MW fraction.

Fluorescence quenching of HAs and FAs of various origin by Cu^{2+}, Co^{2+}, Pb^{2+}, Ni^{2+}, and Mn^{2+} increases as pH increases above *c.* 3 for Cu^{2+} and above *c.* 4 for the other three ions (Figure 8). At pH 6 and 7, Cu^{2+} quenches nearly 80% of FA fluorescence, Co^{2+} and Ni^{2+} are less than half effective, and Mn^{2+} quenches about one-third of FA fluorescence. These findings are in favor of a static quenching mechanism depending primarily on the fraction of FA ligand sites complexing the metal ions, and exclude a dynamic quenching mechanism occurring by collision of FA fluorophores with the metal ions. The greater quenching ability of Cu^{2+} is attributed to its capacity

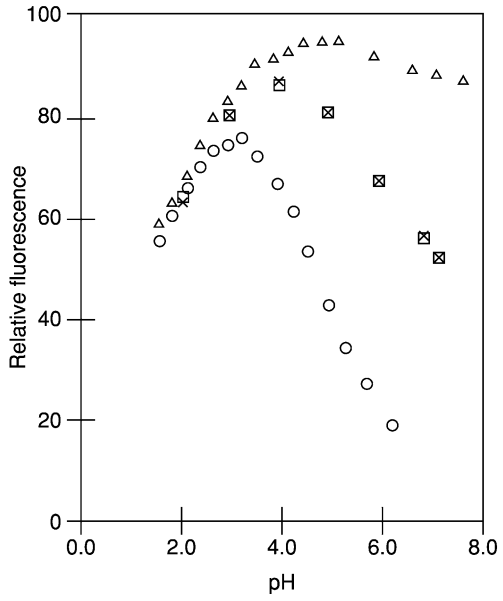


Figure 8 Fluorescence of soil fulvic acid solutions (5×10^{-5} mol l^{-1}) as a function of pH with no metal ion added (triangles) and with threefold molar excess of Cu^{2+} (circles), Co^{2+} (crosses), or Ni^{2+} (squares). Adapted from Saar RA and Weber JH (1980) Comparison of spectrofluorimetry and ion-selective electrode potentiometry for determination of complexes between fulvic acid and heavy-metal ions. *Analytical Chemistry* 52: 2095–2100.

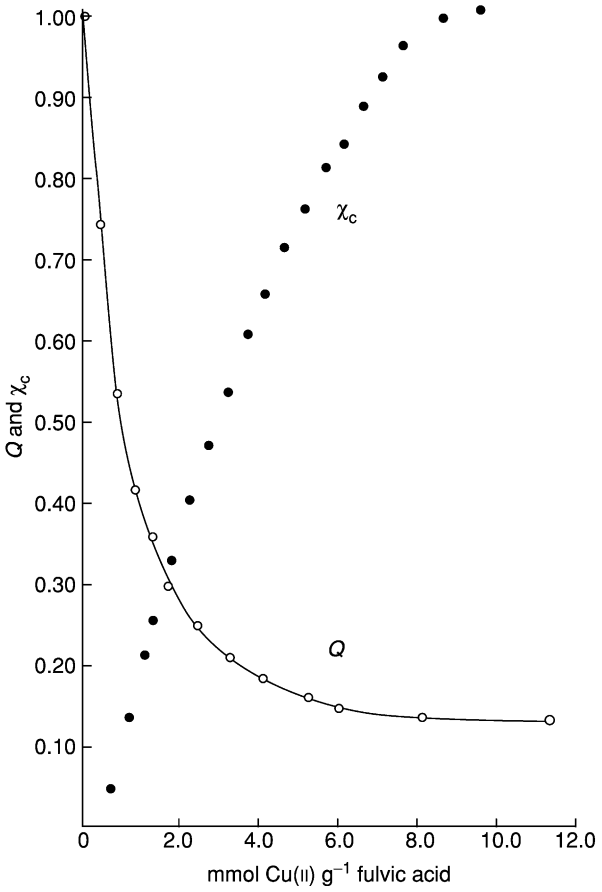


Figure 9 Comparsion of two experimental probes of metal binding to a soil fulvic acid. χ_c, increasing fractions of fulvic acid binding sites occupied by Cu^{2+} as calculated from ion-specific electrode titration; *Q*, relative fluorescence quenching of fulvic acid. Adpated from Underdown AW, Langford CH, and Gamble DS (1981) The fluorescence and visible absorbance of Cu(II) and Mn(II) complexes of fulvic acid: the effect of metal ion loading. *Canadian Journal of Soil Science* 61: 469–474.

to form strong inner-sphere complexes with both strongly acidic carboxyl and weakly acidic phenolic groups in salicylic and phtalic acid-type sites. In contrast, Mn^{2+}, and likely Co^{2+}, would form outer-sphere complexes with FA not involving weakly acidic sites.

Fluorescence quenching kinetics studies of a soil FA added with increasing amounts of Cu^{2+} suggest the occurrence of two mechanisms for binding of Cu^{2+} to FA. During the initial phase, when the Cu^{2+}-to-FA ratio increases up to 0.6, 60% of the total fluorescence quenching and a pH change from 6.4 to 5.8 are measured, indicating the occurrence of a very rapid protolytic reaction (strictly complexation). During the second phase, from a Cu^{2+}-to-FA ratio of 0.6–1.0, the reaction is slower and pH remains relatively stable, thus suggesting charge neutralization as the dominant process. Above a Cu^{2+}-to-FA ratio of 1.0, fluorescence remains almost constant, indicating that Cu^{2+} is no longer specifically bound.

Synchronous-scan fluorescence is a more revealing measurement than emission fluorescence for studying quenching of FA fluorescence by metal ions. Principal component analysis of synchronous quenching spectra obtained by difference before-and-after additions of Cu^{2+}, Pb^{2+}, Ni^{2+}, Co^{2+}, Ni^{2+}, Mn^{2+}, Mg^{2+}, and Al^{3+} at various total concentrations indicates that multiple fluorescent binding sites of FA are involved in divalent metal complexation in solutions at pH 7.5, but not at pH 5. The ion Al^{3+} is shown to compete with Cu^{2+} for FA binding sites, and to displace Mg^{2+} bound to FA. Correlation of synchronous fluorescence and time-resolved fluorescence measurements suggests that a three-component model for FA accommodates both pH dependence and metal ion quenching.

A close inverse relationship exists between the relative fluorescence quenching, *Q*, measured at the maximum emission intensity for a soil FA titrated with Cu^{2+} ions and the fraction of binding sites, χ_c, occupied by the metal ion obtained by Cu^{2+} ion-specific electrode (ISE) potentiometry (Figure 9). The quenching curve obtained for Mn^{2+} has similar characteristics but Mn^{2+}, which is about 100 times more weakly complexed than Cu^{2+}, requires about 100 times the concentration of Cu^{2+} to yield the same degree of quenching. These results confirm the

occurrence of a static, and not collisional quenching mechanism.

Comparative studies by ISE potentiometry, anodic stripping voltammetry (ASV), and fluorescence quenching show that no single site describes adequately Cu^{2+}–FA complexation over the entire pH range, but a mixed mode of coordination would occur, with the dominant binding sites varying with pH and metal-to-ligand ratio. Aromatic amino acidic groups, such as tyrosine and phenylamine moieties, and citrate and malonate groups in FA, are suggested to contribute to Cu^{2+} binding from pH 3 to 7, whereas at higher pHs, polydentate groups such as salicylates and phtalates would be mainly involved in Cu^{2+} complexation by FA.

Quantitative Aspects

Fluorescence quenching titrations of a soil FA with Cu^{2+} have provided a complexing capacity (C_L) value of FA for about 20 mol l^{-1} Cu^{2+} at pH 6 and 7, and a conditional stability constant (K_b) of the complex FA–Cu^{2+} two to three times greater at pH 7 than at pH 5 and 6. Average C_L and K_b values determined by fluorescence quenching are of the same order of magnitude as those determined by polarographic titration. At pH 7 and 6 the final fluorescence intensity is approximately 20% of the initial value for FA, and it is 40% at pH 5, which indicates that most FA fluorophores are directly involved in Cu^{2+} complexation.

With increasing Cu^{2+} concentration, the fluorescence intensity of the shorter EEM wavelength pair of a soil FA decreases more than that of the longer-wavelength pair. The K_b value evaluated assuming a 1:1 complexation stoichiometry for Cu^{2+}-FA binding is slightly larger for the shorter EEM wavelength pair than for the longer one, thus confirming the existence of different types of metal-binding sites for Cu^{2+} in FA.

Synchronous-scan fluorescence spectra of FAs added with increasing amounts of Be^{2+} at pH values from 4 to 7 indicate the presence of a fluorescent component not affected by Be^{2+} and another component resulting from the formation of relatively strong fluorescent complexes with Be^{2+} by different FA binding sites. The calculated C_L are smaller at pH 4 than at higher pH, as expected by protonation of some FA binding sites that become unavailable to complexation at low pH. The K_b values also increase from pH 4 up to pH 6, but decrease slightly at pH 7. A similar analysis applied to complexation of Cu^{2+}, UO_2^{2+}, and Al^{3+} by FA at various pH values and concentrations shows, in the case of Al^{3+}, the presence of one component corresponding to the formation of a fluorescent complex and another component corresponding to a nonfluorescent complex between Al^{3+} and FA. Calculated K_b values indicate that the stability of FA–Al^{3+} complexes decreases markedly at $pH < 4.5$.

With increasing Al^{3+} concentration, the fluorescence polarization increases for the smallest MW fraction of a soil FA and decreases for the largest MW fractions, suggesting that the former FA fraction does not change and the latter fractions change conformation upon complexation with Al^{3+} ions. The K_b values for Al^{3+}–FA complexes can be calculated using the anisotropy values obtained from polarization data. Time-resolved laser fluorescence spectroscopy has also been used to calculate K_b values of complexes formed by soil HAs with trivalent metal ions such as Cm^{3+} and Eu^{3+}.

In general, the information obtained by fluorescence quenching on metal–HS complexation is comparable to that obtained by other techniques, including ASV, ISE potentiometry, and dialysis/atomic absorption spectrometry. Although several limitations and assumptions are inherent to fluorescence data analysis, this method has several advantages over other methods in that it is relatively rapid, no separation is required between bound and free metal ion, there is no need to add supporting electrolyte or buffer or adsorbing material to the sample, neither CO_2 nor O_2 interferes with the measurement, and its sensitivity is enough for application to unmodified, natural organic ligands without preconcentration. A combined ISE potentiometry–fluorescence spectroscopy approach shows the additional advantages of avoidance of errors due to inorganic metal complexes, and applicability to environmental matrices where the metal ISE may not be suitable and to paramagnetic ions for which no ISE is available. The major disadvantage of fluorescence spectroscopy, that is, its efficiency only with strongly binding paramagnetic metal ions such as Cu^{2+}, can be, however, overcome by using fluorescent probes such as the lantanide ion probe, which is particularly promising for studies at environmental concentrations of both the metal and HS.

Sorption/Binding of Fluorescent Organic Pollutants to Soil Organic Components

Both fluorescence polarization and fluorescence quenching methods can be applied with success for molecular and quantitative studies of the binding/sorption of organic pollutants to soil organic components, and especially HS. The major advantage of these methods is that no separation step is required, which eliminates possible errors due to incomplete separation of the free/dissolved from the bound/

sorbed pollutant with consequent potential improvements in reproducibility. Furthermore, both methods are relatively rapid and convenient, and offer good precision and great inherent sensitivity, thus enabling measurements of pollutant concentration at environmental levels.

The fluorescence polarization method is preferentially applied where the pollutant molecule is an efficient fluorophore with respect to the sorbent, i.e., the HS molecule. Alternatively, the fluorescence quenching method can be applied to systems where the fluorescence of the pollutant is affected by that of HS.

Fluorescence Polarization

The fluorescence of perylene in glycerol–water (75/25, w/w) solutions at pH values of 2, 5, and 11 becomes more polarized as a function of added FA (Figure 10), which indicates that perylene binds to FA, forming a larger species that rotates more slowly in solution than does perylene alone. Polarization reaches a maximum value and then remains constant at FA levels above *c.* 3×10^{-6} mol l^{-1}, which suggests that binding is complete at this concentration. The relatively small overall change in polarization suggests that the reaction involves one perylene molecule per one FA molecule, i.e., a 1:1 stoichiometry.

Based on these assumptions, a binding equilibrium constant, K_b, can be calculated for the perylene-FA reaction:

$$\text{pery} + \text{FA} = \text{pery} - \text{FA} \qquad [12]$$

directly from polarization data by derivation of eqn [11], where the Φ_f/Φ_b term can be considered equal to 1:

$$K_b = [\text{pery} - \text{FA}]/[\text{pery}][\text{FA}] = (p - p_f)/\{(p_b - p)\{C_{FA} - [(p - p_f)/(p_b - p)]C_{pery}\}\} \qquad [13]$$

where p_f and p_b are the initial (no FA added) and final polarization values of perylene, and C_{FA} and C_{pery} are the total concentrations of FA and perylene, respectively. The K_b calculated from eqn [13] are 1.2×10^6, 1.8×10^6, and 1.5×10^6 at pH 2, 7, and 11, respectively. These values, which are expected to be considerably larger in water, confirm the existence of extended hydrophobic surface regions in FA available for association to hydrophobic organic pollutants such as perylene and other polycyclic aromatic hydrocarbons (PAHs), and suggest that this process is likely to be significant in natural systems.

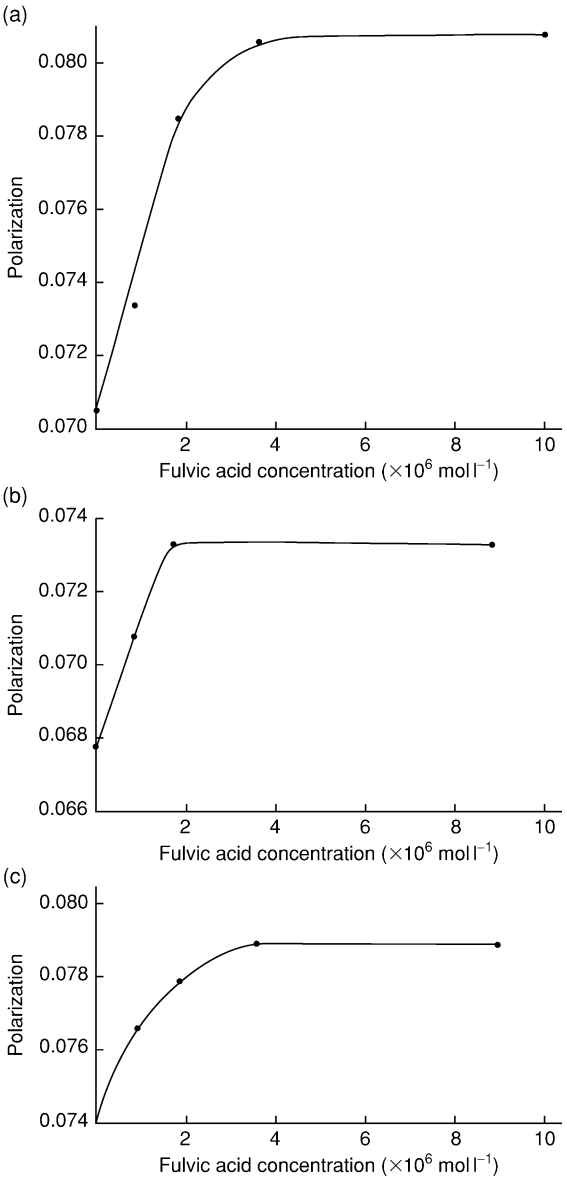


Figure 10 Perylene polarization as a function of added fulvic acid in glycerol–water (75/25, w/w) solutions of perylene (4.9×10^{-7} mol l^{-1}) at pH 2.0 (a), 7.0 (b), and 11.0 (c). Reproduced with permission from Roemelt PM and Seitz WR (1982) Fluorescence polarization studies of perylene–fulvic acid binding. *Environmental Science and Technology* 16: 613–616.

The fluorescence polarization method can be extended to nonfluorescent compounds by either measuring the changes of intrinsic fluorescence polarization of the sorbent HS molecule or performing competitive binding experiments using an adequate fluorophore probe. In the first case, the polarization changes are likely to be small and difficult to measure with precision, thus good-quality instrumentation is required, whereas in the second case, irreversible binding and slow kinetics can complicate the experiment.

Fluorescence Quenching

The rate of binding of benzo(*a*)pyrene (BaP) to a water-dissolved HA at three concentrations has been analyzed by measuring the quenching of BaP fluorescence, i.e., the increase in F_0/F ratio (eqn [9]). At any concentration, quenching is complete within 5–10 min of HA addition (**Figure 11**), which suggests a rapid association of BaP to HA and the existence of only a fast kinetic component.

The K_b values for the association of various PAHs, including BaP, perylene, fenantrene, and anthracene, with water-dissolved HAs and FAs can be derived using the linear Stern–Volmer plots (eqn [9]) obtained by measuring the fractional decrease of fluorescence intensity, F_0/F, as a function of the added HS quencher. Further analysis of fluorescence data indicates the occurrence of a static quenching process, i.e., the formation of a nonfluorescent complex between PAH and HS. The fluorescence quenching method also allows the measurement of the partition coefficients and adsorption coefficients normalized with respect to organic carbon content, K_{OC}, of compounds with large fluorescence efficiencies, such as several PAHs, to soil HAs and FAs under various pH and ionic strength conditions.

Fluorescence quenching and synchronous-scan fluorescence spectroscopy have been used to obtain conditional stability constants for the binding of carbamate pesticides to soil DOM and HA. For example, at pH 6, the binding constant of carbofuran is greater than that of carbaryl and aldicarb. These results can be used to predict the potential transport of pesticides down the soil profile.

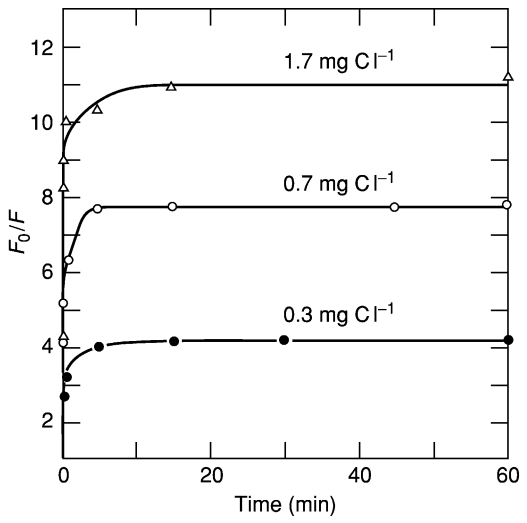


Figure 11 Time course of quenching benzo(*a*)pyrene fluorescence following the addition of humic acid (HA) at various concentrations at time zero. F_0/F is the ratio of fluorescence in the absence of HA to that in the presence of HA. Reproduced with permission from McCarthy JF and Jimenez BD (1985) Interaction between polycyclic aromatic hydrocarbons and dissolved humic material: binding and dissociation. *Environmental Science and Technology* 19: 1072–1076.

The association index (AI), formulated as:

$$\mathrm{AI} = (r_e - r_m)/r_e \qquad [14]$$

where r_e and r_m are, respectively, the hypothetical or 'expected' fluorescence anisotropy value in the absence of interactions and the experimental value of anisotropy, has been used as a quantitative parameter to measure the degree of association between various HAs and the 2,5-diphenyloxazole fluorescent probe as a function of the HA nature, HA and probe concentrations, pH, and ionic strength. Evidence has been obtained of the existence of HA pseudomicelles that are able to sequester hydrophobic species at an extent that is a function of the nature and distribution of HA functional groups, the size and flexibility of HA, and solution polydispersity and conditions. This index has been proposed as a convenient indicator of the general tendency of HAs to associate with hydrophobic species.

Soil Mineral Components

Static and dynamic fluorescence spectroscopy of organic probes sorbed on soil clay minerals can provide important and unique insights into several aspects of specific interfacial and surface behavior of clay particles in dilute colloidal dispersions/suspensions. A probe molecule is selected so that it has a great fluorescence efficiency, is sensitive at small surface loadings, and thus suitable for dilute sample suspensions, and possesses fluorescence properties depending on environmental factors such as pH and degree of molecular aggregation. Furthermore, the adsorption mechanism of the probe by the surface of interest and the specific aspects of the interface/surface that are to be probed must be considered. Common fluorescent probes include some organic cationic dyes, such as acridine orange, methylene blue, proflavine, and R6G, and various PHAs, such as anthracene, 1,2,5,6-dibenzoanthracene, perylene, BaP, pyrene, and various cationic pyrenyl derivatives.

Advantages of the fluorescence method are:

1. The possibility to study aqueous colloidal suspensions noninvasively, i.e., without the need to stop the reaction or separate the solid and liquid phases.
2. The great analytical sensitivity that makes it possible to detect very small concentrations of

sorbed probes in hydrated clay suspensions in conditions approaching those existing in natural environments.
3. The ability of one probe species to provide multiple types of information on the colloid–solution interface under study.
4. The possibility of exploring reaction rates continuously across wide time scales ranging from minutes to microseconds.
5. The study of events occurring on nanosecond time scales using time-resolved fluorescence decay measurements.

However, limitations in using fluorescence spectroscopy to study mineral systems include excessive light scattering by dense colloidal suspensions and the attenuation of fluorescence of probe species sorbed at the mineral–solution interface by quenching cations, mostly Fe^{3+} ion located within the crystal structure or on the surface of clay minerals. The use of fluorescence is therefore restricted to suspensions of clays 'Fe-free' or containing relatively small amounts of structural Fe(III).

Structural Quenching

The fluorescent probes *tris*(2,2′-bipyridine) Ru(III) and *tris*(2,2′-bipyridine) Cr(III) are more strongly quenched by absorption on to Na-saturated Wyoming montmorillonite than on to a synthetic montmorillonite containing Cr(III) in the octahedral layer and hectorite. Furthermore, Fe(III) and Cr(III) in octahedral positions are more effective quenchers than Fe^{3+} and Cr^{3+} residing on the interlayer exchange sites. Similarly, octahedral Cu(II) contained in a synthetic hectorite exhibits a greater quenching efficiency than exchangeable Cu^{2+} for the adsorbed fluorophore [4-(1-pyrenyl)butyl] trimethylammonium (PN4). These effects can be ascribed to the close association of the adsorbed fluorophore with the siloxane ditrigonal cavities on the clay surfaces.

Fluorescence quenching can also be used to determine the location of a fluorescent species at the mineral–water interface and its partitioning into mineral micelles. For example, the inorganic fluorophore Ce^{3+} is shown to occupy preferentially ion exchange sites within the tubules of imogolite. The fluorescent probe pyrene is readily quenched by both polar and nonpolar quenchers when it is located at the water–clay micelle interface, whereas it is not quenched by a polar quencher such as Cu^{2+} when it resides in the nonpolar interior of the micelle. However, pyrene fluorescence can be readily quenched by a nonpolar quencher that can also be dissolved within the interior of the micelle.

Adsorption of Organic Chemicals to Mineral Surfaces

Fluorescence polarization has been used to study the adsorption of several PAHs, including anthracene, pyrene, perylene, 1,2,5,6-dibenzoanthracene, BaP, and chrysene, on to aqueous colloidal kaolin. In all cases a distinct decrease in fluorescence anisotropy is observed as a function of PAH concentration, which implies an RRT considerably greater at small than at large concentrations. These results suggest that, at small concentrations, most PAH molecules are adsorbed on colloidal kaolin, and their motion in solution is hindered. With increasing concentration the surface saturation of the colloid is reached, and additional molecules are no longer adsorbed and remain 'free' to rotate, leading to the decrease in fluorescence anisotropy. Furthermore, adsorption kinetics studies indicate that adsorption occurs within the first few minutes of contact.

Electrofluorescence polarization spectroscopy has been used to measure the polarization of aqueous suspensions of sepiolite, attapulgite, and hectorite particles tagged with a range of dye molecules of differing sizes and charges and oriented in an applied electric field. Data obtained make it possible to infer the different possible binding sites of the dye to the clay. For example, cationic dyes are bound predominantly within the channels of sepiolite rods, whereas no effect is observed with either attapulgite or hectorite.

Surface Acidity

Fluorophores that exhibit pH-dependent fluorescence can be used to probe clay surface acidities and protonation of organic molecules at the clay–solution interface. For example, the emission spectra of quinoline on to Na-hectorite indicate the presence of surface-adsorbed quinolinium ions in the pH range 3–9. The protonation of quinoline can be ascribed to specific acidic sites at the mineral–water interface and/or to the strong ion exchange capacity of quinolinium ions. Furthermore, characteristic emission spectra of diprotonated proflavine are observed when the monoprotonated form is added to suspensions of Na-saturated montmorillonite, Wyoming bentonite, or barasym.

In some cases, fluorescence spectra can be used to infer the presence of both protonated and unprotonated species on the clay surface. For example, aminopyrene sorbed to colloidal silica exhibits both the largely structured emission spectrum of protonated aminopyrene and the broad, featureless emission of neutral aminopyrene, which indicate the presence on the surface of both species, the protonated one on the

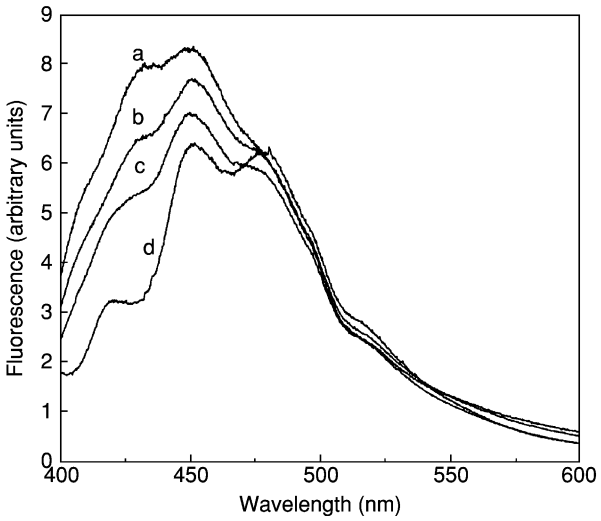


Figure 12 Effect of concentration on the emission spectra of acridine (a, 16 μmol l^{-1}; b, 65 μmol l^{-1}; c, 99 μmol l^{-1}; d, 122 μmol l^{-1}) on hectorite suspension (115 mg l^{-1}) at pH 8.8. Reproduced with permission from Traina SJ and Chattopadhyay S (1996) Spectroscopic methods of characterization surface-sorbed organic contaminants: applications of optical, electronic measurements. In: Sahwney B (ed.) *Organic Pollutant in the Environment*, vol. 8, pp. 173–197. Boulder, CO: Clay Minerals Society.

more acidic sites, and the neutral one on the less acidic sites. Fluorescence emission spectra of acridine sorbed on to Na-hectorite at a pH value well in excess of acridine pK_b (e.g., at pH 8.8) (Figure 12) show that, at very small acridine loadings, the dominant surface species is the acridinium ion, which indicates protonation promoted by the surface even at pH 8.8. At large acridine loadings, the two main components, $\tau_1 = 40$ ns and $\tau_1 = 9$ ns, of the fluorescence decay spectra of acridine adsorbed on hectorite would indicate sorption of both the acridinium ion and neutral acridine, possibly by aggregation of neutral acridine molecules on the surfaces of adsorbed acridinium ions. Regardless of the total acridine concentration, suspensions at low pH values (e.g., 4.5) result in spectra of the acridinium cation only.

Polarity of Mineral Surfaces

The great sensitivity of some fluorescent probes to the surrounding solvent polarity can provide information on the presence of polar and nonpolar regions on mineral colloid surfaces. The significantly smaller value of the ratio of fluorescence emission intensities at $\lambda = 372$ to $\lambda = 391$ nm (I/III ratio) of pyrene adsorbed to an air-dried Na-saturated laponite, with respect to that of aqueous pyrene solutions, indicates that, at small moisture contents, pyrene can be sorbed to low-polarity regions of the clay siloxanic basal planes. However, further dehydration of the clay causes an increase of the I/III ratio above the value of aqueous pyrene, which indicates the existence of an interfacial region that is more polar than bulk water.

The large I/III emission ratios of pyrene, [3-(1-pyrenyl)propyl]trimethyl ammonium (PN3), and 8-(1-pyrenyl)octyl ammonium (PN8), adsorbed on to laponite from water and methanol solutions, confirm that the charged regions of the interlamellar surfaces of clay particles are largely polar. In contrast, the very small I/III ratios observed when the fluorescent probes are sorbed on to clays coated with alkylammonium surfactants indicate that they reside in less polar environments in forms strongly associated to the adsorbed surfactants.

Organic dye molecules can also alter the apparent polarity of clay mineral surfaces. For example, the comparative analysis of the I/III emission ratios of aqueous pyrene alone and added with quinoline in the absence or in the presence of clay particles indicates the direct sorption of pyrene molecular clusters to quinolinium-coated hectorite surface.

Molecular Distribution, Conformation, Aggregation, and Mobility

Excitation spectra and fluorescence lifetime measurements show that adsorbed PN4, PN3, and PN8 cationic probes tend to cluster on the surfaces of hydrated clay minerals with formation of fluorescent dimers, i.e., excimers. The time-dependent decrease of excimer emission from the PN3 probe, which is observed upon its addition to aqueous suspensions of Na-saturated hectorite, laponite, or montmorillonite, is attributed to the initial adsorption of the probe to exterior, nonhomogeneous clay sites in the form of aggregates or hydrophobic micelles exhibiting large excimer fluorescence, followed by migration of the probe to interlayer sites with a decrease in the total amount of fluorescent excimers. The addition of Ca^{2+} ions to the above-described systems increases the excimer fluorescence intensity immediately after adsorption by inducing flocculation of clay particles and increasing the concentration of PN3 probes on the exterior surfaces. Subsequently, PN3 cations migrate into the interlayers of the more flocculated Ca^{2+}-clays, thus decreasing their mobility and excimer emission intensity.

The distribution of pyrenyl cations on clay surfaces can be altered by the presence of other organic cations such as alkylammonium surfactants. For example, PN4 cations can be 'solubilized' into the surfactant hemimicelles and form clusters that are dispersed and more homogeneously distributed on the clay surface. This results in increased PN4 mobility and the corresponding reduction of excimer emission.

Increases of the chain length of the alkylammonium surfactant produce a further decrease in excimer emission.

The effect of pH on the interfacial conformations of polymeric, polyelectrolytic macromolecules, such as pyrene-labeled polyacrylic acid (Py-PAA), can be studied by analyzing the ratio of the Py-PAA monomer emission to the excimer emission both in solution and adsorbed to alumina in aqueous suspension. When Py-PAA is adsorbed on to alumina at pH 8, and the solid is separated from the aqueous phase and then resuspended in a pH 4.8 solution, no change in the excimer emission intensity is observed. This result suggests that the polymer is adsorbed in the uncoiled, linear conformation and remains uncoiled independent of the pH value of the surrounding medium, i.e., no conformational change occurs as a function of pH. In contrast, when Py-PAA is adsorbed on to alumina at pH 4.7, and the particles are separated and then resuspended at pH 8, a decrease in excimer emission intensity is measured. This indicates the occurrence of some uncoiling of the adsorbed polymers. This means that coiled polymers loosely bound to alumina surface at low pH can undergo some conformational reorganization when exposed to a higher pH.

Fluorescence emission spectroscopy shows great promise for studying the molecular organization of other polymers, such as HAs and FAs, at interfacial mineral–solution regions. However, confirmation of fluorescence results is generally required by the use of one or more additional techniques, such as NMR.

Biochemical Transformations on Clay Surfaces

Bacterial degradation of quinoline in suspensions of hectorite has been studied by fluorescence spectroscopy. At pH values well above the pK_b of quinoline, the only fluorescent species in the system is the protonated quinolinium ion sorbed on to hectorite surfaces. The concentration decrease of adsorbed quinolinium can be monitored during its microbial degradation together with the production of the metabolite, 2-hydroxyquinoline. Comparison of degradation rates in the presence and absence of hectorite suggests that desorption of adsorbed quinolinium ion is the rate-limiting step in the bacterial degradation of quinoline.

Ligand-Promoted Mineral Dissolution

Direct evidence can be provided for the formation of fluorescent inner-sphere complexes between Al^{3+} ions and 8-hydroxyquinoline-5-sulfonate at the Al oxide–water interface. Fluorescence measurements show that the concentration of the surface complex is constant during long-term dissolution of the oxide, which indicates that the oxide surface is continuously regenerated during dissolution; this confirms a surface-controlled process of mineral dissolution.

List of Technical Nomenclature

λ	wavelength
λ_{em}	emission wavelength
λ_{exc}	excitation wavelength
$\Delta\lambda$	wavelength difference
Φ_f, Φ_s	fluorescence efficiency, or quantum yield
τ	fluorescence lifetime
q	photon output of the source
A	absorbance
ϵ	molar absorptivity
b	path length of the cell
I	fluorescence intensity
$I_{\parallel}$	fluorescence intensity in the parallel plane
$I_{\perp}$	fluroscence intensity in the perpendicular plane
p	polarization
C	molar concentration
K_b	conditional stability constant
C_L	complexing capacity
K_{oc}	adsorption coefficient normalized to organic carbon content
Q	quenching of fluorescence quantum yield
χ_c	fraction of binding sites occupied
nm	nanometer
AI	association index
ASV	anodic stripping voltammetric
BaP	benzo(*a*)pyrene
DOM	dissolved organic matter
EEM	emission–excitation matrix
EPR	electron paramagnetic resonance
FA	fulvic acid
FCS	fluorescence correlation spectroscopy
FZI	fluorescence zone integration
HA	humic acid

HS	humic substances
IR	infrared
ISE	ion-specific electrode
MW	molecular weight
NMR	nuclear magnetic resonance
PAH	polycyclic aromatic hydrocarbon
PN3	[3-(1-pyrenyl)propyl]trimethylammonium
PN4	[4-(1-pyrenyl)butyl]trimethylammonium
PN8	8-(1-pyrenyl)octylammonium
Py-PAA	pyrene-labeled polyacrylic acid
R6G	rhodamine 6G
RRT	rotational relaxation time
UMC	uncorrected matrix correlation

See also: **Clay Minerals**; **Fourier Transform Infrared Spectroscopy**; **Heavy Metals**; **Organic Matter:** Principles and Processes; Interactions with Metals; **Pesticides**; **Sorption:** Metals; Organic Chemicals

Further Reading

Gauthier TD, Shane EC, Guerin WF, Seitz WR, and Grant CL (1986) Fluorescence quenching method for determining equilibrium constants for polycyclic aromatic hydrocarbons binding to dissolved humic materials. *Environmental Science and Technology* 20: 1162–1166.

Grant CL and Seitz WR (1982) The potential of fluorescence polarization for measuring sorption isotherms of organics. *Soil Science* 133: 289–294.

Kraayenhof R, Visser AJWG, and Wolfbeis OS (2002) *Fluorescence Spectroscopy, Imaging and Probes – New Tools in Chemical, Physical and Life Sciences*. Springer Series on Fluorescence, Methods and Applications. Heidelberg: Springer-Verlag.

Lacowicz JR (1986) *Principles of Fluorescence Spectroscopy*. New York: Plenum Press.

Mobed JJ, Hemmingsen SL, Autry JL, and McGown LB (1996) Fluorescence characterization of IHSS humic substances: total luminescence spectra with absorbance correction. *Environmental Science and Technology* 30: 3061–3065.

Roemelt PM and Seitz WR (1982) Fluorescence polarization studies of perylene–fulvic acid binding. *Environmental Science and Technology* 16: 613–616.

Schulman SG (1985) *Molecular Luminescence Spectroscopy. Methods and Applications*. New York: Wiley.

Senesi N (1990) Molecular and quantitative aspects of the chemistry of fulvic acid and its interactions with metal ions and organic chemicals. Part II. The fluorescence spectroscopy approach. *Analisis Chimica Acta* 232: 77–106.

Senesi N and Miano TM (1993) Potential of fluorescence spectroscopy in the study of interactions of pesticides with natural organic matter. In: Mansour M (ed.) *Fate and Prediction of Environmental Chemicals in Soils, Plants and Aquatic Systems*, pp. 111–121. Boca Raton, FL: Lewis Publishes.

Senesi N, Miano TM, Provenzano MR, and Brunetti G (1991) Characterization, differentiation, and classification of humic substances by fluorescence spectroscopy. *Soil Science* 152: 259–271.

Sharma A and Schulman SG (1999) *Introduction to Fluorescence Spectroscopy*. New York: Wiley.

Traina SJ (1990) Application of luminescence spectroscopy to studies of colloid–solution interfaces. In: Stewart BA (ed.) *Advances in Soil Science,* vol. 14, pp. 167–190. New York: Springer-Verlag.

Traina SJ and Chattopadhyay S (1996) Spectroscopic methods for characterizing surface-sorbed organic contaminants: applications of optical, electronic measurements. In: Sahwney B (ed.) *Organic Pollutant in the Environment,* vol. 8, pp. 173–197. Boulder, CO: Clay Minerals Society.

Valeur B, Brochon J-C, and Wolfbeis OS (2001) *New Trends in Fluorescence Spectroscopy – Applications to Chemical and Life Science*. Springer Series on Fluorescence, Methods and Applications. Heidelberg: Springer-Verlag.

FOLIAR APPLICATIONS OF NUTRIENTS

M Tagliavini and M Toselli, University of Bologna, Bologna, Italy

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

It is well known that certain nutrients in gaseous form are absorbed by plant leaves. While leaves have not developed as extensively as roots as absorbing organs, in the evolution from aquatic to terrestrial plants, they have retained the ability to take up water-soluble nutrients. The nutrition of plants through aboveground organs is generally regarded as a technique to supply nutrients quickly to a target organ. This feature is particularly important for agricultural crops, which perform best when their nutrient demand is fully satisfied throughout the growth cycle.

Since most foliar fertilizers can be mixed with pesticides in the sprayer tank, foliar fertilization is often part of routine cultural practices to complement the soil-nutrient supply. In addition, for many nutrients and under specific circumstances, foliar fertilization provides nutrients to the plant with greater efficiency than soil fertilization and hence represents an important alternative to the latter. Foliar application of nutrients is thus seen as a promising technique in reducing the amount of chemicals in the environment in promoting sustainable agriculture. Several variables influence the effectiveness of foliar-applied fertilizers and may explain the often-reported inconsistency in plant response to foliar nutrition.

Efficiency of Foliar-Applied Fertilizers

In order to produce a positive effect on the mineral nutrition of plants, several requisites must be met:

1. The applied nutrients must be in an available form for leaf absorption;
2. The applied nutrients must reach the target organ (often leaves but also fruits);
3. The target organ should have a sufficient retention area so little is dispersed in the environment;
4. The internal leaf structure, environmental conditions, and nutrient chemistry should be such that nutrients either in ionic or in nonionic form can penetrate the leaf cuticle and enter the apoplast of the cells of the target organ;
5. The given nutrient should be transportable to the target tissues if different from the intercepted organ.

Meeting these requirements, however, does not necessarily improve plant performance. Positive responses to foliar as well as to soil fertilization is expected only when there is a need for any given nutrient: if nutrient availability exceeds nutrient demand, no response to foliar fertilizers is likely.

There are a number of situations in which foliar nutrition can effectively replace or complement (Table 1) soil fertilization. Because the amount of nutrients to be supplied through the leaves is limited by the fact that, above a certain threshold concentration, foliar sprays can be toxic, foliar fertilization can meet plant nutrient demand if the requirement for a given nutrient is low, as in the case of microelements (boron, copper, iron, manganese, molybdenum, and zinc) or when soils already have adequate fertility, so that only small additional amounts are needed. Multiple foliar applications may be needed to ensure adequate nutrient delivery.

There is general agreement that foliar fertilization isan efficient method to prevent or overcome transient nutrient deficiency over the vegetative and

Table 1 Major agronomic and physiology-related conditions promoting the effectiveness of foliar nutrient supply

Objective	*Example*
Prevent or cure transient nutrient deficiency	When nutrient demand exceeds root uptake rate
Bypass soil conditions limiting nutrient availability or uptake and accelerate plant response	Unfavorable pH and conditions for root growth, including soil temperature, moisture, and aeration
Increase the buildup of nutrient reserves for storage and successive remobilization the following year	It applies to perennial plants and to nutrients that are phloem-mobile such as nitrogen
Cure or prevent nutrient deficiencies of specific organs that occur despite optimal root uptake due to problems in within-plant distribution of a given nutrient	It applies to nutrients such as Ca which move mainly through the transpiration stream toward the leaves, leading to calcium-related disorders in fruits, e.g., in apple and in tomato
Supply nutrients to plants with deep root systems when soil-surface application of fertilizers is almost without effect	It applies to trees, especially for nutrients such as K, Ca, or Mg, whose penetration to root zone layer is sometimes difficult and in general to conditions of absence of irrigation or rain

reproductive season for both macro- and micro-nutrients. These situations occur when demand temporarily exceeds root uptake capacity for a given nutrient. In highly productive plants, for example, during specific stages of the reproductive cycle, frequently leaves suffer from a temporary nutrient deficiency. Seeds and fruits withdraw from vegetative parts significant amounts of nitrogen, phosphorus, and potassium (with absolute amounts depending on the crop) at a rate higher than that supplied by root uptake, which often declines as a result of decreasing carbohydrate supply to the root system. Transient nutrient deficiency also occurs in perennial plants during spring, when remobilization of nutrients from storage sites is complete and root uptake is not sufficiently efficient (e.g., due to low soil temperature, excess moisture, or poor root growth) to meet plant requirements: transient deficiency for macronutrients such as nitrogen or micronutrients such as boron often compromises fruit set when occurring during anthesis.

Foliar application is often preferred to soil fertilization when soil constraints such as pH or calcium carbonate limit the availability of nutrients such as zinc, iron, and manganese or when deep-rooted perennial plants are grown in soils with high fixing capacity for nutrients such as potassium. The addition of iron salts to an alkaline or calcareous soil is usually ineffective, as soil mineral iron (often Fe oxides) is poorly available at high pH. For high-value agricultural commodities, the Fe deficiency problem is often overcome by soil application of expensive, synthetic Fe chelates, but foliar application of iron salts may be almost equally effective and represents an economical alternative where growers cannot afford Fe chelates to cure or prevent iron deficiency. It must be stressed, however, that if Fe deficiency persists for long periods, and because Fe mobility in the phloem is very low, repeated foliar applications may be needed during active stages of shoot development to satisfy the Fe requirement of developing leaves. Leaf application of zinc and manganese, often as sulfate salts, is frequently the only way to cure or prevent deficiencies in alkaline soils, because these micronutrients are rapidly fixed when applied to the soil and become unavailable to plants. If applied in high-pH soils, zinc and manganese chelates are, moreover, not stable, as in the case of ethylenediaminetetraacetic acid (EDTA) and diethylenetriaminepentaacetic acid (DTPA), or have little affinity to Mn and Zn, as ethylenediamine-di(*o*-hydroxyphenylacetic) acid (EDDHA). The rationale for preferring foliar application to soil fertilization in optimizing micronutrient nutrition also includes the fact that, in some types of soils, soil supply may increase their availability from deficiency to excess levels, producing undesired effects.

One typical example of the advantage of foliar over soil nutrient application is related to calcium (Ca) nutrition in fruit plants. Because this element moves within the plant almost entirely in the xylem sap via the transpiration stream, it tends to accumulate preferentially in organs whose transpiration rate is high, such as the leaves, with only limited partitioning to fruits. Expanding fruits have a relatively low demand for calcium, but, in some species of marked horticultural interest, low levels of Ca in the fruit can lead to Ca deficiency-related symptoms that compromise fruit quality and storage life, as exemplified by bitter pit in apple and black end-rot in tomato. As these disorders also occur in soils well supplied with Ca and in situations where adequate concentrations of Ca are found in leaves, soil applications are of generally little or no value, and prevention and cure have to rely on direct application of Ca-based products on the fruit.

Application of Foliar Nutrients

The amount of nutrients that can enter the plant through the foliage is a function of the amount of nutrients retained by the leaves, which in turn depends on (1) the total leaf area that can be wetted by a sprayed solution and (2) its nutrient concentration. It is therefore unlikely that, before a crop reaches a desired leaf area index (LAI: the ratio between total crop leaf area and soil area), significant quantities of nutrients can be provided to the plant by foliar nutrient application.

Aqueous-based sprays are generally inefficient in wetting and spreading over waxy leaf and fruit surfaces due to high surface tension (at 20°C, surface tension at the air–water interface is $72.8\,\mathrm{mN\,m^{-1}}$). Because foliar nutrients are often included in standard pesticide sprays, they take advantage of the formulation and tank additives included with the pesticide; when they are applied alone, however, surfactants are often added in the tank to reduce surface tension and increase the wetted leaf area. Different volumes of spray solution are applied under field conditions depending on the crop and cultural practicies: for example, in fruit tree plantations, sprayed volumes often vary from 150 to $1500\,\mathrm{l\,ha^{-1}}$. If low volumes are adopted, e.g., to reduce the number of tank fillings needed to spray a given crop acreage, the concentration of solution needed to provide the same amount of nutrient per unit of area is higher than when high spraying volumes are adopted.

When the sprayed solution reaches the leaf, the impacting of spray droplets results either in reflection or retention. When reflected, a droplet may be intercepted by another leaf or fall to the soil. Reflection depends in part on properties of the wetted surface:

while no significant reflection occurs from both abaxial and adaxial leaf surfaces and fruits of peach (*Prunus persica*), the epicuticular wax of plum (*P. domestica*) fruits are known to reflect marked amounts of sprayed solution. As leaf and fruit surfaces vary widely in morphology and fine structure (Figure 1), leaf retention can vary from 5 to 30 μl cm^{-2}. Reducing surface tension of the sprayed solution, while increasing leaf wetting, may reduce retention and increase spray runoff if high spray volumes are applied: the addition of surfactants therefore does not always ensure high efficiency of leaf-nutrient penetration.

Only a fraction of the nutrient retained in the aqueous solution on a leaf's surface after spraying is immediately absorbed by the leaf before drying; the remainder dries out and forms a deposit on the leaf. Penetration also takes place from the deposit, depending on the physiochemical characteristics of the deposit and relative humidity of the microenvironment. The size and distribution of the deposits are important features affecting the nutrient's chances of being taken up later. Nutrient salts have a different point of deliquescence (POD, relative humidity level that causes dissolution of salt residue; Table 2) but are in general hygroscopic, so, once air humidity is higher than POD, the residue will be resolubilized in water and the nutrients again have a chance to penetrate the leaf. Penetration rate, however, declines over time as nutrient concentration in the water solution outside the leaf also declines as a result of uptake. By the same token, the residue is exposed to the risk of runoff if it rains.

Mechanisms of Penetration of Foliar-Applied Nutrients

Nutrient penetration of the leaf surface greatly depends upon leaf structure and chemical composition (Figure 2). The barrier to penetration of foliar-applied substances is mainly represented by the several layers of the cuticle. The outer layer, or cuticle proper, is made up almost entirely of cutin embedded with epicuticular wax, a highly hydrophobic leaf-surface component whose primary function is preventing uncontrolled water losses. Leaf absorption is regulated by the amount, distribution, and chemical composition of such waxes. The presence of cutin, which contains polar groups such as polyesterified hydroxy fatty acids, makes waxes more hydrophilic. The cuticle inner layers are made up of hemicellulose (polyuronides and glycans) pectinaceous substances, proteins, and polysaccharide microfibrils that may form a continuum from the outer surface to the epidermal cell wall, like a conducting channel for polar solutes. Cutin and occasionally waxes are also present in the inner layers. The cuticle is separated from the

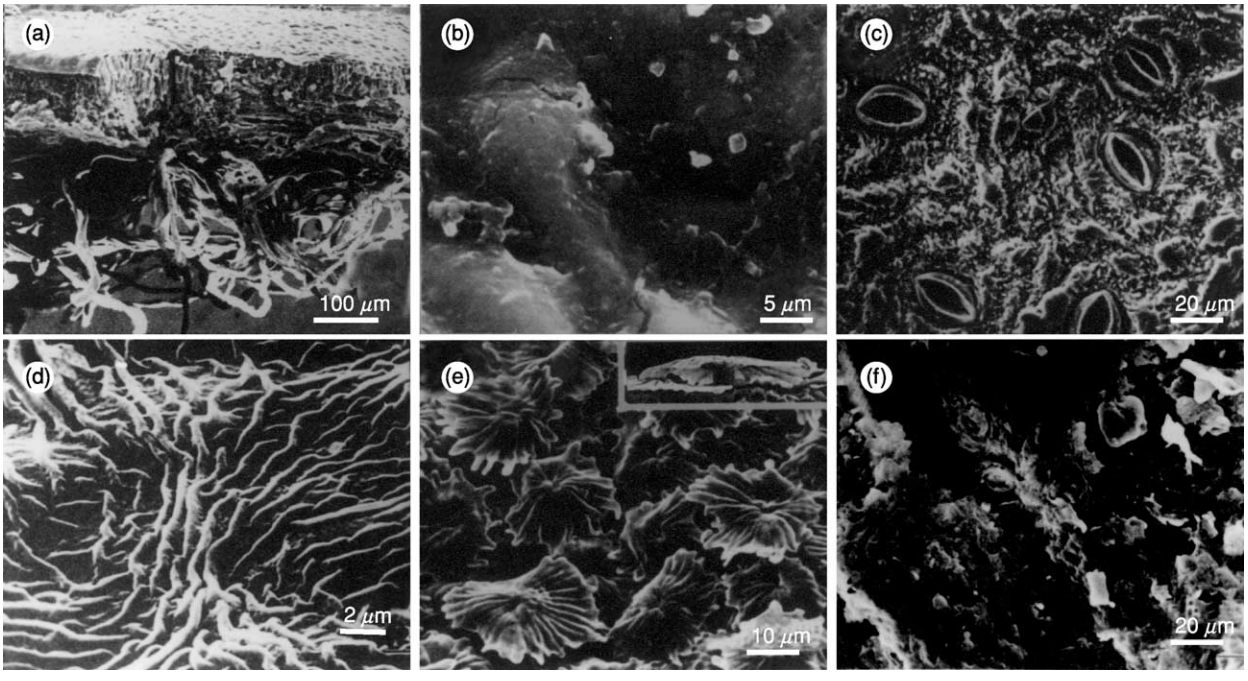


Figure 1 Scanning electron micrographs of the adaxial leaf surfaces of (a) apple and (b) pear; abaxial surfaces of (c) pear, (d) sweet cherry, and (e) olive, with a trichome as inset; and (f) the fruit surface of grapefruit. (Reproduced with permission from Bukovac JM, Cooper JA, Whitmoyer RE, and Brazee RD (2002) Spray application plays a determinant role in performance of systemic compounds applied to foliage of fruit plants. *Acta Horticulturae* 594: 65–75.)

Table 2 Physical properties of selected inorganic salts for foliar nutrition

Salt	*POD (%)*	*Solubility (g kg H_2O^{-1})*	*Molecular weight (g mol^{-1})*
$CaCl_2 \cdot 6H_2O$	33	2790	219
$MgCl_2 \cdot 6H_2O$	33	1670	203
$K_2CO_3 \cdot 2H_2O$	44	1469	174
$Ca(NO_3)_2 \cdot 4H_2O$	56	6600	236
$Mg(NO_3)_2 \cdot 6H_2O$	56	1250	256
NH_4NO_3	63	1183	80
KCl	86	344	75
K_2HPO_4	92	167	174
KH_2PO_4	95	33	136
KNO_3	95	133	101
Calcium propionate $\cdot H_2O$	95	490	204
Calcium lactate $\cdot 5H_2O$	97	31	308
Calcium acetate	100	374	158
$FeCl_3 \cdot 6H_2O$	44	919	270
$Fe(NO_3)_3 \cdot 9H_2O$	54	1500	404
$Mn(NO_3)_2 \cdot 4H_2O$	42	426	251
$MnCl_2 \cdot 4H_2O$	60	1510	180
$ZnNO_3 \cdot 6H_2O$	42	1843	297

POD, point of deliquescence.

Reproduced with permission from Schönherr J (2002) Foliar nutrition using inorganic salts: laws of cuticular penetration. *Acta Horticulturae* 594: 77–84.

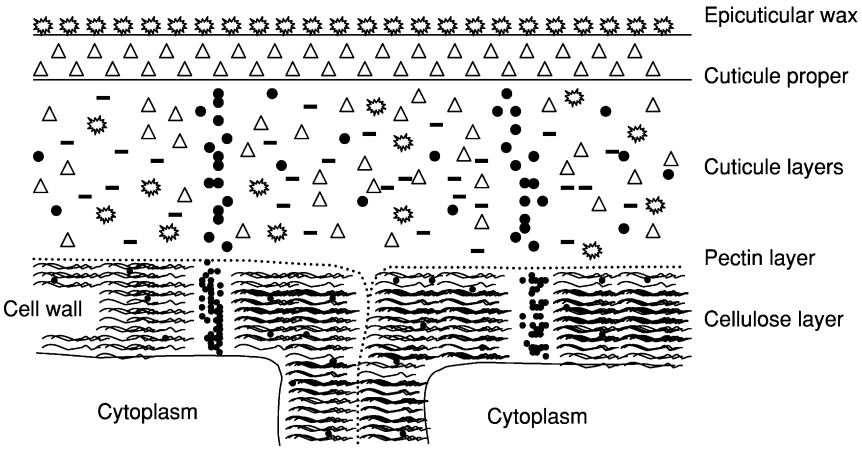


Figure 2 Leaf surface is a mix of wax (jagged symbols), cutin (triangles), pectin (dots), and cellulose (dashed line) arranged in different layers. The hydrophilic fraction (pectin) may form pores that allow a faster penetration of water-soluble compounds.

underlying epidermal cell wall by a pectic layer. Because of its chemical composition, the cell wall does not represent a major barrier to nutrients.

Penetration of substances through the cuticle is a diffusive process. Sprayed water solutions adhering to the cuticles have a concentration higher than the water that saturates the apoplastic space, and this concentration gradient represents the main driving force promoting diffusion into the leaf. The mechanisms of cuticle penetration of inorganic salts differ from those of apolar (nonionic) compounds: most foliar-applied fertilizers are inorganic salts, but nonionic products such as urea and synthetic chelates of metal micronutrients or weak acids such as boric acid are also on the market as foliar fertilizers. The penetration of nonionic compounds is best described as diffusion through the cuticle itself: solute mobility decreases with increasing molar volume and increases with temperature, which enhances the fluidity of amorphous cuticular waxes.

Since inorganic salts are strong electrolytes, when dissolved in water they are highly dissociated and surrounded by water molecules that form hydration shells. Nutrients deriving from inorganic salts are thus highly soluble in water but practically insoluble in lipidic phases such as cutin and cuticular waxes. Unlike nonionic compounds, direct ion diffusion through the cuticle does not occur; hydrated ions require aqueous diffusion paths across the cuticle that are provided by the aqueous pores formed by the hydration of polar groups in the cutin (COOH-, OH-, and ester groups). The radius of aqueous pores

varies depending on plant species and external conditions: mean values of less than 1 nm and densities in the range of 10^{10} pores cm^{-2} are reported.

Aqueous pores swell or shrink depending on humidity levels, being more open for penetration of water and solutes when high levels of humidity cause cuticle swelling. The diffusion of ions through the cuticle is inversely related to the radius of the hydrated ions. While most studies on cuticular penetration have been carried out using astomatous cuticular membranes, cuticular penetration from peristomatal regions is particularly efficient due to the higher permeability of cuticle over stomata in comparison with astomatous areas. In addition, stomata guard and accessory cells often present an abundant density of channel pores, a fact explaining, at least in part, the positive correlation between penetration rate and stomata density.

Nutrients may also enter the leaf directly through the stomata, but the relative importance of stomata for total leaf nutrient penetration is still a matter of debate. In many plant species, trichomes also represent areas where the entry of nutrients is facilitated because of their low cutinization. Accumulation of sprayed nutrients underneath fruit lenticels has also been reported, suggesting that lenticels represent a preferential penetration pathway in some fruits.

Factors Affecting Cuticular Penetration

Cuticular penetration of nutrients is mainly affected by three groups of factors: the genotype of the plant, the environment, and the application techniques. Among genetic factors, leaf anatomy and chemical composition of leaf surface, a peculiarity of different plant species, influence droplet reflection. Generally speaking, reflection is positively related to the level of epicuticular waxes and adversely related to the number of hairs, which accounts for the fact that young leaves are usually more efficient in spray retention and in absorbing nutrients than old leaves. In particular, surface microstructures leading to microroughness is the critical factor in reducing reflection. Leaves of species having stomata only on abaxial leaf surfaces show a preferential uptake of nutrients from this surface rather than from adaxial surfaces.

Environmental factors affect leaf cuticle absorbtion of nutrients in different ways. Air humidity improves cuticular penetration of hydrated ions by reducing the hydrophobic properties of the cuticle surface, by causing cuticle swelling, by delaying droplet drying, and by maintaining deposits in hydrated form and/or redissolving the deposit of salts. As salts differ in their POD (Table 2), in windy and dry areas salts with low POD are preferred for foliar application.

Light also increases rates of leaf nutrient uptake, an effect depending on the high permeability of cuticular membranes over the guard cells of the open stomata. High light intensity, however, is not a favorable condition for foliar uptake, as it may increase cuticle thickness and the amount of cuticular waxes, as demonstrated in different species of *Brassica* and *Prunus* and in some cereals; this might also explain the higher specific absorption rate recorded in spring than in summer. While not affecting ion penetration, increasing temperature, within a certain range, promotes the movement of apolar solutes through cuticle, an effect depending on the fluidity of waxes.

The addition of surfactants to the sprayed solution improves leaf wetting by lowering the surface tension, thereby increasing the contact between the liquid and the leaf surface. A liquid having a surface tension lower than 15–20 $mN\,m^{-1}$ penetrates spontaneously through leaf stomata. Surfactants able similarly to reduce surface tension show varying effectiveness, depending on other intrinsic properties: in the case of $CaCl_2$ (the most widely adopted salt for preventing Ca-related physiological disorders in fruits), wetting agents such as alkyl-polyglucosides at a concentration of 0.2 $g\,l^{-1}$ improve the penetration rate by a factor of 12, whereas protein-based surfactants and sodium EDTA (a synthetic chelate) reduce the penetration of $CaCl_2$. In some cases, surfactants may be toxic also, as has been demonstrated for fluorocarbon polymers and octylphenoxy polyethoxyethanol.

For nonionic foliar fertilizers, the rate of penetration is increased by plasticizers (such as tributyl phosphate and diethyl sebactate), which cause an increase in fluidity of amorphous cuticular waxes, an effect similar to that of increasing temperature.

The penetration velocity of ions reaching the leaf surface is usually highest initially then tends to level off over time. Experimental studies carried out with isotopes, which allow a high accuracy of determination, indicate that the half-time of penetration of $^{45}Ca(NO_3)_2$ (6 $g\,l^{-1}$) in pear leaf disks is 18 h, while most salt-derived Ca is absorbed in less than 100 h. In the case of N, urea is taken up faster than other N salts (i.e., KNO_3) because of its apolar nature. The penetration half-time of urea-^{15}N in fruit tree leaves (such as apple, grapefruit, and red raspberry), is approximately 30–50 h, and as much as 90% of the retained urea-N is found in leaves after 5 days. The effectiveness of foliar penetration also depends on salts: for example, apple leaves absorb P according to the following array: $H_3PO_4 > K_2HPO_4 > NaH_2PO_4 > KH_2PO_4 > Ca(H_2PO_4)_2$.

The pH of the sprayed solution affects the rate of penetration depending on plant species and spray solute. A large range of pH (2–10) has been investigated:

Table 3 Compounds often used for foliar fertilization with correspondent application rates

Nutrient	*Compound*	*Nutrient concentration (%)*	*Compound rate (kg ha^{-1})*	*Compound concentration (g l^{-1})*
B	Boric acid (H_3BO_3)	10	3–12	0.6–1.5
	Polyborate ($Na_2B_4O_7$)	20	3–6	0.6–1
Ca	$CaCl_2 \cdot 2H_2O$	27	2–5	2–3.5
	$Ca(NO_3)_2 \cdot 4H_2O$	15.5	4–8	5–6
Fe	$FeSO_4$	37	0.75–2	0.5–1.5
	FeDTPA/EDTA	4–6	3–5	3
Mg	$MgSO_4 \cdot 7H_2O$	9.6	7.5–15	5–12
	$MgCl_2 \cdot 6H_2O$	25	1.5–4.5	1–3
	$Mg(NO_3)_2$	16	4.5–9	3–6
Mn	$MnSO_4H_2O$	32.5	1–3	0.5–2
N	Urea ($CO(NH_2)_2$)	46.6	2.5–15	2–40
	NH_4NO_3	26–27	4	2–4
	KNO_3	13	7.5	5
	$Ca(NO_3)_2 \cdot 4H_2O$	13	7.5	5
P	NH_4PO_4	26.6	0.5–1	0.5
	KH_2PO_4	22.7	1–1.5	0.6
K	KH_2PO_4	29	10	8–10
	KNO_3	38.7	7.5	5–8
Zn	$ZnSO_4$	58	0.3–2	1–2
	$ZnCl_2$	50	0.4–2.4	1–2
	ZnEDTA	2.5	0.6–1	0.6–0.8

DTPA, diethylenetriaminepentaacetic acid; EDTA, ethylenediaminetetraacetic acid.

low pH values are considered optimal for metal ions (Zn), intermediate values (5.4–6.6) for urea, while high values (7–10) are the optimum for potassium phosphate.

The volume of spray solution (and consequently the concentration of the applied nutrients) can affect the rate of cuticular penetration. Generally, higher volume (i.e., 1000–1500 l ha^{-1}) increases the wetting period and delays the time of deposit formation. Low volume may be preferred when surfactants are added to the spray solution; in this case the nutrients retained will have a higher concentration and the deposits will be more uniformly distributed on the leaf. Experimental data for commercial apple trees show a higher uptake rate of urea-N at a concentration of 2 g of urea l^{-1} compared with 20 g l^{-1} and 40 g l^{-1}. A list of selected compounds suitable for foliar application is reported in Table 3.

Plant Use of Foliar-Applied Nutrients

Once nutrients have penetrated the cuticle and reached the apoplast of the external mesophyll cells, they must cross the plasma membrane, because, with some exceptions, it is within the cytosol that they play their role in plant metabolism. Crossing the plasma membrane for nonelectrically charged molecules such as urea is likely to be a simple diffusion, driven by the higher apoplastic concentration. In the cytosol, the urea is quickly hydrolyzed by urease and the resulting ammonium metabolized in the glutamate–oxoglutarate aminotransferase (GOGAT) pathway.

When a mineral ion derived from foliar application reaches the free space of mesophyll cells, it presumably mixes with nutrients already present derived from internal storage remobilization or root uptake. Anions and cations cross the plasma membrane via transporters or via channels. The uptake of a given ion into the cell lowers its concentration in the apoplast and increases the concentration gradient across the cuticle, resulting in higher nutrient penetration.

Some physicochemical characteristics of the apoplast are very important for entering the cytosol: iron, for example, is mainly present in the apoplast in its oxidized state, Fe(III), while it crosses the plasma membrane only after being reduced to Fe(II), a process greatly affected by pH and the organic acid concentration of the apoplast; if the iron is not reduced, it accumulates in the apoplast. The effectiveness of a foliar-applied iron compound is thus highly dependent upon the conditions affecting leaf uptake. Lowering the pH of a spray solution makes the inactivated iron pools in the apoplast available for uptake and may cause regreening of iron-deficient leaves.

The fate of foliar-applied nutrients that are mobile in the phloem is not limited to the leaves, where the absorption process takes place, as other plant organs such as reproductive or storage structures benefit from foliar supply. Nutrients vary as to their phloem mobility, being higher for nitrogen (as amino acid), K, and P than for Fe, Ca, and Mn. Phloem mobility also depends on plant characteristics: for boron, traditionally considered phloem-immobile, a significant phloem mobility is found in plants using sugar

alcohols, such as sorbitol in apple, for phloem transport of C. The amount of foliar-absorbed nutrients subsequently exported to other plant organs may differ according to the physiological stage of growth and the presence of major sinks. In perennial deciduous plants, when leaves represent sources of N and storage organs represent major sinks, a significant proportion of N (80–87% in cherry and 50–60% in apple and peach) derived from foliar sprays is transported to the rest of the tree before leaf senescence occurs, and most of this can be recovered in winter in the root system. Nitrogen absorbed by leaves from sprays applied in the spring and summer is exported at varying rates depending on the N demand of newly developing leaves and reproductive organs.

List of Technical Nomenclature

μl cm^{-2}	Leaf retention
degrees Celsius	Temperature
mN m^{-1}	Surface tension
percentage relative humidity	Point of deliquescence

See also: **Calcium and Magnesium in Soils**; **Fertilizers and Fertilization**; **Iron Nutrition**; **Macronutrients**; **Micronutrients**

Further Reading

Brown P (2001) Transient nutrient deficiencies and their impact on yield – a rationale for foliar fertilizers? *Acta Horticulturae* 564: 217–223.

Bukovac JM, Cooper JA, Whitmoyer RE, and Brazee RD (2002) Spray application plays a determinant role in performance of systemic compounds applied to foliage of fruit plants. *Acta Horticulturae* 594: 65–75.

Eichert T and Burkhardt J (2001) Quantification of stomatal uptake of ionic solutes using a new model system. *Journal of Experimental Botany* 52: 771–781.

Franke W (1967) Mechanism of foliar penetration of solutions. *Annual Reviews of Plant Physiology* 18: 281–300.

Haynes RJ and Goh KM (1977) Review on physiological pathways of foliar absorption. *Scientia Horticulturae* 7: 291–302.

Mengel K (2002) Alternative or complementary role of foliar supply in mineral nutrition. *Acta Horticulturae* 594: 33–47.

Oland K (1960) Nitrogen feeding of apple trees by postharvest urea sprays. *Nature* 185: 857.

Schönherr J (2002) Foliar nutrition using inorganic salts: laws of cuticular penetration. *Acta Horticulturae* 594: 77–84.

Swietlik D and Faust M (1984) Foliar nutrition of fruit crops. *Horticultural Reviews* 6: 287–355.

Tagliavini M, Millard P, and Quartieri M (1998) Storage of foliar absorbed N and remobilisation for spring growth in young nectarine (*P. persica* var *nectarina*) trees. *Tree Physiology* 18: 203–207.

Toselli M, Thalheimer M, and Tagliavini M (2002) Leaf uptake and tree partitioning of urea-N as affected by concentration and volume of sprayed solution and leaf age of apple trees. *Acta Horticulturae* 594: 591–594.

Weinbaum SA (1988) Foliar nutrition of fruit trees. In: Neumann PM (ed.) *Plant Growth and Leaf-Applied Chemicals*, pp. 81–100. Boca Raton, FL: CRC Press.

Weinbaum SA, Brown PH, and Johnson RS (2002) Application of selected macronutrients in deciduous orchards: physiological and agro technical perspectives. *Acta Horticulturae* 594: 59–64.

FOOD–WEB INTERACTIONS

P C de Ruiter, University of Utrecht, Utrecht, The Netherlands
J C Moore, University of Northern Colorado, Greeley, CO, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Soils harbor a large part of the world's biodiversity. Microorganisms such as bacteria and fungi are by far the most dominant groups of soil organisms, in terms of diversity, numbers, and biomass, but there is also a large variety of micro- and mesofauna such as protozoa (amebae, flagellates, ciliates), nematodes (bacterivores, fungivores, omnivores, herbivores, and predators), microarthropods, e.g., mites (bacterivores, fungivores, predators) and collembola (fungivores and predators), enchytraeids, and earthworms. Trophic interactions among all these soil organisms are known to govern major components in the global cycling of materials, through the decomposition of soil organic matter and the mineralization of nutrients. As these processes determine the availability of nutrients that are limiting to plants, the belowground food web interactions also influence aboveground primary productivity and carbon sequestration. In addition

to their roles in shaping the dynamics and persistence of many populations within the soil community, patterns in the strength of the trophic interactions are known to be crucial for the stability of communities and hence the maintenance of biological diversity.

Trophic Structures in Soil Food Webs

Because of the large diversity of soil organisms, soil food-web descriptions are often constructed by adopting the functional group approach. In this approach, species are aggregated into functional groups according to properties that are important to understand their role in ecosystem processes, such as principal food source, mode of feeding, reproduction rate, defenses against predation, or distribution in the soil profile. Such grouping of species into functional units has been supported by analyses of soil food webs, indicating that the behavior of organisms within functional groups seems more similar than that of organisms within taxonomic units, i.e., taxonomy is not necessarily related to function. An example of a functional group description of a soil food web is given in **Figure 1**, referring to the belowground food web from winter wheat fields in a crop rotation system at the Lovinkhoeve experimental farm in the Netherlands. The soil food-web diagram shows two microbial populations and six general trophic categories of belowground fauna: predators, fungivores, bacterivores, detritivores, herbivores, and omnivores (i.e., consumers of animal prey and other resources), making up five trophic levels: (1) detritus and primary production (roots); (2) primary decomposers and herbivores: bacteria, fungi, phytophagous nematodes; (3) consumers of bacteria and fungi: collembola, fungivorous mites and nematodes, protozoa and bacteriphagous nematodes, and mites; (4) intermediate predators: protozoa, predacious nematodes and predacious collembola; and (5) top predators: predacious mites.

During the last decades of the twentieth century, several large, multidisciplinary research programs provided estimates of the population sizes of functional groups from agricultural soils as well as from native and forest soils (**Tables 1** and **2**). These data show some interesting differences in food-web structure between agricultural soils and native and forest soils. In most agricultural soils, the population sizes in the soil food webs are organized in the form of trophic biomass pyramids (**Table 1**), while in the forest soils high abundances may occur at the higher trophic levels (**Table 2**). Across trophic levels, belowground food webs are organized into compartments of species/functional groups, distinguished on the basis of primary energy source: (1) primary producer (root) channel (I), and (2) the detritus channel (II),

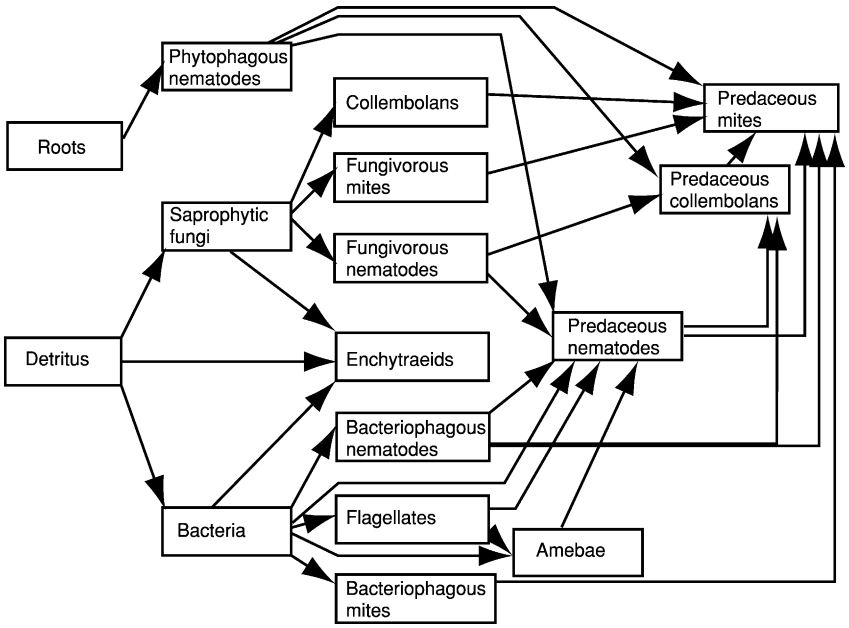


Figure 1 The belowground food web from a winter wheat fields at the Lovinkhoeve experimental farm in the Netherlands. Species are aggregated into functional groups, i.e., based on food choice and life-history parameters. 'Detritus' refers to all dead organic material. Material flows to the detrital pool, though the death rates and the excretion of waste products are not represented in the diagrams, but are taken into account in the material-flow calculations and stability analyses. (Reproduced with permission from De Ruiter PC, Moore JC, Zwart KB *et al.* (1993) Simulation of nitrogen mineralisation in belowground food webs of two winter wheat management practices. *Journal of Applied Ecology* 30: 95–106.)

Table 1 Biomass estimates (kilograms of C per hectare) for the functional groups in agricultural food webs. Values refer to the 0- to 25-cm-depth layer, except for the Horseshoe Bend webs (0–15 cm)

	LH-IF	*LH-CF*	*HSB-NT*	*HSB-CT*	*KS-B0*	*KS-B120*
Microbes	245	228	440	690	740	900
Fungi	3.27	2.12	160	150	1500	2300
Protozoa[a]	20	12	40	50	110	34
Nematodes						
Herbivores	0.4	0.2	0.4	0.5	0.1	0.3
Bacteriovores	0.4	0.3	0.5	1.4	0.5	0.5
Fungivores	0.1	0.1	0.1	0.1	0.2	0.1
Predators	0.1	0.1			0.4	0.4
Arthropods						
Fungivorous collembola	0.38	0.47	0.3	0.1	0.2	0.2
Fungivorous mites	0.05	0.03	1.7	0.5		
Predatory collembola	0.008	0.03				
Predatory mites	0.09	0.06	0.2	0.04	0.2	0.3

[a]Includes amebae and flagellates.
LH, arable land at the Lovinkhoeve (NL); IF, integrated farming; CF, conventional farming; HSB, arable land at the Horseshoe Bend Research Site (Georgia, USA); CT, conventional tillage; NT, no-tillage; KS, Kjettslinge (Sweden); B0, arable land without fertilizer; B120, arable land with fertilizer ($120\,kg\,ha^{-1}\,yr^{-1}$).
(Reproduced with permission from De Ruiter PC, Van Veen JA, Moore JC, Brussaard L, and Hunt HW (1993) Calculation of nitrogen mineralisation in soil food webs. *Plant and Soil* 157: 263–273.)

Table 2 Biomass estimates (for CPER and NS, SS, D, and F in kilograms of C per hectare and for NL in micrograms per gram) for the functional groups in natural (prairie and forest) food webs. Values for CPER and NS/SS/D/F refer to the 0- to 10-cm-depth layer

	CPER	*NL-L*	*NL-F*	*NL-H*	*NS*	*SS*	*D*	*F*
Microbes								
Bacteria	304	290	340	264	27	25	19	62
Fungi	63	64	33	14	437	310	385	426
Protozoa[a]	4	130	11	5	11	22	46	30
Nematodes								
Herbivores	3	0.2	0.1	0.02				
Bacteriovores	6	8	1	0.5	0.7	0.14	0.3	0.4
Fungivores	0.4	2	0.2	0.2	0.2	0.07	0.1	0.1
Predators	1	1	0.2	0.1	0.2	0.07	0.2	0.1
Arthropods								
Fungivorous collembola	0.5	73	4	1	2	1	8	7
Fungivorous mites	4	82	18	3	1	2	4	2
Predatory collembola		176	11	1	0.01	0.1	0.1	0.5
Predatory mites	0.4	62	25	7	0.1	0.3	0.7	0.4

[a]Include amebae and flagellates.
CPER, shortgrass prairie; NL, coniferous in Wekerom in the Netherlands; NL-L, litter layer; NL-F, fragmented layer; NL-H, humus layer; NS, SS, D, F, coniferous forest in north Sweden, south Sweden, Germany, and France, respectively.
(Adapted from De Ruiter PC, Van Veen JA, Moore JC, Brussaard L, and Hunt HW (1993) Calculation of nitrogen mineralisation in soil food webs. *Plant and Soil* 157: 263–273; and Schröter D, Wolters V, and de Ruiter PC (2003) C and N mineralisation in the decomposer food webs of a European food transect. *Oikos* 102(2): 294–308.)

where the latter can be split into IIa the bacterial energy channel and IIb the fungal energy channel. In some agricultural soils, the bacterial energy channel dominates, for example the winter wheat fields at the Lovinkhoeve in the Netherlands; in other soils, the fungal channel dominates, for example the barley fields at the Kjettslinge farm in Sweden. Also the forest soils from the Netherlands are bacterial dominated, while the soils from Sweden, Germany, and France are fungal dominated. Both aspects of trophic structure, the pyramid of biomass and the compartmentalization of interactions and energy flows into energy channels, are intimately related to the stability of the system.

Trophic Interactions and Soil Ecosystem Processes

Major soil ecosystem processes such as the decomposition of the soil organic matter and the mineralization of nutrients are the direct result of feeding and processing of material in the soil food web. Because of

their dominant numbers (Tables 1 and 2), microbes are usually considered to be the most important contributors to the soil processes, but the faunal groups of organisms are also considered to contribute considerably to the soil processes, despite their relatively low densities. The contribution of the fauna to the cycling of energy, matter, and nutrients depends not only on their own population sizes, or rates of energy use, but also on their ability to influence the functioning of the organisms with which they interact. These influences may occur through a variety of mechanisms. Some faunal groups such as microarthropods break up organic matter by feeding on detritus and by ingesting microbes and adhering detrital material, resulting in an increase in the surface area of the organic matter available for microbial attack. The dissemination of microbes is another potentially important influence of the soil fauna on decomposition processes. Nematodes and microarthropods can passively transport bacteria, fungi, and protozoa in the gut or on the cuticle across regions of soil that are otherwise impenetrable to the microbiota or at greater rates of dispersion than would otherwise be effected. Also in this way, faunal groups can act as catalysts by bringing primary decomposers into contact with substrate. By feeding on microbes, the fauna may affect microbially mediated degradation of the soil organic matter by altering the composition of the microbial community structure through selective grazing, or by affecting microbial growth and metabolic activities by altering nutrient availability. The predatory fauna may also affect ecosystem processes indirectly by decreasing bacterial-feeding nematode populations and subsequently allowing bacterial populations to increase.

Food Web Models

In order to analyze the role of soil food webs and functional groups of soil organisms in soil ecosystem processes, a Detrital Food Web Model has been developed wherein C and N mineralization rates are derived from the feeding rates among the trophic groups using basic elements of the life histories and behaviors of the organisms, and the laws of thermodynamics and conservation of matter. The model was originally developed to analyze N flow rates in the food web from the shortgrass prairie at the Central Plains Experimental Range (Colorado, USA); later it was applied to other soil food webs (see legends to Tables 1 and 2).

The model calculates the feeding rates by assuming equilibrium, i.e., annual growth rates of the populations balance the annual natural death rates and the death rates due to predation:

$$F_j = \frac{d_j B_j + P_j}{a_j p_j} \quad [1]$$

where F_j is the feeding rate of group j, d_j is the specific death rate of group j, B_j is the annual mean population size of group j, P_j is the death rate of group j due to predation, a_j is the assimilation efficiency of group j, and p_j is the production efficiency of group j. If a predator is considered to feed on more than one prey type, then both the preference of the predator for a given prey and the relative population sizes of the prey types are taken into account:

$$F_{ij} = \frac{w_{ij} B_i}{\sum_{k=1}^{n} w_{kj} B_k} \quad [2]$$

where F_{ij} is the feeding rate by group j on prey i, w_{ij} is the preference of predator j for prey i relative to other prey types, and n is the number of potential prey types. The model starts calculating the feeding rates of the top predators, because their losses only consist of natural death. Next, the feeding rates of the groups one trophic level below the top predators, whose losses consist of natural death and death due to predation by the predators, are calculated. This procedure is continued working back throughout the web to the lowest trophic levels. C and N mineralization are derived from the feeding rates:

$$C_{ij} = a_j(1 - p_j)F_{ij} \quad [3]$$

$$N_{ij} = a_j\left(\frac{1}{CN_i} - \frac{p_j}{CN_j}\right)F_{ij} \quad [4]$$

where C_{ij} is the C mineralization rate by group j feeding on group i, N_{ij} the N mineralization rate by group j feeding on group i, and CN_i and CN_j are the C to N ratios of group i and j, respectively. The model requires data for population sizes, specific death rates, energy conversion efficiencies, preference weighing factors, and C to N ratios. Population sizes are to be established for the particular sites, while most values for the physiological parameters can be derived from the literature, sometimes in combination with adjustments based on the specific environmental conditions. The model provides estimates of the C and N mineralization per trophic interaction, per trophic group and of the community as a whole. In some cases the calculated overall mineralization rates can be compared with observed rates.

The outcome of the calculations in terms of their contribution to N mineralization rates (Tables 3 and 4) show a similar pattern to the population sizes

Table 3 N mineralization rates (kilograms per hectare per year) of the functional groups in the different food webs. Values refer to the 0- to 25-cm-depth layer, except for the Horseshoe Bend webs (0–15 cm)

	LH-IF	*LH-CF*	*HSB-NT*	*HSB-CT*	*KS-B0*	*KS-B120*
Microbes						
Bacteria	52	40	97	137	37	20
Fungi	1	0.8	15	13	19	24
Protozoa[a]	57	35	115	144	54	16
Nematodes						
Herbivores	0.2	0.1	0.1	0.1	0.004	0.06
Bacteriovores	0.9	0.9	0.9	2	2	2
Fungivores	0.08	0.06	0.05	0.03	0.3	0.2
Predators	0.06	0.07			0.3	0.3
Arthropods						
Fungivorous collembola	0.2	0.2	0.1	0.03	0.4	0.4
Fungivorous mites	0.02	0.01	0.6	0.2		
Predatory collembola	0.01	–				
Predatory mites	0.03	0.02	0.08	0.01	0.1	0.2

[a]Include amebae and flagellates.

(Reproduced with permission from De Ruiter PC, Van Veen JA, Moore JC, Brussaard L, and Hunt HW (1993) Calculation of nitrogen mineralisation in soil food webs. *Plant and Soil* 157: 263–273.)

Table 4 N mineralization biomass estimates (for CPER and NS/SS/D/F in kilograms C per hectare per year and for NL in micrograms per gram per year) for the functional groups in natural (prairie and forest) food webs. Values for CPER and NS/SS/D/F refer to the 0- to 10-cm depth layer

	CPER	*NL-L*	*NL-F*	*NL-H*	*NS*	*SS*	*D*	*F*
Microbes								
Bacteria	45	−120	−20	−10	−11	−24	−14	−32
Fungi	8	150	30	5	2	13	46	30
Protozoa[a]	12	300	20	10	10	37	71	61
Nematodes								
Herbivores	1	0.1	0.1	0.1				
Bacteriovores	13	30	3	1	0.3	0.1	0.1	0.2
Fungivores	0.2	2	0.1	0.1	0.01	0.00	0.00	0.00
Predators	0.4	2	0.4	0.1	0.01	0.00	0.00	0.00
Arthropods								
Fungivorous collembola	0.2	0.1	0.1	0.1	0.2	0.2	0.7	0.7
Fungivorous mites	0.8	10	3	0.3	0.06	0.2	0.2	0.3
Predatory collembola	0.003	–	–	–	0.00	0.00	0.03	0.2
Predatory mites	0.1	100	30	10	0.03	0.1	0.2	0.1

[a]Include amebae and flagellates.

(Adapted from De Ruiter PC, Van Veen JA, Moore JC, Brussaard L, and Hunt HW (1993) Calculation of nitrogen mineralisation in soil food webs. *Plant and Soil* 157: 263–273; and Schröter D, Wolters V, and de Ruiter PC (2003) C and N mineralisation in the decomposer food webs of a European forest transect. *Oikos* 102(2): 294–308.)

(Tables 1 and 2). While in agricultural soils, bacteria are usually the most important contributors to N mineralization; in forest soils the model indicates N immobilization by bacteria, due to the low N content of soil organic matter in these soils (Table 4). The calculated N mineralization rates per functional group confirm the expectation that the contribution of faunal groups can be high relative to their contribution to the soil biomass. For example, while microbes may constitute more than 90% of the total soil biomass, the contribution of the soil fauna to N mineralization can exceed that of the microbes (Figure 2).

For some soil food webs, the calculated overall mineralization rates can be compared with observed rates. This verification shows that the model calculates C and N mineralization rates close to the observed rates for the winter wheat fields at the Lovinkhoeve farm and for the barley fields at the Kjettslinge farm (Table 5). For the fields at the Horseshoe Bend research site, the calculated C mineralization rates are close to the observations, but the calculated N mineralization rates are higher than the observed rates. The verification of C and N mineralization rates in the coniferous forest layers in Wekerom in the Netherlands reveals that the

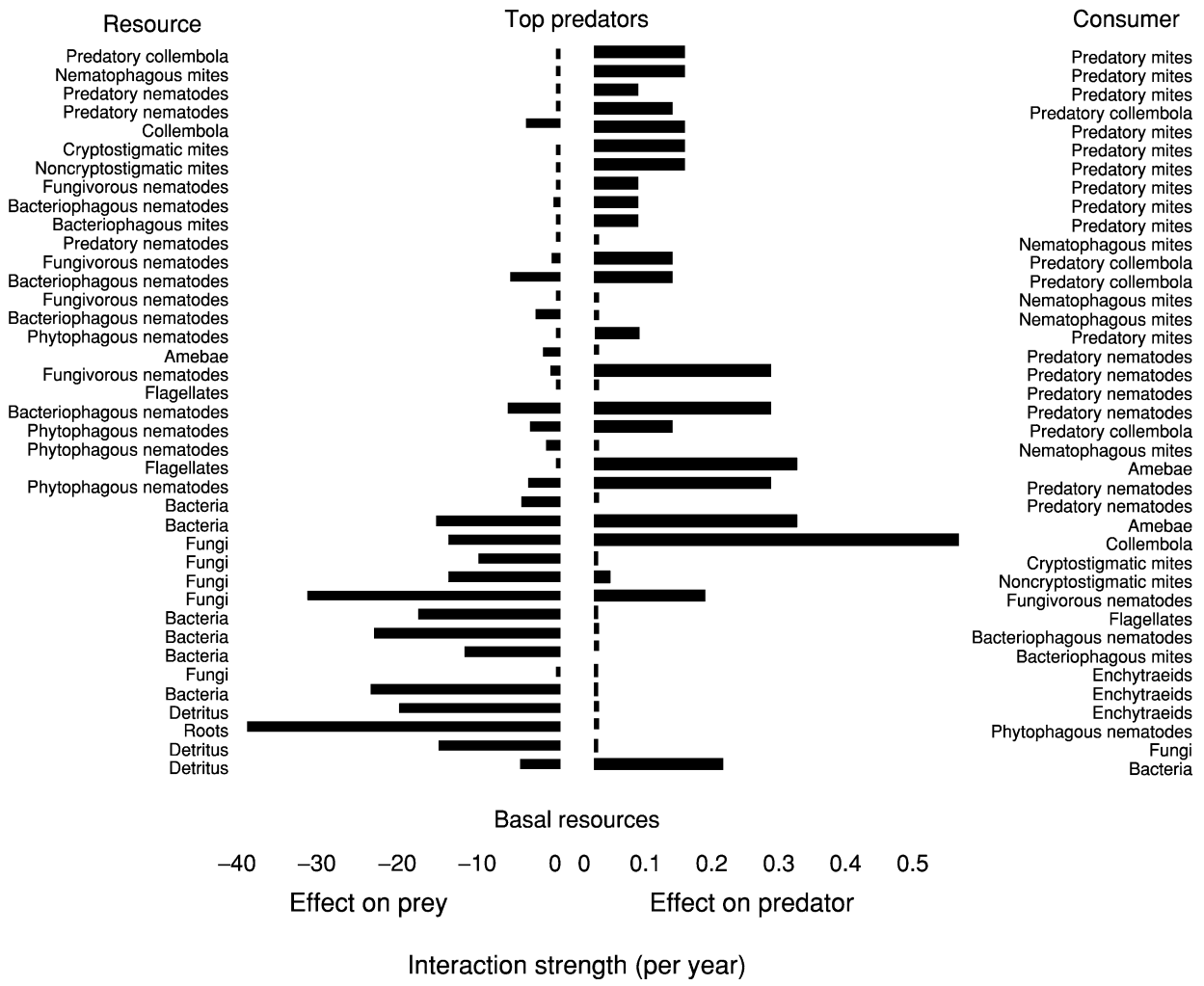


Figure 2 Interaction strengths (per year) along trophic level in the food web from a winter wheat field at the Lovinkhoeve experimental farm in the Netherlands, serving as a representative example. (Reporduced with permission from De Ruiter PC, Neutel AM, and Moore JC (1995) Energetics, patterns of interaction strengths, and stability in real ecosystems. *Science* 269: 1257–1260.)

calculated rates show a similar pattern in differences between the litter, fragmented, and humus layer. For the litter layer, the model underestimates C mineralization by a factor of 2, while for the fragmented and humus layer the model calculates rates close to the observed rates. The calculated N mineralization rates in the litter layer accounts for 85% of the observed rate, while for the fragmented and humus layers this is 50% and 30%, respectively.

Despite the fact that the calculated rates are sometimes close to the observed rates, the model estimates always have to be treated with caution. First, because the model results rely on parameter values that are relatively uncertain, for example the C to N ratio of the substrate for the microbes and the microbial death rates. Second, the indirect effects are only partly taken into account. The model formulations address the stimulatory effects of grazing on microbial growth rates, but other mechanisms behind the indirect effects are at most implicitly taken into account by using observed biomass estimates as input parameters. Despite these weaknesses, the Detrital Food Web Model is regarded as a useful tool in analyzing and understanding the role of the soil food web in soil mineralization processes.

Trophic Interactions and the Dynamics and Stability of Soil Populations and Food Webs

Trophic interactions are likely to affect the distribution and abundance of organisms in fundamental ways, since the success of populations is largely a function of benefits derived from the acquisition of energy (and nutrients) and losses derived from predation. Food web descriptions of the soil community

Table 5 Comparison of the overall C and N mineralization rates (kilogram per hectare per year) as calculated by the food web model with reference values of C and N mineralization published for the different webs

	C mineralization		*N mineralization*	
	Simulated	*Reference value*	*Simulated*	*Reference value*
LH-IF	1890	1690	119	97[a]
LH-CF	1270	1430	78	70[a]
HSB-NT	3970	4800	242	160[b] 90[c]
HSB-CT	4800	5200	300	200[b] 100[c]
KS-B0	1990	1700	120	70[d] 80[e]
KS-B120	1660	1700	70	80[d] 90[e]
CPER	1690	–	87	–
NL-L	25	50	0.7	0.85
NL-F	5	5	0.1	0.2
NL-H	2	2	0.02	0.06
NS	770	–	<1	–
SS	1500	–	30	–
D	2800	–	100	–
F	2600	–	60	–

[a]Obtained from *in situ* measurements.
[b]Obtained from an N budget for 1978–1979.
[c]Obtained from an N budget for 1979–1980.
[d]Obtained from a simulation model.
[e]Obtained from an N budget.

therefore provide a way to analyze the dynamics and persistence of the various populations in the context of the stability of the community as a whole. Central in analyses of the role of trophic interactions in community stability are the interaction strengths. Interaction strengths refer to the per capita – in this case per biomass – effects upon one another. The interaction strengths can be derived from the population sizes and energy flow rates (i.e., the feeding rates; see eqns [1] and [2]) by assuming Lotka-Volterra equations for the dynamics of the functional groups:

$$\dot{X}_i = X_i\left(b_i + \sum_{j=1}^{n} c_{ij}X_j\right) \qquad [5]$$

where X_i and X_j represent the population sizes of group i and j, respectively, b_i is specific rate of increase or decrease of group i, and c_{ij} is the coefficient of interaction between group i and group j. Mathematically, interaction strengths are defined as the entries of the Jacobian community matrix (α_{ij}) being the partial derivatives near equilibrium: $\alpha_{ij} = (\partial \dot{X}_i / \partial X_j)*$. Values for the interaction strengths can be derived from the equilibrium descriptions by equating the death rate of group i due to predation by group j in equilibrium, $c_{ij}X_i^*X_j^*$, to the mean annual feeding rate, F_{ij} (eqn [2]) and the production rate of group j due to feeding on group i, $c_{ji}\ X_j^*X_i^*$, to $a_j\ p_j\ F_{ij}$. With equilibrium population sizes, X_i^*, X_j^*, assumed to be equal to the observed annual mean population sizes, B_i, B_j, the effect of predator j on prey i is:

$$\alpha_{ij} = c_{ij}X_i^* = -\frac{F_{ij}}{B_j} \qquad [6]$$

and the effect of prey i on predator j is:

$$\alpha_{ji} = c_{ji}X_j^* = \frac{a_j p_j F_{ij}}{B_i} \qquad [7]$$

Estimates of the interaction strengths obtained this way for the soil food webs reveal patterns along trophic position, characterized by relatively strong top-down effects at the lower trophic levels and relatively strong bottom-up effects at the higher trophic levels (**Figure 2**). The patterns of interaction strengths are important to the community stability as is indicated by a comparison between the stability of community matrix representations of seven soil food webs (from the prairie and arable soils), using the empirically based values of interaction strengths ('real' matrices) and that of matrices in which these values are randomized. The comparison shows that matrices including the realistic patterns of interaction strengths have a much higher level of stability than their randomized counterparts (**Figure 3**).

The stabilizing patterns of the interaction strengths are the direct result of patterns in the energetic properties of the food webs such as the population sizes

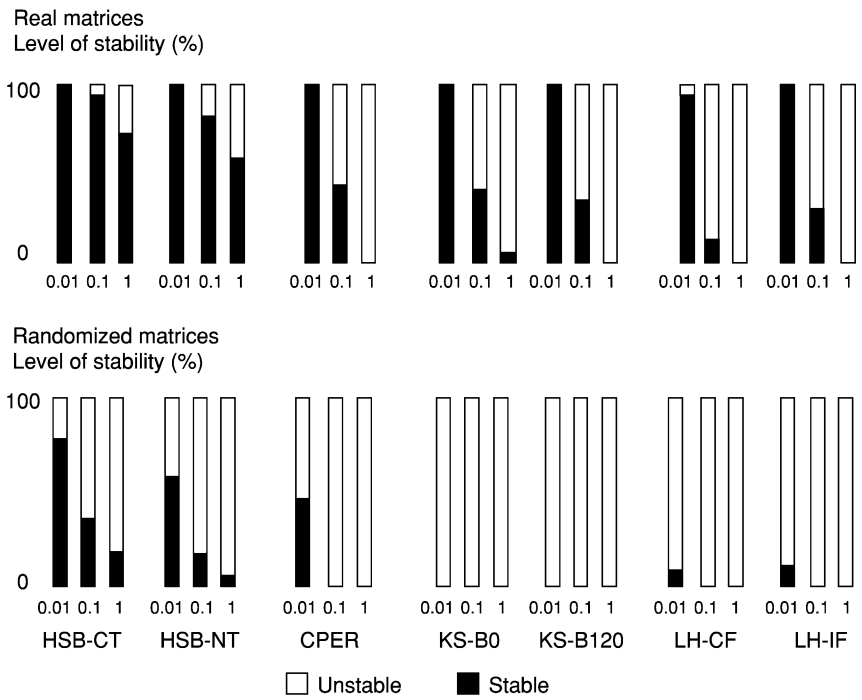


Figure 3 The effects of the patterning of interaction strengths on the stability of seven soil food webs from prairie and arable land. The black fraction in the bars denotes the level of stability (percentage) based on 1000 model runs. Stability of the community matrices is established by evaluating the signs of eigenvalues of the matrices; when all real parts are negative, the matrix is stable and the food web is considered to be locally stable. HSB, Horshoe Bend site, arable land; CT, conventional tillage; NT, no tillage; CPER, shortgrass prairie; KS, Kjettslinge; B0, arable, no fertilizer; B120, arable, with fertilizer; LH, Lovinkhoeve, arable; CF, conventional farming; IF, integrated farming. (Reproduced with permission from De Ruiter PC, Neutel AM, and Moore JC (1995). Energetics, patterns of interaction strengths, and stability in real ecosystems. *Science* 269: 1257–1260.)

(biomasses) and feeding rates (eqns [6] and [7]). Therefore, when the soil biology is looked at in terms of trophic interactions in the soil food web, the structure of the community and the dynamics of the soil populations become inextricably interrelated with soil ecosystem processes and functioning.

Further Reading

Andrén O, Lindberg T, Boström U, Clarholm M, Hansson A-C, Johansson G, Lagerlöf J, Paustian K, Persson J, Petterson R, Schnürer J, Sohlenius B, and Wivstad M (1990) Organic carbon and nitrogen flows. *Ecological Bulletin* 40: 85–125.

Berg MP, De Ruiter PC, Didden WAM, Janssen MPN, Schouten AT, and Verhoef HA (2001) Community food web, decomposition and nitrogen mineralisation in a stratified Scots pine forest soil. *Oikos* 94: 130–142.

Coleman DC, Reid CPP, and Cole CV (1983) Biological strategies of nutrient cycling in soil ecosystems. *Advances in Ecological Research* 13: 1–55.

de Ruiter PC, Van Veen JA, Moore JC, Brussaard L, and Hunt HW (1993) Calculation of nitrogen mineralization in soil food webs. *Plant Soil* 157: 263–273.

de Ruiter PC, Neutel AM, and Moore JC (1994) Modelling food web and nutrient cycling in agro-ecosystems. *Trends in Ecology and Evolution* 9/10: 378–383.

de Ruiter PC, Neutel AM, and Moore JC (1995) Energetics, patterns of interaction strengths, and stability in real ecosystems. *Science* 269: 1257–1260.

Hendrix PF, Parmelee RW, Crossley Jr, DA, Coleman DC, Odum EP, and Groffman PM (1986) Detritus food webs in conventional and no-tillage agroecosystems. *Bioscience* 36: 374–380.

Hunt HW, Coleman DC, Ingham ER, Ingham RE, Elliott ET, Moore JC, Rose SL, Reid CPP, and Morley CR (1987) The detrital food web in a shortgrass prairie. *Biology and Fertility of Soils* 3: 57–68.

Macfayden A (1963) The contribution of the microfauna to total soil metabolism. In: Doeksen J and van der Drift J (eds) *Soil Organisms*, pp. 3–17. Amsterdam: North Holland.

Moore JC and De Ruiter PC (1991) Temporal and spatial heterogeneity of trophic interactions wthin below-ground food webs: an analytical approach to understand multidimensional systems. *Agriculture, Ecosystems and Environment* 34: 371–397.

Moore JC and Hunt HW (1988) Resource compartmentation and the stability of real ecosystems. *Nature* 333: 261–263.

Moore JC, Walter DE, and Hunt HW (1988) Arthropod regulation of micro- and mesobiota in belowground detrital food webs. *Annual Review of Entomology* 33: 419–439.

Moore JC, de Ruiter PC, and Hunt HW (1993) Influence of productivity on the stability of real and model ecosystems. *Science* 261: 906–908.

Schröter D, Wolters V, and de Ruiter PC (2003) C and N mineralisation in the decomposer food webs of a European forest transect. *Oikos* 102: 294–308.

Verhoef HA and Brussaard L (1990) Decomposition and nitrogen mineralization in natural and agro-ecosystems: the contribution of soil animals. *Biogeochemistry* 11: 175–211.

FORENSIC APPLICATIONS

W F Rowe, George Washington University, Washington, DC, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Forensic scientists encounter soil both as trace evidence and as a matrix in which other physical evidence may be found. Trace evidence is the term used by forensic scientists for the minute pieces of physical evidence that can be transferred between the perpetrator and the victim of a crime, between the perpetrator and the crime scene, and between the victim and the crime scene. Trace evidence is used to establish an association among the members of the triad: perpetrator, victim, and crime scene. To give a specific example, suppose traces of soil are found on the clothes of a suspect in a rape case. This soil will be removed in the forensic science laboratory and compared with known soil samples from the crime scene. If the soil from the clothes is indistinguishable from the known sample a positive association between the suspect and the crime scene has been established. Other types of trace evidence include glass, paint, hair, fibers, gunshot residue, explosive residue, ignitable liquids, and biological fluids (such as blood and semen).

The transfer of trace evidence is governed by the Locard Exchange Principle, which states that when two surfaces come into contact there is a mutual exchange of trace evidence between them. Such evidentiary transfers may be primary transfers (i.e., evidence is transferred from surface A to surface B and later recovered from surface B) or secondary transfers, which involve a transfer of evidence from surface A to surface B and from surface B to surface C. Tertiary and even higher-order transfers of trace evidence can occur; however, these transfers present serious interpretive problems for the forensic scientist, because the ultimate source of the trace evidence may be impossible to identify.

Collection of Evidentiary Soil Samples

Known soil samples are collected at crime scenes, as well as any alibi locations. At scenes of crimes, the crime scene investigator collects soil samples from the following locations: (1) the areas immediately surrounding victims' bodies, (2) areas adjacent to any shoe or tire impressions, and (3) any visibly disturbed areas. If the composition of the surficial soil horizons varies within the crime scene areas, representative samples of each soil type are obtained. Because criminal suspects may claim that soil on their clothing or vehicles originated from locations other than the crime scene (e.g., homes or places of work) elimination samples are also obtained from these alibi locations. Clearly the case against a suspect is strengthened if the soil from suggested alibi locations does not match the questioned soil samples. Generally approximately 100 g of soil is sufficient for subsequent laboratory analyses. However, if the soil at a site contains a substantial amount of coarse material, the mass of the known sample is increased so that an adequate sample of the fine fraction is available for subsequent laboratory analyses. Samples are collected from only the top 5–7 cm of the surficial soil horizon. The soil samples are placed in clean containers such as baby-food jars or Ziplok-style plastic bags. The evidence containers are then sealed (with evidence tape or a heat sealer) and marked for identification. Chain-of-custody forms are also initiated for the known soil samples. The locations where the known soil samples have been collected are documented in standard ways: through notes, sketches, and photographs.

The collection of the questioned soil sample is a two-stage process. First, soil from the crime scene is transferred to clothing or vehicles; then the transferred soil is collected at the crime scene, by a crime scene technician, or at the forensic science laboratory. The initial transfer at the crime scene may not produce a representative sample of the soil at that location; for example, clothing rarely retains

coarse material such as gravel. Similarly the fines in a shovel load of soil tossed into the open bed of a pickup truck may be blown or washed away before the crime scene investigator has a chance to collect this evidence.

At the forensic science laboratory, trace evidence such as soil is collected in a variety of ways. Hand-picking with a pair of tweezers and a low-power magnifier is used if there is only a small amount of trace evidence, as well as collection using tape lifts or adhesive lifters. Neither of these methods is feasible if an item of clothing has substantial quantities of soil adhering to it, but soil can be scraped gently from an item of clothing on to a sheet of glossy paper. Ideally, scraping should be done in a draft-free, clean room so that air currents do not carry away fine particulates. If an adherent soil layer is thick the laboratory analyst has to consider the possibility that the layer consists of multiple layers deriving from more than one location. Therefore, every effort is made to retain the stratigraphy of the sample and to avoid mixing soil from different layers. A vacuum cleaner with a special filter attachment can also be used to collect soil samples; however, vacuuming is likely to destroy evidence of stratigraphy.

Vehicles require special attention. The crime scene investigators collect soil samples from the vehicle's tires and undercarriage. The interior of the vehicle is also carefully searched, with particular attention being paid to the pedals and floor mats. The vehicle's trunk (boot) is also examined for traces of soil; digging tools with adherent soil may have been placed there.

Laboratory Analysis of Forensic Soil Samples

The methods of soil analysis used in forensic science are predicated on the size of the sample and the use to which the analytical results will be put. The aim of forensic soil analysis is to associate a soil sample taken from an item of clothing or from a vehicle with a specific location. In order to do this, the methods of analysis chosen have to be able to discriminate between soil samples from different locations. Moreover, a method for comparing soil samples has to be practical (i.e., feasible with common laboratory instrumentation), inexpensive, accurate, and applicable to small soil samples. A survey of published materials on forensic soil analysis reveals that this is an area that has yet to be affected by standardization of methodology. Neither the Scientific Working Group on Material Analysis Techniques (SWGMAT) nor American Society for Testing and Materials (ASTM) Committee E30 has promulgated any recommended analytical methods. Consequently, there is considerable variation among forensic science laboratories in their analyses of soil samples, but the methods employed can be placed in three broad categories: (1) methods for comparing soil organic content, (2) methods for comparing particle-size distributions of soil mineral fractions, and (3) mineralogical analysis.

Soil Organic Content

Soil Color

Soil color has been found to be the property of soil that most reflects its pedogenic environment and history. Soil organic matter and iron oxides contribute most to soil color. Organic matter darkens soil, while iron oxides produce a range of soil colors that are dependent on the oxidation state of the iron. The most widely used method for determining soil colors is comparison of soil samples with the color chips in the Munsell soil color charts. In the Munsell color system, color is expressed in terms of hue (basic color), value (lightness or darkness), and chroma (intensity of basic hue). Color determinations are applied to air-dried bulk soil samples; but more elaborate color-determination schemes have also been advocated. Because most of the color of a bulk soil sample is due to the clay fraction, which contains clay particles intimately bound to soil humus (forming the clay–humus complex) and clay particles coated with iron oxides, some forensic soil analysts have suggested basing color determinations on the clay fraction of the soil. Other forensic soil analysts have proposed determining the colors of the soil sample after air-drying, wetting, organic matter removal, iron oxide removal, and ashing. The effectiveness of these treatments in allowing color discrimination of soils from different locations depends on the geographic origins of the soils being compared. A study conducted in the UK by the Home Office Central Research Establishment has found that combining the colors of air-dried samples, moistened samples, and ashed samples results in a high level of discrimination of soils from different locations. On the other hand, a similar study conducted by the Japanese National Research Institute of Police Science has found that color determinations after a different regimen of treatments (air-drying and moistening of bulk samples; organic matter and iron oxide removal from clay fractions) provides the highest discriminatory power. In the latter study, Munsell colors were converted into x-, y-, and z-coordinates: the x-coordinate represents the hue, with basic hues from 2.5YR to 5GY, assigned

numerical values from 1 to 10; the *y*-coordinate represents the value; and the *z*-coordinate represents the chroma. Color comparisons between different soil samples are based on the Euclidean distance d^2 between the color coordinates of the soil samples. Pairs of soil samples for which $d^2 \leq 1$ were considered indistinguishable. Using this approach, 97% of 190 pairs of soil samples from different locations were determined to be different.

Percentage Organic Content

The soil organic content can be determined directly by hydrogen peroxide digestion. Properties related to soil organic content can also be examined. Soil pH is commonly measured, although it is known to undergo seasonal changes. Soil polysaccharide content is compared spectrophotometrically. A Fourier transform infrared (FT IR) spectrometric method for the comparison of soil organic fractions has also been developed. In this technique, a potassium bromide disk is prepared from a dried, sieved soil sample. After an FT IR spectrum is obtained, the potassium bromide disk is ground up and placed in a muffle furnace, where the soil organic fraction is burned away. The potassium bromide containing the soil mineral fraction is then reground and pressed into a disk. A second FT IR spectrum is obtained and subtracted from the first to obtain an infrared spectrum of the soil organic fraction.

Soil Particle-Size Distributions

In most forensic science laboratories that conduct soil analyses, particle-size distributions are determined by passing a carefully weighed soil sample through a nest of sieves. Both manual shaking and mechanical shaking are used. Mechanical shakers give more reproducible results than manual shaking; however, experimental parameters such as the torque applied to the stack of sieves in the mechanical shaker and the sample weight can affect the results. In general, obtaining reproducible results with sieving requires careful attention to experimental conditions. A wet sieving method applicable to small soil samples has been developed and is claimed to give more reproducible results than dry sieving.

The UK Home Office Central Research Establishment has conducted a study to develop statistical methods for the comparison of soil samples. Five mock crime scenes were used: at scenes I and II, soil was sampled adjacent to a pair of shoe prints in two gardens; at scenes III and IV, soil was sampled adjacent to pairs of plants in the same two gardens; and, at scene V, soil was sampled adjacent to tire tracks in a rural area. The following analyses were performed: the colors of dry and ashed (850°C for 30 min) soil samples were determined; the pH of each sample was measured; the total polysaccharide content of each sample was determined spectrophotometrically; and particle-size distributions were determined for each sample using both conventional dry sieving and a Coulter counter for the silt fractions. The mock crime scenes were uniform with respect to most measures except total polysaccharide content. The particle-size distributions were compared using an index of variability (IV) and the maximum difference (MD). IV is the difference in the total silt or sand content of two samples, while MD is the maximum difference in the particle-size distributions. From replicate soil samples, the means and standard deviations of each variable were calculated and univariate confidence intervals were determined. Because IV and MD are not independent variables, better confidence intervals were constructed by assuming that IV and MD have a bivariate normal distribution with correlation coefficient ρ. The variables were transformed by a coordinate rotation to produce two normally distributed independent variables. Confidence contours were constructed for the new variables. As a test of this model, 153 pairs of nonduplicate soil samples from the mock crime scenes were compared using a 95% confidence contour; 97% of the pairs were correctly identified as different.

An experiment conducted at the University of Strathclyde, Scotland, UK, illustrates the discriminating power of particle-size measurements when combined with percentage organic matter. A field was divided into 100 cells each measuring 3 m × 3 m. Samples of soil weighing 400 g were collected from the surficial horizon in each cell. The percentage organic content was determined for each of six 1.5-g aliquots of soil from each cell. The dried mineral fraction from each soil aliquot was then sieved. Three variables were selected for comparison: the median value of the particle-size distribution, the percentage of particles in the 90- to 250-μm size range, and the percentage of soil organic matter. Soil samples were selected at random from 20 of the cells and compared with the soil samples from all 100 cells using the *z*-statistic.

Suppose that the soil sample is to be characterized by a variable X having variance σ^2. Measurements of X on m soil samples taken from cell i (one of the c control cells) yield a mean of $<x_i>$. Measurements of X on n soil samples from an unknown cell yield a mean of $<y>$. The test statistic for the two means being equal is:

$$z_i = \left| \frac{<x_i> - <y>}{\sigma\sqrt{\frac{1}{n} + \frac{1}{m}}} \right| \quad \text{for } i = 1 \ldots c \qquad [1]$$

The most probable cell for the origin of the soil is the control cell for which the probability $P(z_i)$ of obtaining at least the value of z_i is maximum. If the soil has been characterized by N independent variables $X_1 \ldots X_N$, the most probable cell for the origin of the soil will be that corresponding to:

$$\max\left[\prod_{j=1}^{N} P_j^N(Z_i)\right] \quad i = 1 \ldots c \qquad [2]$$

Of the 20 randomly selected blind samples, 19 were assigned to the correct control cell using the aggregated P-value. For the one random sample incorrectly assigned, the correct cell had the second-highest aggregated P-value.

The z-statistic method is robust and applicable to many forensic problems. However, the three variables selected in this study may not be the best for soils from another location, and the appropriate variables to use may not be evident *a priori*. Another problem is that statistical significance cannot be assigned to the aggregated P-values.

A study conducted by the Japanese National Research Institute of Police Science has examined the effectiveness of particle-size distributions in distinguishing soil samples from different locations. In this study the soil samples were dry-sieved and the weight percentages of the soil fractions in the following three size categories determined: <0.05 mm, 0.05–0.2 mm, and 0.2–2 mm. Pairs of soil samples were compared by computing the Euclidean distance d^2 between the samples. The threshold value of d^2 to be used to declare paired soil samples to be different was determined from replicate measurements on three of the soil samples. Using this approach, 95.9% of 2628 sample pairs were correctly distinguished as different. When the particle-size distributions were combined with color determinations, 99.5% of the soil samples could be correctly discriminated.

Soil Mineralogical Analysis

In the USA many forensic laboratories have relied on density gradient separations as a useful method for the comparison of the mineral contents of soil samples. Typically, glass tubes are filled with a series of liquids having decreasing densities. The tubes are allowed to stand for several hours so that interdiffusion of the layers can produce a linear density gradient. The samples to be compared are added to the tops of separate density gradient columns and their mineral particles allowed to settle. Mixtures of bromoform and bromobenzene or bromoform and acetone are used to construct the density gradients. The mineral fractions separated in density gradient tubes are recovered and identified by polarized light microscopy or other techniques. X-ray diffractometry has also been used widely for mineralogical analysis. The clay mineral fractions can be isolated from soil samples by sedimentation and their X-ray diffraction patterns can be obtained from oriented mounts. Standard methods for the identification of the clay minerals (dehydration, glycolation, etc.) are used. FT IR can also be used for mineralogical analysis; it is best applied to mineral fractions obtained from density gradient separations.

Soil as a Matrix

A wide variety of evidentiary materials may be found in contact with soil. The bodies of murder victims may be buried; particulate trace evidence such as hairs and textile fibers may be on a buried body or may have been buried seaparately; molecular trace evidence such as ignitable liquids or explosive residues may be dispersed in soil after an arson fire or after an explosion; and deoxyribonucleic acid (DNA) evidence in body fluids and tissues may also have been in contact with soil. The effect of soil on these categories of physical evidence is the subject matter of the newly emerging field of forensic taphonomy.

Buried Remains

The interaction of soil and buried human remains is complex. Decomposing remains contribute to the overall organic content of the soil at the burial site so that the presence of a grave may be inferred by the elevated organic content and darker color of a localized region of a soil horizon. Studies conducted at the University of Tennessee's Anthropological Research Center (also known as the 'Body Farm') have shown that decomposing human remains elevate the temperature of the soil immediately surrounding them and contribute large quantities of aliphatic carboxylic acids to the soil solution. It has been proposed that the relative abundance of certain of these aliphatic acids be used to estimate postmortem interval.

It has been widely recognized for years that burial in soil retards decomposition. This retardation is largely due to the reduction in insect activity. Insects, especially flies, are the main agents for the skeletalization of human remains. While adult flies may feed on carrion, it is mainly fly larvae (maggots) that consume decaying soft tissue. However, even a thin layer of soil may prevent oviposition by most flesh and carrion flies. There are, nevertheless, a few species of flies capable of digging down through soil to

deposit eggs on a corpse. Some species of carrion beetle may dig several feet down to reach remains.

The ultimate fate of buried human remains depends on general climatic conditions. Remains in well-drained soils in the temperate climatic zones will eventually be completely skeletalized. If the soil pH is low, even hard tissues such as bone or tooth enamel will ultimately be destroyed; only a soil stain may remain to indicate the presence of a burial. Even bodies in coffins or vaults may be completely destroyed if water infiltrates these enclosures. On the other hand, water can act to preserve buried remains. If the burial site is very damp, adipose tissue in the body may be converted to a light-colored, cheesy mass called adipocere. Adipocere is composed mainly of long-chain aliphatic carboxylic acids. Its formation may be due to the direct hydrolysis of fats by groundwater; however, there exists evidence (primarily from studies of adipocere in Egyptian mummies) that adipocere may also be produced by the action of microorganisms such as members of the genus *Clostridium*. Adipocere may coat the body, discouraging or completely preventing insect attack on the remains. If the soil is very wet and filled with decaying vegetation (e.g., in peat bogs), the soil solution will have a high concentration of tannins and a low oxygen level. Under such anaerobic conditions, skin may be converted to leather and microbial decomposition (putrefaction) of soft tissue may be prevented. 'Bog bodies' may be preserved for hundreds or thousands of years with internal organs more or less intact and available for examination upon autopsy.

The geologic principles of stratification provide a basis for the understanding of buried crime scene evidence. The excavation of graves generally interrupts and mixes soil horizons. Visible evidence of the mixing of the surficial soil horizon with an underlying soil horizon is one of the indicators of the presence of a grave. Graves can also be located using geophysical prospecting techniques such as infrared thermography and ground-penetrating radar. Once a grave has been found, the crime scene investigators have to exercise care during the exhumation of the remains to retain the stratigraphy within the grave. This stratigraphy provides important evidence about the circumstances surrounding the digging of the grave. For example, murderers may spread layers of lime, calcium carbonate, or calcium hydroxide over the remains in the belief that these chemicals will hasten decomposition or reduce the odor of decomposition. The stratigraphy of the burial can be affected by bioturbation, sedimentation, and compression–depression. Scavenging animals may dig down into the infill of the grave to reach the decaying remains; small rodents may also burrow through the grave. Compression–depression of the grave infill leads to slumping of the surface of the grave. The backfill can be tamped down as the perpetrator replaces the soil dug out of the grave. The grave infill may also settle as the soil settles and the remains decompose.

This settling (along with desiccation of the surficial soil horizon after rains) results in cracking of the soil at the edges of the grave. The surface of a grave may also show a secondary depression, caused by an initial, localized expansion, which is the result of the abdomen of the corpse filling with putrefactive gases followed by collapse when the putrefactive gases ultimately escape. If the body is buried facedown a secondary depression may not form. Shallow holes left by scavenging animals may resemble secondary depressions produced by the process of bloating and collapse. Areas of compaction may also be observed within a grave if the victim was buried alive. In struggling to escape the victim may displace soil upwards or sideways. Clearly such evidence would be of the greatest importance in a subsequent murder trial.

When a grave is excavated, investigators have to take care to preserve any impression-type evidence. Shoe impressions may be observed around and within the grave; digging implements such as shovels may leave tool marks on the walls of the grave and on clods of soil removed from the grave; and fabric impressions from the digger's clothing may be left around or within the grave, particularly if the implement used to dig the grave required the digger to kneel while using it.

Trace Evidence in Soil

Contact with soil can alter trace evidence in a number of undesirable ways. The evidence might be completely destroyed or it might be so altered that it can no longer be associated with its original source. The alteration or destruction of trace evidence is caused mainly by soil microorganisms; for example, the restriction endonucleases found in soil microorganisms quickly cut DNA strands into small fragments. Soil microbes can also metabolize the hydrocarbons in ignitable liquids that can be used as arson accelerants. Even if some of the hydrocarbons in the ignitable liquid are detectable in the soil sample, it is sometimes impossible to categorize correctly the type of ignitable liquid used (e.g., light petroleum distillate, gasoline, kerosene, and so forth). In some soils *n*-alkanes are degraded more rapidly than aromatic hydrocarbons, while in others the reverse is true.

The enzyme systems of soil microorganisms can also break down hairs and textile fibers. Among the natural textile fibers, cotton is very rapidly destroyed by contact with soil. As a general rule, in most

climates buried cotton fibers are completely destroyed within a few months. Lignified cellulosic fibers (e.g., flax and hemp) are broken down more slowly. The rapid destruction of cellulosic fibers in soil is easily understood: most soils contain microorganisms that use plant debris as sources of energy and of carbon for growth. Buried proteinaceous fibers such as hair, wool, and silk can survive for decades, even in a well-watered soil. Hair and wool fibers, however, can be degraded by soil fungi; fungal hyphae can penetrate hair and wool fibers and grow within the fibers, ultimately completely consuming them. Cellulosic and proteinaceous fibers are more likely to survive if they are buried close to metal artifacts containing copper (e.g., zippers, snaps, button loops, brads, and tacks). Copper ions are well known to have bactericidal effects. Some synthetic fibers also deteriorate rapidly in contact with soil: buried rayons and cellulose acetates are destroyed almost as rapidly as cotton. On the other hand, other synthetic fibers such as nylons and polyesters deteriorate very slowly, presumably because soil microorganisms are unable to elaborate enzymes capable of breaking down the primary polymer structure of these fibers.

See also: **Acidity**; **Archeology in Relation to Soils**; **Bacteria:** Soil; **Clay Minerals**; **Fourier Transform Infrared Spectroscopy**; **Microbial Processes:** Environmental Factors; **Organic Matter:** Principles and Processes; **pH**

Further Reading

Chaperlin K and Howarth PJ (1983) Soil comparison by the density gradient method – a review and evaluation. *Forensic Science International* 23: 161–177.

Dudley RJ (1975) The use of colour in the discrimination between soils. *Journal of the Forensic Science Society* 15: 209–218.

Dudley RJ (1976) The particle size analysis of soils and its use in forensic science – the determination of particle soil size distributions within the silt and sand fractions. *Journal of the Forensic Science Society* 16: 219–229.

Dudley RJ and Smalldon KW (1978) The objective comparison of the particle size distribution in soils with particular reference to the sand faction. *Medicine, Science and the Law* 18: 278–282.

Dudley RJ and Smalldon KW (1978) The evaluation of methods for soil analysis under simulated scenes of crimes conditions. *Forensic Science International* 12: 49–60.

Graves WJ (1979) A mineralogical soil classification technique for the forensic scientist. *Journal of Forensic Science* 24: 323–339.

Hochrein MJ (2002) An autopsy of the grave: recognizing, collecting, and preserving forensic geotaphonomic evidence. In: Haglund WD and Sorg MH (eds) *Advances in Forensic Taphonomy: Method, Theory, and Archaeological Perspectives*, pp. 45–70. Boca Raton, FL: CRC Press.

Murray RC and Tedrow J (1981) *Forensic Geology.* Englewood Cliffs, NJ: Prentice Hall.

Murray RC and Solebello LP (2002) Forensic examination of soil. In: Saferstein R (ed.) *Forensic Science Handbook,* vol. I, 2nd edn, pp. 615–633. Upper Saddle River, NJ: Prentice Hall.

Palenik SJ (1988) Microscopy and microchemistry of physical evidence. In: Saferstein R (ed.) *Forensic Science Handbook*, vol. II, pp. 161–208. Englewood Cliffs, NJ: Prentice Hall.

Sugita R and Marumo Y (1996) Validity of color examination for forensic soil identification. *Forensic Science International* 83: 201–210.

Sugita R and Marumo Y (2001) Screening of soil evidence by combination of simple techniques: validity of particle size distribution. *Forensic Science International* 122: 155–158.

Wanogho S, Gettinby G, Caddy B, and Robertson J (1985) A statistical method for assessing soil comparisons. *Journal of Forensic Science* 30: 864–872.

Wanogho S, Gettinby G, Caddy B, and Robertson J (1987) Some factors affecting soil sieve analysis in forensic science. 1. Dry sieving. *Forensic Science International* 33: 129–137.

Wanogho S, Gettinby G, Caddy B, and Robertson J (1987) Some factors affecting soil sieve analysis in forensic science. 2. Wet sieving. *Forensic Science International* 33: 139–147.

FOREST SOILS

J R Boyle, Oregon State University, Corvallis, OR, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Forest soils, where soil formation has been influenced by forest vegetation, are generally characterized by deeply rooted trees, significant 'litter layers' or O horizons, recycling of organic matter and nutrients, including wood, and wide varieties of soil-dwelling organisms (Figure 1). There are also soils now covered with forest vegetation, often plantations, on lands that were not naturally forested. These soils are probably undergoing processes that give them 'forest soil-like' characteristics, e.g., litter layers from trees, woody organic residues from deep roots, and associated soil microbe and fauna populations. Like other soils, forest soils have developed, and are developing, from geological parent materials in various topographic positions interacting with climates and organisms. Forest soils may be young, from 'raw' talus, recent glacial till or alluvium, or 'mature,' in relatively stable landscape positions. Just as forest vegetation of the world varies greatly, so do forest soils, e.g., they are shallow, deep, sandy, clayey, wet, arid, frigid, or warm. The following are representative of variability in properties of forest soils, complied for the forest soils of the Douglas fir region of northern California, Oregon (Figure 2), Washington, and southwestern British Columbia: soil depths, 1–2.5 m; nitrogen contained in forest floors, 170–1300 kg ha^{-1}; total N in soils, 1460–22 500 kg ha^{-1}. There are forest soils classified in all the orders of the US soil classification system, reflecting the wide range of conditions in which forests occur.

Forest soils have been studied by many generations of soil scientists. Some studies have been focused mainly on ecologic characteristics, for example on surface organic layers in forests in Denmark, which introduced the terms 'mor' and 'mull'; while other investigations have dealt with nutrients, water supplies, soil organisms (especially mycorrhiza-forming fungi), fertilizer additions, and other impacts of forest management.

Because human land uses have often appropriated the 'best' soils and landscape positions for agriculture, many forest soils are less than optimum in properties that control fertility and potential vegetation productivity. This is illustrated by the designation of many forest soils as 'steep, stony lands' during early soil surveys in the US Lake States. Of course, there are also forest soils with high productive potentials, i.e., with porous and well-aerated root zones, good nutrient-supplying and -retaining capacities, excellent capacities to store plant-available water, and characteristics amenable to robust populations of soil microbes and fauna. And, in many areas of the world today, forests grow on lands once used for

Figure 1 Beech (*Fagus sylvatica*) forest in Germany.

Figure 2 Douglas-fir (*Pseudotsuga menziesii*) forest landscape in Oregon, USA.

Figure 3 *Pinus radiata* plantation in New Zealand. (Photo by B. Dyck.)

intensive agriculture, e.g., parts of northern and central Europe, the northeastern and southeastern USA, and New Zealand (Figure 3).

Many forest soils are shallow and low in organic matter and water-holding capacity, on steep, rocky landscapes, in mountainous areas of the world, e.g., some of the Norwegian and Swedish mountains, the Alps, the Andes, the Sierra Nevada and Rocky Mountains of North America, and in arid environments, e.g., in Mediterranean countries, parts of Chile, Mexico, and the ponderosa pine regions of the western USA. In large areas of Canada, the US Lake States, Norway, Sweden, Finland, and Siberian Russia, forests grow on soils that are shallow, often very wet, on peatlands of several types, and on some soils with permafrost. These forests are slow-growing, due in part to limited nutrient- and water-supplying capacities, shallow rooting potentials, and cold temperatures of the soils.

Roots and Mycorrhizae

Tree roots can penetrate to depths of 10 m or more, depending on the impermeable or porous nature of

subsoils and/or bedrock. Deep root systems provide anchorage and stability for trees and provide access to nutrient and water sources far below normal 'soil.' Roots at depth proliferate fine roots and mycorrhizae in thin deposits of soil materials in fissures, providing access to water and nutrient supplies that are not accounted for in evaluations of 'standard' soil profiles. Nutrient uptake from deep roots, transport to and incorporation in plant tissues, subsequent deposition of the tissues as litter, and litter decomposition can contribute to 'upward' cycling of nutrient elements from deep in a soil or the regolith to surface soil horizons.

Mycorrhizae, symbiotic root-fungus structures essential for all forest trees, provide, via their proliferations of fungal hyphae, greatly expanded surfaces for sorption of water and nutrients, and, in some cases, they provide some inhibition of invasions of pathogens. In addition, mycorrhizae facilitate plant uptake of nutrients, especially N and P, from complex compounds not traditionally considered to contain plant-available elements. There is also substantial evidence that mycorrhizae, and other rhizosphere organisms, can facilitate root uptake of nutrients from primary minerals; these processes involve 'dissolution' of mineral structures and nutrient release due to both acidity and complexing characteristics of rhizosphere compounds from fungi, and perhaps from other microbes.

In some forest ecosystems, tree roots represent 30–50% or more of annual primary productivity. Annual or more frequent death and decay of vast proliferations of fine roots in forest soils provide significant additions of organic matter and carbon sequestration, contributing to processes that sustain soil fertility and provide vast stores of carbon. Where there are significant quantities of deep roots, including large, lignaceous roots, these 'ecosystem functions' are difficult to quantify, while being of great significance; decomposition and carbon release processes can be very slow – storage of carbon and nutrients in deep root-derived organic matter is very long-term.

Litterfall and Litter: Organic Matter Recycling in Forest Soils

Forest vegetation adds litter and organic matter to soil surfaces and within soils, because of fauna-mediated incorporation of litter and to the deaths of fine roots and mycorrhizae. These additions are recurrent seasonally and annually in most temperate forests and are continuous and/or episodic throughout the year in many tropical forests. While forest vegetation remains, this litterfall recycles nutrient elements via decomposition processes. In some systems, from 30 to 60% or more of annual tree requirements for nutrients may be met by this recycling. Litter organic matter, interacting with mineral soil particles, and in processes of comminution and decomposition by soil fauna and microbes, contributes to building and maintenance of soil structure and accompanying porosity, and water- and nutrient-holding capacity.

The plant sources, amounts, and compositions of litterfall change during ecosystem succession as forests develop during their regrowth after disturbance of a previous forest or after planting. These variations and dynamics depend on the nature and intensity of site manipulation and management, e.g., elimination of vegetation that competes with desired trees will limit the variety of litter during a forest growing cycle. Because of this repeated and continual litter recycling, whether from complex plant communities or from forest trees in monoculture stands, some forest ecosystems can be self-maintaining for nutrients and soil organic matter for the many years that trees grow and, perhaps, indefinitely. Other forest ecosystems require inputs of some nutrients to sustain desired growth of the trees, e.g., pine plantations on very old, deeply weathered sandy soils of Florida require P fertilizer for satisfactory commercial forest growth.

Accumulated litter layers provide habitats for proliferation of roots and mycorrhizae, and for many organisms, including fauna and microbes that carry out decomposition. Due to the variety in and composition of litter, in any single forest ecosystem decomposition processes are very complex, involve differing groups of organisms, and proceed at different rates (Figure 4). Early stages of decomposition for much litter include comminution – chopping and shredding – of individual plant materials. These processes may be carried out by centipedes, beetles, springtails, worms, slugs, and any of a vast variety of soil fauna, depending on the ecosystem. Of course fungi and bacteria are the ultimate decomposers. In some forest litter layers, nets and mats of fungal hyphae bind together litter components to create carpet-like forest floors. In some tropical forests, litter that reaches the forest floor is already being decomposed and has a very short residence time on the soil surface, creating a rather bare mineral soil surface.

Forest soil scientists recognize three broad categories of litter layers or humus forms: mull, moder, and mor. Mull humus is litter that is rather rapidly comminuted and decomposed, with organic matter being rapidly incorporated into A horizons by soil fauna, with earthworms often playing major roles in decomposition and incorporation processes. Mulls contain what are designated as L and F, or Oi and Oe, horizons. Mor humus types, often forming in coniferous forests with cool, moist – or very dry – forest floor

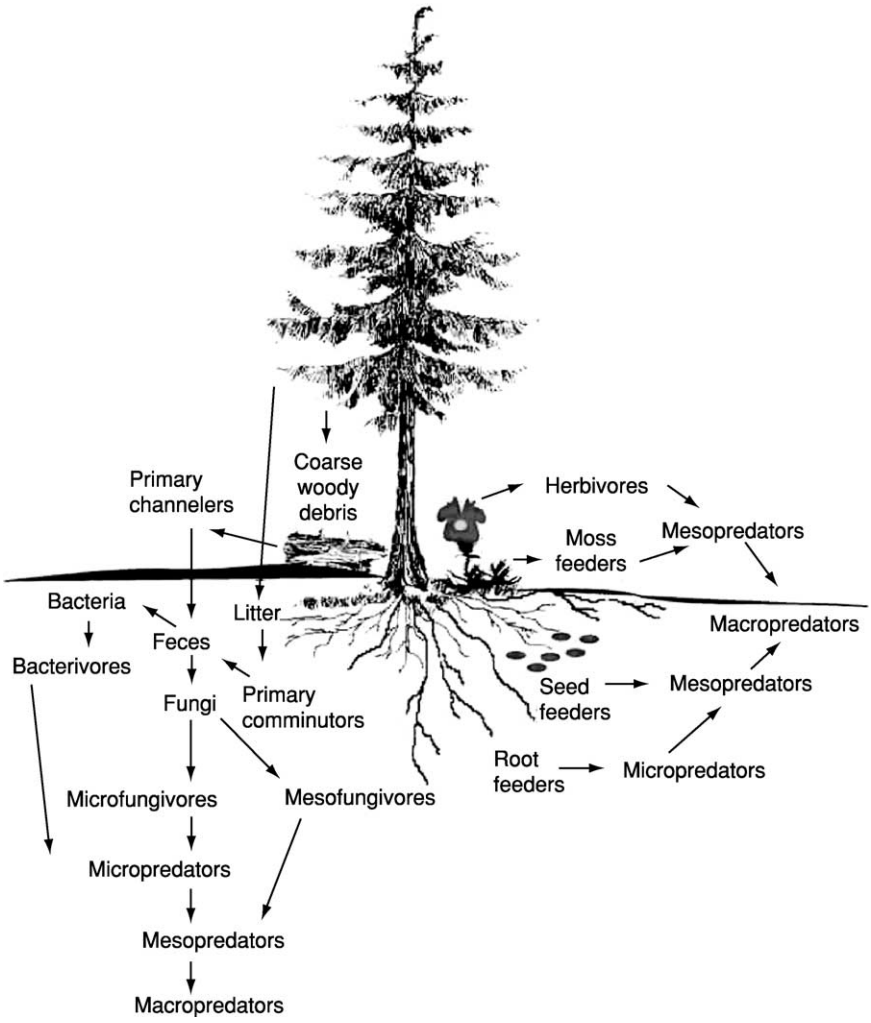


Figure 4 Roles of soil fauna in forest ecosystems. (From a drawing by Andy Moldenke.)

conditions result from litter that is slow to decompose and be incorporated with mineral soil materials; fungi predominate as final decomposers. Mor litter layers, with L, F, and H (Oi, Oe, and Oa) layers and very thin or absent A1 horizons, can accumulate to thicknesses of 15–30 cm or more. Moder humus forms are intermediate in characteristics, with significant accumulations of slowly decomposing materials – L/Oi layers – above F/Oe horizons and an A1 or Ah horizon. Within each of these broad categories of humus forms are numerous variations, depending on the characteristics of the forest ecosystem.

Biogeochemistry and Hydrology

Studies of potential nutrient supplies from forest soils and of nutrient element states and dynamics in forest ecosystems have long been conducted as site-quality classifications, as nutrient-cycling studies, and, since the late 1960s, as research of forest ecosystem biogeochemistry. For many forests, attention has focused on nitrogen, because it is very often the nutrient most limiting for forest growth. Studies of nitrogen biogeochemistry in forests have included: total ecosystem supplies; rates of inputs from precipitation and symbiotic dinitrogen fixation; stores and dynamics in litter and soil organic matter; ammonification and nitrification; sequestration and leaching rates; and denitrification. For some tree species and systems, there has been research of supplies and deficiencies of phosphorus, sulfur, and potassium (e.g., in coarse outwash sands, in glaciated landscapes in the state of New York, in the USA). And, for some systems, researchers have identified limiting quantities of boron (e.g., for *Pinus radiata* in New Zealand) and other micronutrients.

With broad ecosystem and watershed studies, scientists have documented interactions among system hydrology, vegetation, soils, and nutrient elements, especially nitrogen in nitrate form. Removals and

suppression of vegetation followed by accelerated decomposition of organic matter, interruption of transpiration flows of water, and alterations of stream water chemistry in forested watersheds have been documented, with significant transfers of nitrate nitrogen to streams in some forest systems, while other systems retain nitrogen rather well after disturbances. These research results are very relevant for sustainable management of forest ecosystems and interacting aquatic systems, and are important components of scientific knowledge of forest–soils interactions and implications for management.

Within the hydrologic cycle and watershed context of forest lands, forest soils contribute to moderation of flows of water, e.g., influencing intake and percolation, limiting runoff, providing filtration, storage, and transmission and delayed flows to streams. Because of litter layers, high porosities associated with soil fauna activities, root proliferation and depth, and many macropores, forest soils have very high infiltration and percolation rates, rates that far exceed rainfall intensities in many ecosystems. These and other attributes of forest ecosystems provide valuable watershed values for the millions of people dependent on forested landscapes for water supplies. The scientific discipline of forest hydrology encompasses these and other aspects of forest–soil–water interactions.

Disturbance Effects on Forest Soils

Because 'crop rotations' for forests are of such long duration, e.g., decades to more than 100 years, it is important to consider effects of both natural and man-created disturbances to forest soils when contrasting forest soils with agricultural soils. Disturbances in forest ecosystems influence forest soils and associated dynamics of water, nutrients, and vegetation. Disturbances include windthrow (or blowdown), fires, landslides or other mass movements, and disease or insect outbreaks that defoliate and/or kill many trees. Winds that blow down trees or patches of forests (in some cases hundreds to thousands of hectares) often result in excavation, the tipping up of large root masses with accompanying soil disturbances. This mixes soil horizons, exposes fresh materials to weathering processes, and often creates pit-and-mound microrelief, adding to soil spatial variability, creating numerous microhabitats for subsequent vegetation and contributing to challenges for researchers sampling soils.

Forest fires, with their extreme ranges of intensities, durations, and extents, alter soils indirectly, via depositions of ash, charcoal, and partially burned litter, and directly, depending on temperatures imparted, in effects on soil organisms, root systems, organic matter, chemical characteristics, and physical properties. Losses of vegetation and forest floor organic matter alter system biogeochemistry and hydrology. Nutrient elements, especially nitrogen, may be volatilized in fires and lost from the forest. Residual ash materials provide a pulse of nutrients that may stimulate plant regrowth, or ash may be carried off-site by postfire winds or overland flows of water. With reduced interception, storage, and transpiration capacities owing to removal of vegetation and forest floors, infiltration, percolation, storage, and runoff of water will change; effects vary greatly depending on ecosystem and fire characteristics and postfire weather. Plant succession following a fire can result in rapid uptake and storage of nutrient elements, will contribute to rebuilding forest floors and soil organic matter, and often includes nitrogen-fixing plants that can replenish, to some extent, nitrogen loss during a fire. Because of the complex and variable interactions among fire characteristics and ecosystems, very few generalizations about fire–soil interactions are valid for all forest soils, presenting significant challenges for managers of postfire landscapes, e.g., to replant or not; is erosion protection or nutrient replenishment needed?; what vegetation will result from successional processes?

Diseases in forests (e.g., root rots; needle cast fungi) damage and kill trees, including root systems, contributing dead organic matter on soil surfaces and within soils, and influencing all related ecosystem dynamics of nutrient biogeochemistry, soil structure, and hydrology. Similarly, insect epidemics alter forests and influence soils, e.g., those of defoliating insects (e.g., gypsy moths, Douglas-fir tussock moths) may cause significant, unusual depositions of foliar organic matter to forest floors, with attendant nutrient dynamics and other system effects; large-area tree death due to bark beetles leaves 'standing dead' trees that are fire hazards as well as potential soil organic matter.

Human Effects on Forest Soils

Human activities of global, regional, watershed, and site-specific scales influence properties and processes of forest soils. These influences have been well researched by forest-soil scientists, ecologists, chemists, aquatic biologists, and forest engineers, and are only listed here rather than discussed in any detail. Effects of acid rain and related pollutants, including in some cases heavy metals, deposited in forests from the atmosphere have been well documented. Effects include changes in soil acidity, interactions with aluminum chemistry, leaching of cation nutrients, alterations of nitrogen status, influences on root growth, transfers

of chemicals to waterways, and effects on soil and aquatic biota.

Forest-management activities include use of heavy machines in forest harvests, site preparation, removal of vegetation and alteration of plant communities, and uses of fertilizers, herbicides, and pesticides. Physical site manipulations such as mechanical removal of competing vegetation and forest floor organic matter, burning, and scarification can alter soil physical properties such as porosity, infiltration capacity, percolation rates, and susceptibilities to erosion. Burning, to eliminate or reduce competing vegetation and forest floor and litter horizons for preplanting site preparation, affects soil properties and processes, as do draining and bedding to alter water tables in relation to rooting zones. Fertilizer additions and additions of herbicides and pesticides affect soil chemistry, organic matter, and biota. These management manipulations can influence forest floors and litter horizons, affect soil porosity, penetrability for roots, hydrology, gas exchanges, chemical properties, organic matter accumulations and dynamics, and soil biota, including potential symbionts of nitrogen-fixing plants.

Forest plantations of single tree species have very different ecosystems compared to many native forests, and create intensive and usually simplifying interactions with soils. These interactions may or may not influence soil properties essential for sustained productivity, e.g., nutrient element biogeochemistry, organic matter dynamics, soil biota, and dynamics of soil structure and porosity. These influences can result when vegetation is extremely simplified by use of site manipulations and herbicides to eliminate plants that would compete for site resources with the desired trees. In the most simplified plantations, with a single tree species as the sole source of organic matter recycled to soils, litter horizons and soil biota will be simplified, resulting in less-complex and -resilient nutrient-cycling dynamics. It is important that forest managers understand the implications of site manipulations for soil properties and processes essential for long-term productivity. With appropriate management, forest soils in such systems can be maintained or enhanced in their productive capacities.

'Ecosystem Services' of Forest Soils

In the context of the concept of 'ecosystem services,' i.e., beneficial interactions of forest ecosystems with the biosphere, forest soils provide several properties and functions. Among these are the vast reservoirs of stored carbon and the myriad habitats for macro-, meso-, and microorganisms contributing to biodiversity. Watershed and hydrologic functions of forest-soil ecosystems include stabilization of steep slopes and moderation of water flows to streams. Forest managers must be cognizant of these benefits and of the effects of management on altering critical soil properties and processes that influence these and other ecosystem services.

Criteria and Indicators for Soils: Sustainable Forest Management

In the context of criteria for sustainable world ecosystems and societies, the Montreal Process 'Criteria and Indicators' for forests and forestry include protection against accelerated soil erosion and maintenance of soil organic matter, as well as attendant soil processes and properties. Soil scientists have developed a variety of indicators of soil quality, i.e., measurements or estimates of soil characteristics considered essential for sustaining forests. These recognize the important functions of soils in influencing carbon and energy fixation in forest growth, nutrient cycling, hydrology, water quality, and biodiversity. Quantification of soil quality or its changes due to management is difficult because of complex, variable processes. The challenge is to identify relatively simple, outcome-oriented indicators for soils, e.g., bulk density, aeration porosity, and organic matter content. In addition to development and testing of indicators, there is a need for fine-scale assessments of soil properties linked to watershed-scale areas, because local, site-specific changes in aggregate over a large area have off-site and cumulative effects. It is also important to stratify land areas based on potential risks to soils imposed by management practices; this can provide a framework for considering management alternatives. Soil-quality indicators can serve as one basis for adaptive forest management designed to meet sustainability objectives for soils, water, and ecosystem productivity.

Summary: Soil Conditions for Forest Productivity

For forests, including aquatic ecosystems, soils provide numerous essential conditions: water-holding, -transmitting, and -supplying capacity; porosity and aeration; physical tilth for root growth; nutrient-supplying and -holding capacities; habitats for symbiotic microorganisms and for myriad soil organisms essential for decomposition and maintenance of structure. Perennial presence of forest vegetation and recycling of organic matter via litterfall and litter decomposition are essential components for productive, sustainable forest ecosystems (Figure 5). A simplified

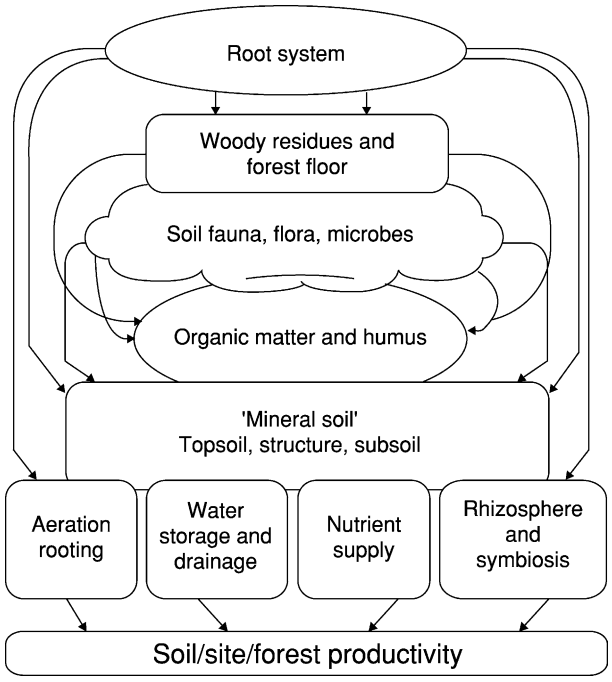


Figure 5 Forest productivity – a view from the bottom.

set of principles for managing forest soils includes these: maintain surface and incorporated organic matter; maintain topsoil – prevent erosion; maintain soil structure for water, air, nutrients, and organisms; and maintain significant land areas of forest-soil ecosystems for all the benefits they provide for the biosphere and human societies.

See also: **Carbon Cycle in Soils:** Dynamics and Management; **Classification of Soils**; **Factors of Soil Formation:** Biota; **Fauna**; **Humification**; **Mycorrhizal Fungi**; **Organic Matter:** Genesis and Formation; **Organic Residues, Decomposition**; **Productivity**; **Sustainable Soil and Land Management**; **Watershed Management**

Further Reading

Attiwell PM and Leeper GW (1987) *Forest Soils and Nutrient Cycling*. Melbourne, Australia: Melbourne University Press.

Boyle JR and Powers RF (eds) (2001) *Forest Soils and Ecosystem Sustainability*. Papers from the Ninth North American Forest Soils Conference, 1998, Tahoe City, Calif. Amsterdam, The Netherlands: Elsevier.

Burger JA and Kelting DL (1998) Soil quality monitoring. In: *The Contribution of Soil Science to the Development of and Implementation of Criteria and Indicators of Sustainable Forest Management*, pp. 17–52. SSSA Special Publication No. 53. Madison, WI: Soil Science Society of America.

Fisher RF and Binkley D (2000) *Ecology and Management of Forest Soils*, 3rd edn. Chichester, UK: John Wiley.

Green RN, Trowbridge RL, and Klinka K (1993) *Towards a Taxonomic Classification of Humus Forms*. Forest Science Monograph 29. Bethesda, MD: Society of American Foresters.

Heilman PE, Anderson HW, and Baumgartner DM (1979) *Forest Soils of the Douglas-Fir Region*. Pullman, WA: Washington State University Cooperative Extension Service.

Lutz HJ and Chandler RF Jr. (1946) *Forest Soils*. New York: John Wiley.

McFee WW and Kelly JM (eds) (1995) *Carbon Forms and Functions in Forest Soils*. Papers from the Eighth North American Forest Soils Conference, 1993, Gainesville, Fla. Madison, WI: Soil Science Society of America.

Powers RF, Tiarks AE, and Boyle JR (1998) Assessing soil quality: practicable standards for sustainable forest productivity in the United States. In: *The Contribution of Soil Science to the Development of and Implementation of Criteria and Indicators of Sustainable Forest Management*, pp. 53–80. SSSA Special Publication No. 53. Madison, WI: Soil Science Society of America.

Ramann E (1893) *Forstliche Bodenkunde und Standortslehre*. Berlin, Germany: Springer-Verlag.

Schoenholtz S, Van Miegoret H, and Burger JA (2000) A review of chemical and physical properties as indicators of forest soil quality: challenges and opportunities. *Forest Ecology and Management* 138: 335–356.

Switzer GL and Nelson LE (1972) Nutrient accumulation and cycling in loblolly pine (*Pinus taeda* L.) plantation ecosystems: the first 20 years. *Soil Science Society of America Proceedings* 36: 143–147.

Tamm O (1950) *Northern Coniferous Forest Soils*. (Translated by Anderson ML.) Oxford, UK: Scrivener Press.

Weetman G and Webber B (1972) The influence of wood harvesting on the nutrient status of two spruce stands. *Canadian Journal of Forest Research* 2: 351–369.

Youngberg CT (ed.) (1965) *Forest-soil Relationships in North America*. Proceedings of the Second North American Forest Soils Conference, Oregon State University, 1963. Corvallis, OR: Oregon State University Press.

FOURIER TRANSFORM INFRARED SPECTROSCOPY

D Peak, University of Saskatchewan, Saskatoon, SK, Canada

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Fourier transform infrared (FTIR) spectroscopy is a form of vibrational spectroscopy that is useful in the study of a variety of soil chemical processes. In the mid-infrared (mid-IR) range, vibrations arise from many environmentally important molecules such as organic acids, soil organic matter, mineral phases, and oxyanions. It is possible to utilize FTIR spectroscopy as a quantitative analytical method and also as a tool to determine bonding mechanisms in solids and on surfaces. Molecular vibrations can be related directly to the symmetry of molecules, and so it is often possible to determine precisely how a molecule is bonding on surfaces or as a component in a solid phase from its infrared spectrum. Many experimental methods exist for probing samples of various states and at different spectral regions.

The Nature of Molecular Vibrations

The infrared region of the electromagnetic spectrum lies between 10 and 10 000 cm^{-1}, which can be expressed as a frequency of 10^{12}–10^{14} Hz. This frequency overlaps with the frequency of molecular vibrations. When infrared radiation is the same frequency as a molecular vibration, it can induce a transition to a higher energy state. This transition will only occur if it induces a dipole. A true molecular vibration is defined as a movement of the atoms of a molecule that do not change the center of mass and that involve no rotation. Vibrations that obey this definition are known as the 'normal modes.' For a nonlinear molecule, the normal modes may be due to symmetric or asymmetric stretching, and in-plane or out-of-plane bending. Not all of a molecule's normal-mode vibrations will be active when subjected to infrared energy. Instead, the symmetry of a molecule will affect the position and number of infrared-active peaks.

Infrared Peaks of Interest

In the mid-IR range (400–4000 cm^{-1}), there are several extremely important and useful classes of bonds that can be studied. FTIR spectroscopy has been used extensively by both geologists and soil scientists to study clay minerals and other inorganic soil components. The structure of minerals can substantially affect their infrared spectra, and entire textbooks are available with tables of peak positions and figures containing spectra of reference minerals. Water has an intense infrared absorbance in the mid-IR region, which typically is considered a limitation of the technique for environmental research. However, the strong absorbance of water bands has been exploited by researchers who investigated the effects of humidity changes on self-supported films of clay minerals.

Many organic compounds also have diagnostic peaks in the infrared region, and FTIR has been used to study the nature of soil organic matter as well as the adsorption of organic acids. Many metals and metalloids form negatively charged oxyanion molecules in aqueous solution, and FTIR spectroscopy has extensively been used to study their adsorption and precipitation reactions. Figure 1 contains infrared spectra from selected experiments using FTIR to study reactions of: (1) natural organic matter (NOM), and (2) inorganic minerals in soils. Infrared absorbance bands in Figure 1a have been labeled to show how changes in functional-group makeup can give rise to new infrared peaks. Also notice how many of the peaks in the NOM disappear or shift in position when NOM is adsorbed on to a mineral surface. This is due to chemical bonding with carboxylic and phenolic NOM groups and the iron oxide surface. In Figure 1b, a mineral transformation is monitored using FTIR spectroscopy. The initial γ-Al_2O_3 solid phase is not stable in water, and over the 2.5-month experimental period it transforms into a crystalline gibbsite $Al(OH)_3$ phase.

Infrared Spectral Analysis

A detailed analysis of the theoretical infrared spectrum from first principles is not always required. It is possible simply to compare unknown samples to those of spectra with known molecular configurations to determine the structure of an unknown. This is referred to as spectral fingerprinting. This technique relies on one's ability to synthesize reliably reference compounds of known chemical structure or obtain their spectrum from a library. This approach is reasonably straightforward in the case of organic molecules, as large commercial software spectral libraries are available. In the case of unknown inorganic spectral identification, however, suitable standards can often be difficult to synthesize, and aqueous

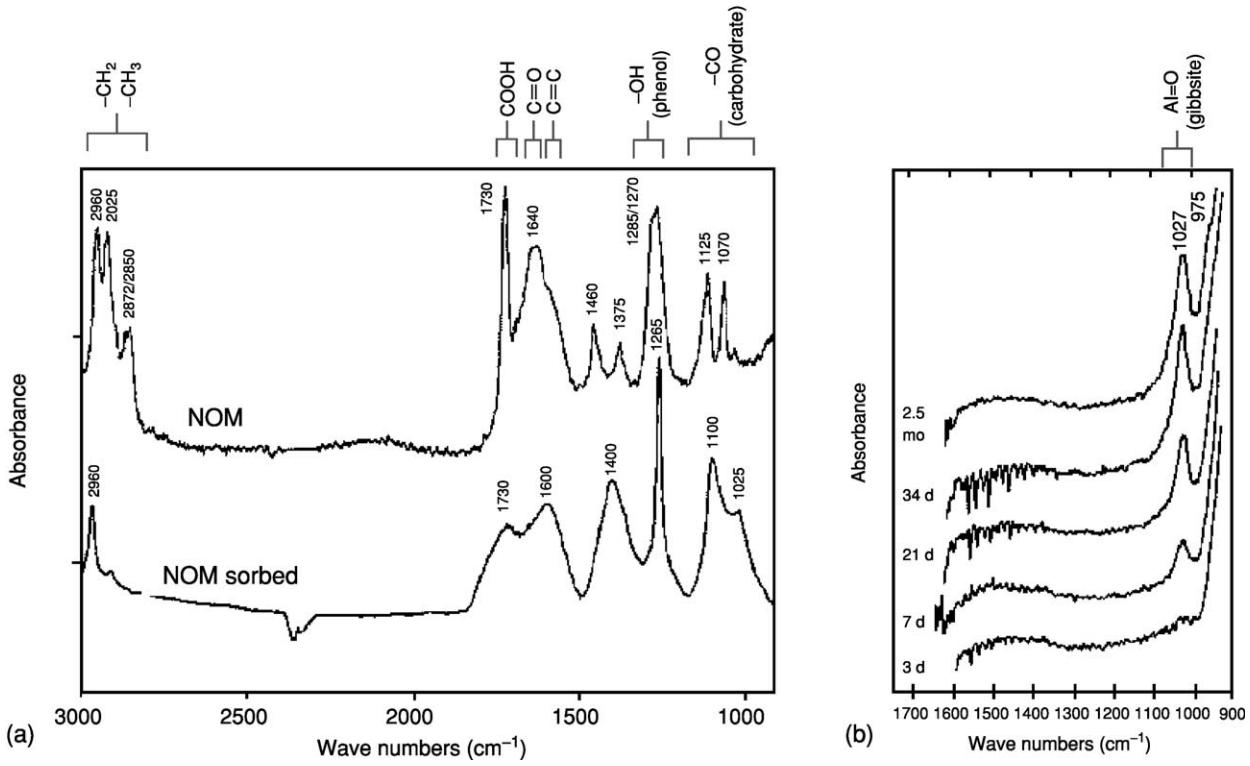


Figure 1 (a) Infrared spectra of natural organic matter in water compared with natural organic matter (NOM) adsorbed on an iron oxide. Reproduced with permission from Gu B, Schmitt J, Chen J, Liang L, and McCarthy JF (1994) Adsorption and desorption of natural organic matter on iron oxide: mechanisms and models. *Environmental Science and Technology* 28: 28–46. (b) Infrared spectra of γ-Al_2O_3 aged from 3 to 75 days. The peak at 1027 cm^{-1} is due to transformation of the initial Al_2O_3 into a crystalline gibbsite phase. Adapted from Wijnja H and Schulthess CP (1999) ATR-FTIR and DRIFT spectroscopy of carbonate species at the aged γ-Al_2O_3 – water interface. *Spectrochimica Acta. Part A Molecular and Biomolecular Spectroscopy* **55**(4): 861–872, with permission.

standards are often present as mixtures of species. For these reasons, assigning peaks based on rules from molecular symmetry is often necessary.

Symmetry rules assign a molecule to a unique point group based upon the arrangement of its atoms in space. A common example of how infrared spectra may be related to molecular symmetry is the case of sulfate. FTIR spectroscopy has been used by many researchers to study sulfate bonding in mineral structures and as surface adsorption complexes. The clear relationship between the molecular symmetry of sulfate species, its point group, and its infrared spectrum is an example of the utility of this approach. In the case of aqueous sulfate, there are two normal modes that are accessible to spectroscopic investigation. They are the nondegenerate symmetric stretching ν_1, and the triply degenerate asymmetric stretching ν_3 bands. As a free anion in solution, sulfate has tetrahedral symmetry and belongs to the point group T_d. For this point group, only one broad peak, due to the ν_3 stretch, is observed at approximately 1100 cm^{-1}. If sulfate is covalently bonded to one or more cations, the symmetry is lowered, a ν_1 peak becomes infrared active, and the ν_3 band splits into more than one peak. When sulfate is bound as a monodentate complex, the ν_3 band splits into two peaks, while the ν_1 band becomes fully active at approximately 975 cm^{-1}, and C_{3v} symmetry results. If sulfate forms a bidentate binuclear (bridging) complex, the symmetry is further lowered to C_{2v}, and the ν_3 band splits into three bands between 1050 and 1250 cm^{-1}, while the ν_1 band is usually shifted to a higher wave number. The relationship between molecular symmetry and the number and position of infrared peaks are detailed in Figure 2. Sulfate is typical of tetrahedral oxyanions, and so the relationship also may be readily extended to other tetrahedral oxyanions of environmental relevance (such as chromate, selenate, phosphate, and arsenate).

Similar rules may be derived for molecules of other point groups also, making the symmetry approach very powerful in the assignment of molecular configuration from FTIR spectra. However, one must be aware that infrared peaks may arise from many potentially occurring molecular vibrations. Of particular concern in environmental science is the fact that interactions with protons will substantially affect the infrared spectrum of sulfate, not only when

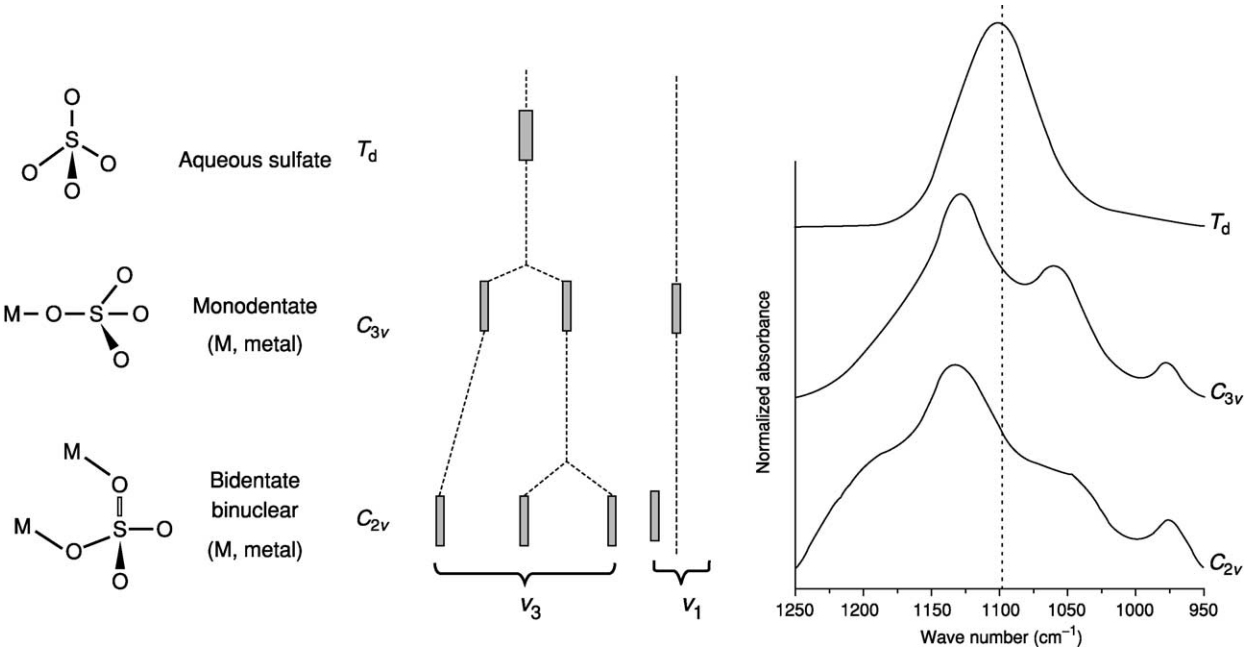


Figure 2 Relationship between molecular symmetry and infrared spectra for aqueous, monodentate, and bidentate sulfate species. Adapted from Peak D, Ford RG, and Sparks DL (1999) An *in situ* ATR-FTR investigation of sulfate bonding mechanisms on goethite. *Journal of Colloid Interface Science* 218: 289–299.

covalently bound to sulfate (as bisulfate or sulfuric acid), but also when hydrogen bonds are formed with waters of hydration. To improve peak assignment in such cases, experiments are usually repeated in D_2O as a solvent. Substitution of D for H in surface complexes or mineral structures will shift the position of infrared peaks, while peaks that do not involve protons are unaffected. It is also possible to calculate theoretical infrared spectra using molecular modeling software. The absolute position of peaks may not be correct, but trends in how species are affected by the presence of protons or metals are still quite helpful in peak assignment.

FTIR as a Quantitative Tool

The infrared absorbance of a molecular vibration obeys the Beer–Lambert law. This means that FTIR spectroscopy, if properly calibrated, is a quantitative technique. The relationship is as follows:

$$A = \epsilon b C \qquad [1]$$

where A is absorbance, ϵ is the molar absorptivity coefficient, b is the thickness of the sample (or pathlength), and C is the concentration of the molecule in the sample. Note that ϵ is specific to a given vibration, and that A is typically reported in terms of peak height or integrated absorbance from a fitted peak. Changes in sample configuration for different types of FTIR experiments can make quantification difficult.

Instrumentation

FTIR actually refers to a class of spectrometers that are distinguished from older, double-beam infrared units by their use of a single beam and an interferometer and detector to collect a spectrum over the entire wavelength of interest simultaneously. This raw data, called an interferogram, is an oscillation over time. This interferogram is then converted to a function of frequency, as in a traditional infrared spectrum, by use of a mathematical procedure called Fourier transformation. FTIR units have largely replaced older double-beam units because they are have better spectral resolution, allow for much more rapid data collection, and have an improved signal-to-noise ratio.

Common Experimental Techniques

Several different FTIR experimental techniques are commonly employed by environmental researchers. These are transmission mode experiments with pressed disks, diffuse reflectance experiments with powder samples, and attenuated total reflectance studies with aqueous solutions and hydrated suspensions. The primary differences in the commonly used techniques are: (1) the amount of sample preparation that is required, (2) the spectral range that is accessible to the technique, and (3) the detection limits. Each technique may be particularly well suited to

certain research questions, and more than one technique is sometimes required.

Transmission-mode Experiments

Transmission-mode experiments are the both the simplest and most common type of FTIR experiment. Samples are diluted in some medium and placed perpendicular to the infrared beam. A detector measures the amount of infrared beam that is transmitted through the sample. To obtain the spectrum of only the sample, the spectrum of the pure medium is first collected and used as a reference. A spectra of the sample in the medium is then collected and ratioed to the reference to obtain the spectrum of the sample. A typical experimental configuration for transmission-mode experiments is shown in Figure 3.

Solid samples are typically diluted to 1–2% wt in KBr salt and then made into a thin disk with a press. The final product is a thin, clear disk that is placed in a sample holder and analyzed. For transmission mode experiments to be successful, water must be removed from the system, so samples are heated to remove as much water as possible prior to dilution. Care must be taken with amorphous and unstable solids to avoid transformations that may occur at elevated temperatures. Organics can also often be studied in transmission mode by simply placing them in a thin liquid cell. The medium chosen as a solvent should have a relatively low infrared absorbance and not have infrared peaks that overlap with those of the organic of interest.

Transmission-mode experiments can provide quantitative information more easily than some other experiments because Beer's Law is directly applicable. The path length through the sample must be known, and molar absorptivity can be learned by collecting several spectra of samples with known concentrations to construct a calibration curve. Once the path length and molar absorptivity are known, concentration and absorbance can be determined.

This technique is quite useful for soil scientists who are studying mineral transformations or who need quantitatively to determine the mineral composition in unknown samples. However, the technique involves substantial sample modification that may change the structure of both solid phases and sorbed surface species. Particularly in mixed systems, peaks of minor components may be masked by other species of higher concentration.

Diffuse Reflectance Infrared Fourier Transform Experiments

The diffuse reflectance technique was developed partly to address some of the shortcomings of the pressed-disk method of sampling, and also to extend FTIR spectroscopy to additional types of samples. Many substances, particularly powders, scatter light in all directions in a diffuse manner. The diffuse reflectance technique isolates the portion of the infrared light that is diffusely scattered from that which is either transmitted or specularly reflected and collects it as a diffuse reflectance spectrum. A typical experimental configuration for transmission-mode experiments is shown in Figure 3.

DRIFT spectroscopy is conducted only on powder samples. Samples may be run neat (with no dilution),

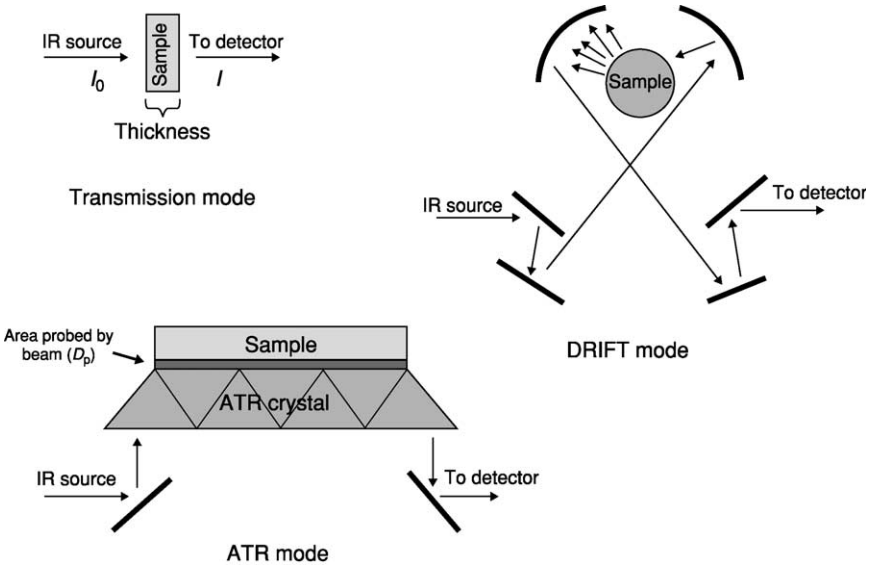


Figure 3 Common sampling methodologies for transmission-mode, diffuse reflectance (DRIFT), and attenuated total reflectance (ATR) Fourier transform infrared experiments. I_0, energy prior to the sample; I, energy after the sample; D_p, depth of penetration through the sample.

or diluted with a nonabsorbing matrix such as KBr. Samples are typically ground with a mortar and pestle to avoid large particles with flat surfaces that may distort the spectrum by exhibiting specular reflectance. Samples are not pressed into disks as with transmission experiments, but are simply packed into a sample holder. This makes the collection of DRIFT spectra somewhat simpler than transmission-mode experiments. It also avoids changes in sample structure that may occur when subjected to elevated pressure. However, quantitative DRIFT studies are more difficult than transmission-mode IR, because steps in sample preparation can substantially affect a diffuse reflectance spectrum. Specifically, the particle size, the sample homogeneity, and the degree of packing will influence the results.

Regardless of the care taken in sample preparation, DRIFT data differ from data obtained in a transmission-mode experiment. Peak intensities at higher wave number are typically compressed, and all peak shapes tend to be rather rounded compared with sharp peaks from pressed disks. A mathematical correction can be applied to the raw spectra to compensate for these anomalies and make DRIFT data directly comparable with transmission-mode references. This conversion is known as the Kubelka–Munk equation, and it can usually be performed by the software used to control the FTIR.

The primary advantage of DRIFT over transmission-mode studies is that sample preparation does not require pressing disks. This means samples can be prepared more quickly and with less effort. However, samples are still dried and diluted in salts, so sample preparation may still affect the chemical species in an environmental sample. The primary drawback of DRIFT is that it trades ease of sample preparation and ability to observe weaker bands by analysis of neat samples for additional complexity in quantitative analysis.

Attenuated Total Reflectance Experiments

The attenuated total reflectance (ATR) method is a recent advancement in FTIR that allows spectra to be collected in the presence of water. Water is an extremely strong infrared absorber, and traditional sampling methods require drying to remove interferences from H_2O. The ATR technique circumvents this limitation by directing the infrared beam through a long crystal. The beam is redirected through the crystal at an angle and bounces multiple times internally and then exits to the detector. The infrared beam produces an evanescent wave on the surface of the crystal, and hydrated samples in close contact with the crystal can be analyzed by comparing the reflectance of the sample with that of the bare crystal. A typical experimental configuration for ATR-mode experiments is shown in **Figure 3**.

Since the ATR technique presents a substantial deviation from a transmission-mode experiment, experimental complications that affect the adherence of ATR data to the Beer–Lambert law are to be expected. For example, the effective path length of the beam is required for quantitative research and is related to the number of reflections through the crystal and the depth of penetration into the sample. The number of reflections is related to the ATR crystal's physical properties such as angle of incidence of the crystal, and the crystal's physical length and thickness. The depth of penetration into the sample is also related to the crystal's physical dimensions, but is also a function of both the crystal and the sample's refractive index and the wavelength of light that is passed through the crystal. ATR crystals may have 30°, 45°, or 60° effective angles of incidence, and may be made of a variety of materials. The most common and widely used material for ATR crystals is ZnSe, but diamond, amorphous material transmitting IR (AMTIR), germanium, and silicon are also available. Different crystal materials may be preferred due to their refractive index, their useful spectral range, or their chemical sensitivity.

The fact that sample penetration depth changes with wavelength presents a serious problem in attempting quantitative experiments with the ATR method. For this reason, ATR data are properly reported as percentage reflectance (%R) rather than percentage transmittance (%T), as the intensity of light transmitted is not a constant throughout the spectrum with ATR, but instead is a function of wavelength. Fortunately, a mathematical correction can be made to ATR data that accounts for the wavelength dependence of penetration depth. This correction is typically available in the instrument's data-collection software.

With the ATR technique, it is possible to analyze a wide range of samples. Solids can be pressed into contact with the ATR crystal, while liquid samples and hydrated suspensions can be placed into a sample holder that has a trough with a volume of a few milliliters. This flexibility in sampling has made it possible to conduct adsorption studies of organic acids and oxyanions *in situ* with FTIR. In fact, the ability to monitor the mechanism of adsorption on mineral surfaces is the primary application of ATR-FTIR in environmental science.

Two experimental approaches are commonly utilized in adsorption studies: analysis of reacted suspensions of soil solids, and analysis of adsorption on to solids deposited on to the ATR crystal. Both experimental approaches have merits and drawbacks. It is

faster and easier to conduct macroscopic adsorption experiments in the laboratory and then use ATR-FTIR spectroscopy to analyze representative samples than it is to conduct all experiments *in situ* with deposited solids. This also provides the opportunity to know the exact surface loading and reaction conditions of the sample prior to analysis. However, using wet pastes or slurries presents some experimental difficulties also. To obtain a spectrum of the adsorbed species, one must separate the solid from the solution via filtration, resuspend the solid in the supernatant at a constant solid-to-solution ratio for all samples, and then collect a spectrum of the supernatant, the unreacted solid, and the reacted solid. By subtracting both the supernatant and the unreacted solid spectra from the reacted spectrum, the result is the spectrum of the adsorbed species.

In contrast, the deposition method places a thin coating of a mineral on the ATR crystal that is constant throughout the experiment. A reference spectrum of the ATR crystal, the mineral deposit, background electrolyte, and water is collected initially. Then, reactant is added to the solution at a low (micromolar) concentration. The reactant is concentrated at the mineral–crystal interface so that adsorbed species are above detection limits. Since the solid concentration at the surface is constant and the solution-phase reactant is not detected, a high-quality spectrum of the absorbed ion can be obtained. Due to improved spectral quality, results from deposition studies are typically much easier to fit and interpret.

The primary advantage of the ATR method for environmental science research is that it allows environmental samples to be analyzed with normal amounts of water present and at ambient pressures. The results from ATR are therefore much more representative of chemistry in natural systems. However, ATR experiments are more difficult to compare with traditional transmission studies, difficult to conduct reliably in a quantitative manner, and the experimental procedure is more time-consuming than other methods.

Further Reading

Farmer VC (1974) *Infrared Spectra of Minerals*. London, UK: Mineralogical Society.

Johnston CT, Farmer W, and Aochi Y (1996) Fourier transform infrared and Raman spectroscopy. In: Sparks D (ed.) *Methods of Soil Analysis, part 3. Chemical Methods*, 3rd edn, pp. 269–321. Book Series no. 5. Madison, WI: Soil Science Society of America.

Nakamoto K (1986) *Infrared and Raman Spectra of Inorganic and Coordination Compounds*. New York: John Wiley.

Noggle J (1998) *Physical Chemistry*. New York. NY: Harper-Collins.

Shriver DF, Atkins PA, and Langford CH (1994) *Inorganic Chemistry*. New York, NY: WH Freeman.

Suarez DL, Goldberg S, and Su C (1998) Evaluation of oxyanion adsorption mechanisms on oxides using FTIR spectroscopy and electrophoretic mobility. *ACS Symposium Series. American Chemical Society* 715: 136–178.

Wilson EB, Decius JC, and Cross PC (1980) *Molecular Vibrations: The Theory of Infrared and Raman Vibrational Spectra*. New York, NY: Dover Publications.

FRACTAL ANALYSIS

Y Pachepsky, USDA Agricultural Research Service, Beltsville, MD, USA
J W Crawford, SIMBIOS, University of Abertay, Dundee, Scotland, UK

Published by Elsevier Ltd.

Introduction

Geometric irregularity is an intrinsic property of soils. Pores, particles, surfaces, roots, and living organisms in soils have a wide range of sizes and shapes. Soil surface is rough, and soil map contours have complex shapes. Geometric irregularity in soils is easy to perceive and observe, but quantifying it has long presented a daunting challenge. Such quantifying becomes imperative because the geometric irregularity is both a cause and a reflection of soil spatial variability that strongly affects soil functioning in ecosystems and soil management.

Geometric quantities in soils are measured with different measurement scales, i.e., lengths, areas, or volumes within which curves or surfaces are assumed to have a geometric shape, composed of straight lines, circular arcs, parts of planes, or sphere surfaces. Measurement scale is often called 'support,' 'resolution,' or simply 'scale.' Because actual curves and surfaces are irregular within the measurement scale, total length, area, or volume does not remain

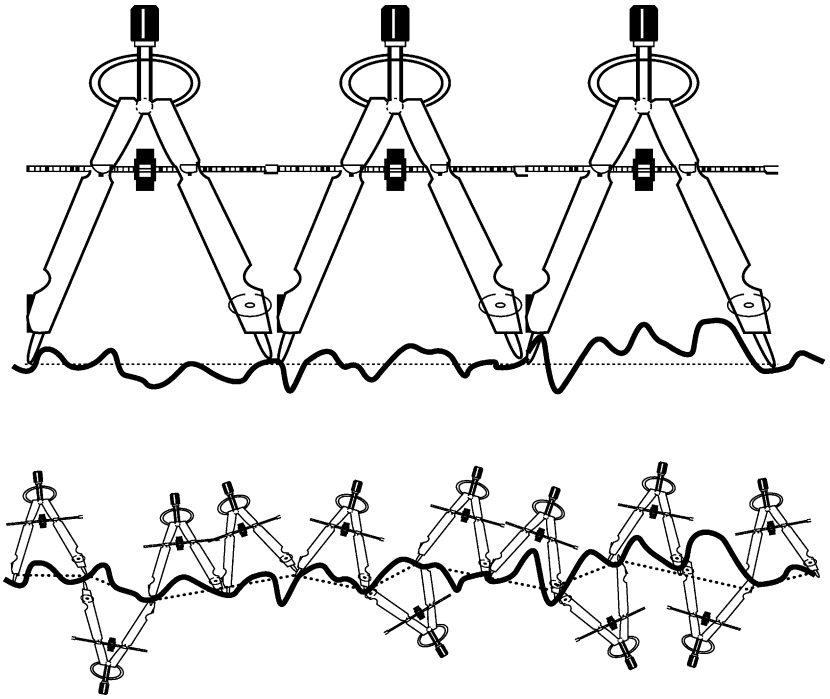


Figure 1 Measurements with traditional geometric tools ignore surface irregularity within the observation scale. Measured length of a line (dotted) is the sum of distances between the compass dividers. The larger divider step gives the total length of 105 length units, whereas the smaller divider step results in 120 length units.

constant as the measurement scale changes (Figure 1). A decrease in measurement scale usually causes an increase in total length, area, or volume. Thus soil geometric properties appear to be scale-dependent.

One class of geometric objects that have dependence of total length, area, or volume on the measurement scale consists of objects that reveal similar geometric shapes when observed at different measurement scales. Such objects were termed 'fractals' by B.B. Mandelbrot, who suggested that fractals rather than regular geometric shapes such as segments, arcs, circles, and spheres are more appropriate to approximate irregular natural shapes that have hierarchies of ever-finer detail. This started the expansion of applications of fractal geometry, which had became very popular during recent decades, because of its promise to relate features of natural objects observed at different scales.

Fractal Objects

Geometric fractal objects reveal similar features at different scales because they are constructed iteratively, by repetition of the same form- or shape-changing operation at different measurement scales. If this operation does not change from one scale to another, then the result is an ideal geometric fractal. An example of such construction is given in Figure 2 for the fractal object 'Menger sponge.' The initial shape, or initiator, is a cube. The shape-changing operation, or generator, divides the cube into 27 smaller, equal cubes and extracts seven cubes, one from each facet and one from the center. Figure 2 shows the first two stages of the iterative application of this operation. A Menger sponge is self-similar because it can be broken down into arbitrary small pieces, each of which is a small replica of the entire structure. If the Menger sponge of a certain size is reduced three times, then 20 copies of the reduced sponge can be pasted together to give back the original Menger sponge; a nine-times reduction requires 400 copies; a 27-times reduction requires 8000 copies, etc. In all self-similar objects, there is a relationship between the reduction factor R and the number of pieces N into which the object is divided. Those relationships follow a power law:

$$N = R^{-D_{ss}} \qquad [1]$$

where the exponent D_{ss} is called the self-similarity, or similarity, dimension. The value of D_{ss} can be found by equating logarithms of both parts in Eqn [1], i.e., $D_{ss} = -\log(N)/\log(R)$. For the Menger sponge, the reduction factor is one-third $D_{ss} = -\log(20)/\log(1/3) = -\log(400)/\log(1/9) = -\log(8000)/\log(1/27) = 2.727$. The reduction factor refers to a single iteration; the total reduction after i iterations is $(1/3)^i$.

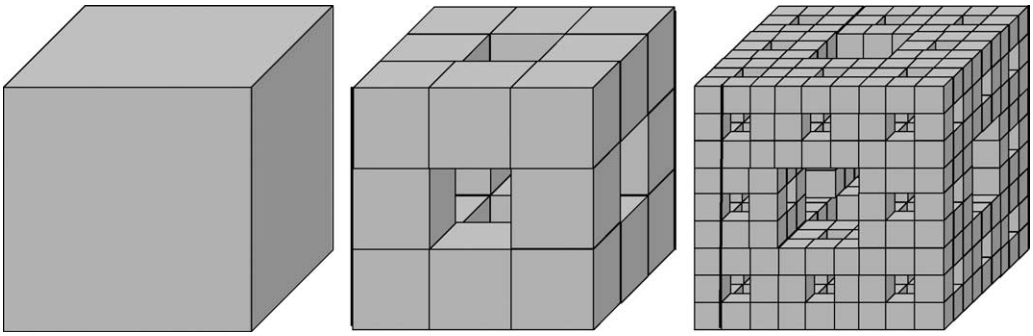


Figure 2 The first two iterations of the Menger sponge construction.

The self-similarity dimension receives a geometric meaning when a fractal object is used as the model of a real system, and geometric measures of this object are equated to geometric measures of the system. In this case, the measurement scale is the reduced length, area, or volume of the initiator. For example, let empty parts of the Menger sponge be used to simulate pores in a porous medium at different measurement scales. Let V_0 be the volume of the initiator. The measurement scale is V_0R^3, the number of similar pieces at this scale is N, and their total volume is:

$$V = NV_0R^3 = V_0R^{3-D_{ss}} \qquad [2]$$

As the measurement scale becomes finer and R decreases, the remaining solid volume decreases, approaching zero, and the pore volume approaches V_0. Because the fractal object represents remaining mass in this case, the dimension D_{ss} is interpreted as the mass fractal dimension. Depending on the scale-dependent measure of interest represented by the specific fractal object or on a physical process simulated by the generator, D_{ss} is interpreted as being the surface fractal dimension, fragmentation fractal dimension, etc. Eqns [1], [2], and similar relationships hold over a range of scales, i.e., they are scale-invariant within this range. They are interchangeably referred to as 'scaling relationships,' 'fractal scaling laws,' or simply 'scaling.'

A common way to reveal a fractal scaling is to plot the measurement scale against the property of interest in double-logarithmic, or log-log, scale. The presence of a linear relationship indicates that a fractal model may be appropriate to simulate the scale dependence in the property. The fractal dimension can then be found from the slope of the regression line. For example, Eqn [2] in log-log scale can be transformed into:

$$\log V = \log V_0 + (3 - D_{ss})\log R \qquad [3]$$

and the fractal dimension D_{ss} equals 3 minus the slope of the regression line.

The iterative construction of ideal fractals continues ad infinitum, whereas natural objects have a minimum and maximum observable scales, or lower and upper cutoff scales beyond which Eqn [1] breaks down. The value of the measure for the initiator, for example the V_0 value in Eqn [2], is not known if the fractal scaling is only valid within a specific scale range. Thus, four values – fractal dimension, upper and lower scale cutoff, and the initiator measure – are parameters of a fractal model in soil studies. Usually, the range of scales between cutoffs does not exceed two orders of magnitude if scales are expressed in units of lengths. The larger the fractal dimension the more rugged is the surface, the more irregular is the line, and the more complex is the pore space.

Cutoffs typically coincide with changes in processes causing or reflecting scaling; for example, pore-space fractal scaling inferred from water retention data in the capillary range is limited by the water content at wilting point and the water content at which the macropores become emptied. The meaning of the initiator measure is not currently well understood, although concepts of topothesy, crossover length, and lacunarity have been explored to interpret this fractal parameter.

Fractal dimensions of geometric fractals have limits reflecting space-filling properties. Straight line, plane, and Euclidean space have dimensions one, two, and three, respectively. A fractal line has its fractal dimension between one and two, i.e., it does not fill the whole plane although the larger the fractal dimension is the more rugged is the line and the closer it is to occupy a part of the plane. Similarly, a fractal surface has its fractal dimension between two and three. Fractal dimensions of surface and volume fractals theoretically cannot exceed three.

Ideal geometric fractals as in **Figure 2** cannot be found in real porous media in the same way as cannot be found ideal geometric lines, spheres, cylinders, etc. Applying random changes in the generator procedure

results in constructing stochastic fractals that look much more like fractals in nature. Because random changes are applied, various realizations of stochastic fractals with the same initiator and generator differ. Stochastic fractals retain the important feature of dependence of geometric properties on a measurement scale, only this feature is preserved in a statistical sense. This means that Eqn [1] will not hold exactly for any realization of stochastic fractals at any measurement scale. Nevertheless, a plot of the measurement scale versus the property of interest in log-log scale is close to linear for each realization of a stochastic fractal, and statistical hypotheses of equality between average slope values and average intercept values for any two realizations cannot be rejected. Realizations of stochastic fractals are statistically similar.

Data on surface roughness, topography, and distributions of properties along spatial transects do not follow the self-similarity law (Eqn [1]), but demonstrate another, anisotropic type of scale-dependence, 'self-affinity.' Self-similar fractals are isotropic, which means that, in two-dimensional xy space, $f(x,y)$ is statistically similar to $f(Rx,Ry)$ where R is a reduction factor. Self-affine fractals are anisotropic, which means that in the two-dimensional xy space, $f(x,y)$ is statistically similar to $f(Rx,R^{H}y)$ where H is called the Hurst exponent. That is, changing the horizontal measurement scale R times corresponds to changing the vertical measurement scale R^{H} times. High values of the Hurst exponent indicate some memory or autocorrelation in the data. Low values suggest an anticorrelation or self-correcting response. Self-affine surfaces are observed in three dimensions, so that the vertical scale has to be decreased R^{H} times and scales in horizontal directions have to be decreased R times to obtain a statistically similar surface. The fractal dimension of this surface is $D_s = 3 - H$. A line formed as a cross-section of any fractal self-affine surface and a vertical plane also displays self-affine properties. This line will have a fractal dimension D_L, related to the Hurst exponent H as:

$$D_L = 2 - H \qquad [4]$$

The value of D_L is easy to interpret with elevation data. Any two consecutive height variations are likely to have opposite signs when the fractal dimension D_L is larger than 1.5 at any particular scale. Any two consecutive height variations are likely to have the same sign when the fractal dimension D_L is less than 1.5 (**Figure 3**).

Multifractals present yet another example of self-similar geometric objects that have been successfully applied to simulate properties of soils and other

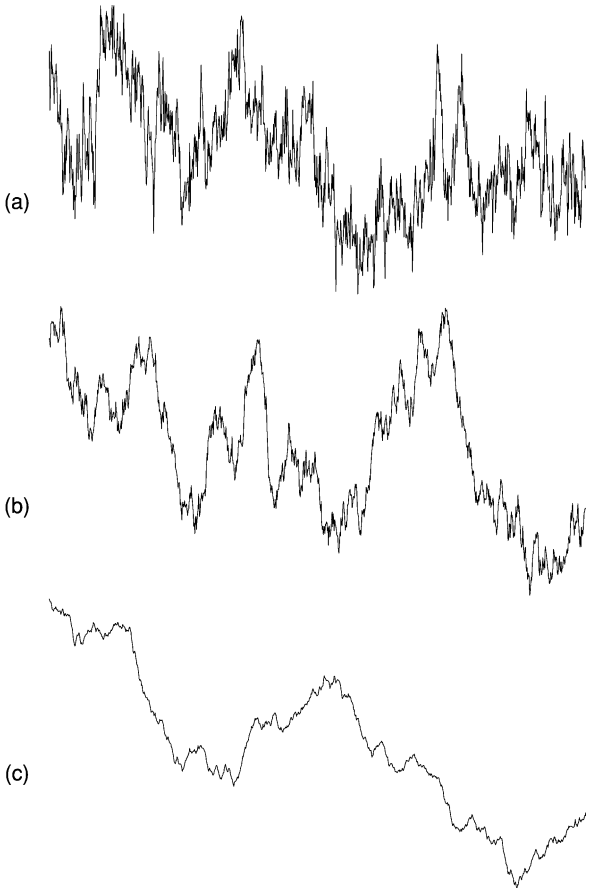


Figure 3 Self-affine fractal lines: (a) fractal dimension $D_L = 1.8$; (b) $D_L = 1.5$; (c) $D_L = 1.2$.

natural systems. Unlike the Menger sponge, in which the similarity of structure persists across iterative steps, multifractals are generated by cascade processes that create more heterogeneity as the resolution decreases. An example of a cascade process is shown in **Figure 4**. A unit-length segment is the initiator. The reduction factor is 2, and random split of each segment from the previous step into two new segments occurs as the generation progresses. The random-split variable ranges from 0 to 1, with a mean value of 0.5. The distributions of lengths are monofractal if the distribution of the random-split variable does not change from one scale reduction to another. However, if the variance of the random-split variable decreases as the generation continues, the distributions become multifractal. Distinguishing between monofractal and multifractal distributions can be based on considering statistical moments about zero:

$$E[W^{q}] \cong \frac{\sum_{i=1}^{m} W_i^{q}}{m} \qquad [5]$$

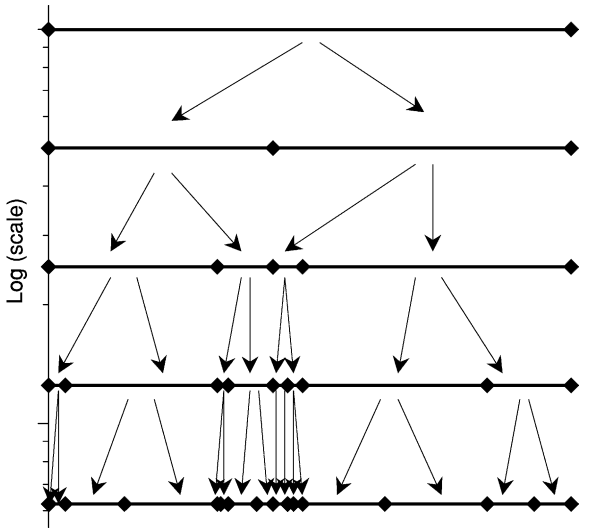


Figure 4 Generating multifractal-size distribution with a cascade process.

where q is the order of the statistical moment, W is the scale-dependent measure, and m is the total number of measurements. Multiscaling requires that:

$$E\left[W_R^q\right] = R^{K(q)}E[(W_0)^q] \qquad [6]$$

where K is the scaling exponent, W_0 is the measure at the measurement scale zero, and W_R is the same measure at the scale changed R times. If $K(q)$ is proportional to the moment order q, then the scaling is called 'simple scaling.' If $K(q)$ is a nonlinear function of q, scaling is interpreted as multiscaling or multifractal.

Various other types of fractal objects can be generated that are suitable to simulate scaling in soil properties. Scaling exponents and fractal dimensions have different meanings for those objects, because the construction algorithms are different.

Physical Models of Emerging Fractal Scaling

Many physical processes have been shown to generate fractal scaling. Some of them are relevant to soil development and functioning, and the correspondent fractal models are being applied to soils. One example is the scale-invariant fragmentation that leads to dependence between the number of fragments and their size:

$$N_r = kr^{-D_F} \qquad [7]$$

where r is the fragment radius, N_r is the number of fragments having radii less than r, k is a constant, and D_F is the fragmentation fractal dimension. The incomplete fragmentation process has the fractal dimension that can be related to density-size scaling of soil aggregates and appears to be the mass fractal dimension.

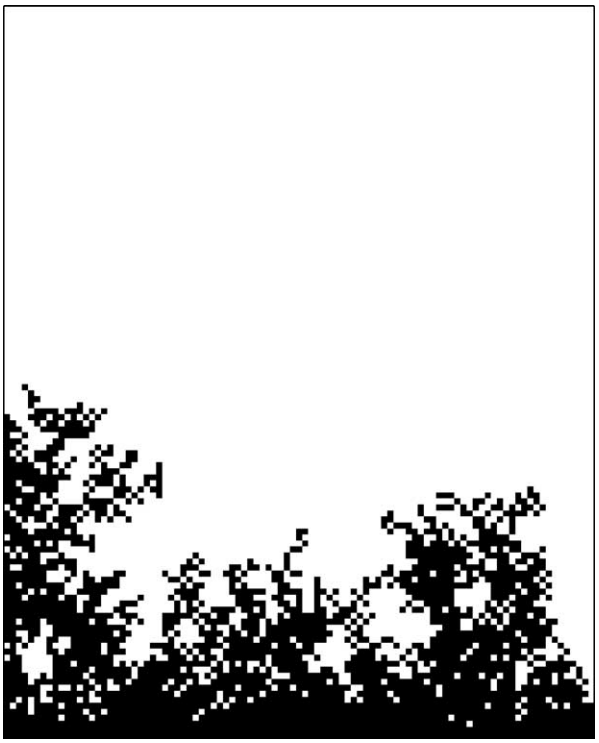


Figure 5 Diffusion-limited deposition of particles on a surface.

Agglomeration on a substrate can also generate fractal structures. Particles may arrive at the surface by random diffusion or with ballistic energy. They may either just stick together as they touch each other or there is a chemical reaction involved which may change the local concentration gradient and sticking probability. Allowing particles to walk randomly until they strike the surface or another particle, which is already motionless, builds the fractal coating. An example of diffusion-limited deposition on a substrate is shown in Figure 5. Clusters are formed, and the fractal dimension of such clustering can be measured by log-log plotting the total number of particles as a function of distance from the surface. The fragmentation dimension appears to be close to the surface fractal dimension computed for pores if particles have a fractal-size distribution following the fragmentation law (Eqn [7]) and their sedimentation is simulated as a ballistic process of agglomeration on a substrate (Figure 6). The number of particles in a cross section of fractal agglomerate on a surface can be interpreted as a solute concentration, and the agglomerate can be viewed as a model for the distribution of solute particles in a porous medium (Figure 7).

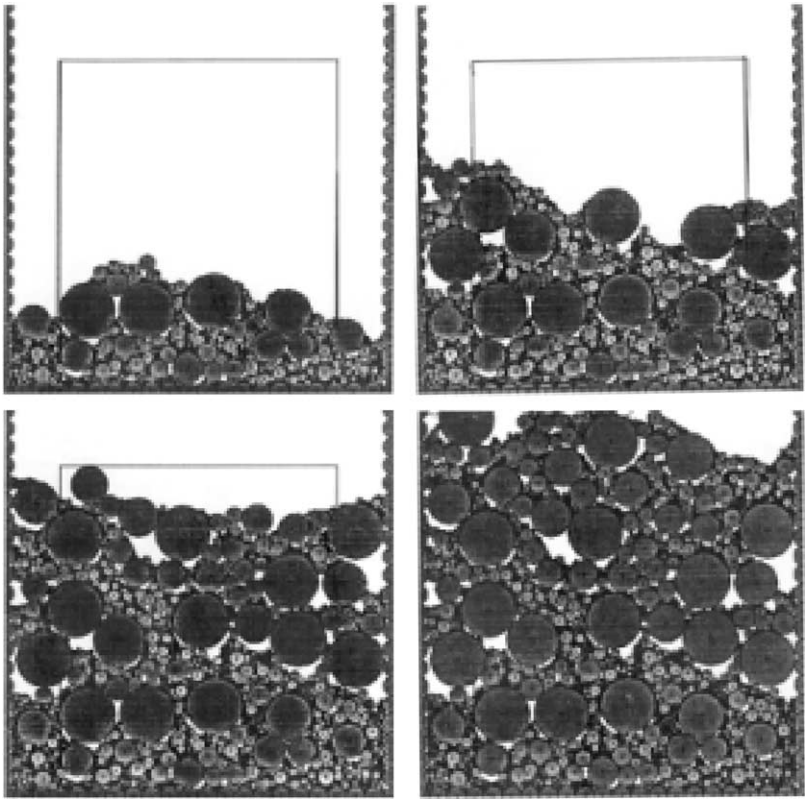


Figure 6 Consecutive stages of particle sedimentation simulated with a ballistic deposition algorithm; particle-size distribution was derived from the fractal fragmentation model. Reproduced with permission from Ghilardi P, Kai AK, and Menduni G (1993) Self-similar heterogeneity in granular porous media at the representative elementary volume scale. *Water Resources Research* 29: 1205–1214.

Random motions of large numbers of particles can cause fractal scaling in the particle ensemble characteristics. In particular, the fractal Brownian motion can be an underlying generator of data sequences that demonstrate self-affine scaling with the Hurst exponent $H \neq 0.5$ and $D_L \neq 1.5$ in Eqn [4]. Values of $H < 0.5$ are often related to the presence of correlated structural units in soil structure at different scales.

Measuring Fractal Parameters of Soils

Both direct and indirect measurements of geometric quantities are in use in soil studies. Indirect, or proxy, geometric measurements are very common. Nongeometric values are converted into geometric values using a physical model thought to be approximately valid in soils. For example, capillary pressures are converted into radii using capillary models, masses of adsorbed molecules are converted into area values using a molecular monolayer model, sedimentation time is converted into particle radii using the Stokes law, etc. With both direct and proxy measurements, methods of determining fractal parameters differ mostly in the way of changing the measurement scale.

Direct Geometric Measurements

The divider method is the first method applied to estimate the fractal dimension of self-similar lines in nature. The compass dividers are set to a specific stride length (measurement scale) as shown in Figure 1, and then the line is walked along. The fractal dimension is one minus the slope of the log(length)-log(scale) plots. Figure 8 shows boundaries of the fingers appearing at the wetting front in layered model soil systems that have been measured with various stride lengths, and the values of fractal dimension were successfully related to the distance of the finger penetration.

The box-counting method is based on covering the studied area with grids of different mesh sizes Δ and, for each mesh size, counting the number of grid squares N_B that include the studied line. The dependence between N_B and Δ follows a power-law relationship:

$$N_B \propto \Delta^{-D_{L,B}} \qquad [8]$$

if the line is fractal; and $D_{L,B}$ is the Kolmogorov dimension which approximates the fractal dimension of the line. The box-counting method is applicable to complex outlines and to branching lines (Figure 9).

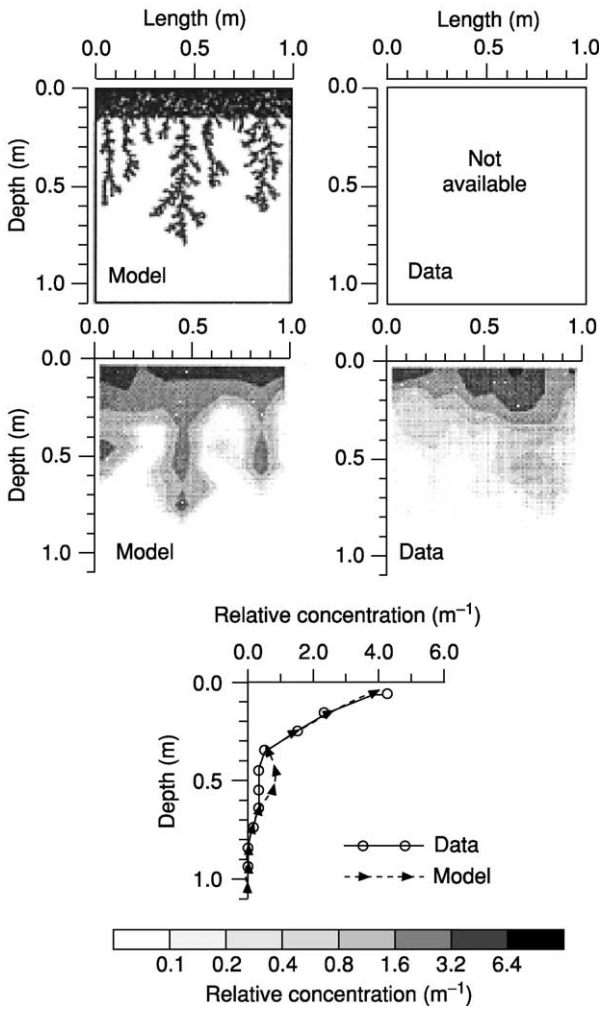


Figure 7 Comparison between the diffusion-limited aggregation model predictions and experimental data on bromide transport in layered soil. No comparison can be made at a microscopic scale of 0.2 cm representing the grid at which particles have traveled in simulations. At the field observation scale of approximately 10 cm, random distributions of solute concentrations in simulations and in experiment are difficult to compare. Mean concentrations averaged horizontally across the 1.4 × 1.4-m plot compare satisfactorily. Reproduced with permission from Flury M and Flühler H (1995) Modeling solute leaching in soils by diffusion-limited aggregation: basic concepts and application to conservative solutes. *Water Resources Research* 31: 2443–2452.

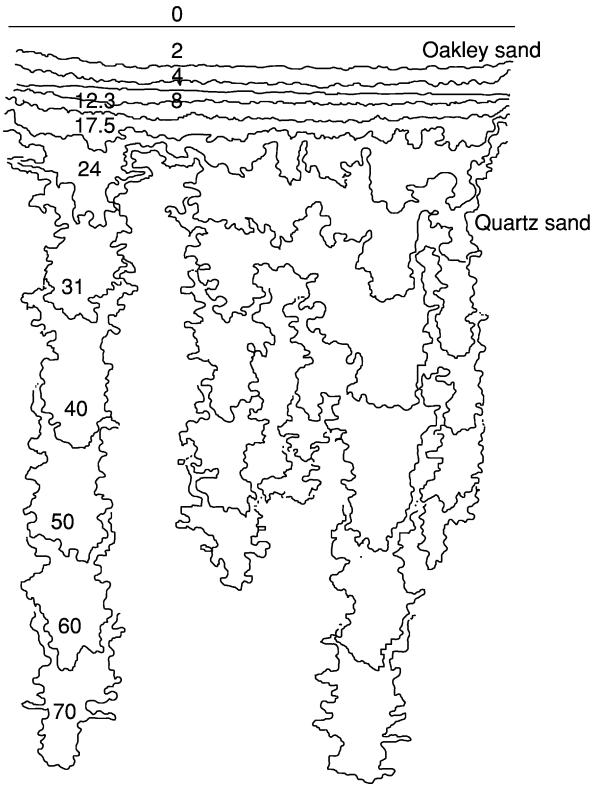


Figure 8 Infiltration and movement of water in quartz sand beneath Oakley sand. Each contour outlines the wetted area. Numbers at the contours show the time from the beginning of infiltration (minutes). Measurement of the contour lengths has been made with the divider. The divider stride length was varied from 1 to 18 cm for each of the contours, and the line fractal dimensions were computed from relationships between the measured contour length and the divider stride used as the measurement scale. Values of fractal dimensions were in the ranges from 1.007 to 1.09 and from 1.31 to 1.42 before and after the front reached the boundary between layers, respectively. Reproduced with permission from Chang WL, Biggar JW, and Nielsen DR (1994) Fractal description of wetting front instability in layered soils. *Water Resources Research* 30: 125–132.

Fractal dimensions can also be computed for lines defined as boundaries in images originating, for example, from soil thin sections or from remote sensing data. There is more uncertainty in defining fractal parameters with data on lines from imagery, because lines are defined by separation of two phases. Separation of two phases is based on a subjectively selected threshold between two ranges of brightness, and a noise which is always present in imagery can result in a rougher apparent boundary. Box-counting is applicable to such data. Progressive coarsening of images is yet another way to observe the line at different scales and to compute the Kolmogorov dimension.

Erosion- and dilation-based techniques are also used with lines from images. Dilation adds any background pixel that touches the line, while erosion removes any pixel that touches the background. Performing each of these operations M times and counting pixels that are affected as a function of M produces a plot whose slope in log-log coordinates gives the Minkovsky dimension that approximates the fractal dimension D_L.

Self-affine lines require different techniques to estimate the Hurst exponent and the fractal dimension. The Korcak method consists in making a cumulative plot of the number of intercepts as a function of

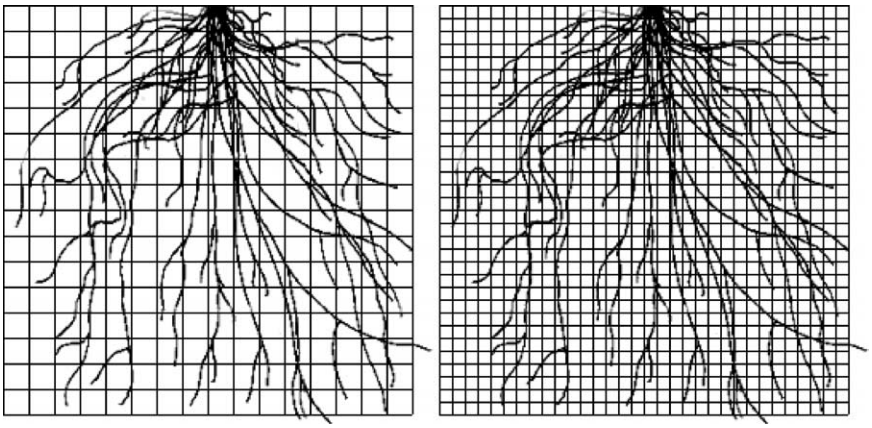


Figure 9 Box-counting to estimate fractal dimension of the root system for a maize plant grown in boxes 1.2 m long, and 0.05 m wide. Roots cross 198 grid cells in the 16×16 grid and 554 cells in the 32×32 grid. From those two measurements, the Kolmogorov dimension is $D_{L,B} = -\log[N_B(16)/N_B(32)]/\log[(1.2/16)/(1.2/32)] = -\log(198/554)/\log(2) = 1.484$. Reproduced with permission from Eghball B, Settimi JR, Maranville JW, and Parkhurst AM (1993) Fractal analysis for morphological description of corn roots under nitrogen stress. *Agronomy Journal* 85: 287–289.

lengths for any arbitrarily selected horizontal line. A log-log plot of the number of distances that exceed a value x as a function of x has a slope that is exactly 1 less than the fractal dimension of the profile. The semivariogram method is based on quantifying the spatial structure in a self-affine line using the dependence between variability in measurements and the distance between the measurement points. The semivariogram γ is computed as:

$$\gamma(h) = \frac{1}{2m_h} \sum [Z(x) - Z(x+h)]^2 \qquad [9]$$

where Z is the observed point on the self-affine line, e.g., elevation, h is the distance between observation points, interpreted as a measurement scale in this case, and m_h is the total number of pairs of points separated by the distance h. For a self-affine line:

$$\gamma \propto h^{2H} \qquad [10]$$

and the log-log plot of h versus γ provides the value of H and D_L from Eqn [4], as shown in Figure 10.

The power-spectrum method is based on Fourier analysis. Squared coefficients of the Fourier expansion, or magnitudes, are plotted against frequencies in log-log scale. The slope β of the linear regression line is related to the fractal dimension of the line as $D_L = 2 + \beta/2$. This method can also be applied to measure the fractal dimension of surfaces D_s, as $D_s = 3 + \beta/2$ (Figure 11).

Roughness-based methods are also applied to estimate fractal parameters of lines and surfaces. In the absence of a spatial trend, variances of the property in question are computed for different measurement scales. This is best done by placing many windows of the same width w (or area) on the data, either systematically or randomly, and calculating the mean value of the standard deviation for multiple windows. The procedure is repeated with windows

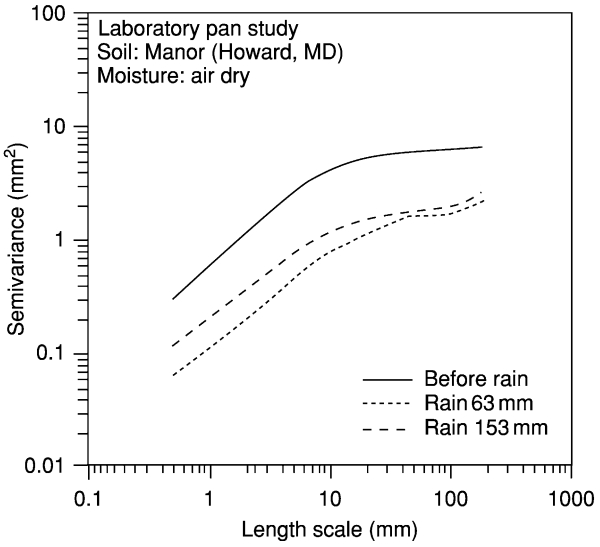


Figure 10 Changes in the semivariogram of soil surface elevation and in surface fractal dimensions caused by simulated rainfall. The tilted laboratory pan was filled with surface soil (coarse loamy, micaceous, mesic Typic Dystrochrept), and the surface topography of the central pan area of $0.54 \times 0.6\ m^2$ was digitized with laser scanner before and after rain events. Surface fractal dimensions are 2.51, 2.83, and 2.68 before rain, after 63 mm of rain, and after 155 mm of rain, respectively. The lower and the upper cutoff values of the fractal scaling in this case are 0.6 mm and 10 mm, respectively. Reproduced with permission from Huang C and Bradford JM (1992) Applications of a laser scanner to quantify soil microtopography. *Soil Science Society of America Journal* 56: 14–21.

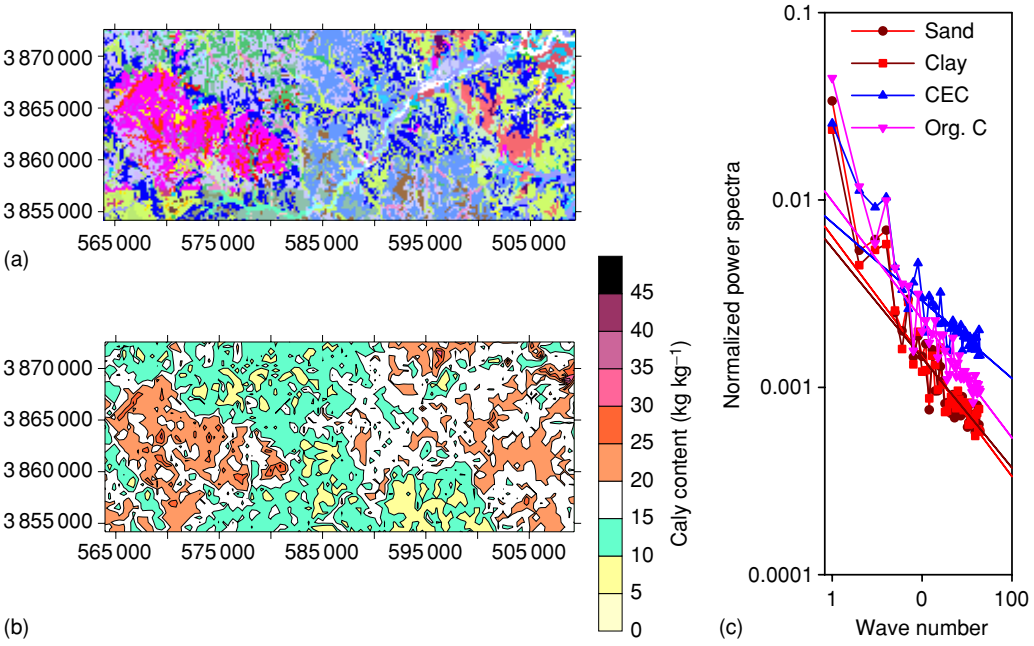


Figure 11 Application of the power-spectrum method to estimate surface fractal dimensions of soil basic properties in an area within the Little Washita watershed, OK, USA: (a) 44 individual soils defined from the MIADS data set for 200 m × 200 m pixels within the area; total number of pixels was 93 × 225; (b) map of clay content defined for each soil as the average value over the range given by soil survey; values of sand and organic carbon contents, and cation exchange capacity were defined for each pixel analogously; (c) power spectra developed for soil basic properties using their pixel values; the upper cutoff of fractal scaling is approximately 7 km, the lower cutoff is the pixel size of 200 m for all shown properties; estimated surface fractal dimensions are 2.68, 2.71, 2.79, and 2.68 for clay content, sand content, cation exchange capacity (CEC), and organic carbon (Org. C) content, respectively.

of different sizes, and the standard deviations are plotted against the window size in the log-log scale. The slope of the regression line added to 2 or 3 gives an estimate of the fractal dimension for the line or surface, respectively. One implementation of roughness-based methods consists in removing the linear trend in data within the windows and computing the root-mean-square (RMS) of residuals, or 'roughness,' as:

$$\mathrm{RMS} = \frac{1}{n_\mathrm{w}} \sum_{i=1}^{n_\mathrm{w}} \sqrt{\frac{1}{m_\mathrm{i} - 2} \sum_{j=1}^{m_\mathrm{i}} \epsilon_\mathrm{ij}} \qquad [11]$$

where n_w is the total number of windows of length w in the data, m_i is the number of points in the ith window, and ϵ_ij is the residual of the linear regression in the jth data point in the ith window. The window size serves as a measurement scale.

The area-perimeter, or slit-island, method can be used to find fractal parameters of complex outlines that can be seen in soil thin sections. This method does not require perimeter measurements with different step length but, instead, requires measurements of area and perimeters of the outlines. The fractal dimension of the irregular outlines, D_L, representing either particles or pores, is included in the relationship between area A and perimeter P:

$$P \propto A^{\mathrm{D_L}/2} \qquad [12]$$

Results of applying this relationship to thin-section images of soils under different long-term management are shown in Figure 12.

The number of visible macropores N_M that have radii larger than r often scales as $N_\mathrm{M} \propto r^{-\mathrm{D_M}}$, and fractal parameters of this scaling can be found from a log-log plot.

Proxy Measurements

Data on adsorption of molecules of different sizes reveal probing irregularity of surfaces at different scales, as shown in Figure 13. Assuming that molecules form a monolayer, one can use the molecule radius r as a measurement scale and relate the number of molecules N_m needed for monomolecular coverage of a fractal surface to the molecule radius as:

$$N_\mathrm{m} \propto r^{-\mathrm{D_s}} \qquad [13]$$

where D_s is the surface fractal dimension. The area of the monomolecular coverage of spherical particles is $A_\mathrm{m} = 4\pi r^2 N_\mathrm{m}$ and it changes with the radius of the

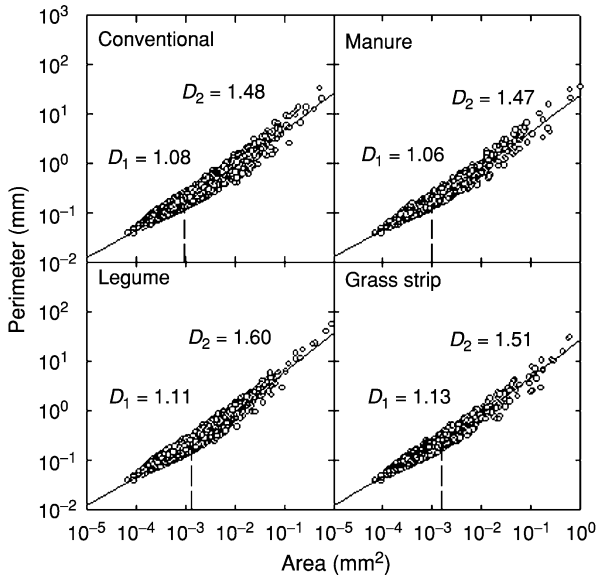


Figure 12 Fractal dimensions of pore outlines from thin section images for the Comly silty loam soil in plots where nitrogen was supplied from conventional fertilizer sources, manure, legumes, and in uncultivated plots under grasses where no nitrogen has been supplied. All plots were included in the experiment for 11 years. Small pores with areas less than 10^{-9} m^2 had fractal dimension D_1 between 1.06 and 1.12; pores with areas exceeding 10^{-9} m^2 had fractal dimension D_2 between 1.42 and 1.51. The value of D_2 is larger in samples from the legume plots than in samples from other plots. The cutoff between two linearity ranges corresponded to the boundary between vuggy and rounded pores. Reproduced from Pachepsky YA, Yakovchenko V, Rabenhorst MC, Pooley C, and Sikora LJ (1996) Fractal parameters of pore surfaces as derived from micromorphological data: effect of long-term management practices. *Geoderma* 74: 305–325.

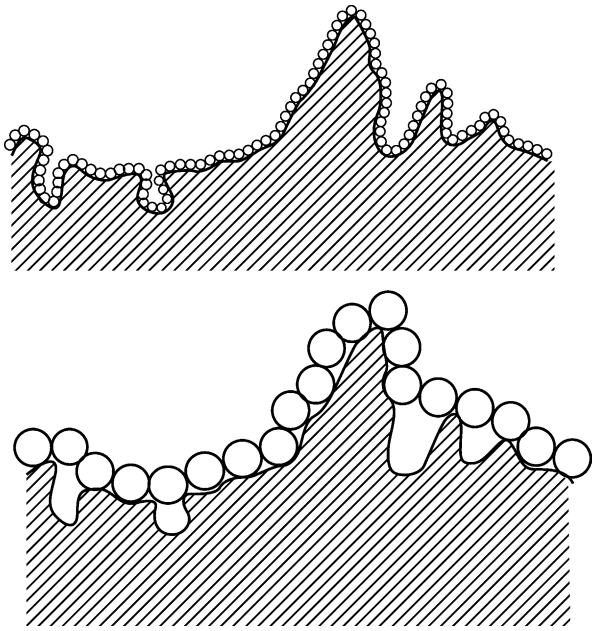


Figure 13 The molecular accessibility of a rough surface depends on the molecule size: larger molecules contact a surface that is much smaller than that contacted by smaller molecules.

molecule as $A_m \propto r^{2-D_s}$. The value of $D_s > 2$ and an increase in r is followed by a decrease in the area which is faster than r^2 because of the appearance of large caverns, as shown in Figure 13. The value of N_m can be estimated from the classic Brunauer–Emmett–Teller (BET) equation or from its modifications for fractal surfaces.

Isotherms of adsorption on fractal surfaces can be used to estimate fractal parameters when adsorption data for a single adsorbate are available in the range of relative pressures close to saturation. A fractal analog of the Frenkel–Halsey–Hill isotherm equation is:

$$n \propto \left(-\ln \frac{p}{p_0}\right)^{D_s-3} \qquad [14]$$

Here n is the adsorption and p/p_0 is the relative pressure. The measurement scale is the mean radius of the liquid–vapor interface, r_i, which is calculated from the Kelvin equation:

$$r_i = \frac{2\sigma V_m}{RT(-\ln \frac{p}{p_0})} \qquad [15]$$

where R is the gas constant, T is the temperature (Kelvin), σ is the surface tension of the liquid adsorbate, and V_m is the molar volume of the liquid adsorbate. Eqn [14] was derived by assuming the sequential filling of pores from small to large, that is, capillary condensation-type local adsorption isotherm. Eqn [14] is used with data obtained for $p/p_0 > 0.7$. A different model for calculating the fractal dimension D_s from the adsorption isotherm data reflects the idea that a surface is smoothed by the adsorbate, and the surface of the adsorbent covered with an adsorbate is smaller than the surface of the uncovered adsorbent. The area A_i of the 'condensed adsorbate–vapor' equilibrium interface is obtained as:

$$A_i = \frac{RT}{\sigma}\int_n^{n_{max}} (-\ln \frac{p}{p_0})\mathrm{d}n \qquad [16]$$

Here n_{max} is the maximum adsorption, n is the adsorption corresponding to the value of A_i. The fractal scaling law relates the area to the mean radius of the interface (Eqn [15]), and the resulting equation is:

$$A_i \propto \left(-\ln \frac{p}{p_0}\right)^{D_s-2} \qquad [17]$$

Both Eqn [14] and Eqn [17] are represented by straight lines in log-log coordinates, and slopes give values of the surface fractal dimension as shown in Figure 14. A caution has to be exercised in interpreting

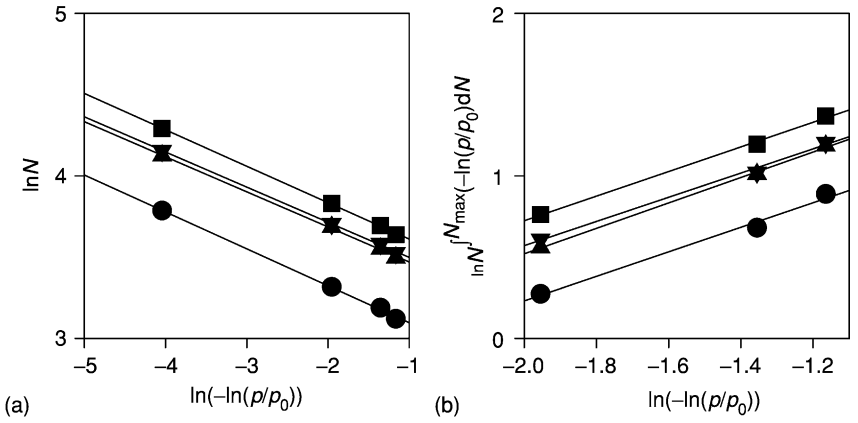


Figure 14 Fractal scaling in water vapor adsorption measured in samples of Udic Argiboroll, Moscow region, from four plots where a long-term experiment on grassing arable land had been carried out for 12 years: (a) Eqn [14], (b) Eqn [17] applied. Circles, no harvesting, no fertilization; squares, no harvesting, 60 kg NPK annually; triangles, harvested, no fertilization; inverted triangles, harvested, 60 kg NPK annually. Values of the surface fractal dimension D_S are in the range 2.75–2.85. Removal of carbohydrates has resulted in an increase in D_S. Differences in management practices did not affect values of D_S in the scale range studied, whereas the monolayer capacity was affected. Reproduced from Hajnos M, Korsunskaia LP, and Pachepsky YA (2000) Fractal dimensions of soil pore surfaces in a managed grassland. *Soil and Tillage Research* 55: 63–70.

values of fractal dimensions obtained with adsorption because of specific interactions of the adsorbate molecules with soil-pore surfaces.

The mercury porosimetry and water desorption are both used to estimate mass or surface fractal dimension of soil from the scaling relationship:

$$\frac{dV}{dr} \propto r^{2-D} \qquad [18]$$

Here V is the pore volume not yet filled by the intruding mercury or occupied by water, r is the mean pore radius serving as the measurement scale and estimated from the pressure values with the Laplace equation for capillaries. Selection of the geometric model of fractal medium determines whether D is an estimate of the surface or the mass fractal dimension.

Light, UV, and X-ray scattering have been applied to study fractal dimensions of organic substances in soils. Measurements with the single wavelength take advantage of the scaling dependence of the intensity of the scattered radiation I on scattering vector Q:

$$I(Q) \propto Q^{-s} \qquad [19]$$

where the scaling exponent s is equal to the mass fractal dimension D_m if the object is the mass fractal, and $s = 6 - D_s$ if the object of the study is the surface fractal. Measurements with multiple wavelengths relate the turbidity τ of a suspension to the wavelength λ as the measurement scale:

$$\tau \propto (\lambda)^{S-4+\delta} \qquad [20]$$

where S is the scaling exponent, δ is the value specific to the structure of organic molecules, $\delta = -0.2$ for humic substances. Values of S less (or greater) than 3 imply the presence of the mass (surface) fractal. Log-log plotting wavelength versus turbidity results in straight lines with slopes related to fractal dimensions, as shown in Figure 15.

Measurements of mass are used to estimate the number of particles in fragmentation scaling (Eqn [7]). The number of particles in the radius range r_i to r_{i+1} is found as $M/C\rho\bar{r}^3$, where M is the total measured mass of particles, ρ is the particle density, C is the constant reflecting the deviation of particle shapes from sphere, and $\bar{r}$ is the characteristic radius of particles in the range r_i–r_{i+1}:

$$\bar{r} = \begin{cases} \dfrac{D_F}{3 - D_F} \dfrac{r_{i-1}^{3-D_F} - r_i^{3-D_F}}{r_i^{D_F} - r_{i-1}^{D_F}}, D_F \neq 3 \\ \dfrac{3\ln(r_{i-1}/r_i)}{r_i^3 - r_{i-1}^3}, D_F = 3 \end{cases} \qquad [21]$$

Transport Processes

The relationship between parameters of the transport and fractal parameters of soil-pore surface or pore space depends strongly on specifics of the fractal medium model. Such detailed models are rarely available in soil studies, and data on transport are rarely used to infer fractal parameters of soils. Instead, the transport of solutes, colloids, and microorganisms in soils is viewed as a fractal process in which either distances of particle movement or waiting times for

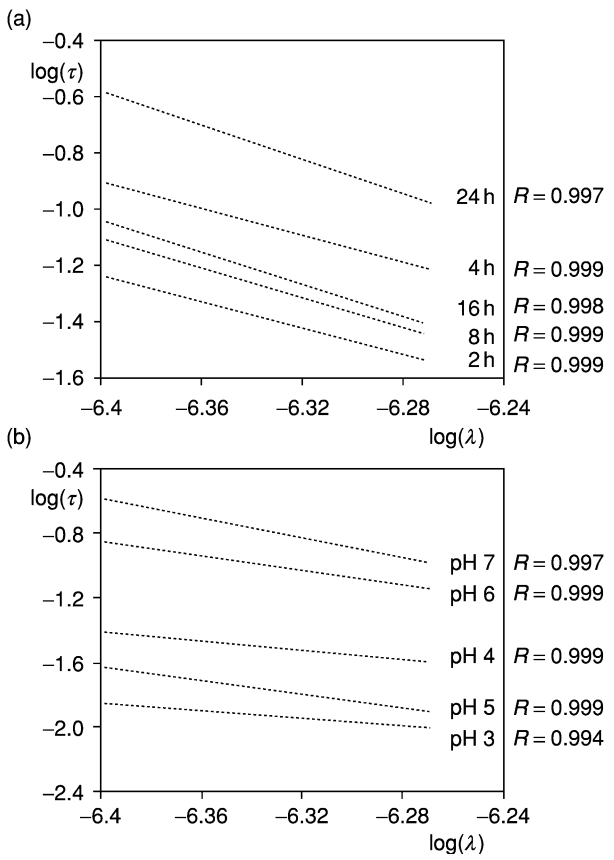


Figure 15 Fractal scaling in dependencies of turbidity τ on wavelength λ in aqueous suspensions of humic acid from Elliot silt loam: (a) at pH 7 after various equilibration times; (b) at various pH values after 24 h equilibration. Fractal dimensions decrease from 2.08 to 1.45 as pH grows from 3 to 7 and morphology of the humic particles becomes less irregular. Reproduced with permission from Senesi N, Rizzi FR, and Acquafredda P (1996) Fractal dimension of humic acids in aqueous suspensions as a function of pH and time. *Soil Science Society of America Journal* 60: 1773–1780.

particles to move have power-law distributions reflecting fractal-like scaling in time or in space. This type of distribution manifests itself in a dependence of dispersivity and diffusivity on the distance from the source of particles participating in transport. Using fractional derivatives and equations of fractional kinetics shows a promise in simulating fractal transport.

Uncertainty in Measuring Fractal Parameters

While it is true that soil is not made up of spheres and straight lines, it is also true that soil is far from being precisely fractal. Rather it is hoped that fractal models may lead to a more accurate description of soil in process models than methods of classic geometry. However, relationships of fractal geometry can be only approximately true in soils. Different methods of characterizing scaling lead to different results even for ideal geometric fractals. Values of fractal parameters obtained by different methods should never be compared for soils. Using proxy measurements creates additional uncertainty, since ideal shapes are assumed to convert nongeometric to geometric measures.

A computational uncertainty also exists in selecting the range of measurement scales in which the fractal scaling is valid. Cutoff values have to be included in the fractal parameter estimation procedure along with slope and intercept of the linear regression used to estimate the fractal dimension and the initiator measure.

The Use of Soil Fractal Parameters

The major contribution that fractal geometry can make to a description of soil is a general, simple, and succinct representation of complex structure by a small number of fractal parameters. Fractal parameters are complementary to other soil parameters. Changes in fractal parameters along with changes in other soil parameters reflect effects of soil management and ecosystem functioning on soil properties. Relationships between structure-dependent different soil properties can be derived from a fractal model when it is selected. Relations between hydraulic conductivity and water retention give one such example. Fractal analysis can help to suggest an iterative process that has caused the observed 'property-size' distribution. Cutoffs of the fractal scaling range indicate the change in processes causing the scaling or the change in the performance of a measurement device. They are useful in defining the representative elementary volume or area.

The presence of fractal scaling provides simple ways to simulate soil structure and structure of soil cover. Multiple realizations of such simulations create synthetic data essential to study the effect of soil variability on the performance of new measurement techniques for determining soil properties. It has been found in several research fields that fractal features arise as a result of underlying chaos. Growing interest in the predictability of soil behavior in a changing environment may induce the search for chaos in soils in order to relate it to observed fractal scaling in soil properties. Some time series in soil exhibit scaling properties and can be analyzed using fractal models. Data-intensive sensor technologies such as remote sensing imagery and laser altimetry are becoming more common in soil studies. Fractal-based data analysis techniques can be used to compress these observations into a small number of parameters for comparison and classification purposes.

Fractals should never be considered as an ultimate model of heterogeneity in the soil system. Rather they provide a balance between accuracy and clarity that may aid us in gaining insight into sources and consequences of the observed soil complexity. Eventually, once a greater insight into key processes is obtained, we expect causes of the apparent fractal scaling to be revealed and quantified.

List of Technical Nomenclature

β	Slope of the regression line of wave number versus the power spectrum
γ	Semivariogram
Δ	Grid mesh size (m)
ϵ_{ij}	Residual of the linear regression in the jth data point in the ith window
λ	Wavelength (m)
σ	Surface tension of the liquid adsorbate ($N\,m^{-2}$)
τ	Turbidity (m)
A	Pore outline area (m^2)
A_i	Area of the 'condensed adsorbate–vapor' equilibrium interface
A_m	Area of the monomolecular coverage of spherical particles (m^2)
D_F	Fragmentation fractal dimension
D_L	Line fractal dimension
$D_{L,B}$	Kolmogorov dimension
D_m	Mass fractal dimension
D_S	Surface fractal dimension
D_{ss}	Self-similarity dimension
E	Statistical moment about zero
H	Hurst exponent
h	Distance between observation points (m)
I	Intensity of the scattered radiation ($mol\,m^{-2}\,s^{-1}$)
K	Scaling exponent
M	Number of erosion–dilation operations
m_i	Number of points in the ith window
N	Number of pieces into which the object is divided
n	Adsorption ($mol\,kg^{-1}$)
N_B	Number of squares N_B through which any part of line passes
N_M	Number of visible macropores
N_m	Number of molecules needed for monomolecular coverage of a surface
n_m	Maximum adsorption ($mol\,kg^{-1}$)
N_r	Number of fragments
n_w	Total number of windows of length w
P	Pore outline perimeter (m)
p	Partial pressure ($N\,m^{-2}$)
p_0	Partial pressure at saturation ($N\,m^{-2}$)
Q	Scattering vector (m^{-1})
q	Order of the statistical moment
R	Reduction factor
R	Gas constant ($J\,mol^{-1}\,K^{-1}$)
r	Radius of fragment, particle, molecule, pore (m)
r_i	Mean radius of the liquid–vapor interface
RMS	Root-mean-square of residuals
s, S	Scaling exponent
T	Temperature (K)
V_0	Volume of the initiator (m^3)
V_m	Molar volume of the liquid adsorbate ($m^3\,mol^{-1}$)
W	Scale-dependent measure
w	Window length (m)
x, y	Cartesian coordinates (m)
Z	Observation on a self-affine line

Further Reading

Anderson AN, McBratney AB, and Crawford JW (1998) Applications of fractals to soil studies. *Advances in Agronomy* 63: 1–76.

Avnir D, Farin D, and Pfeifer P (1985) Surface geometric irregularity of particulate materials: the fractal approach. *Journal of Colloid and Interface Science* 103: 112–123.

Baveye P, Parlange J-Y, and Stewart BA (eds) (1998) *Fractals in Soil Science*. Boca Raton, FL: CRC Press.

Falconer KJ (1997) *Techniques in Fractal Geometry*. Chichester, UK: John Wiley.

Giménez D, Perfect E, Rawls WJ, and Pachepsky YA (1997) Fractal models for predicting soil hydraulic properties: a review. *Engineering Geology* 48: 161–183.

Mandelbrot BB (1982) *The Fractal Geometry of Nature*. San Francisco, CA: W. H. Freeman.

Molz FJ and Liu HH (1997) Fractional brownian motion and fractional gaussian noise in subsurface hydrology: a review, presentation of fundamental properties, and extensions. *Water Resources Research* 33: 2273–2286.

Pachepsky YA, Crawford JW, and Rawls WJ (eds) (2000) *Fractals in Soil Science*. Amsterdam, the Netherlands: Elsevier.

Perfect E and Kay BD (1995) Applications of fractals in soil and tillage research: a review. *Soil and Tillage Research* 36: 1–20.

Russ JC (1994) *Fractal Surfaces*. New York: Plenum Press.

Senesi N (1996) Fractals in general soil science and in soil biology and biochemistry. In: Stotzky G and Bollag J-M (eds) *Soil Biochemistry*, vol. 9, pp. 415–472. New York: Marcel Dekker.

Turcotte DL (1996) *Fractals and Chaos in Geology and Geophyisics*. Cambridge, MA: Cambridge University Press.

FREEZING AND THAWING

Contents
Cycles
Processes

Cycles

B Sharratt, USDA Agricultural Research Service, Pullman, WA, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Soil freezing and thawing are of economic importance and can be an environmental concern in cold regions, because these processes influence erosion, runoff, and greenhouse gas emissions. Soil freezing and thawing are typically confined to latitudes greater than 40° in the Northern and Southern Hemispheres and are largely governed by atmospheric conditions. Soil freezing and thawing are also influenced by surface characteristics such as amount and orientation of crop residues, depth and type of tillage, and snow cover. Soil and crop residue management can therefore influence the occurrence of freezing and thawing. Management practices that result in little residue cover, soil consolidation, or a thin snow pack generally hasten soil freezing in the autumn and soil thawing in the spring. In cold regions, plastic and bituminous mulches are applied to soil to enhance solar radiation absorption and thus expedite soil thawing or prevent soil freezing.

Global Significance

Soil freezing and thawing are confined to latitudes greater than approximately 40° in the Northern and Southern Hemispheres. Soils in tropical and subtropical regions rarely freeze, except at high elevations. The latitudinal zone of ephemeral frost, between the latitudes of approximately 40° and 60°, encompasses regions of seasonally frozen soil. These soils remain frozen for 1 to more than 180 days, go through a single to as many as 100 freeze–thaw cycles, and completely thaw during some portion of the annual cycle. Ephemeral frost occurs in regions typified by continental, marine, and highland climates. Permafrost typically occurs at latitudes greater than 60°; soils at these high latitudes are some of the least productive in the world due to factors such as low fertility and poor profile development. Dry and cold growing seasons also limit crop adaptation and production in permafrost regions. Although seasonal freezing and thawing occur in permafrost (often referred to as the active layer), such events are generally singular during the annual cycle.

Freezing and thawing are important processes that affect the quality and productivity of soil resources. These processes modify soil properties such as thermal conductivity, electrical conductivity, hydraulic conductivity, infiltration, density, shear strength, penetrability, porosity, aggregate stability, and structure. Changes in these properties influence soil water content and availability, nutrient availability, erosion, trafficability of soils, rooting density and distribution, and growth and productivity of agricultural crops. Freezing and thawing also affect the population and diversity of soil microorganisms and fauna. These organisms can be beneficial for enhancing soil aggregation and hastening the decomposition of organic

material in soil. Many soil organisms are able to survive subfreezing temperatures by supercooling and entering freeze-resistant stages or by migrating to regions within the soil profile that remain unfrozen.

Water resources are affected by freezing and thawing as a result of redistribution of water within the soil profile or by changes in the rate of infiltration of water into soil. Redistribution of water in the profile occurs in response to water potential gradients that arise as soils freeze. Water migrates from regions of high water potential (a large amount of liquid water in the soil matrix) to regions of low water potential (a small amount of liquid water in the matrix) in the soil profile. This phenomenon is an important mechanism by which recharge occurs in the upper portion of the soil profile due to water migration from wet, unfrozen subsoil to the advancing freezing front. Frozen soil impedes infiltration and thereby enhances runoff. Runoff may carry sediment and nutrients that degrade water resources. Freezing and thawing can also influence the availability and quality of water resources due to the creation of large vertical fractures within the soil profile. These fractures, created by freezing-induced shrinkage stresses, allow for rapid infiltration of surface water and recharge of groundwater. Water flowing through these fractures may be laden with chemicals that contribute to the degradation of groundwater resources in cold regions.

Soil freezing can influence the air quality in cold regions. Freeze-drying of soil during winter results in the breakdown of aggregates into particulate matter; the finest fraction of this particulate matter is subject to transport by wind. Fine particulate matter in the atmosphere, resulting from wind erosion, adversely affects human health. Freezing and thawing also influence greenhouse gas emissions. During winter, emission of these gases is constrained by the relatively impervious frozen soil layer. As soils thaw during spring, however, emission of greenhouse gases is accentuated by either saturated conditions in the top soil (favors nitrous oxide emission) or a burst in microbial activity (favors the emission of carbon dioxide).

Guiding Principles

The frequency and rapidity of soil freezing and thawing are largely governed by atmospheric conditions. Solar radiation and air temperature determine to a great extent the quantity of energy available to modify soil temperatures. Soil freezing is generally of greater risk at locations with lower seasonal insolation or high elevation. In addition, soil surface characteristics also affect the frequency and rapidity of soil freezing and thawing. Surface characteristics such as soil roughness and crop residue cover influence the rate of heat exchange between the soil and atmosphere and thus soil freezing and thawing. Soil and crop residue management is therefore important in determining the frequency and duration of soil freezing and thawing.

Freezing and thawing of soil subject to tillage and residue management practices in cold regions are governed by the heat and water balance of the snow pack, residue layer, and soil (Figure 1). The change in the physical state of the soil is largely driven by atmospheric conditions. Atmospheric conditions such as solar radiation, wind, humidity, air temperature, and precipitation determine to a large extent the amount of energy and water available for heating and evaporative processes at the exchange surface (snow, residue, or soil). Net heat flow must be balanced by heat storage or phase change within a freezing or thawing snow–residue–soil system. In addition, net water flow must be balanced by an increase or decrease in water stored within the system. Heat and water flow processes are intimately related in a freezing or thawing snow–residue–soil system. Within the

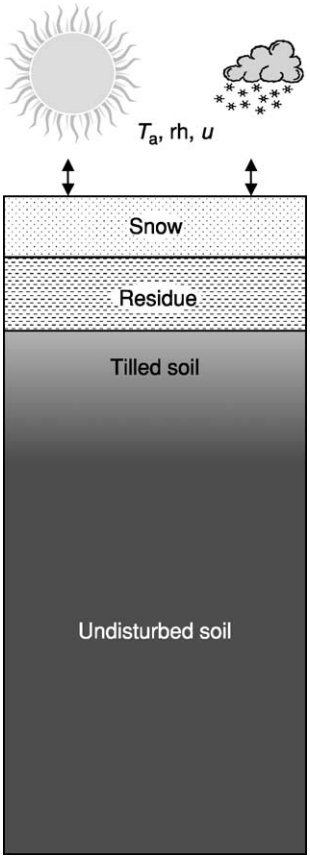


Figure 1 Exchange processes within a snow–residue–soil system governed by solar radiation, air temperature (T_a), relative humidity (rh), wind speed (u), and precipitation.

snow pack, water and water vapor influence heat transfer according to:

$$\delta(k_{sp}\delta T\delta z^{-1})\delta z^{-1} = C_i\delta T\delta t^{-1} + \rho_l L_f \delta w_s \delta t^{-1} + L_s(\delta\rho_v\delta t^{-1} + \delta q_v\delta z^{-1}) - S \quad [1]$$

where k_{sp} is thermal conductivity of snow, T is temperature, z is depth, C_i is volumetric heat capacity of ice, t is time, ρ_l is density of liquid water, L_f is latent heat of fusion, w_s is volumetric liquid water fraction, L_s is latent heat of sublimation, ρ_v is water vapor density, q_v is water vapor flux, and S is an energy source or sink. Equation [1] indicates that any difference in heat flow between the upper and lower boundary of the snow pack (the term on the left side of the equality sign in eqn [1]) must be accounted for by heat storage (first term on the right side of eqn [1]), latent heat associated with phase changes (second, third, and fourth terms on the right side of eqn [1]), or an energy source or sink such as solar and long-wave radiation (fifth term on the right side of eqn [1]).

In the residue layer, heat flow is influenced by water flow according to:

$$\delta(k_r\delta T\delta z^{-1})\delta z^{-1} = C_r\delta T\delta z^{-1} + L_v E - S \quad [2]$$

where k_r is the combined conduction–convection heat transfer coefficient for crop residue, C_r is volumetric heat capacity of crop residue, L_v is latent heat of vaporization, and E is rate of evaporation from the surface of residue elements to void spaces between elements. Equation [2] indicates that any difference in heat flow between the upper and lower boundary of the residue layer (term on the left side of the equality sign in eqn [2]) must be accounted for by heat storage (first term on the right side of eqn [2]), latent heat associated with evaporation (second term on the right side of eqn [2]), or an energy source or sink (third term on the right side of eqn [2]).

In the soil profile, water flow in response to water potential gradients creates an additional complexity in describing heat flow. This complexity arises due to thermal advection that accompanies water flow. Heat flow within a freezing and thawing soil is described by:

$$\delta(k_s\delta T\delta z^{-1})\delta z^{-1} = C_s\delta T\delta t^{-1} - \rho_i L_f \delta\theta_i\delta t^{-1} + L_v(\delta\rho_v\delta t^{-1} + \delta q_v\delta z^{-1}) - C_l\delta(q_l T)\delta z^{-1} - S \quad [3]$$

where k_s is thermal conductivity of soil, C_s is volumetric heat capacity of soil, ρ_i is density of ice, θ_i is volumetric ice content, C_l is volumetric heat capacity of water, and q_l is liquid water flux. Equation [3] indicates that any variation in heat flow within the soil profile (term on the left side of the equality sign in eqn [3]) must be accounted for by heat storage (first term on the right side of eqn [3]), latent heat associated with phase changes (second, third, and fourth terms on the right side of eqn [3]), advection associated with soil water flow (fifth term on the right side of eqn [3]), or an energy source or sink (sixth term on the right side of eqn [3]).

The composition and structure of the uppermost exchange surface will govern heat and water flow processes at the surface and therefore influence the state of the soil system. Soil and crop residue management can modify heat and water flow processes at the upper surface by altering snow cover, residue layer thickness and orientation, or depth of the tilled soil layer. In addition, soil amendments can influence heat and water exchange processes at the soil surface and thus freezing and thawing.

Tillage Practices

Tillage is most often used to prepare seedbeds and to control weed infestations. Tillage can also be used to alter the physical properties at the soil surface to control erosion. In cold regions, tillage practices that hasten soil warming during spring are most desirable. Soil warming can be accomplished by reducing the amount of crop residue on the soil surface or by roughening and darkening the soil surface. Removing crop residue or roughening the soil surface maximizes solar radiation absorption and thus the amount of energy available to warm soil. For more than a quarter of a century, no-tillage has been advocated for reducing wind and water erosion on agricultural lands. This practice, however, retards soil warming during spring due to the insulating effect of crop residue on the soil surface. Therefore, no-tillage is seldom a viable and sustainable option in cold, humid regions.

Tillage practices in common use throughout the world include no-tillage, strip tillage, ridge tillage, moldboard plowing, and various forms of reduced tillage using chisels, sweeps, and disks. The soil microclimate created by any one tillage practice will be quite different from that created by other tillage practices due to the effect of tillage implements on crop residue cover and soil disturbance. Any type of tillage practice results in partial burial of crop residues as well as reduction of soil density within the tilled zone. A reduction in residue cover removes the protective insulation at the soil surface and thereby enhances the likelihood of more frequent and rapid soil freezing and thawing. The rapidity of soil freezing and thawing as influenced by tillage practice is

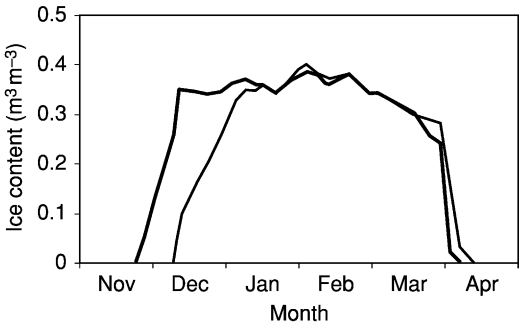


Figure 2 Ice content during winter of a soil subject to autumn chisel plow (thick line) and no tillage. (Adapted from Grant RF, Izaurralde RC, and Chanasyk DS (1995) Soil temperature under different surface managements: testing a simulation model. *Agricultural and Forest Meteorology* 73: 89–113 with permission.)

Table 1 Maximum depth of frost penetration during a winter with little snowfall at three locations

	Frost depth (cm)		
Tillage	*Minnesota*	*Oregon*	*Sweden*
Moldboard plow	74	26	61
No tillage, no residue	76	26	65
No tillage, residue	70	15	51

Sources of information include: Benoit GR, Young RA, and Lindstrom MJ (1988) Freezing-induced field soil water changes during five winters in west central Minnesota. *Transactions of the ASAE* 31: 1108–1114; Pikul JL, Zuzel JF, and Greenwalt RN (1986) Formation of soil frost as influenced by tillage and residue management. *Journal of Soil Water Conservation* 41: 196–199; Thunholm B and Hakansson I (1988) Influence of tillage on frost depth in heavy clay soil. *Swedish Journal of Agricultural Research* 18: 61–65.

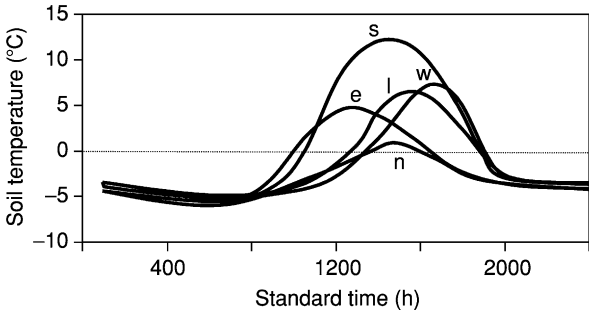


Figure 3 Near-surface soil temperature on various aspects of ridged and level (l) soil surfaces during a clear spring day in Minnesota. (n, north; s, south; e, east; w, west.)

portrayed in Figure 2 by the change in ice content of soil during the autumn and spring. More rapid formation or disappearance of ice is apparent in a soil subject to autumn tillage versus no tillage. Therefore, no-tillage retards cooling of the soil during autumn and warming of the soil during spring as compared to other tillage practices.

Cultivation of soil (either with or without crop residue on the soil surface) reduces the thermal conductivity within the tilled layer and thus the rapidity of heat loss and frost penetration during winter. The maximum depth of frost penetration can be reduced by several centimeters as a result of tilling a bare soil (Table 1). The extent to which tillage influences soil frost penetration, however, will vary across regions and landscapes as a result of the spatial variability in snow cover. Snow cover tends to mask any difference that soil surface characteristics (roughness, residue cover) may have on heat exchange processes. Indeed, as little as 15 cm of snow is sufficient to insulate the soil from atmospheric extremes. Changes in soil thermal conductivity associated with tillage will also influence the rate of soil thaw during spring. In the absence of snow cover, tilled soil will thaw at a slower rate during spring than a bare, nontilled soil. Complete thaw of the soil profile, however, will occur several days earlier in the spring for a tilled versus nontilled soil. Thus, although the rate of thaw is retarded by tillage, reduced frost penetration during winter will result in an earlier thaw of a tilled soil profile versus an untilled soil profile during spring.

Strip-and-ridge tillage systems have been used as a management strategy to modify the soil microclimate. Strip tillage implies row cultivation after harvest, resulting in alternate bands of cultivated soil and residue-covered soil. The width of the band of cultivated soil may vary, but little additional advantage in modifying soil microclimate is achieved for band widths greater than 20 cm. The principal advantage of this strategy is to bolster solar radiation absorption within the cultivated row while retaining residue on the soil surface to control erosion. The additional energy absorbed can expedite thaw by several days and increase temperatures by several degrees within the cultivated rows during spring. Ridge slope and aspect will influence radiation absorption and thus the frequency and duration of soil freezing and thawing. Ridges with a southerly exposure offer the advantage of absorbing more radiant energy than any other ridge aspect in the Northern Hemisphere. Ridges with a south-facing aspect thaw sooner and attain a higher daytime temperature than most other ridge aspects. East-facing aspects thaw earliest on clear days, although daytime temperatures are suppressed during midday as compared to west-facing aspects and level surfaces. Soil thaw and warming are most frequently delayed on northerly aspects (Figure 3).

Crop Residue Management

Crop residue is an effective material that can minimize the erosion of soil by wind and water. Residues

reduce the forces of wind and water that would otherwise act upon loose particles at the soil surface. Residues also provide thermal protection to plants from winter temperature extremes. Residues insulate the soil surface from both winter and summer atmospheric extremes by impeding the movement of heat and water vapor between the soil and atmosphere. Crop residue, therefore, retards heat loss from the soil during winter as well as hinders warming of soil during summer.

Crop-residue management includes strategies that alter the amount and orientation of crop residue on the soil surface. The amount and thickness of residue lying on the soil surface determine to a large extent the amount of thermal insulation afforded by the residue layer. For example, an increase in the thickness of the residue layer will enhance the resistance for heat and water vapor exchange and therefore sensible heat transfer between the soil and atmosphere. Percentage residue cover is also important in determining the proportion of solar radiation intercepted by the residue and the soil surface. Radiant energy absorbed by residue is largely dissipated as sensible heat to the atmosphere, thereby resulting in little transfer of heat to the soil. Residue orientation refers to the architecture of the residue canopy; orientation can range from standing stubble to residue lying prostrate on the soil surface. Standing stubble dissipates wind energy at the soil surface and thus minimizes the effectiveness of heat and water vapor transfer by convection from the soil to atmosphere. In cold and windy regions, stubble is important in retaining or trapping snow. Standing stubble also influences the interception of solar radiation; taller stubble generally traps more incident radiation and thus reduces the proportion that is reflected from the residue surface.

Residue management strategies commonly employed to alter soil microclimate or protect the soil from erosion include cutting stubble at various heights above the soil surface at the time of harvest, burning residue in the autumn or spring, and removing residue from the row (row cleaning) at the time of sowing. Perhaps the most effective management strategy for controlling erosion and moderating the soil microclimate has been to cut stubble at various heights above the soil surface. The winter thermal regime of the soil can be dramatically altered by stubble height, particularly in regions where snow is blown by strong winds. Taller stubble retains or traps more snow on the soil surface; the additional snow cover better insulates the soil from winter atmospheric extremes. Taller stubble effectively reduces frost penetration and hastens thawing in the spring (Table 2).

Burning residue is an effective strategy commonly employed in the autumn before the first snowfall and in the spring after thaw. The ash-covered or bare surface resulting from burning the residue offers little thermal protection to the soil. Soils subject to residue burning, therefore, rapidly freeze and thaw.

Table 2 Maximum frost depth and day of complete soil thaw during a cold, dry winter in Minnesota

Stubble height (cm)	*Frost depth (cm)*	*Day of thaw*
0, no residue	114	18 May
0	102	14 May
30	75	2 May
60	40	17 April

Adapted from Sharratt BS (2002) Corn stubble height and residue placement in the northern US Corn Belt: I. Soil physical environment during winter. *Soil and Tillage Research* 64: 243–252, with permission from Elsevier.

Removing residue from the seed row either prior to or during sowing is new technology developed to optimize the soil thermal regime in early spring. Little is known, however, concerning the effect of this strategy on the frequency and depth of soil freezing and thawing. Nevertheless, row cleaning results in a narrow band of bare soil that will thaw and warm more quickly in spring. Daily temperatures can rise 2°C as the width of the soil band increases from 0 to 20 cm. Wider bands of bare soil have no effect on soil temperature. The physical properties of residue also influence heat and water transport between the soil and atmosphere. Little is known, however, about the extent to which soil microclimate can be altered by changing residue properties. Residue color has been exploited in an attempt to alter solar radiation absorption of a soil–residue system. Black residue absorbs approximately 15% more radiant energy than natural-colored residue, but little of this additional energy is utilized in soil thawing or warming.

Soil Amendments

Soil amendments are often used in cold regions to modify the thermal environment of plants. Plastic and bituminous mulches are common amendments used to bolster soil temperatures or expedite snowmelt. These mulches maximize radiation absorption during the day as well as minimize soil heat loss during the night. Plastic mulches not only have the capability to moderate soil temperatures, but they can also modify the soil moisture regime by reducing evaporative loss from the soil. The color, thickness, and composition of the plastic material will influence the amount of solar radiation transmitted to and absorbed by the soil surface. Although these properties (color, thickness, composition) influence radiation absorption, color is perhaps the most widely exploited for promoting soil warming. Black plastic

effectively absorbs solar radiation due to its low reflectivity, but little of the absorbed radiant energy is transmitted to the soil beneath the plastic (the energy is dissipated as sensible heat from the surface of the plastic to the atmosphere). White plastic transmits and absorbs little solar radiation owing to its high reflectivity. Clear plastic allows a greater proportion of solar radiation to be transmitted to the soil surface than either black or white plastic; therefore soil temperatures are typically greater under clear plastic than black plastic mulch. Near-surface soil temperatures can be enhanced by as much as 10–20°C using plastic mulch in the early growing season. The largest degree of moderation occurs during the daytime, but temperatures can be raised by several degrees during the nighttime.

Bituminous mulch has also been used to moderate soil temperatures in cold regions. These mulches are frequently used to hasten melting of snow and warming of soil during early spring. The low reflectivity of bituminous mulch results in high absorption of solar radiation. The high thermal conductivity of bituminous material and the good contact of the material with the soil allow for efficient transfer of absorbed radiant energy from the mulch to the soil. Soil temperatures can be bolstered by several degrees using bituminous mulch, thereby delaying the occurrence of soil freezing in autumn and accelerating the occurrence of soil thawing in spring.

List of Technical Nomenclature

θ_i	Volumetric ice content in soil
ρ_i	Density of ice
ρ_l	Density of liquid water
ρ_v	Water vapor density
C_i	Volumetric heat capacity of ice
C_l	Volumetric heat capacity of water
C_r	Volumetric heat capacity of crop residue
C_s	Volumetric heat capacity of soil
E	Rate of evaporation from the surface of residue elements to void spaces between elements
k_r	Combined conduction–convection heat transfer coefficient for crop residue
k_s	Thermal conductivity of soil
k_{sp}	Thermal conductivity of snow
L_f	Latent heat of fusion
L_s	Latent heat of sublimation
L_v	Latent heat of vaporization
q_l	Liquid water flux
q_v	Water vapor flux
S	Energy source or sink
T	Temperature
t	Time
W_s	Volumetric liquid water fraction in snow pack
z	Depth

See also: **Conservation Tillage**; **Crop-Residue Management**; **Freezing and Thawing:** Processes; **Mulches**; **Thermal Properties and Processes**; **Zone Tillage**

Further Reading

Benoit GR, Mostaghimi S, Young RA, and Lindstrom MJ (1986) Tillage–residue effects on snow cover, soil water, temperature and frost. *Transactions of the American Society of Agricultural Engineers* 29: 473–479.

Flerchinger GN and Saxton KE (1989) Simultaneous heat and water model of a freezing snow-residue-soil system. I. Theory and development. *Transactions of the American Society of Agricultural Engineers* 32: 565–571.

Miller RD (1980) Freezing phenomena in soils. In: Hillel D (ed.) *Applications of Soil Physics*, pp. 254–299. New York: Academic Press.

Sharratt BS (2002) Corn stubble height and residue placement in the northern US Corn Belt. I. Soil physical environment during winter. *Soil and Tillage Research* 64: 243–252.

Sharratt BS and Flerchinger GN (1995) Straw color for altering soil temperature and heat flux in the subarctic. *Agronomy Journal* 87: 814–819.

Waggoner PF, Miller PM, and DeRao HC (1960) *Plastic Mulching: Principles and Benefits*. Connecticut Agricultural Experimental Station Bulletin 634. New Haven, CT: Connecticut Agricultural Experiment Station.

Williams PJ and Smith MW (1991) *The Frozen Earth: Fundamentals of Geocryology*. New York, NY: Cambridge University Press.

Wilson C (1969) *Climatology of the Cold Regions Northern Hemisphere. II*. Monograph I-A3b. Hanover, NH: US Army Cold Regions Research and Engineering Laboratory.

Processes

G N Flerchinger, USDA Agricultural Research Service, Boise, ID, USA
G A Lehrsch, USDA Agricultural Research Service, Kimberly, ID, USA
D K McCool, USDA Agricultural Research Service, Pullman, WA, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Seasonally frozen soil strongly influences runoff and erosion on large areas of land around the world. In many areas, rain or snowmelt on seasonally frozen soil is the single leading cause of severe runoff and erosion events. As soils freeze, ice blocks the soil pores, greatly diminishing the permeability of the soil. This is aggravated by the tendency of water to migrate to the freezing front, causing elevated ice content and frost heave.

Soil freezing and thawing also play a role in a variety of other environmental processes. Frost heave poses significant problems for structures, roads, and plant roots. Soil freezing and thawing can create stress fractures and alter soil physical properties, including pore continuity and aggregate stability; these alterations can influence soil hydraulic properties and erodibility long after the soil is thawed. Water migration associated with soil freezing can strongly influence solute movement. Artificial freezing of the soil has been used to create a barrier in order to isolate contaminants within the soil.

Freezing and thawing of the soil are controlled by the complex interactions of heat and water transfer at the soil surface governed by meteorological and environmental conditions at the soil–atmosphere interface. Different types of frost may form, depending on soil moisture content, rate of freezing, ground cover, and soil physical characteristics. Soil permeability, erodibility, and frost heave depend largely on the type of frost formed.

Heat and Water Relations During Freezing and Thawing

Due to negative water potentials, soil water exists in equilibrium with ice at temperatures below the normal freezing point of bulk water and over the entire range of soil-freezing temperatures normally encountered. When ice is present in the soil, the soil matric potential is strongly influenced by the temperature. As temperature at the freezing front decreases, more and more water freezes, water potential becomes more negative, and liquid water content continues to drop, creating a gradient in water potential and liquid water content. This drop in liquid water content at the freezing front has a similar effect to drying of the soil, and water migrates from moist regions to the freezing front. This often results in elevated ice content, ice lenses, and frost heave.

When ice is present, soil water potential is a function of temperature. This relation is expressed by the Clausius–Clapeyron equation as:

$$\phi = \pi + \psi_{\mathrm{m}} = L_{\mathrm{f}}\left(\frac{T - T_{\mathrm{frz}}}{T}\right) \qquad [1]$$

where ϕ is total water potential, π is soil water osmotic potential, ψ_{m} is soil matric potential, T is absolute temperature, T_{frz} is the freezing point of bulk water (typically 0°C or 273.16 K), and L_{f} is energy, termed the latent heat of fusion, required to freeze water. Thus, when ice is present in the soil, heat and water flux through the soil are tightly coupled, i.e., the matric potential and therefore liquid water content are defined by the temperature and osmotic potential. The relation between matric potential and liquid water content defined by the moisture-release curve is typically assumed valid for frozen conditions.

Darcy's equation can be used to describe steady-state, one-dimensional water flux through the soil:

$$q_{\mathrm{l}} = -K\left(\frac{\partial(\psi_m + \psi_{\mathrm{g}})}{\partial z}\right) \qquad [2]$$

where K is the unsaturated conductivity, ψ_{m} is soil water matric potential, ψ_{g} is gravitational potential, z is depth within the soil, and $\partial(\psi_{\mathrm{m}} + \psi_{\mathrm{g}})/\partial_z$ is the gradient in soil water potential. The transient mass balance equation for water, including the effects of freezing and thawing within the soil, can be written as:

$$\frac{\partial\theta_{\mathrm{l}}}{\partial t} + \frac{\rho_{\mathrm{i}}}{\rho_{\mathrm{l}}}\frac{\partial\theta_{\mathrm{i}}}{\partial t} = -\frac{\partial q_{\mathrm{l}}}{\partial z} + U \qquad [3]$$

Terms on the left-hand side of the equation represent: the time rate of change of liquid water content, and the time rate of change of ice content. Terms on the right-hand side are: the gradient in water flux (i.e., the net flux of water into a layer of soil), and a source/sink term for water. Here, θ_{i} and θ_{l} are volumetric ice and water content, and ρ_{i} and ρ_{l} are the density of ice and water. Eqn [3] states that the net liquid water flux into a soil layer must equal the combined change in ice and water content of the soil. When the net flux is equal to zero, any change in liquid water content must be offset by a change in ice content, adjusted for the difference in density. Although this change in density can result in expansion of the soil matrix, it is not the primary cause of frost heave.

Neglecting vapor flux terms, the equation used for describing one-dimensional heat flow within a partially frozen soil is:

$$C_s \frac{\partial T}{\partial t} - \rho_i L_f \frac{\partial \theta_i}{\partial t} = \frac{\partial}{\partial z}\left[k_s \frac{\partial T}{\partial z}\right] + S \qquad [4]$$

The terms in eqn [4] represent: energy stored in the soil due to a temperature increase; latent heat required to freeze water; net thermal conduction into a layer; and a source term for heat added to the soil. Here, C_s is the volumetric heat capacity of the soil, $\partial T/\partial t$ is the time rate of change of temperature, L_f is the latent heat of fusion, and k_s is soil thermal conductivity. This equation states that the net heat transfer into a volume of soil by thermal conduction and source terms is offset by a change in temperature and a change in ice content of the soil.

Water Migration and Frost Heave

From eqn [1], as the temperature drops below freezing and ice begins to form, the water potential becomes more negative. This creates a gradient in water potential and causes moisture movement toward the freezing zone. Water movement to the freezing zone is described by eqn [2]. If water movement to the freezing zone is sufficient, ice lenses occur, causing the soil matrix to expand. In the case of vertical frost penetration from the soil surface, ice lenses can cause the soil to heave upwards if the pressure associated with freezing exceeds overburden pressures.

The extent of water migration and ice accumulation in the freezing zone is controlled primarily by the rate of freezing front advance in relation to the unsaturated hydraulic conductivity. When the soil is frozen rapidly, there is little opportunity for water to migrate to the freezing front and soil water is essentially frozen in place. Similarly, if the unsaturated conductivity is low, water migration to the freezing front will be slow. Very dry and/or coarse-textured soils have relatively low unsaturated conductivities and exhibit much less frost-related water movement and frost heave than moist, fine-textured soils. Excessive water migration and frost heave are experienced most often when a very moist, fine-textured soil is frozen relatively slowly.

Freezing Dynamics

Soil temperature and water dynamics during soil freezing are shown for a silt loam soil in Figure 1. Continuous measurement of near-surface soil ice content is problematic; however, liquid water content during soil freezing can be measured quite accurately with time domain reflectometry (TDR). Therefore, for illustrative purposes, total water (liquid plus ice) and liquid water content plotted in Figure 1 were simulated by the simultaneous heat and water

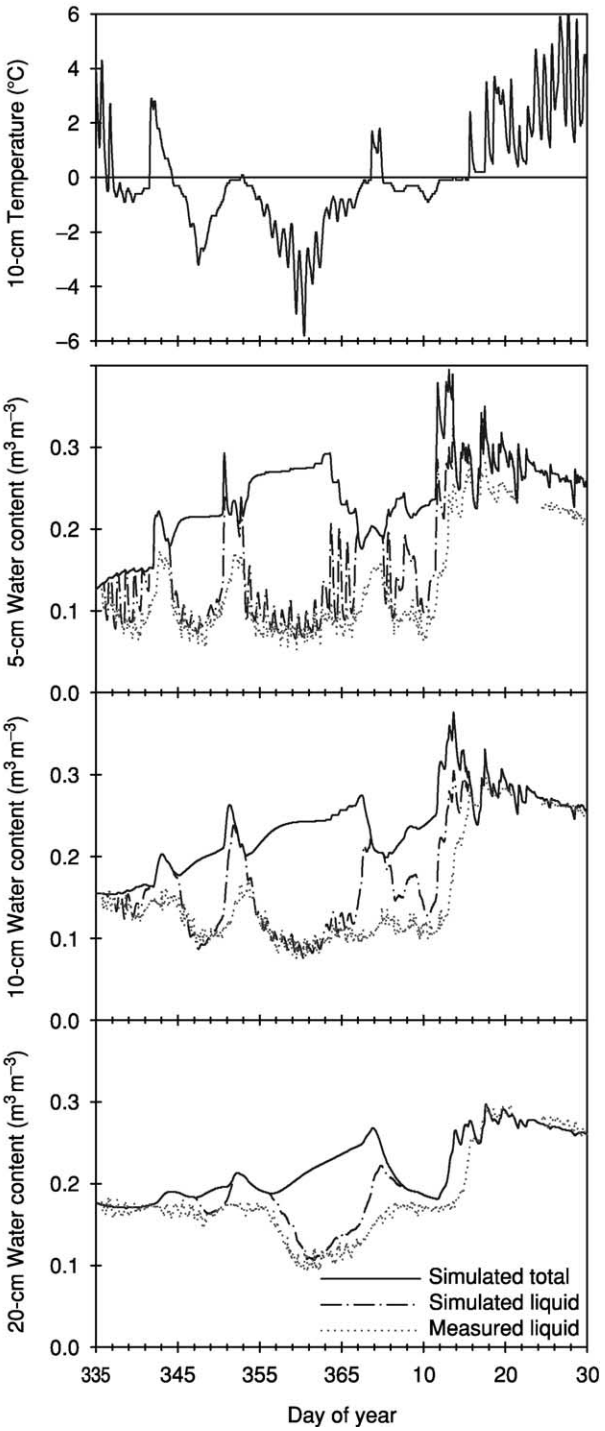


Figure 1 Simulated 10-cm soil temperature, as well as total water content and simulated and measured liquid water content of a silt loam soil for the 5-,10-, and 20-cm depths. (Adapted from Flerchinger GN (2002) Coupled soil heat and water movement. In: *Encyclopedia of Soil Science*. New York: Marcel Dekker, Inc., with permission.)

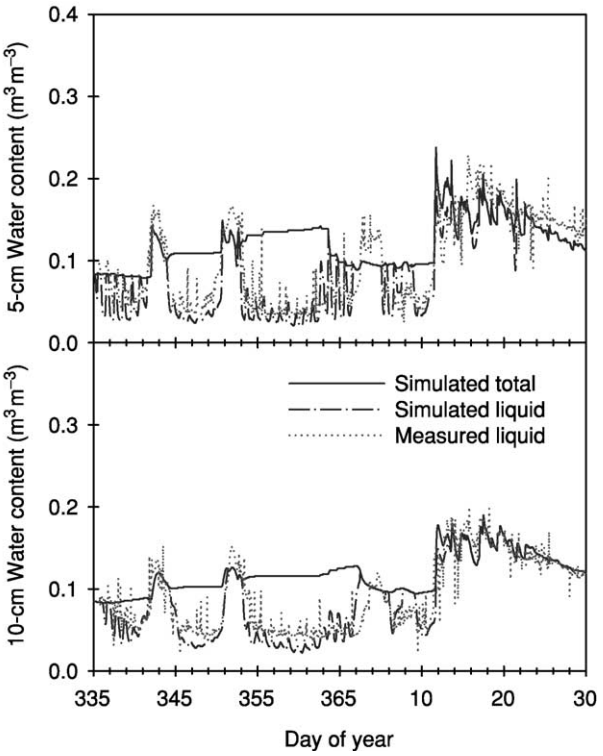


Figure 2 Simulated total water content and simulated and measured liquid water content for a loamy sand soil for the 5- and 10-cm depths. (Adapted from Flerchinger GN (2002) Coupled soil heat and water movement. In: *Encyclopedia of Soil Science*. New York: Marcel Dekker, Inc., with permission.)

(SHAW) model; measurements of liquid water content are plotted for comparison.

The 10-cm soil temperature plotted in Figure 1 shows several freeze–thaw cycles. Frozen conditions can be observed from the water content plotted for the 5-, 10-, and 20-cm depths by separation of the simulated liquid water content line from the total water content line; the difference between the two lines is ice content. Accumulation of ice and an increase in total water content due to water migration to the freezing front can be observed at all three depths plotted. After 9 mm of rain and snowmelt on days 341 and 342, simulated water content above 20 cm was decreasing on day 343 due to drainage. After initiation of soil freezing on day 344, direction of flow reversed, and simulated water flow above 10 cm was upward toward the freezing front. Total water content of the 5-cm depth began to increase on day 344, while liquid water content continued to decrease. As the frost front advanced and the 10-cm depth began to freeze, water migration into the 5-cm soil layer ceased, and the 10-cm total water content began to increase. Subsequently, the 20-cm depth began to freeze on day 347.

Soil water content is shown in Figure 2 for a loamy sand soil, which is a much more coarsely textured soil than the silt loam. Soil water dynamics for the loamy sand are considerably less responsive to freeze–thaw processes than the silt loam soil. Due to the low unsaturated conductivity of the loamy sand, there is much less moisture migration to the freezing front than for the silt loam. As a result, increase in total water content is much smaller.

Freeze–Thaw Impacts on Infiltration

Rain and/or rapid snowmelt on impermeably frozen soil is the leading cause of severe flooding and erosion in many areas of the world. Soil freezing can dramatically reduce the soil's infiltration capacity. Ice blocks the soil pores, resulting in large runoff events from otherwise mild rainfall or snowmelt events.

The permeability of frozen soil is affected by the occurrence, depth, and ice content of the soil, which is dependent on the interrelated processes of heat and water transfer at the soil surface and within the soil profile. Soil freezing and thawing can also alter soil physical properties or structure that impact infiltration. Changes in aggregate stability or stress fractures caused by freezing affect soil structure and pore continuity and thus affect infiltration even after the soil is thawed.

The type of frost formed influences soil permeability after freezing. Soil frost may be divided into four types: granular, honeycomb, stalactite, and concrete. Granular frost is usually found in woodland soils containing organic matter. It consists of small frost crystals, which aggregate around soil particles, but remain separate from each other. Honeycomb frost is commonly found in highly aggregated organic soils and has a loose porous structure, which resembles a honeycomb. Both granular and honeycomb frost typically have high infiltration rates. Stalactite frost often forms in bare soil, which is saturated at the surface. This type of frost consists of loosely fused columnar ice crystals and absorbs water rapidly because of its open porous structure. Concrete frost usually forms in bare, fine-textured, agricultural soils where upward migration of moisture is significant. It is characterized by a complex formation of many thin ice lenses and leaves the ground very hard, much like concrete. Depending on water content, concrete frost can be almost impermeable.

Frozen soil infiltration rates decrease dramatically with soil water content. Water is held less tightly in the large pores, therefore the largest pores that contain water at the time of freezing are the first to freeze. When these larger pores, which conduct water more readily, are blocked by ice, the tortuosity of flow paths increases and permeability and infiltration are severely reduced. Infiltration into a frozen, relatively

dry silt loam can typically decrease from 1.0 cm h^{-1} to 0.01 cm h^{-1} as the water content increases to near saturation.

Ice lenses that form due to water migration from the unfrozen subsoil toward the freezing front have an added effect on infiltration. These lenses, typically formed in wet soils having a high proportion of silt, are often a barrier to infiltration. Melting of ice lenses during infiltration influences the temporal variation in infiltration. If ice lenses melt during an infiltration event, infiltration rate can nearly return to unfrozen infiltration rates.

Tillage and surface characteristics that alter porosity and water and heat transport processes can influence infiltration. Tillage processes that create macropores typically have a positive effect on infiltration. Unless the soil is extremely wet, macropores are not filled with water at the time of freezing and remain open for infiltration. However, tillage is likely to have little effect on infiltration if the freezing front descends below the depth of tillage.

Knowledge of frozen-soil infiltration processes lags considerably behind nonfrozen processes. Accurate quantitative descriptions or algorithms of frozen soil infiltration are lacking, partly due to experimental difficulties in measuring infiltration into frozen soil and characterizing the ice content and structure within the frozen soil. With the exception of expensive laboratory techniques such as nuclear magnetic resonance (NMR), there are no quantitative means of directly measuring ice content and structure in the soil, which is the single most important factor affecting infiltration potential upon freezing. Ice content can be computed as the residual between liquid water content measured by TDR and total water content measured by neutron probe or gravimetric samples; however, the sampling volumes of these techniques are dramatically different, making accurate measurement of ice content difficult.

Our understanding of frozen-soil infiltration processes is further hampered by the fact that ice content, pore blockage, and infiltration rate change as water infiltrates into frozen soil. Introducing water into frozen soil causes freezing of the infiltrating water, thawing of the ice contained within the soil, or both. Thus, there is no steady-state infiltration rate analogous to that in unfrozen soil. An approach to circumvent this problem is to use an alternate fluid that remains viscous at subfreezing temperatures. Fluids such as ethylene glycol and air have been used as test fluids for characterizing infiltration of frozen soils. Measured permeability for these alternate fluids can be related to hydraulic conductivities by accounting for differences in density and viscosity.

Various approaches exist for estimating infiltration of frozen soils. Depending on the level of sophistication, adjustments for frozen conditions may be based on: simply whether the soil temperature is below freezing; the amount of ice present in the soil; or the available porosity remaining in the frozen soil. Very simple approaches use essentially a simple on/off switch for accounting for frozen-soil effects, in which the curve number or hydraulic conductivity is set to an arbitrary value to cause reduced infiltration when the soil is frozen. Slightly more sophisticated methods use an adjustment factor to hydraulic conductivity based on antecedent water content or ice content of the soil. Many detailed approaches for estimating infiltration in frozen soils assume the hydraulic conductivity and water retention characteristics are the same for frozen and unfrozen soils. Thus, hydraulic conductivity for infiltration is based on the unsaturated hydraulic conductivity computed from the available porosity (total porosity less volumetric ice content).

Freeze–Thaw Impacts on Soil Erodibility

As soil freezes, water migration to the freezing front can cause ice lenses to form. When the ice lenses melt, the soil often cannot reabsorb all of the excess water, particularly if an impermeably frozen layer still exists below the thawed layer. This supersaturated state results in soil that is extremely weak and susceptible to erosion. However, after drainage and consolidation, soil strength returns. Thus, partially thawed and thawed but unconsolidated soil is highly susceptible to erosion.

Shear strength of a soil is indicative of its resistance to erosion. Specifically, it is defined as the resistance to deformation by the action of tangential (shear) stress. Soil shear strength is made up of cohesion between particles and resistance of particles sliding over each other due to friction or interlocking. Cohesion is composed of true cohesion and apparent cohesion. True cohesion is a function of soil mineralogy and results from chemical bonds between particles. Apparent cohesion, however, is determined by water tension within the soil and is strongly influenced by water content. As the soil thaws at a high water content, soil strength due to apparent cohesion is nil.

Shear stress, τ, caused by water flowing over the soil surface, is defined as:

$$\tau = \gamma R S \qquad [5]$$

where R is the hydraulic radius of the flow, γ is density of water, and S is the slope of the channel of surface. The minimum amount of shear stress required to initiate particle movement is termed the

critical shear stress, τ_{cr}, which is a measure of the soil's shear strength. Soil detachment is computed as:

$$D = \alpha(\tau - \tau_{cr})^{\beta} \quad [6]$$

where α is referred to as soil erodibility and β is a fitted exponent. Studies on thawing soils have shown that moisture content has some effect on τ_{cr}, but the major effect is on α, the soil erodibility. The erodibility of a soil thawed at 5-cm water tension can be approx. 14 times that of the same soil thawed at 45 cm of tension. However, the thawed soil can regain its strength upon drying after tensile forces related to water tension are restored. This strength can be restored in a matter of hours if an impermeable soil layer thaws and allows drainage to occur or if strong evaporative conditions exist. Thus, shear strength can fluctuate dramatically, and timing of rainfall or snowmelt during the first hours of thawing can make a big difference to the erosion that occurs.

Freeze–Thaw Impacts on Aggregate Stability

Aggregate stability, a measure of an aggregate's resistance to breakdown when subjected to external forces, is an important soil property, because soil susceptibility to water and wind erosion increases as aggregate stability decreases, in general. Moreover, many soil physical and hydraulic properties, such as surface-sealing rate, infiltration rate, and hydraulic conductivity, are influenced by aggregate stability. In addition, on medium-textured soils with unstable surface aggregates, crusts can form that hinder or, in some cases, prevent the emergence of seedlings of sown crops.

Mode of Action

As the soil temperature drops below freezing, ice crystals form in the soil matrix, forming first in the soil pores. Once a crystal forms, water flows to the crystal due to a potential gradient, enlarging it, which exerts pressure on nearby aggregates. If those aggregates are constrained and cannot move away from the expanding ice crystal, the pressure exerted upon them can fracture the aggregates directly or develop planes of weakness that can, upon subsequent wet-sieving, cause the aggregates to fracture.

Factors Affecting the Impact

Antecedent water content is the greatest single factor that determines how an aggregate responds to freezing. Aggregate stability decreases, often linearly, with increasing water content at freezing. This relationship appears to hold for many soils ranging in texture from sandy loams to silty clays. As soil water contents increase, more water is available to form ice crystals or ice lenses. Moreover, wetter soil has more water-filled pore space and thicker water films surrounding soil particles, increasing the area through which unsaturated water flow can occur. Also, in wetter soil, tortuosity is less, thereby shortening the water's flow path and speeding its movement to the ice crystal, increasing the latter's rate of expansion. To preserve aggregation and reduce erosion in temperate regions, it is often recommended that producers minimize autumn soil water contents near the soil surface, whenever possible.

The rate at which soil freezes determines in large measure the impact of freezing on soil structure, principally by affecting water redistribution within the soil. If the upper, moist horizons of a soil freeze quickly, that is, if the air temperature decreases sharply, water in those horizons is essentially frozen in place. Consequently, ice lenses, even if they form, do not thicken appreciably and do not compress nearby soil. On the other hand, if a relatively wet soil freezes slowly, water moves to an ice lens and freezes there, thickening the lens, and causing structural deterioration and frost heave.

Soil texture and organic matter also influence aggregate response to freezing. Soils containing high proportions of sand are easily weakened or even fractured. On the other hand, soils with a lot of clay are better able to withstand pressures exerted by nearby ice-crystal enlargement, probably because of additional bond strength provided by more or stronger clay bridges that form between silt and/or sand particles within the aggregate. Organic matter, known to increase the stability of unfrozen aggregates with diameters greater than 0.25 mm, also strengthens aggregates that are later subjected to freezing stresses, providing the water content at freezing is not too great. Elasticity provided by organic matter may enable aggregates from medium-textured (or finer) soils frozen at relatively low water contents to withstand ice-lens expansion pressures before fracturing. Available data indicate that organic matter contents of 3% or more are particularly beneficial for soils that are medium-textured or finer.

The number of freeze–thaw cycles that an aggregate experiences also determines how stable the aggregate is after freezing, with responses being somewhat soil-dependent. In the past, aggregate stability was thought to decrease as freeze–thaw cycles increased, beginning with the first freeze–thaw cycle and continuing monotonically thereafter. Most early research subjected initially air-dried aggregates to many freeze–thaw cycles, often five to ten or more. Recent research, however, has demonstrated that aggregates that have not been air-dried between

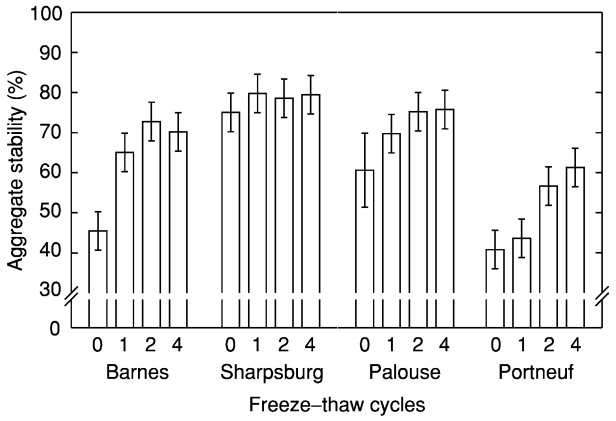


Figure 3 Aggregate stability of four soil types measured after freeze–thaw cycles. (Bars represent 95% confidence intervals.) (Reproduced from Lehrsch GA (1998) Freeze–thaw cycles increase near-surface aggregate stability. *Soil Science* 163: 63–70.)

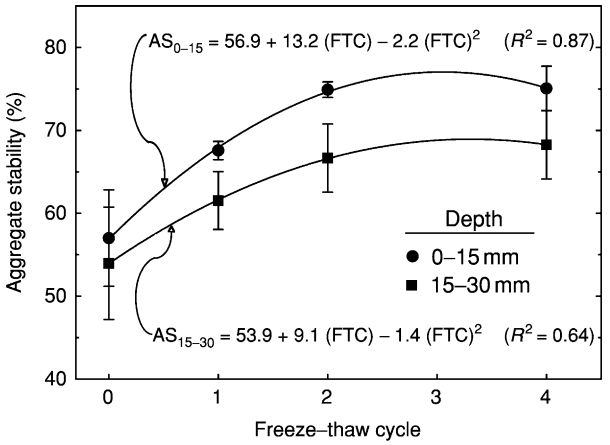


Figure 4 Effects of freeze–thaw cycles on aggregate stability at different soil depths. (Bars represent 95% confidence interval; AS denotes aggregate stability; FTC denotes number of freeze–thaw cycles.) (Reproduced from Lehrsch GA (1998) Freeze–thaw cycles increase near-surface aggregate stability. *Soil Science* 163: 63–70.)

sampling and analysis often increase in stability with the first two or three freeze–thaw cycles (Figure 3). The imposition of many freeze–thaw cycles does indeed decrease aggregate stability; what was seldom recognized was that just a few cycles could increase the stability of aggregates from medium- and fine-textured soils. It has been postulated that ice formation in interaggregate pores and the initiation and early enlargement of ice lenses increase particle-to-particle contacts. Migration of water to the enlarging lens then dries the soil matrix surrounding the ice lens, positioning polysaccharides on soil particle surfaces, gathering and arranging clay domains at points of contact between soil particles, and/or precipitating slightly soluble bonding agents such as $CaCO_3$, silica, or iron oxides at contact points. These processes help aggregates reform and increase in strength, after thawing.

It is particularly interesting that aggregate stability often increases with the first few freeze–thaw cycles but then decreases as more and more freeze–thaw cycles accrue (Figure 4). Opposing forces may be responsible for such phenomena. A force serving to strengthen aggregates may be the result of the precipitation of slightly soluble bonding agents at points of contact between soil particles during the first few cycles. An opposing force that weakens aggregates may be due to ice-lens formation, compression of nearby aggregates, and development of fracture planes. The strengthening process may well occur, and be dominant, for the first two or three freeze–thaw cycles, until most bonding agents have been precipitated from the soil solution. As freeze–thaw cycles continue to accrue, however, more and more fracture planes may be formed in aggregates near where the ice lenses formed. This persistent weakening process may then begin to play the dominant role, decreasing aggregate stability after two or three cycles, as suggested by the fitted curves in Figure 4 and commonly reported in the literature.

Aggregates that are constrained from moving about either in a sample or in a soil profile are weakened more by freezing than are unconstrained aggregates. Aggregates below the soil surface are less stable after freezing than aggregates at or near the soil surface, regardless of the number of freeze–thaw cycles to which they are subjected (Figure 4). Freezing and thawing processes thus affect aggregate stability and soil structure, in addition to water and heat flow through soil.

Soil Freezing and Thawing Effects on Solute Migration

Solutes directly affect water movement during freezing by altering potential gradients. Solutes in the soil solution add an osmotic component to the total water potential. This additional component affects soil water redistribution within the profile as a soil freezes at a given temperature. Specifically, frost heave is greater, that is, more and thicker ice lenses are formed where the soil solute concentration is low, rather than high.

As previously noted, temperature gradients cause soil water to migrate to the freezing front. Solutes in the soil water are carried to the freezing front via convection. As water freezes at the front, solutes are excluded and accumulate in the unfrozen water contained

in thin films around soil particles, within soil pores, or in relatively concentrated brine pockets, which may be entrapped within the ice lens itself. If the concentration of a solute in the unfrozen soil solution exceeds its solubility limit, that solute may be precipitated at particle-to-particle contact points, potentially increasing aggregate stability. Upon thawing, the solute transport to the freezing front via mass flow often results in increased concentrations of soluble constituents in the soil water near where the ice lens had formed.

The concentrated brine can also migrate through the frozen soil under certain climatic conditions. When air temperatures at the soil surface are very low, a temperature gradient induces these trapped brine pockets to migrate downward through the ice and frozen soil, into deeper but warmer portions of the soil profile.

See also: **Energy Balance**; **Polar Soils**; **Swelling and Shrinking**

Further Reading

Flerchinger GN and Saxton KE (1989) Simultaneous heat and water model of a freezing snow-residue-soil system. I. Theory and development. *Transactions of the American Society of Agricultural Engineers* 32(2): 565–571.

Formanek GE, McCool DK, and Papendick RI (1984) Freeze–thaw and consolidation effects on strength of a wet silt loam. *Transactions of the American Society of Agricultural Engineers* 27(6): 1749–1752.

Fuchs M, Campbell GS, and Papendick RI (1978) An analysis of sensible and latent heat flow in a partially frozen unsaturated soil. *Soil Science Society of America Journal* 42(3): 379–385.

Iskandar IK, Wright EA, Radke JK, Sharratt BS, Groenevelt PH, and Hinzman LD (eds) (1997) *Proceedings of the International Symposium on Physics, Chemistry, and Ecology of Seasonally Frozen Soils*. CRREL Special Report 97-10. Hanover, NH: US Army Cold Regions Research and Engineering Laboratory.

Kok H and McCool DK (1990) Quantifying freeze/thaw-induced variability of soil strength. *Transactions of the American Society of Agricultural Engineers* 33(2): 501–506.

Lehrsch GA (1998) Freeze–thaw cycles increase near-surface aggregate stability. *Soil Science* 163: 63–70.

Miller RD (1980) Freezing phenomena in soils. In: Hillel D (ed.) *Applications of Soil Physics*, pp. 254–299. New York: Academic Press.

Perfect E, van Loon WKP, Kay BD, and Groenevelt PH (1990) Influence of ice segregation and solutes on soil structural stability. *Canadian Journal of Soil Science* 70: 571–581.

Stahli M, Jansson PE, and Lundin LC (1999) Soil moisture redistribution and infiltration in frozen sandy soils. *Water Resources Research* 35(1): 95–103.

VanKlaveren RW and McCool DK (1998) Erodibility and critical shear of a previously frozen soil. *Transactions of the American Society of Agricultural Engineers* 41(5): 1315–1321.

FUNGI

K Ritz, Cranfield University, Silsoe, MK45 4DT, UK

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Fungi are ubiquitous in soils, but comprise a varying proportion of the microbial biomass in different systems. They tend to dominate in soils of high organic matter and constitute a smaller proportion in intensively managed mineral soils. They are involved in a plethora of functional roles, encompassing biological, chemical, and physical interactions, and are of great ecological and economic significance. The study of fungi is termed mycology (after *mukēs*, Greek for 'fungus').

Classification

The 'true fungi' belong to the kingdom of Fungi and are eukaryotic, exclusively heterotrophic organisms with cell walls that contain chitin or chitosan as a major constituent, and typically have a thread-like hyphal growth-form, although unicellular forms are common in the yeasts. There are four phyla, the Chytridiomycota, the Zygomycota, the Ascomycota, and the Basidiomycota, plus an informal group denoted the mitosporic fungi (formerly the Fungi Imperfecti or Deuteromycota), which lack a sexual phase and are not ascribed to a formal taxonomic position. Only the Chytridiomycetes produce motile zoospores. Representatives from all taxonomic groupings are commonly found in soils, where fungal biodiversity,

as for all soil microbes, is typically several orders of magnitude more than aboveground or aqueous systems.

In an informal sense, organisms studied by mycologists cross the eukaryotic kingdoms of Protozoa and Chromista as well as the Fungi. Within the Protozoa, the cellular slime molds (phyla Dictosteliomycota, Acrasiomycota) and plasmodial slime molds (phyla Myxomycota, Plasmodiophoromycota) form fruiting bodies with a resemblance to some fungal types; the ameboid phagocytotic cells of these organisms are common in soils and are implicated in nutrient cycling. Within the Chromista, the Oomycetes form mycelia but contain some cellulose in their walls and produce motile zoospores. This group includes some important soil-borne plant pathogens within the order Pythiales, such as *Phytopthora infestans* (potato blight) and *Pythium* spp. (damping off).

Biology

Morphology and Form

A fundamental characteristic of the filamentous (eucarpic) fungi is their growth-form, which consists of a thread-like hypha which grows by apical extension and periodic branching to form a mycelium that permeates the environment in which the fungus is growing (**Figure 1**). The diameters of individual hyphae vary according to age, species, and nutritional conditions but are typically 3–10 μm. In extreme cases, up to 2 km of hyphae per gram of soil have been measured in some soils; a typical cultivated arable soil contains several meters per gram. Mycelia are indeterminate structures that grow in three dimensions and therefore conceptually are difficult to size. In soil they may occupy almost any volume, and the largest organisms on the planet are probably mycelia of *Armillaria bulbosa*, individual clones of which are reported to extend continuously over tens of hectares in North American forests (**Figure 1g**). Hyphae are generally compartmentalized by cross-walls (septa) which occur at varying intervals depending on species (**Figure 1b**). Septa may be complete or perforated to varying degrees and play a role in regulating the passage of materials, including organelles, between adjacent compartments. If mycelia are fractured, septal pores become occluded, which prevents leakage of cytoplasm from the open ends of broken hyphae. Hyphae may aggregate to form higher-order structures such as strands (cords; **Figure 1e**) or rhizomorphs, which enable long-distance foraging (decameters in the case of woodland basidiomycetes). Hyphal aggregation and differentiation can produce morphologically complex macrostructures such as the intricate fruiting bodies (basidiocarps) of some Basidiomycetes. Those fungi that do not form hyphae (predominantly yeasts) have a holocarpic growth form, comprising individual cells that replicate by budding or binary fission. Some fungi are dimorphic and can switch between holocarpic and eucarpic growth forms depending on environmental conditions (e.g., *Mucor hiemalis*).

Physiology and Reproduction

All fungi are obligate heterotrophs, i.e., they utilize fixed (organic) C sources as substrate. Across the kingdom an extremely wide range of compounds can be utilized, ranging from C_1 compounds (methylotrophic yeasts) through to lignin and industrially produced polyphenols – almost any naturally occurring macromolecule occurring in soil can be degraded by some fungi. A concomitantly wide range of enzymes and metabolic capabilities are expressed. The range of compounds utilizable by any particular species is usually reflected by their normal environmental niche, and different types vary widely in their capacity to utilize different substrates. Respiration can be aerobic or anaerobic, and obligate and facultative types of both forms occur. Some fungi obtain their nutrition from other living organisms via parasitic or mutualistic associations, whilst others are saprotrophs. Many soil fungi are very efficient scavengers of nutrients and grow sparsely but effectively under the oligotrophic conditions that prevail in many soils. A wide variety of secondary metabolites are produced by soil fungi, including many volatile organic compounds (VOCs) which may act as signals or substrates as they diffuse through pore networks. If fungal spores or mycelial fragments are added to nonsterile soils, they rarely germinate. This phenomenon, known as fungistasis (or mycostasis) is probably due to a lack of available nutrients (the extant microflora normally having depleted the microbially available nutrient pool), the presence of inhibitory (antibiotic) compounds produced by the existing microflora or a combination of both.

Reproduction can be sexual or asexual, and a variety of mating systems occur. Many fungi have genetic systems that prevent mating between genetically identical cells or mycelia. If hyphae encounter each other, they may fuse (anastomose). If they are somatically compatible, the fusion results in the formation of larger mycelial networks, with attendant functional consequences in terms of spatial exploration, activity, translocation, and combativeness. Somatic incompatibility, however, results in a rejection response,

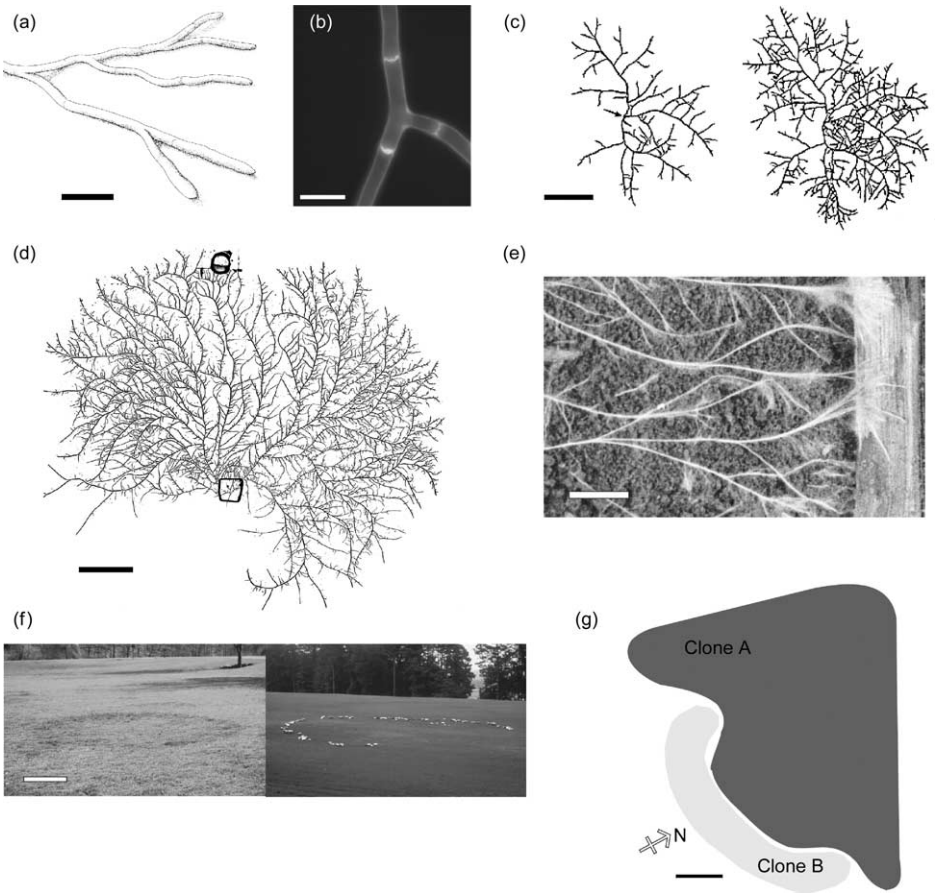


Figure 1 Aspects of mycelia of soil fungi at different scales. (a) stylization of fungal hyphae, showing branching habit and morphology of tips. Scale bar 50 μm; (b) micrograph of hyphae of *Rhizoctonia solani* showing branch and septa. Scale bar 50 μm; (c) mycelium of *Trichoderma viride* 44 h (left) and 48 h (right) after germination, grown on spatially uniform nutrients. Scale bar 100 μm; (d) mycelium of *T. viride* grown on spatially nonuniform nutrients (point of inoculation is denoted by a square, localized nutrient source denoted by a circle). Scale bar 5 mm; (e) colonizing fans at tips of foraging cords of *Agrocybe gibberosa* when a piece of straw (to right of image) is encountered. Scale bar 5 mm; (f) fairy ring in grassland, resulting from enhanced mineralization of soil nutrients by subterranean mycelium, expressed as enhanced plant growth (left image) and following basidiocarp emergence (right image). Scale bar 1 m; (g) mapped extent of two clones of *Armillaria bulbosa* in a Michigan forest. Scale bar 100 m. ((a) is adapted from Jennings DH and Lysek G (1999) *Fungal Biology*. Oxford, UK: Bios Scientific, with permission. (b) is courtesy of Karl Ritz. (c) is reproduced with permission from Ritz K and Crawford JW (1990) Quantification of the fractal nature of colonies of *Trichoderma viride*. *Mycological Research* 94: 1138–1142. (d) is adapted from Crawford JW *et al.* (1993) Quantification of fungal morphology, gaseous transport and microbial dynamics in soil: an integrated framework utilising fractal geometry. *Geoderma* 56: 157–172, with permission. (e) is reproduced with permission from Robinson CH *et al.* (1993) Resource capture by interacting fungal colonies on straw. *Mycological Research* 97: 547–558. (f) is reproduced with permission from Eric Nelson, Cornell University. (g) is adapted from Smith ML *et al.* (1992) The fungus *Armillaria* bulbosa is among the largest and oldest living organisms. Nature 356: 428–431 © 1992 Macmillan Magazines, with permission.)

and the cytoplasm in fused compartments is destroyed. Fungal propagules generally take the form of asexual or sexual spores, sclerotia (dense, heavily pigmented hyphal aggregations), or hyphal fragments. Some species sporulate very profusely (e.g., *Penicillium*), others have never been known to produce any form of spore (e.g., *Rhizoctonia solani*). In soil, propagules are dispersed by mass flow in soil water, carried by animals internally or externally to their bodies, or via physical (often faunally mediated) soil movement.

Biotic Interactions

In the soil environment, no organism grows in isolation, but always in the context of a wider microbial community; organisms have evolved under such circumstances over millennia. It is not surprising therefore that there are extensive interactions between soil fungi and all other classes of organism, including other fungi. At present we only have a limited understanding of the degree and nature of such interactions, since the majority of microbiological research has been

on the growth and physiology of microbes grown in isolation, in simple combination, and often *in vitro*.

Fungal Associations with Other Microbes

Fungal associations with algae or cyanobacteria, in the form of lichens, play a role in soil genesis via rock weathering. In soils proper, when fungal mycelia meet, there can be a variety of outcomes from such interactions. If they are somatically compatible, they will fuse to form a larger mycelium (see above); other outcomes include an intermingling without fusion, chemically mediated repellence, physical exclusion by the development of insulated barriers, or death and replacement of one mycelium by the other. There are complex hierarchies of combativity between different species and strains, further dependent on environmental factors such as temperature, water, and nutritional status. Fungal communities in soil are therefore in a state of perpetual dynamics and interaction. There are also many examples of specifically hyperparasitic fungi (e.g., *Arthrobotrys oligospora*, *Coniothyrum minitans*) and bacterial mycoparasites (e.g., *Pseudomonas* spp.). Mycoviruses do not have an extracellular phase, and since they do not necessarily affect the fungal phenotype they may be more prevalent than was hitherto thought.

Fungal Associations with Plants

Plants are the primary autotrophs in terrestrial systems and therefore responsible for the deposition of fixed C into soils, which provides the majority of substrate for the heterotrophic soil biomass. Biotic associations between vascular plants and fungi are extensive.

Mutualism Mutualistic associations between plant roots and fungi are termed 'mycorrhizas' (Figure 2a–e) and first evolved almost simultaneously with the emergence of land plants. Such associations are extremely common, and the natural state for the majority of root systems is to be infected to some extent by mutualistic fungi. There are four major types of mycorrhizal association that differ anatomically and physiologically and have variable host ranges (Table 1). In the majority of cases, the fungal partner derives carbon from the host, and the plant benefits from an enhanced uptake of nutrients such as P or N and water, mediated by extensive external mycelia. Mycorrhizal infection may also confer resistance against root pathogens. Given that most soil microbes are C-limited, it is likely that mycorrhizal hyphae constitute a significant proportion of the fungal biomass in vegetated soils, since they are intimately associated with a readily available source of C via their host plants. Mycorrhizas strongly affect the competitive ability of plants (often mediated via nutrient acquisition) and thus play a role in governing the community structure of vegetation. Mycorrhizal networks can infect a number of plants of the same or different species simultaneously and thus form an interconnected network, providing the possibility for nutrient transfer between plants via interconnecting hyphae. The ecological significance of such phenomena is unclear.

Parasitism Plant pathogenic fungi cause very significant economic losses to crops in both temperate and tropical agricultural and forestry systems. They also occur in natural systems and play a role in mediating plant community structure, although relatively little is known about the extent of such regulation.

Soil-borne plant pathogens occur across all fungal taxonomic groups and vary in their host specificity; important examples include *Gaeumannomyces graminis* (take-all of wheat; Figure 2f), *Rhizoctonia solani* (root and stem rots, including potato tubers, with a wide host range), *Pythium* spp. (damping-off), *Armillaria mellea* (honey fungus, with a wide host range of trees), *Fusarium oxysporum* (a wide variety of forms, which affect specific plants – e.g., *F. cubense* causes wilt in banana).

The primary site of infection is normally roots, although some pathogens attack the hypocotyls or stem bases of plants on or near the soil surface, and there are a range of tuber diseases. Many pathogens show strong taxes toward roots, mediated via soluble or volatile exudates, or electrical charges (electrotaxis). Oomycete zoospores show negative geotaxis and a tendency to swim against the direction of water flow, which keeps them in upper soil layers where roots are more abundant, as well as electrotaxis toward roots and wound sites. The pathogenic mode of action takes many forms between different pathogens, from necrotrophic attack and wholesale degradation of tissue, through cortical invasion, root hypertrophy, to the blocking of vascular tissue and hence interference with water and solute transport (wilts). There is often a proliferation of mycelia on root surfaces prior to penetration via localized aggregations of hyphae (infection cushions).

Control of root pathogens is often based upon use of resistant cultivars, crop rotation to mitigate build-up of pathogen populations, and appropriate soil management such as tillage practice, organic matter incorporation, or liming. Persistent growth of the same cereal crop can result in so-called take-all decline, where disease severity is reduced after several seasons, thought to be due to the build-up of microbial communities that are antagonistic to *Gaeumannomyces*. Chemical control is rarely feasible due to

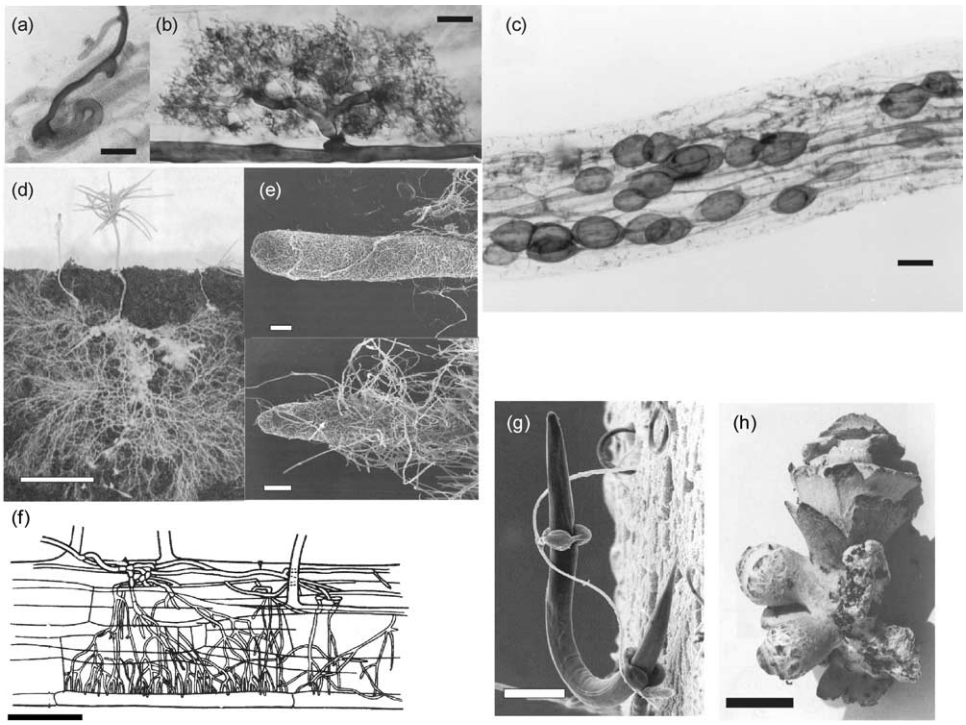


Figure 2 Mutualistic associations involving soil fungi. (a) infection point of arbuscular mycorrhizal (AM) fungus on root of *Plantago lanceolata* (plantain). Scale bar 50 μm; (b) arbuscule in root cell of *Allium porrum* (leek). Scale bar 3 μm; (c) intraradical mycelium and vesicles of AM fungus proliferating in root cortex of *P. lanceolata*. Scale bar 200 μm; (d) external mycelium of *Suillus bovinus* on *Pinus sylvestris* (pine) in microcosm system. Scale bar 5 cm; (e) ectomycorrhizal colonization of roots of *Betula alleghaniensis* (yellow birch) by *Pisolithus tinctorius*. Top: thin mantle of mycelium formed on root surface; bottom: compact mantle entirely enveloping roots, and development of external mycelium. Scale bar 100 μm; (f) infection of *Triticum aestivum* (wheat) roots by *Gauemannomyces graminis* (take-all). Scale bar 250 μm; (g) nematode trapped by constricting rings of *Arthrobotrys anchonia*. Scale bar 100 μm; (h) basidiomata of *Leucoagaricus gongylophorus* removed from nest of *Atta cephalotes* (ant). Scale bar 2 cm. ((a, c) Reproduced with permission from Karl Ritz. (b) Reproduced with permission from Brundrett MC *et al.* (1983) A new method for observing the morphology of vesicular-arbuscular mycorrhizae. *Canadian Journal of Botany* 62: 2128–2134. (d) Reproduced with permission from Harley JL, Smith S, and Reid DJ (1996) Mycorrhizal Symbiosis, 2nd Edition, with permission from Academic Press. (e) Reproduced with permission from Massicotte HB *et al.* (1990). Structure and ontogeny of *Betula alleghaniensis-Pisolithus tinctorius ectomycorrhizae. Canadian Journal of Botany* 68: 579–593. (f) Adapted from Skou JP (1981) Morphology and cytology of the infection process. In: Asher MJC and Skipton PJ (eds) *The Biology and Control of Take-all.* London, Academic Press, with permission. (g) Reproduced with permission from George Barron. (h) Reproduced with permission from Fisher PJ *et al.* (1994) *Leucoagaricus basidiomata* from a live nest of leaf-cutting ant *Atta cephalotes. Mycological Research* 98: 884–888.)

the buffering capacity of soil against bioactive compounds and the rapid degradation of such compounds by the soil microflora.

Fungal Associations with Fauna

Fungal associations with soil fauna are extremely common, including a wide range of mutualistic and parasitic relationships. For example, predatory, nematode-trapping species (e.g., *Arthrobotrys* and *Dactylella* spp.) have evolved an extensive variety of mechanisms to procure their prey, including sticky hyphae, branches, nets, and knobs, which are sufficiently adhesive to prevent escape of nematodes prior to digestion by the fungus (analogous to the insectivorous sundew plants), and hyphal rings (lassoes), some types of which contain inflating trigger cells, which grip the bodies of prey to prevent escape (Figure 2g). Other fungi (e.g., *Verticillium* spp.) infect nematodes via spore attachment and hyphal penetration of cysts, juveniles, or adults. Virtually all insects studied to date appear to have a range of parasitic and mutualistic (principally in their guts) fungi associated with them, including during larval and pupal stages.

Fungal hyphae and spores are also attacked or grazed by representatives from most faunal groups, including protozoa, nematodes, worms, insects, and mites. Subtle interactions occur. For example, there is a wide variation in the degree of palatability of different fungi to different fauna, which results in preferential grazing of mycelia by soil fauna, analogous to aboveground herbivory. Soil fauna play a strong role in dispersing fungi through soil, either as spores attached to their bodies or via egestion in faeces at sites remote from where the fungus was

Table 1 Basic characteristics of the four mycorrhizal types

Mycorrhizal type	*Habit*	*Host range*	*Examples*	*Notes*
Arbuscular (AM)	Endotrophic; form arbuscules and sometimes vesicles in root cortices	*c.* 90% of vascular plants; exceptions include members of Brassicae	Exclusively order Glomales within Zygomycetes (e.g., *Acaulospora, Gigaspora, Glomus*)	Obligate mutualists; can only be cultured in association with plants (including root-organ tissue culture systems)
Ecto- (ECM)	Ectotrophic; form extensive mycelial sheath around root	Mainly woody plants	Basidiomycetes (e.g., *Amanita, Suillus*); Ascomycetes (e.g., *Melanospora*); Zygomycetes (*Endogone*)	May also grow saprotrophically in absence of host plant, and in pure culture
Ericaceous	Endotrophic; extensive intracellular coils formed inside host cortex	Ericaceae	*Hymenoscyphus*, *Oidiodendrum*	May enhance availability of N to host by degradation of soil organic N
Orchidaceous	Endotrophic; short-lived intracellular coils in host root cells	Orchidaceae	*Rhizoctonia, Marasmius*	Often obligate mutualism required for host plant, to enable seedling development; achlorophyllous orchids are solely dependent on fungus for substrate

ingested. Some Basidiomycetes (e.g., *Termitomyces*, *Acromyrmex*, *Leucoagaricus* spp.; Figure 2h) are 'farmed' by termites and ants, maintained within the insect colonies in essentially pure culture on imported plant material.

Fungal Contributions to Soil-System Function

Nutrient Cycling

Soil fungi play a crucial role in nutrient cycling in terrestrial systems, due to the primary role they play as decomposers, mediated by a large repertoire of catabolic activities. They are particularly involved in cycling C, N, and P, but have roles in most of the other soil elemental cycles. Fungi are primarily involved in the decomposition of plant material, the lignocellulosic components of which are relatively recalcitrant to bacteria. There is typically a succession of fungal (and other microbial) species on new substrate when it enters soil, based upon the initial decomposition of simpler compounds and a subsequent degradation of more complex polymers. Cellulases are produced by a wide variety of fungi across the taxonomic classes; these include the brown rots, so-called because their activity results in a darkening of wood as the cellulose is decomposed and lignin (which they are not able to decompose) remains. Within fungi, polyphenol oxidases (mainly laccases), involved in the decomposition of lignin, are largely confined to Basidiomycetes and Ascomycetes. These include the white rots, species which normally also have strong cellulase activity and render wood a light color during decay.

As well as invoking the enzymatically mediated degradation of polymers, many fungi also produce a variety of compounds such as organic acids or siderophores, which solubilize or immobilize essential or toxic metals, with consequences for both their own growth and the growth of proximal organisms. Fungi also act to immobilize nutrients in soil through sequestration in mycelia, which are further released following death and lysis or attack by pathogens or grazers. Eucarpic fungi play a significant role in physically transporting nutrients through the soil fabric via translocation within mycelia, thus short-circuiting other diffusion and mass-flow transport mechanisms, both in saprotrophic and mutualistic (mycorrhizal) contexts.

Soil Architecture

Eucarpic fungi influence soil structural dynamics via a number of mechanisms and on different spatial scales. Hyphal networks enmesh soil particles together, rather like the bars in reinforced concrete. Sparse mycelia in nutrient-poor mineral soils are unlikely to be particularly influential in this respect, but the dense mycelia that can form where substrate is abundant, for example in forest litter layers, can be quite resistant to physical disruption. These mechanisms operate on scales of millimeters to centimeters. The variety of

compounds exuded by hyphae as they grow bind soil particles together via adhesive or charge-based mechanisms, on scales from nanometers (charge) to millimeters (adhesive). Such binding results in increased aggregation of soil and an associated structural generation. Degradative enzymes prevalent near hyphal tips may also decompose organic matter binding soil particles together, resulting in structural degeneration. Fungi produce large quantities of a diverse array of hydrophobic proteins (hydrophobins), which serve to insulate hyphal walls. This is important due to the large surface area-to-volume ratio that arises from the mycelial growth-form. However, these compounds also coat soil particles and can strongly influence the water sorptivity and repellency characteristics of the soil. This in turn affects the hydraulic properties on larger scales. It has recently been suggested that glomalin, a glycoprotein produced in abundance by arbuscular mycorrhizal (AM) fungi and which is resistant to decomposition, may play a key role in effecting mechanisms of adhesion and soil stabilization. As hyphae extend through soil, they may also cause physical restructuring of soil, by rearranging particles, and orientate clay lamellae parallel to their walls (**Figure 3**).

Whilst fungi certainly affect soil structural dynamics, the physical architecture of the soil also affects fungal growth in that it provides the physical framework through which fungi must grow in order to locate resources. The eucarpic growth-form is particularly effective for growth in such spatially structured environments, where complex pore networks prevail and water and substrate are distributed in a heterogeneous manner within the three-dimensional matrix. Hyphae are able to grow across the surfaces of pore walls but also bridge wider pores; since mycelia form an interconnected network, materials can be transported within them from zone to zone, potentially supplementing growth factors where they are in demand in particular regions. Translocation within most soil fungi is as yet poorly understood. There is evidence that different materials can be simultaneously transported in opposite directions within mycelia. Most substrate in soils is distributed patchily, and path lengths between resource depots can be extended greatly by the tortuosity of pore networks. Many species show foraging strategies which balance explorative and exploitative growth strategies relative to the distribution of resources encountered in their normal habitats. Such strategies are highly developed in the cord-forming Basidiomycetes. Physical exclusion mechanisms, where pore diameters are too small to permit penetration by hyphae, probably operate; the extent to which hyphae can physically separate soil particles to allow penetration, akin to plant roots, is largely unknown. Physical protection of organic matter and organisms (including hyphae) from degradation or predation via such spatial exclusion also occurs. Hyphae are generally poor at growing through water-filled pores and hence are less prevalent in waterlogged soils. There are thus strong interactions between soil structure, water, and mycelial growth.

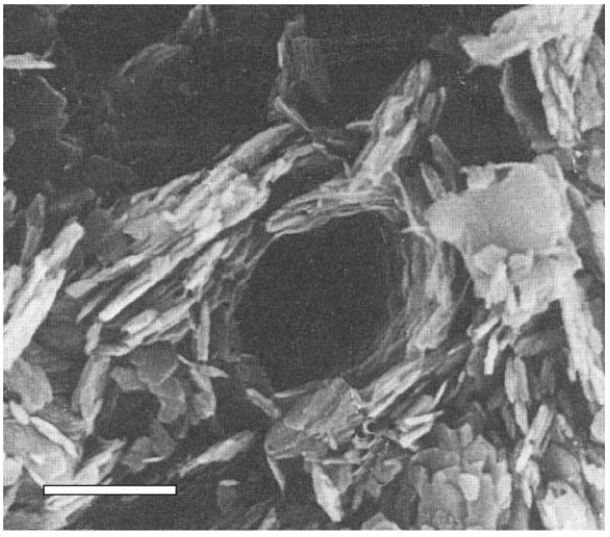


Figure 3 Orientation of kaolinite clay platelets by hyphae of *Chetomium* sp. Scale bar 2 μm. (Reproduced with permission from Dorioz JM *et al.* (1993) The role of roots, fungi and bacteria on clay particle organization. An experimental approach. *Geoderma* 56: 179–194.)

Methods of Study

A comprehensive range of methods has been developed and applied to study the presence, abundance, activity, and distribution of fungi in soils. 'Quantification' of soil fungi presents a particular challenge, since there is a diversity of functionally important components, ranging from total hyphal length through mass of cytoplasm and number of propagules. Methods must be carefully matched to the nature of the study being undertaken.

Isolation Methods

A wide diversity of fungi can be isolated from soils by traditional methods of dilution plating and enrichment growth on broad-scale or specific gel-based media supplemented with a variety of nutrient sources, ranging from oligotrophic (based on silica gels, which are devoid of organic compounds) through to C-rich media such as Czapek-Dox medium (an agar gel based on 3% sucrose). Since relatively low dilutions of soil are often necessary compared with those appropriate for bacteria (typically 10^{-3}

in contrast to 10^{-5}), media may be supplemented with antibacterial compounds to inhibit contaminating overgrowth. Many soil prokaryotes are apparently not cultivable *in vitro*, but the extent to which the 'noncultivability syndrome' may apply to soil fungi is as yet unclear. Whilst such plating methods may be appropriate to determine fungal richness (i.e., a basic range of cultivable species present), meaningful interpretation of 'colony-forming units' (CFUs) by such means is often dubious. This is because colonies can form from hyphal fragments as well as spores, and the number of CFUs will therefore be strongly affected by the propensity for, and degree of fragmentation of, mycelia during sample preparation – it is difficult to relate this to the probable status of any mycelia in the soil. In addition, the expression of colonies on media can be dependent on the presence or absence of other species within the same plate. Hyphal fragments or (larger) spores can be directly extracted from soils by micromanipulation, on density gradients or by elutriation and sieving. 'Baiting' methods are based on the insertion of sterile substrates such as gel-filled perforated tubes, cotton strips, or hair into undisturbed soil, followed by removal and isolation of fungi colonizing these substrata. If an air gap surrounds the baits, it is assumed that any fungi entering them were in an active hyphal state in the soil.

Direct Observation

Measurement of total hyphal length can be made by direct observation of thin films of dispersed soil using a microscope. It is normally necessary to stain hyphae using colored (e.g., aniline blue, acid fuchsin) or fluorescent (e.g., Mg ANS, an 8-anilino-1-naphthalene sulfonic acid salt; Fluorescent Brighteners – FB24, also known as Calcofluor W M2R) stains to render them visible against the complex background of soil particles. Metabolically active hyphae can be distinguished from inactive forms by use of vital stains such as fluorescein diacetate (FDA) or tetrazolium dyes. By measuring hyphal diameters as well, estimates of biovolume and thence biomass can be made using conversion factors. Soil films can be produced by smearing small volumes of suspension on to glass slides or pulling suspensions down on to membrane filters. Soil thin-sections, produced by impregnation of undisturbed soil with hard-curing resin, and cutting and polishing the resultant blocks with lapidary (jeweller's) pastes, enable the *in situ* observation of fungal features (Figure 4). Computerized image processing and analysis can be used to automate identification and measurement procedures, but such methods rely on a high degree of contrast between hyphae and background being obtained. It is rarely possible to identify fungi by these methods, since most hyphae do not have a sufficiently distinct morphology. Infection of roots by AM fungi is normally assessed by direct microscopic observation of roots cleared of cytoplasm by treatment with alkali, and stained using dyes such as trypan blue or chlorozol black E.

Molecular Methods

Complementary to cultivation and direct-observation techniques are so-called molecular analyses, which focus on biochemicals that are more or less specific to fungi. These methods circumvent many of the problems associated with *in vitro* cultivation of fungi and offer a wide range of scales of resolution in quantitative analysis. Total fungal biomass in soils can be estimated by direct extraction and measurement of biomarker compounds such as chitin or ergosterol. The effectiveness of this approach relies on such compounds being specific to fungi and present in a constant proportion relative to the amount of biomass. Few compounds meet these criteria strictly. For example, chitin is also present in arthropods and ergosterol concentrations can vary significantly depending on the age, nutritional status, and species of fungus. More potential is shown by phospholipid fatty acids (PLFAs); these membrane-bound compounds vary in composition with reasonable consistency between different organisms, can be readily extracted from soils and have wide-ranging PLFA profiles determined by gas chromatography and mass spectrometry. Indicator types associated with fungi include 18:2ω6, as a general indicator, and 16:1ω5, which is apparently specifically associated with AM fungi. Immunological techniques based on polyclonal or monoclonal antibodies, particularly via the enzyme-linked immunosorbent assay (ELISA) can be used to detect the presence of specific fungi in soil samples. Analysis of nucleic acids directly extracted from soils is revolutionizing soil ecology, including mycology. With burgeoning knowledge of the specific nucleic acid sequences associated with particular fungi, especially in relation to ribosomal DNA (rDNA; for fungi typically 28S, 18S, and 5.8S, and associated internally transcribed spacer, ITS, regions), it is possible to prescribe probes that detect fungal DNA with virtually any degree of specificity, from the broadest of scales to strain-specific. Such probes can be labeled with fluorescent, radioactive, magnetic, or enzyme tags which can enable their direct detection in soil-based systems. Application of the polymerase chain reaction (PCR) can enable the detection of extremely low concentrations of specific DNA

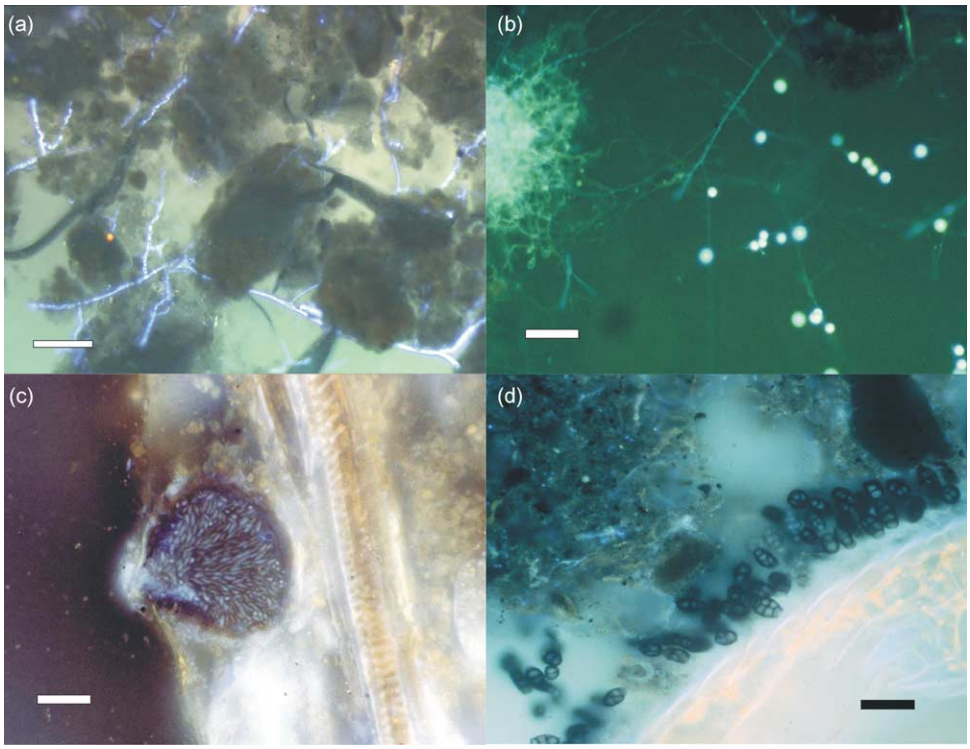


Figure 4 Fungal features visualized in thin sections of undisturbed soil. (a) Hyphae of *Rhizoctonia solani* predominating in soil pores. Scale bar 100 μm; (b) unidentified mycelium in grassland soil showing extensive branching in soil pore, and associated sporophores. Scale bar 50 μm; (c) perithecium of unidentified fungus embedded in cortex of root in thin-section of arable soil (note orientation of orifice through which spores are released into soil pore). Scale bar 50 μm; (d) accumulation of fungal spores in vicinity of decomposing root (located in lower-right quadrant of image). Scale bar 50 μm. (Reproduced with permission from Karl Ritz, David Crabb, Kirsty Harris, Scottish Crop Research Institute.)

sequences in complex environmental samples and can also be used in a broad-scale or specific sense. Many techniques exist to analyze PCR products, with cloning and sequencing providing the greatest resolution in that it literally provides the individual base-sequences of the products. It is considered that fungal DNA may be more resistant to extraction from soils than prokaryotic DNA, although there is little firm evidence for this.

Activity

Specific assessment of fungal activity in soils is difficult, since it will invariably occur in combination with the activity of the rest of the biomass. One approach is to measure bulk respiration in the presence and absence of metabolic inhibitors specific to prokaryotes such as streptomycin; assuming the inhibitors are actually specific and fully effective, the difference in respiration is attributable to fungi. More effective procedures are based on measuring the incorporation of stable or radioactive isotopes into fungal marker compounds, such as the incorporation of ^{14}C-labeled acetate into ergosterol. Since RNA is produced to a greater extent in metabolically active cells and hyphae, analysis of directly extracted RNA, in a similar manner to that used for DNA (see Molecular Methods section) offers great potential. However, RNA is difficult to extract from soils and mycological applications to date have been relatively limited.

Exploitation of Soil Fungi

Whilst soil fungi are unquestionably vital to the functioning of terrestrial ecosystems, they also play an important and increasing role in biotechnology and environmental management. Culinary exploitation and cultivation has occurred for millennia, with mushroom-growing industries present in most continents. The Périgord truffle (*Tuber melanosporum*) remains resistant to routine, large-scale cultivation and is consequently one of the most expensive epicurean foods on a weight basis, retailing at potentially thousands of dollars per kilogram. On a contrasting scale, the production of single-cell protein in large-scale fermentors from *Fusarium graminearum* is a successful and established industry. Many fungi isolated from soil provide a range of important primary and secondary metabolites which are

exploited industrially. These include various organic acids (e.g., *Aspergillus*, *Rhizopus* spp.), polysaccharides (e.g., *Aureobasidium* spp.), and lipids (e.g., *Fusarium* spp.), as well as antibiotics (e.g., *Penicillium*, *Aspergillus*, *Sordaria* spp.), other pharmaceuticals (e.g., *Penicillium*, *Tolypocladium* spp.), and agrochemicals (e.g., *Strobilurus* spp.). Several thousand fungally derived antibiotics have been identified, but relatively few have found clinical use, since most are too toxic to humans. Fungi that are antagonistic to pathogens (see above) have actual and potential use as biocontrol agents against other fungi, nematodes, and weeds. Establishment of plants in horticultural or restoration scenarios can be compromised by the absence of suitable mycorrhizal hosts; inoculation with appropriate fungal partners can be effective and economically viable. Use of fungi in bioremediation of soils contaminated by industrial pollutants such as phenolics, hydrocarbons, and heavy metals has potential – claims are often made for Basidiomycete species that show strong phenol oxidase activity, such as *Phanerochaete chrysosporium* – and, in combination with plants (phytoremediation), metal-solubilizing fungi may offer effective soil cleanup mechanisms. Increasing use of soil fungi is likely given further research and understanding of the biology of these highly diverse and versatile organisms.

See also: **Bacteria:** Soil; **Biodiversity**; **Fauna**; **Microbial Processes:** Environmental Factors; **Rhizosphere**

Further Reading

Cairney JWG and Burke RM (1996) Physiological heterogeneity within fungal mycelia: an important concept for a functional understanding of the ectomycorrhizal symbiosis. *New Phytologist* 134: 685–695.

Carlile MJ, Watkinson SC, and Gooday GW (2001) *The Fungi*. London, UK: Academic Press.

Davet P and Rouxel F (2000) *Detection and Isolation of Soil Fungi*. Translated by Thanikaimoni K. Enfield, NH: Science Publishers.

Dighton J (2003) *Fungi in Ecosystem Processes*. New York, NY: Marcel Dekker.

Domsch KH, Anderson T-H, and Gams W (1981) *Compendium of Soil Fungi*. London, UK: Academic Press.

Gadd GM (2001) *Fungi in Bioremediation*. Cambridge, UK: Cambridge University Press.

Garrett SD (1981) *Soil Fungi and Soil Fertility*. Oxford, UK: Pergamon Press.

Griffin DM (1972) *Ecology of Soil Fungi*. London, UK: Chapman and Hall.

Lemke PA, Esser K *et al*. (eds) (1994 *et seq*.) *The Mycota: A Comprehensive Treatise on Fungi as Experimental Systems for Basic and Applied Research*. Berlin, Germany: Springer-Verlag.

Rayner ADM (1991) The challenge of the individualistic mycelium. *Mycologia* 83: 48–71.

Smith SE and Read DJ (1997) *Mycorrhizal Symbiosis*. London, UK: Academic Press.

Wainwright M (1988) Metabolic diversity of fungi in relation to growth and nutrient cycling in soil – a review. *Transactions of the British Mycological Society* 90: 159–170.

G

GEOGRAPHICAL INFORMATION SYSTEMS

J Böhner, Universität Göttingen, Göttingen, Germany
T Selige, Technische Universität München, Munich, Germany
R Köthe, Scilands, Gesellschaft zur Bearbeitung digitaler Landschaften, Göttingen, Germany

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

The term 'geographic information system' (GIS) has become synonymous with a wide range of computer applications, technologies, and scientific methods related to the use of geospatial information. Although the origin of GIS can be traced back to early electronic mapping tools, such as the Canada Geographic Information System (CGIS) developed in the mid-1960s, limiting the definition of GIS to a spatial visualization tool would be to disregard its current scientific and technological standards, as well as the huge expansion of GIS applications in science, politics, and economics. Driven by the rapid growth of computer technologies on the one hand and globally increasing demands on spatial information on the other, GIS has evolved from an electronic mapping facility into a fast-growing spatial science technology in less than 40 years. Due to the amalgamation of computer-assisted cartography with database technologies and growing capabilities of analyzing data across different layers in an object-oriented programming environment, current GIS provides a universal management technology for capturing, analyzing, modeling, and displaying spatial data.

Apart from its technical dimension, the term GIS also describes an emerging methodical discipline which has become an established part of the curriculum, especially in geography. However, the close connection between geography and GIS is largely based on a chance semantic coincidence and less the result of methodical commitments to computer-based spatial analyses that took place during the so-called quantitative revolution of geography in the early 1970s. The evolution of GIS was fostered by the special applications, ideas, and GIS-based solutions of scientists, federal and national agencies, governmental departments, and commercial providers, as well as the huge user community. A multitude of alternative terms for GIS bear witness to the rapid spread of GIS technology, ranging from 'market-analysis information system' to 'image-based information system'. Since the early 1990s, these have included the term 'soil information system.'

Indeed, the notion of quantitative pedology, i.e., the modeling of soil formation as a function of state factors, initially seemed perfectly suited to the new GIS technology, especially since key functional features enabled overlay and modeling operations across apparently disparate spatial data sets. Although these optimistic assessments of 'intelligent technologies' soon had to be revised and, particularly in soil sciences, were followed by some disillusionment, mainly on account of the sheer complexity of soil, three GIS applications to soils have now become established. These take GIS back to its original purpose as a useful instrument in: (1) data inventory and management, (2) data analysis and mapping, and (3) modeling and decision support. Before discussing these in more detail, the following will summarize major GIS applications to soils to provide an overview of elementary features, basic principles, and methods of GIS.

Elementary Features and Basic Principles of GIS Applications to Soils

Basic principles of soil-related GIS applications are best shown using a three-step scheme of GIS evolution. Based on the long-standing development of the Canadian GIS family, with its strong commitment to environmental issues, the three-stage scheme is also appropriate for showing how soils are integrated into a GIS context within the above fields of application. A structured scheme, organized along a hierarchical sequence of activities and GIS applications, is given in Figure 1. Here, the assumption is of a research process ultimately targeted at supporting management decisions.

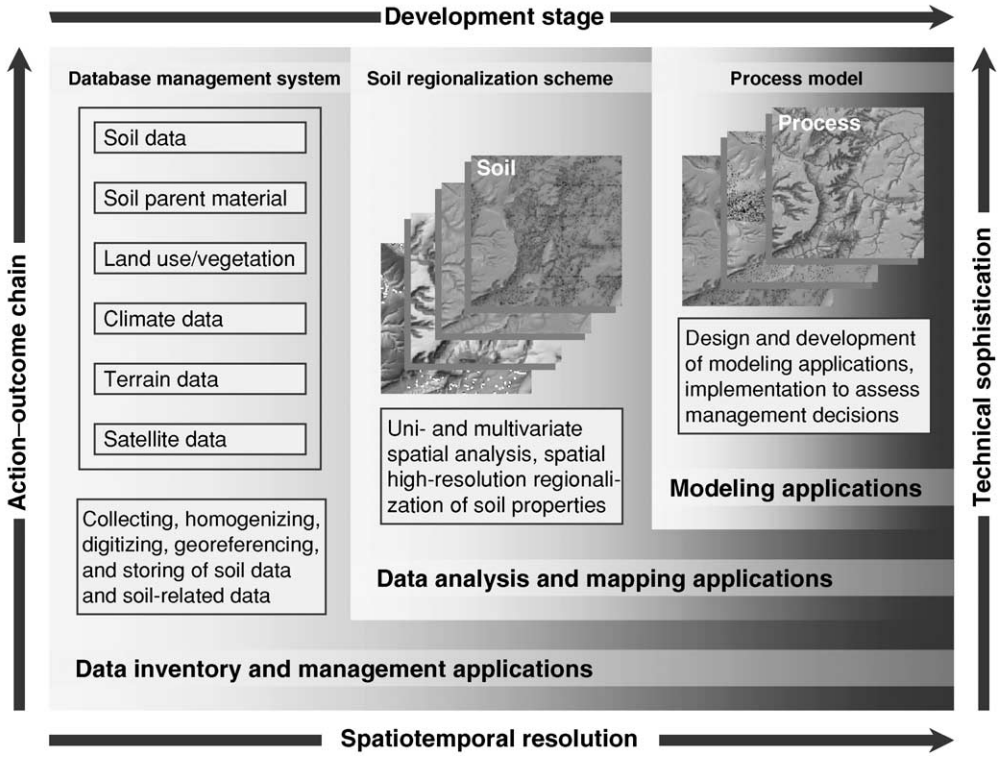


Figure 1 Conceptual model of the role of geographic information systems (GIS) and GIS applications to soils.

Taking the input data level as the essential and most crucial resource of a GIS, the initial phase of GIS evolution is characterized by data inventory and management applications. Assembling an appropriate database involves collecting, homogenizing, digitizing, georeferencing, and structured storing of data. In purely technical terms, the work consists of the design and implementation of a database management system suitable for fast data queries or searching. Although distinctions between soil and other intimately related environmental layers such as vegetation or climate are technically difficult, proper digital representation of soil and its three-dimensional variability above all requires different statistical data scales, ranging from nonquantitative nominal (e.g., soilscape) to metric entities (e.g., the organic matter content of a certain horizon). Since soils have in the past been entered into GIS as part of ambitious national or federal mapping programs that based their inventories on available soil maps, analogous soil maps have so far constituted a major data source. This leads to a distinct dominance of vector data representations, commonly consisting of discrete soil entities such as genetic soil types and local profile descriptions. Regularly spaced, continuous raster data representations of soil properties in contrast are still less common. Despite much progress in GIS-based data assembly and data management, digital maps and attribute tables therefore remain simple metaphors of the traditional soil map. The latter already represents a condensed and highly abstracted result of a knowledge-based research process, drawing on all the skills of soil scientists and cartographers at their current state of the art. Since all subsequent tasks and operations remain restricted by these academically essential but hardly operational issues, proper soil data representation demands at least some additional assembly of metadata information (e.g., site-specific descriptions of profile settings, scale issues, map conventions) if a solid scientific evaluation of all subsequent research results is to be enabled.

While inventory and management applications reflect the more technical dimension of GIS, the second stage of GIS evolution emphasizes data analysis and mapping applications. Since the term 'spatial analysis' includes a wide range of different methods that greatly vary in sophistication and complexity, a differentiation according to uni- and multivariate analysis to complement corresponding statistical methods is suggested. In this case, univariate spatial analyses would comprise all applications that analyze spatial (neighborhood) dependencies within one layer, e.g., autocorrelation, routing, or geostatistical analyses. Examples for soil-based applications include neighborhood analyses for delineating soil

associations, or the geostatistical analyses frequently used to estimate a continuous metric soil layer (e.g., horizon depth) from random field data. The multivariate option could thus denote all kinds of analyses performed across different layers, leading to the production of a new spatial data set. This latter key functionality of a GIS is often utilized for soil regionalization purposes and will be discussed below in the Data analyses and mapping applications section. Although most current hybrid or extended GIS are capable of performing both kinds of spatial analyses using both vector and raster data, functionality and quality of analysis results still depend on which data model is used and reflect the former preferences and roots of software packages. Particularly in soil science, where spatial analyses and mapping are based on two differing views of soil, either regarding it as a spatially discrete genetic entity or a composition of spatially continuous layers, a GIS software package still needs to be chosen in accordance with the type of spatial analyses to be carried out.

In the third and most developed phase of GIS, modeling and decision-support applications are emphasized. This results from the user's desire to understand spatial patterns and processes or to undertake more complex modeling operations, required for instance to assess objectively environmental management or political decisions. Due to increasing computing capacities and advanced object-oriented programming environments, current GIS provides sufficient modeling capabilities to render this most prominent application of GIS. This in turn substantiates the use and view of GIS as a decision-support system. While soil still forms the actual research object or predictant variable in GIS-based analysis and mapping, the modeling stage integrates soil and its closely related environmental layers as predictor variables. Their aim is to model more complex climatic processes or to run scenarios of land use change in an attempt to predict the effects of management decisions.

Throughout these stages of GIS development, soils evolve from a rather static object of inventory to an essential, highly dynamic, reactive and controlling subject. In view of the spatial resolution required by certain applications, large-scale inventories satisfactory for purposes of national or federal information are replaced by a high-resolution regionalization of soil properties capable of supporting complex modeling applications. The importance of GIS also changes; it is best described as a transition process, beginning with its use as a transaction-processing system and technical tool for data management. GIS then goes on to become an array of scientific analytical methods capable of supporting the basic demands of spatial information science, with the final stage as a decision-support system, with potential uses within a range of applications.

Data Inventory and Management Applications

In view of society's growing demands for environmental services, a huge number of international, national, and federal inventory and mapping programs have been set up from the early pioneering stage of GIS in the 1960s. Initially conceived against the urgent need for increased agricultural production, more recent aims include the maintenance and qualitative enhancement of natural and rural (agro-)environments. Due to the dual role of soil as a production factor and an interacting and controlling layer within the atmosphere–biosphere–geosphere system, the assemblage of soil data was of particular interest in the context of developing ambitious agroenvironmental instruments capable of supporting sustainable development. The momentum of this development is reflected in an increasing number of soil information systems. Although they are mostly embedded in primary land or environment information systems, they distinctly show the technical and scientific specialization required for handling and providing digital soil and soil-related information.

On a global scale, the broad-resolution *Digital Soil Map of the World* (DSMW), a digital derivative of the Food and Agriculture Organization of the United Nations (FAO/UNESCO) soil map of the world, is considered the most detailed, globally consistent soil data set. The original $1:5\,000\,000$ map represents a generalization of more detailed data derived from about 11 000 soil maps reviewed from various countries, which vary widely in reliability, detail, precision, scales, and methodologies. In line with the analogous source, both vector- (scale $1:5\,000\,000$) and raster (spacing 5×5 arc min)-based derivatives include soil types, texture, profile depth, and surface slopes, as well as derived soil properties provided by interpretation programs and related data files on agronomic and environmental parameters. Apart from the maps of classification units contained in the *World Soil Reference Base* (WSRB) and tables on special soil analyses for every country of the world, an additional soil database was created for global environmental studies, including data on soil-moisture storage capacity, soil-drainage classes, and effective soil depth. A corresponding data set of water-holding capacities in a $1^\circ \times 1^\circ$ grid (latitude, longitude), drawn from original sources to meet the requirements of general circulation model (GCM) experiments, can be quoted as another example of

using soil data in the context of global environmental studies.

Despite its broad spatial resolution and distinctly limited information content, features that are unlikely to be improved by the current SOTER revision (soil and terrain database of the world), the compilation of the global soil map and its digital derivative did contribute much to academic debates on the global harmonization of soil classification concepts and may also have influenced the development of many continental, national, and federal soil information systems. Prominent examples of well-established projects, such as the National Soil Information System in the USA (NASIS), the Canadian Soil Information System (CanSIS), the Australian Soil Resource Information System (ASRIS), or the European Soil Information System (EUSIS), form a suitable basis for a steady, coordinated, and harmonized inventory of soil and soil-related data in accordance with essential environmental tasks. Indeed, precise manuals of procedure setting out requirements, inherent concepts, and methods essential for a rationalized inventory of georeferenced, digital soil databases have given rise to an increasing availability of plausible and consistent soil data accessible for many purposes.

Using the ambitious EUSIS project as an example, Figure 2 shows the construction of a soil information system database. The relational structure consists of topological, semantic, and geometric data sets, organized along a hierarchical chain of discrete soil entities (objects). The topological tables contain information on the distribution of soil bodies within soilscapes and soil horizons within soil bodies (organization tables), as well as information on the nature of limits, separating different soil-scapes within a soil region (limits table). The properties of soils, mostly attached to soil bodies and horizons, are given in the semantic attribute data sets, whereas positions and shapes are captured by the geometric-attribute data sets. To address the needs of soil information users at different scales, EUSIS is organized along a scale-dependent structure. Spatial soil information is assembled and provided at global 1 : 5 000 000 scales compatible with the SOTER database of the FAO, at intermediate 1 : 1 000 000 scales covered by the *Soil Geographical Database of Europe* (SGDBE) and at 1 : 250 000 scales. Larger scales (1 : 50 000, 1 : 5 000) are covered by regional to local authorities. To ensure homogenous data contribution and assembly, EUSIS partly links up with existing national and federal soil information systems (e.g., Lower Saxony Soil Information System, NIBIS, Germany). Special working groups in the European Soil Bureau (1 : 1 000 000, European Soil Database working group; 1 : 250 000, working group) coordinate data contribution and management in accordance with the relevant scales.

Despite clear differences between individual soil information systems, particularly in terms of subdividing soil categories and the match between soil taxonomy and soil map units, most systems reflect the EUSIS structure and are conceptualized along traditional map scale conventions. Since the genetic soil category or soil type, although somewhat modified, is still employed as the major response unit, the common practice of direct, rule-based delineation of pedotransfer functions from these discrete entities yields no more than a coarse estimate, with clear limitations in terms of spatial distribution, resolution,

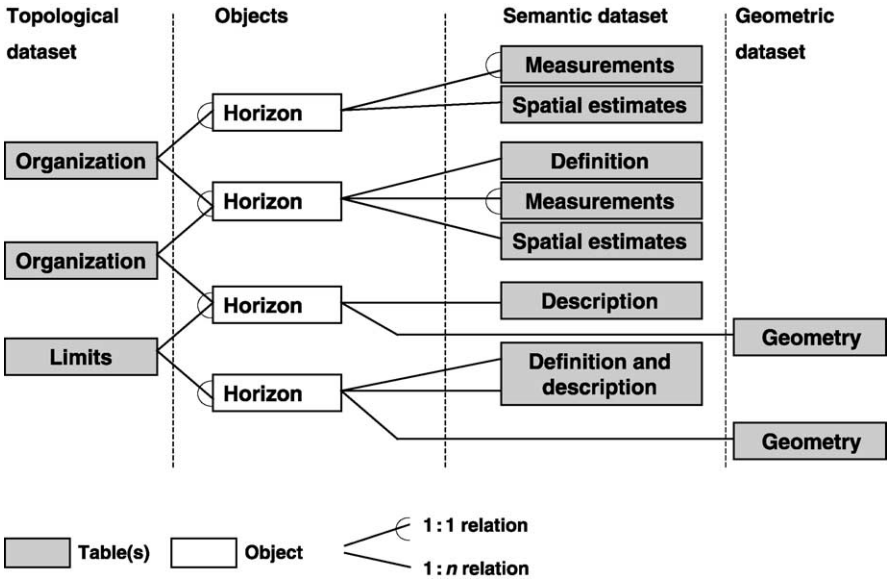


Figure 2 General structure of the European Soil Information System (EUSIS) database.

and reliability. Though soil information systems technology stresses its developments by catchwords such as 'multiuser access' or 'just-in-time performance,' its continued focus on discrete mapping units may well satisfy classification and pedogenetic criteria but does not necessarily meet the needs of prospective uses. Despite the global increase in reliable soil data, the current global soil inventory is not in a position to meet the increasing demands posed by soil protection, sustainable soil management, or other major political issues. These ambitious aims demand a spatially flexible database of different soil properties, which may be obtained through further GIS analyses.

Data Analyses and Mapping Applications

GIS-based data analyses and mapping operations have been widely developed for the purpose of yielding high spatial-resolution estimates of different soil properties. Apart from univariate interpolation techniques performed on random field data, for example, multivariate regionalization approaches in quantitative soil pattern analyses are the most frequently used, describing the spatial variability of soils in terms of a set of determining predictant variables and thus providing insight into the processes and factors involved. The term 'regionalization' is preferred in this context, since it comprises the somewhat misleading terms of 'upscaling,' 'downscaling,' 'interpolation,' and 'extrapolation' and also covers the major contemporary approaches of soil-mapping support, roughly to be divided into a predominantly vector-based approach and a continuous raster-based approach.

The knowledge-based discrete approach (or concept map approach) is still the most frequently used instrument of support in soil mapping. In accordance with Jenny's factorial theory of pedogenesis, a set of state factors such as, e.g., parent material, terrain, climate, or vegetation is statistically compared with the spatial soil pattern by utilizing overlay and intersection algorithms as key functional operations of GIS. The smallest common geometries of state factors are identified as a surrogate for the site-specific process structure, to enable a rule-based assignment of soil parameters and infer estimations of spatial soil patterns with a certain probability. Since state factors within a landscape system are intimately related, a landscape component can thus be assigned to typical patterns of landform elements, soil types, or vegetation sequences at each level of the spatial hierarchy. Adequate capture of covariance structures allows data gaps to be filled (e.g., in the ASRIS project performed on the basis of adjacent, properly mapped soil distributions), or infers spatial soil pattern at finer resolutions commensurate with terrain or vegetation information. By integrating sophisticated complex analytical methods such as fuzzy logic or the use of neuronal nets, characterized by their reduced reliance on expert knowledge in computing concept maps, descriptive rule-based models and expert systems can yield many supporting features of an improved and extended spatial definition of soil patterns. However, the discrete approach has distinct limits, particularly resulting from the finite number of predictant entities which may neither represent the spatial soil diversity in the area concerned nor the process scopes relevant to soil. Also, the classification of continuous, process-controlling geofactors (e.g., climate, relief) represents a critical task, since it involves an a priori assumption that the limits inferred are actually relevant for the distribution of soil characteristics and its spatial definition.

In view of these deficiencies, the continuous soil regionalization approach is sometimes preferred, particularly where metric soil properties are required for further use, e.g., in process models. Based on the definition of soil as a composition of continuous layers, a set of continuous terrain parameters has been used as predictor variables in the context of using statistical and geostatistical procedures for inferring spatial soil estimates from metric point data. Results indicate significant correlations between soil and terrain variables, particularly when complex, hydrologically based terrain indices, such as the Terrain Wetness Index (TWI; see **Figure 3**) are analyzed. Other results confirm that much of the variation apparent in soil is a response to near-ground runoff and associated translocation processes. To ensure consistent applicability of a continuous regionalization scheme to other climatic regions, recent applications have extended the continuous approach by integrating spatial high-resolution climate and terrain variables to complex process parameters. **Figure 3** exemplifies the Mass Balance Index (MBI) for the parameterization of slope transfer processes, integrating the universal soil-loss equation (USLE) *R*-factor and different catchment area parameters (catchment area, catchment slope). Since terrain parameters based on nonlinear discharge models tend toward strongly artificial variability, particularly in hydrologic homogenous areas (e.g., large valley grounds), the SAGA (system for an automated geoscientific analysis) Terrain Wetness Index (STWI) applies a slope-dependent, iterative modification scheme to the catchment area size (**Figure 3**) which keeps it suitable for the regionalization of hydromorphic soil properties. Different statistical and geostatistical regionalization methods performed on a set of terrain and process parameters were evaluated with regard to their routine application in the context of soil mapping. Universal kriging

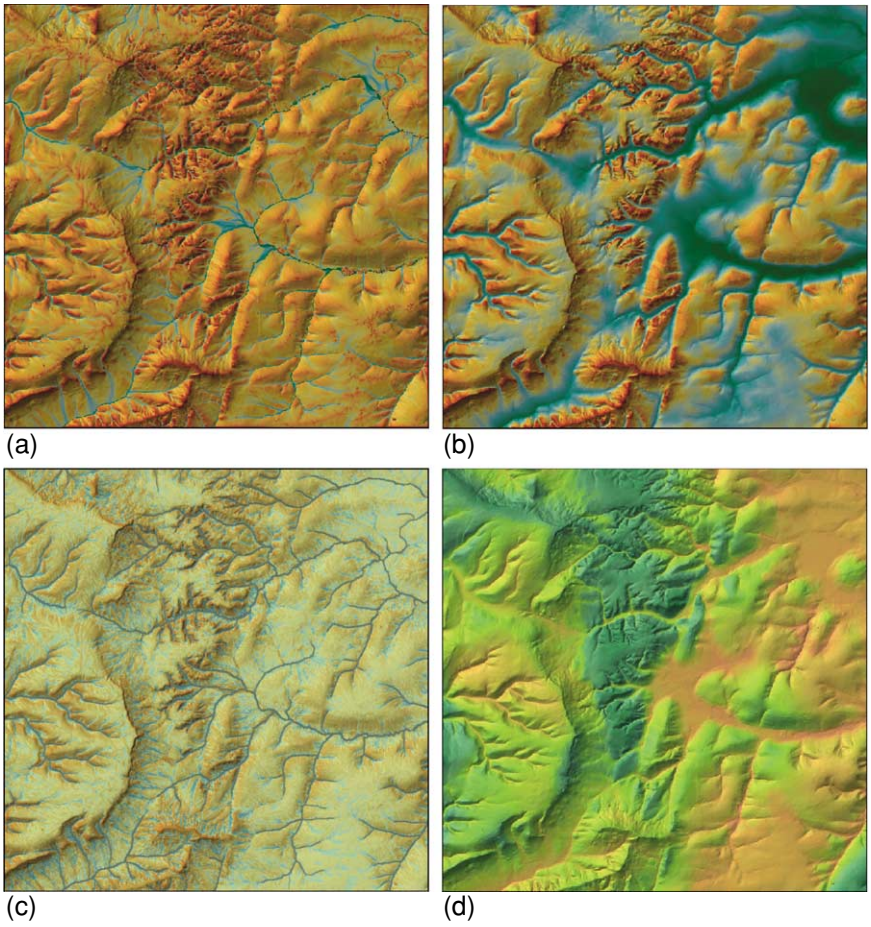


Figure 3 (a) Terrain wetness index. (According to Moore ID, Gessler PE, and Peterson GA (1993) Soil attribute prediction using terrain analysis. *Soil Science Society of America Journal* 57: 443–452); (b) SAGA (system for an automated geoscientific analysis) terrain wetness index; (c) mass balance index; and (d) silt content of the soil surface layer (red, high silt content; green, low silt content). (Adapted from Böhner J, Köthe R, Conrad O *et al.* (2003) Soil regionalization by means of terrain analysis and process parameterization. In: Micheli E, Nachtergaele F, and Montanarella L (eds) *Contributions to the International Symposium Soil Classification 2001,* Ispra, Italy. EUR 20398 EN. European Soil Bureau, Joint Research Centre.)

was assessed as a valuable tool for soil mapping and further modeling support. Results are exemplified for the regionalization of the topsoil silt content obtained by universal kriging (Figure 3). The strong emphasis on process throughout the regionalization scheme reflects the basic assumption that no matter which strategy is used for soil regionalization, a reliable spatial prognosis of soil properties can only be achieved if the underlying spatial data directly represent soil-related processes rather than being substituted by discrete factor combinations. Distinct limits in applicability, often the main focus of criticism, are due to the comparably high demands on soil measurements or profile data. In view of the growing availability of very-high-resolution remote sensing and particularly radar data, capable, for example of picturing topsoil moisture distribution and thus topsoil subscale (random) variability, this difficulty is likely to be overcome in the medium term. Nevertheless, a proper point database remains an essential basis for continuous soil regionalization.

Modeling and Decision-Support applications

Since the above soil regionalization approaches empirically detect predictable set structures within spatial soil variations and are thus occasionally termed 'models' themselves, this term is really reserved for the type of process modeling applications described below. The integration of soil as a reacting and controlling body into process models expands the capabilities of quantitative, GIS-based simulation, with ensuing benefits to a range of applications, particularly in the context of management decision support. Apart from hydrologic models, with their specific demands on horizontal and vertical soil data distribution, a wide range of applications

exist in the context of soil erosion risk assessments. They can roughly be divided into empirical and physical approaches, with both requiring pedophysical values or at least a parameterization of certain soil properties.

Possibly the most prominent approach for estimating water-induced soil erosion, the USLE, is now predominantly carried out on a GIS basis on account of its comparatively simple structure. The USLE consists of an empirical equation which calculates the mean annual soil-erosion rates on agricultural slopes as a function of six variables (R, precipitation factor; L, slope length factor; S, slope factor; C, vegetation cover and tillage factor; K, soil-erodibility factor; P, erosion-protection factor). The parameterization of the topsoil erodibility in the K-factor considers soil particle size distributions (sand and silt percentages), organic matter content, and classes of permeability of the soil surface layer, e.g., entailed or delineated from soil taxonomy profiles. Comparable demands on the parameterization of topsoil properties are also valid for the empirically based wind-erosion equation (WEQ), predicting the potential mean annual soil loss at the field scale by determining the influence of several primary variables characterizing soil erodibility, wind erosivity, and the actual land-use situation.

Due to the limited transferability of empirically based erosion risk assessments, more sophisticated, process-based approaches have recently been developed capable of simulating water and wind erosion in a physically consistent manner. Their major advantage of wide applicability is countered by the disadvantage of extended data requirements on the spatiotemporal resolution and the quality of input parameters. This is particularly valid for soil data, which

Figure 4 Spatial distribution of the erosion–accumulation balance for the event of 13 and 14 March 1994 at the Barnham test site, UK. Positive values (green) indicate accumulation; negative values (yellow and red) indicate erosion (dark green, forests; gray, settlements and Honington airport). Average wind direction of the storm was west (240–300° direction angle). Model domain covers 5 × 5 km. Reproduced from Bohner *et al.* (2003) The Weels model: methods, results and limitations. *Catena* 52: 289–308, with permission from Elsevier.

have to be available either as physical soil attributes or as delineated parameterizations. With respect to its prospective use in case studies, management decision support, and particularly in context with precision farming applications, major requirements concern the spatial resolution and thus demand a suitable procedure for the regionalization of metric topsoil properties.

Emphasizing this, Figures 4 and 5 exemplify the results of physically based wind-erosion simulations using the wind erosion on European light soils (WEELS) model, performed on spatial high-resolution estimations of soil particle distribution and surface-layer organic matter content. Apart from the soil-erodibility factor, soil attributes are essential inputs for the temporal high-resolution simulation of topsoil moisture contents, since they essentially determine the process of wind erosion. The WEELS model also considers various climatic parameters, annual changes in land use, crop cycles, and crop phenology, as well as the influence of the terrain. Due to its high temporal resolution (hourly) the WEELS model allows the simulation of wind-erosion events (Figure 4 shows an event at the Barnham test site, UK) as well as long-term estimates of erosion and accumulation rates. The spatial distribution of erosion and accumulation rates for the Barnham test site (Figure 5) was derived from a 29-year model run (1970–98). The option of long-term simulations enables the assessment of different land-management scenarios and their susceptibility to wind-erosion processes. The result of land-use change scenarios for a test area in Grönheim in the Cloppenburg Geest area (the western part of Lower Saxony) has shown the possibility of minimizing erosion risks through sustainable land-use strategies. This can underline the capabilities of GIS-based modeling and decision support in the context of environmental policy and management.

Figure 5 Modeled spatial distribution of the mean annual erosion–accumulation balance at the Barnham test site, UK (1970–98). Due to changing wind directions and annually varying vegetation cover, the orientation of the soil erosion rate gradient may differ among the individual fields (dark green, forests; gray, settlements and Honington airport). Model domain covers 5 × 5 km. Reproduced from Bohner *et al.* (2003) The Weels model: methods, results and limitations. *Catena* 52: 289–308, with permission from Elsevier.

Summary

The evolution of GIS in terms of technical maturity and functionality offers many possibilities for application to soils, ranging from data assembly and inventory support to complex modeling. The proliferation of GIS within the scientific community has fostered its frequent use in spatial data analyses, particularly in the context of soil-mapping support and soil regionalization. Due to their increasing ability to analyze data across different layers, GIS constitute suitable instruments for developing soil-related process models with wide potential applications for instance in decision-support systems. However, the clear technical advantages cannot detract from the fact that the most crucial resource of a GIS, the database itself, still tends to be a digital derivative of a traditional soil map. It therefore comprises individual scientific field experience as well as mapping conventions, neither of which can be easily reproduced in a digital format. The persistance of the genetically well-founded, discrete soil entity as the basic information unit and particularly the direct delineation of (coarse estimated) pedotransfer functions is hardly capable of meeting the increasing demands for soil information in terms of spatial resolution and reliability, and thus to a certain extent even violates the capabilities of current GIS. A major alternative, the spatial high-resolution regionalization of available soil profile data – an alternative which would deliver more-reliable soil estimates despite its being transcribed from an older classification system – above all requires an appropriate soil profile database. Though soil profile data are often compiled as part of developing national or international soil information systems, a systematic compilation and homogenization of available profile data on a national or international level is currently unavailable. This is therefore suggested as one of the most important future tasks in the further development of national and international soil information systems. Clearly this is no easy feat, but a rewarding task in view of the expected increase in knowledge on soil-formation processes. Rather than GIS users, it will, however, require GIS-using soil scientists.

See also: **Environmental Monitoring**; **Erosion:** Water-Induced; Wind-Induced; **Remote Sensing:** Organic Matter; **Spatial Variation, Soil Properties**; **Statistics in Soil Science**

Further Reading

Böhner J, Köthe R, Conrad O *et al.* (2003) Soil regionalisation by means of terrain analysis and process parameterisation. In: Micheli E, Nachtergaele F, and Montanarella L (eds) *Soil Classification 2001* EUR20398 EN. Luxembourg: Office for Official Publications of the EU.

Böhner J, Schäfer W, Conrad O, Gross J, and Ringeler A (2003) The WEELS model: methods, results and limitations. *Catena* 52: 289–308.

Bullock P, Jones RJA, and Montanarella L (eds) (1999) *Soil Resources of Europe*. EUR18991EN. Luxembourg: Office for Official Publications of the EU.

Crain IK and MacDonald CL (1984) From land inventory to land management. *Cartographica* 21: 40–46.

FAO–UN (1970–78). *Soil Map of the World*, scale 1 : 5 000 000, vols I–X. Paris, France: UN Educational, Scientific, and Cultural Organization.

Finke P, Hartwich R, Dudal R *et al.* (1998). *Georeferenced Soil Database for Europe – Manual of Procedures*, version 1.0. EUR18032EN. Luxembourg: Office for Official Publications of the EU.

Gessler PE, Moore ID, McKenzie NH, and Ryan PJ (1995) Soil landscape modelling and spatial prediction of soil attributes. *International Journal of Geographical Information Systems* 9: 421–432.

Longly PA, Goodchild MF, Maguire DJ, and Rhind DW (1999) *Geographical Information Systems*, vol. 1, *Principles and Technical Issues*. New York: John Wiley.

Maguire DJ, Goodchild MF, and Rhind DW (1991) *Geographical Information Systems – Principles and Applications*, vol. 1, *Principles*, pp. 9–20. Harlow, Essex: Longman.

Moore ID, Gessler PE, Nielsen GA, and Peterson GA (1993) Soil attribute prediction using terrain analysis. *Soil Science Society of America Journal* 57: 443–452.

Morgan RPC, Quinton JN, Smith RE *et al.* (1998) The European soil erosion model (EUROSEM): a process-based approach for predicting sediment transport from fields and small catchments. In: *Earth Surface Processes and Landforms*, vol. 23, pp. 527–544. New York: Kluwer.

Schmidt J, von Werner M, and Michael A (1996) *Erosion 2D/3D – ein Computermodell zur Simulation der Bodenerosion durch Wasser*. Leipzig: Sächsisches Landesamt für Landwirtschaft.

Sinowski W (1995) *Die dreidimensionale Variabilität von Bodeneigenschaften*. Forschungsverbund Agrarökosysteme München (FAM), vol. 7. Aachen: Shaker-Verlag.

Wischmeier WH and Smith DD (1978) *Predicting Rainfall Erosion Losses – A Guide to Conservation Planning*. Agricultural Handbook No. 537. Madison, WI: USDA.

Woodruff NP and Siddoway FH (1965) A wind erosion equation. *Proceedings of the Soil Science Society of America* 29: 602–608.

GERMINATION AND SEEDLING ESTABLISHMENT

A Hadas, The Volcani Center, Bet Dagan, Israel

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Seed germination and seedling establishment are critical phases in the plant growth cycle, since they influence and determine species survival in natural habitats and the onset and yields of agricultural crops. A dormant seed combines an embryonic plant and stored materials to be used during germination and seedling establishment until the onset of photosynthesis by the established seedling. 'Germination' is a general term describing the sequences of complex processes involved in initiating an array of metabolic activities that lead eventually to the renewal of growth of the dormant seed embryo and, ultimately, to seedling establishment. The various processes known to occur during these two phases take place in parallel or in serial sequences. Their times of initiation and of transition from one sequence to another are triggered by events endogenous to the seeds (seed-development processes of the parent plants) and depend greatly on the prevailing environmental conditions (aeration, temperature, available water and nutrients, soil mechanical impedance) and seed-development conditions.

Seed Germination and Environmental Conditions

Proper seed germination and stand establishment depend strongly on the environmental conditions (moisture, thermal and aeration regimes in the soil, light). Where favorable environmental conditions prevail, other factors may decide the success or failure of seed germination and stand establishment (seed-development processes on the parent plants, seed dispersion, and depth of seed burial). Seeds are self-contained units, owing to the storage materials contained in the seeds. Therefore, the environmental requirements for germination and seedling establishment are considered to be fewer and simpler than those required by full-sized plants.

Water Requirements

Water uptake by seeds is a prerequisite for proper germination and seedling establishment. The amount of water required by a seed for germination itself is considered to be very small, but the rate of uptake and the total amounts taken up depend greatly on the seed and soil properties with respect to water – the water potential differences between the seed and the soil – and are controlled by the water conductivity of the soil and the soil–seed contact zone. Whether or not the amount of water taken up will suffice for germination and seedling development depends on the water energy status of the seed, and on the soil-water potential. There is a certain minimal level of seed hydration, termed the 'critical hydration level,' below which the seed will not germinate; instead it may enter secondary dormancy. Seeds that enter secondary dormancy become more susceptible to pests and diseases. These seeds may germinate later in the season or during the coming seasons. Thus the uniform germination, establishment, and stand of crops are impaired. The viability of the seeds that reenter secondary dormancy diminishes with time. Species and cultivars differ markedly in conditions they require for germination, and these differences are attributed to the differences among the water regimes and other soil physical conditions of the soil to which the plants were adapted or which they encountered during germination.

Temperature Requirements

Temperature affects the soil properties, with respect to water, that influence seeds, air and water regimes, and the biological activity of seeds themselves. Soil temperature varies greatly, because of both diurnal and seasonal thermal processes, and is dependent on the constituents of the soil (water, solids, organic materials, and air), its texture, structure, color, and layering. For germination to occur, the temperature of the seed environment should fall within a favorable, species-specific range. Within that interval lies the optimal temperature at which optimal germination is observed. The temperature limits (minimum and maximum temperatures) below or above which no germination will occur are 3–5° and 30–43° for wheat cultivars; and 8–12° and 40–44° for corn cultivars. The corresponding optimal temperatures are 15–30° and 26–32° for wheat and corn cultivars, respectfully. Planting during spring depends on the soil temperature and, as an example, corn will be planted only when the soil temperature at the planting depth, during most of the diurnal cycle, is slightly above its minimal temperature for germination. Favorable temperature ranges, specific germination-enhancing periodicities of diurnal or seasonal temperature variation, induction of secondary dormancy, and the combined effects of water stress and temperature vary among

species. Germination is greatly affected by the interactions among temperature, soil-water potential, and water flow in the soil, and by variations in the Q_{10} factors of effective rates of seed biological activity. Each of the processes that occur in plants has its own rate response to temperature, commonly designated as a coefficient, Q_{10}, which is defined as the number of times that process rate increases with a 10° rise in temperature. The plant response to interactions between temperature and processes involved in water uptake, respiration, or cell division and growth are extremely complicated and unpredictable. Adverse effects of soil-water stress on germinating seeds intensify as temperature rises and may persist beyond the germination stages, into the seedling growth and emergence stages. Under field conditions, soil temperature is characterized by a diurnal temperature fluctuation, with amplitudes as great as 10° or more. If seeds respond to a single process, the observed rate will differ from that measured under constant temperature conditions.

Aeration Requirements

The total air content in the soil and the rate of gaseous exchange greatly affect the soil biological activity and the availability of oxygen to germinating seeds and developing seedlings. However, the effects of the interactions between the total air content, its constituents and composition in the soil, and the rate of gaseous exchange are complex and difficult to define, unless a definite knowledge of the interrelationships among complex diffusion processes (in air-filled pores and across water films) that control the oxygen supply and the dissipation of respiratory and decomposition by-products (CO_2, N_2, NO_2, H_2S, ethylene, methane) is attained. Low oxygen availability reduces or even prevents germination in most species. Oxygen supply to support metabolic activity in germinating seeds becomes decisive at a very early stage of germination, namely at the beginning of respiration and utilization of storage materials, and especially when growth is initiated during seedling development. Oxygen requirements increase with soil temperature, and under light and/or water stress. Very often a conflict develops between oxygen supply and water supply to germinating seeds, because of the very low solubility and diffusivity of oxygen in water. Thus, oxygen supply is greatly impaired as the thickness of the water films around the seeds and the hydrated seed coat increases, especially with seeds that have a swollen, hydrated mucilaginous cover. Low CO_2 concentrations may stimulate germination, either solely or at times in combination with ethylene. Soil seals or crusts and compacted soil layers may have deleterious effects on gas exchange and, in turn on seed germination, and seedling development and establishment.

Light Requirements

Sensitivity to light is observed in germinating seeds of various species, which fall into four classes: those that are insensitive to light, like most agricultural crops (e.g., anemone, oregano); those that require a short exposure to light during their germination (e.g., nasturtium, compass lettuce); those that will germinate in total darkness (tulip, gladiolus, Nigella); and those that will germinate under full exposure to light. The sensitivity of light-sensitive seeds increases as they take up more water. Germination of light-sensitive seeds is impaired when they are shaded by a dense plant canopy or when light is prevented from reaching the seeds because of burial under the soil surface or in cracks in the soil.

Soil Mechanical Constraints

Soil particles of various sizes and origins form a matrix that exhibits a degree of resistance under mechanical stress, described as 'soil mechanical strength.' Soil strength depends on the soil constituents, density, moisture content, and structure; it increases with increasing soil bulk density, and decreases with soil-water and organic matter contents. Silty soils with low organic matter content tend to deform plastically and to compress easily, and to form surface seals under the impact and slaking action of raindrops or under instantaneous flooding by rain or irrigation. These thin, rather dense seals may restrict gaseous exchange and water infiltration into the soil around the seeds, mechanically impede or obstruct germination and seedling emergence, or there can be a combination of these effects. Wheat seedlings are affected by the interaction between crust strength, water content, and rainfall. Adverse effects on seed germination and seedling establishment, similar to those of soil seals, are caused by mechanical compaction of the soil surface.

Water Uptake by Seeds and Seedlings

Water uptake by a germinating seed is an essential step toward rehydration of its tissues. The initiation of the array of metabolic processes in the seed and the minute amounts of water required for germination depend on the seed genome and the constituents of its individual parts. The various organs (embryo, cotyledons) and tissues differ in their internal physical structure, biochemical properties, and composition; therefore, they differ in their water retention, distribution, and swelling. In most seeds the seed embryos are minute (the embryo weight varies between 0.005

and 0.02 of the total dry seed weight). Storage tissues constitute the major volume of the seeds, and the observed amount of water taken up by those seeds is that which is taken into these tissues under the water regime that exists during the germination period of these seeds.

Water uptake by dry seeds during germination is characterized by three phases: the initial phase – the imbibition phase – is characterized by a saturation kinetics pattern; the second phase – the transition phase – is characterized by a low-to-negligible water uptake rate; and the third phase – the growth phase – is characterized by a rapid, exponential increase in the water uptake rate, accompanied by the emergence of the radicle. The first two phases are observed in dead, inert, and viable seeds alike, whereas the growth phase is unique to viable, germinating seeds. Water uptake rates during these phases are controlled by one or more of the following factors: (1) the seed properties with respect to water (the seed's water content–water potential characteristics; the seed's diffusivities to water, which range between 1.7×10^{2} to $1.5 \times 10^{-6}\,m^2\,day^{-1}$); (2) the soil-water properties (soil-water characteristics, diffusivities to water of the soil around the seed, which range from 4.0×10^{4} to $5.0 \times 10^{7}\,m^2\,day^{-1}$); and (3) the seed–soil interface properties with respect to water (seed–soil contact area, impedance to water flow across the contact zone). Once the third phase starts, the radicle emerges, the roots elongate, and the seedling starts to take up water with its radicle and developing roots, while the hypocotyl or mesocotyl emerges and elongates toward the soil surface. Water uptake by the radicle or the roots is controlled by one or more of the following factors: (1) the soil-water properties; (2) the root anatomy (root and/or radicle structure and tissue arrangement); (3) rate of elongation, radicle or root–soil contact, and hydraulic impedance to water flow across the seed–root contact zone.

The Imbibition Phase

The imbibition phase starts with the entry of water into the seed. This water is distributed in crevices, cracks, and flaws in the seed cover, and is absorbed by the seed tissues. Water uptake rate measurements taken during this phase have shown these rates to be: (1) temperature-dependent; and (2) accompanied by increases in respiration rate and in light sensitivity in some seed species. These observations suggest that water uptake during imbibition is not a 'passive' process, as it is usually taken to be, but becomes an active one at an early stage of this phase. The end of this phase is marked by an asymptotic approach to a final water gain, or hydration level, which depends on ambient soil-water potential, soil conductivity to water, seed–soil contact, and seed composition.

The Transition Phase

During the transition phase, known also as the 'pause phase,' the seed-water content, respiration rate, and apparent morphology remain almost unchanged. Nevertheless, a variety of metabolic processes are activated, and differences in the activity levels of these processes and in their order of occurrence have been observed among seeds of different species and among seeds that have reached different hydration levels. Therefore, any adverse environmental conditions that lead to redrying of the seeds, e.g., by subjecting them to water stress, or reduce the seed–soil contact area, and thus influence their hydration levels, may impair, retard, or even inhibit germination. If no damage had resulted, no secondary dormancy been induced, and no inhibitory processes been blocked, the germination of these seeds upon rewetting would be enhanced because of the high concentration of unused metabolites accumulated prior to drying. The duration of the transition phase influences the initiation time and the extent of radicle growth. Seeds have been observed to reach the transition phase and to remain in it for long periods that may extends for days, weeks, or more before germination. Toward the end of the transition phase, the embryonic cells of the radicle start to divide.

The Growth Phase

The growth phase starts with an increase in the respiratory rate, initiation of cell division, and extension of the embryonic radicle and ends with radicle protrusion. The common definition of germination states that a seed has germination once its radicle has broken out of the seed coat and started to elongate. From that stage, with germination completed, the seedling development and growth have commenced, and water is taken up solely by the elongating roots.

Differentiation Between the Seed Germination Phases

The definition of and the differentiation among the phases are based on an arbitrary partitioning of the continuous sequential order of processes in the germinating seed and the developing seedling. In reality, all the phases are interdependent, and the interrelationships among them suggest that each phase depends greatly on the preceding phases. The end of the third phase (marked by radicle protrusion), termed 'the growth phase,' cannot logically be separated from

the subsequent seedling growth toward its protrusion through the soil surface and its establishment. The end of the transition phase depends on whether the amount of water finally taken up at the end of imbibition has reached the required minimal level. A generally accepted concept is that to germinate a seed must reach a minimal water content – the 'critical hydration level.' That 'hydration level' does not reflect the water distribution among the seed components, nor is it an absolute value, since it depends on the water uptake rate, variations in soil-water potential, conductivity to water and temperature, and seed adaptability to variations in these environmental factors. The concept of critical hydration level fails in cases of partial seed wetting, especially when the wetted seed volume includes only the embryo and the adjacent storage tissues, and an affirmative concept is the 'critical water potential,' taken as the external water potential at or below which seeds cannot reach their critical hydration level. These concepts are applied respectively, when water uptake is evaluated according to its final amount of water taken up or in terms of seed–soil or seed–substrate water potential equilibrium states.

Physical Principles of Water Uptake by Seeds and Seedlings

Water movement within the soil toward the seed or radicle–soil contact interfaces, across them and into the seed or into the root, can be described by the general flow equation combining the Darcy flow equation and the continuity equation based on the principle of conservation of matter:

$$\frac{\partial\theta}{\partial t} = \frac{\partial}{\partial S}\left[k(\psi)\frac{\partial\psi}{\partial S}\right] = \frac{\partial}{\partial S}\left[D(\psi)\frac{\partial\theta}{\partial S}\right] \qquad [1]$$

For water movement in the soil and to the seed, it is simpler to use the right and the left sides of this equation, and the proper initial and boundary conditions. The initial conditions describe the initial state of the system under consideration, whereas the boundary conditions express the conditions that are imposed on or those that will exist once the processes to be analyzed are started. For a planted seed that starts to imbibe water from the soil around it, the initial conditions are:

$$\theta = \theta_{\text{initial}}$$
$$t < 0$$
$$0 < r \leq a$$

The boundary conditions that will exist once imbibition starts are:

$$D_{\text{seed}}(\partial\theta_{\text{seed}}/\partial r) = (LD_{\text{seed}}/a)[\theta_{\text{final}} - \theta_{\text{t}}]$$
$$t \geq 0$$
$$r = a$$

where θ, θ_{initial}, θ_{final}, θ_{t}, θ_{seed}, are the volumetric water contents of the soil, the initial and final values (cubic centimeters per cubic centimeter), at time t, respectively; θ_{seed} is the volumetric water content of the seed (cubic centimeters per cubic centimeter); ψ is the soil-water potential (kilopascals); $k(\psi)$ is the soil capillary conductivity (centimeters per day); $D(\theta)$, D_{seed} are the diffusivity coefficients of water in the soil and the seed, respectively (centimeters squared per day); r, a are the radial distance from the center of the seed and the seed radius, respectively (centimeters); and L is the seed–soil interface impedance to water flow (per day) (Figure 1). The rate of flow depends on: the water potential gradients (the driving force that induces water movement) in the soil, across the seed–soil contact zone, and in the seed tissues; and on the respective conductivities to water of the soil, the seed and radicle–soil contact zone and the seed or radicle tissues. In order to simplify the solutions to the above equation and the initial and boundary conditions, the number of variables has to be reduced. The choice to use the right side of the equation given above (where use of diffusivity to water in the soil is made) simplifies the calculations by reducing the number of variables and makes the measurements required to determine the flow parameters for use in the equation practicable.

Under normal conditions water will tend to move from the higher (in the soils) to the lower (in the seed

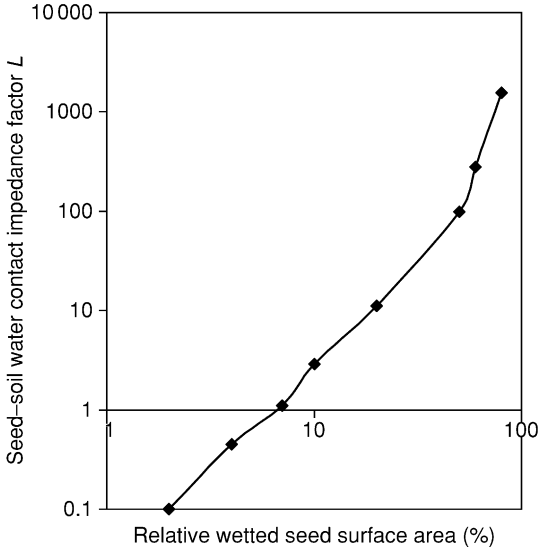


Figure 1 Seed–soil water contact impedance factor L as a function of the relative wetted seed surface area.

or radicle) water potential. The energy status of soil water termed 'soil-water potential' depends on the soil texture, structure, water content, soil solution composition, and solutes concentration. The matric potential is the soil-water potential component that arises from the interaction of soil water with the matrix of the solid particles in which it is contained. The osmotic potential is the soil-water potential component that arises from the interaction of soil-water solutes' composition and concentrations with the active membranes of the seed cells. Soil-water potential per unit of soil is given as work per unit of soil volume, or soil weight. At the beginning of the imbibition phase, these gradients are very large and the uptake rate is high, but, as the seed gains water, the gradient diminishes and the rate of water uptake decreases. As water moves into the seed, the water content in the soil next to the seed surface decreases; the matric and the osmotic components of the soil water decrease also (due to reduction of water content and an increase in solute concentration caused by ion exclusion). Consequently, the conductivity and diffusivity to water, through the soil and across the seed–soil contact, decrease, which tends to make the impedance to water flow increase. But, on the other hand, as the seed water content increases, the seed swells and the seed–soil contact improves, which tends to cause a decrease in the flow impedance. The combined effect leads to the net water flow rate into the seed.

Specific Effects of Matric and Osmotic Soil-Water Potential Components

Both the matric and the osmotic components of the soil-water potential are directly involved in soil-water movement to germinating seeds. Seeds immersed in solutions and in saturated soils respond equally to equal changes in these two components, provided their membranes are intact and fully active. But in soils, small reductions in soil-water matric potential have been observed to affect germination, and seedling development and establishment to a greater extent than equal or even greater reductions in the soil-water osmotic potential (Figure 2). This difference in response is because the changes in the seed–soil contact area and soil-water content that accompany small changes in the soil-water matric potential lead to proportionally greater changes in the soil conductivity and diffusivity to water.

The matric potential may affect seed swelling during germination and radicle or hypocotyl elongation, by its direct contribution to the effective soil mechanical stress. Under normal conditions the stresses induced in the soil are too small to confine seeds, to impair their swelling and germination, or to restrict seedling development. But poor germination observed in or next to compacted soils, or when seeds are entrapped in shrinking soil, can be attributed to greater mechanical constraints imposed by the soil, which are partly due to the matric potential.

Seed-Water Potential

The seed-water potential results from the osmotic water potential derived from the composition of the seed cells and the concentration of their constituents, and the turgor water potential component derived from the cell membranes and elasticity of the wall structure. In dry seeds the water potential is very low (negative in relation to free water), but as the seed imbibes water its water potential increases and becomes less negative. By changing the concentrations of their constituents and by modifying the activity and selectivity of their membranes, the seed cells can regulate their water potential relative to that of the soil in contact with the seed. These changes require energy inputs that develop from respiration and use of storage material. The lower the soil-water potential becomes, the greater effort and the faster the rate of depletion of the material stored in the seed for growth and establishment. Consequently, germination rate and final stand establishment will be impaired.

Seed and Soil Conductivities and Diffusivities to Water

Reported values of seed diffusivity to water (1.5×10^{-5} to $1.6 \times 10^{3}\,m^2\,day^{-1}$) are lower than those reported for soils (4.0×10^{4} to $5 \times 10^{7}\,m^2\,day^{-1}$) by several orders of magnitude. These observations suggest that under moist or wet soil conditions the rates of water uptake by seeds are not limited by soil-water movement rates toward them. Observed and calculated rates of water movement toward seeds indicate that for seeds that maintain lower water potentials than those in the soil, the soil could provide water to the seeds faster than the experimentally observed rates. These calculations strongly suggest that water uptake is controlled by the low permeability of the seed coat as well as the low seed diffusivity to water.

Seed-Coat Diffusivity to Water

In general, seed coats are nonuniform in shape and roughness, and in their ability to transmit water. Most seed coats are semi- or impermeable to water but they may have an opening (micropyle, hilum, chalaza). The few measured values of diffusivity to water of permeable, saturated seed coats (ranging from 9×10^{-2} to $3 \times 10^{2}\,m^2\,day^{-1}$) were found to be equal to or lower than those for the wet seed bulk (2.4×10^{2} to $1.6 \times 10^{3}\,m^2\,day^{-1}$). These values of seed-coat

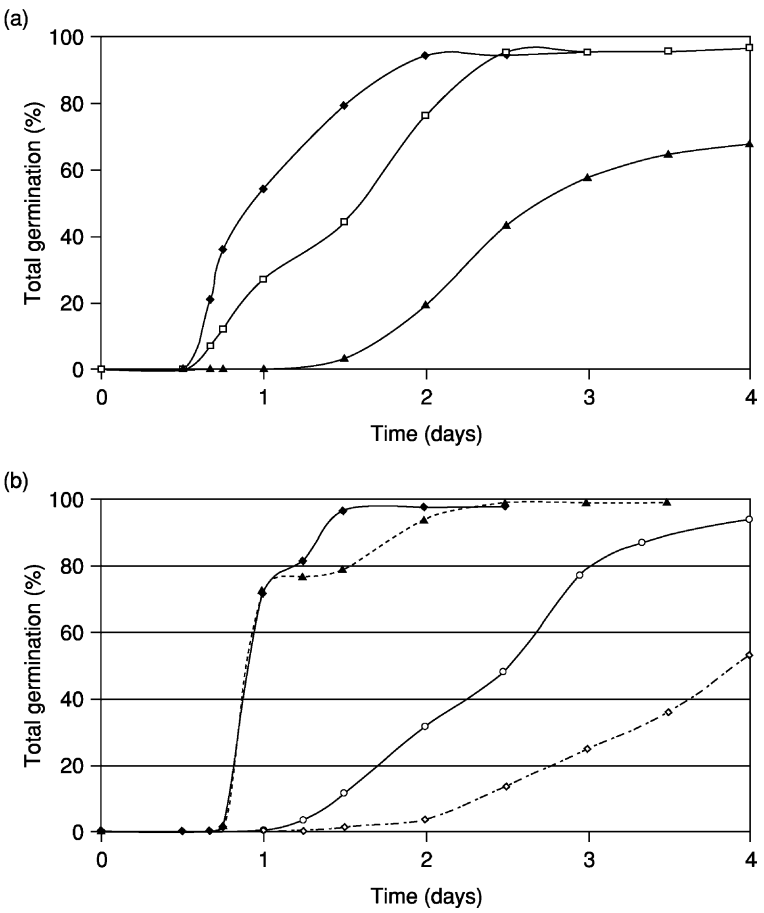


Figure 2 (a) Total germination of chickpea seeds as a function of time and matric potential: rhomboids, 100 J kg^{-1}; squares 500 J kg^{-1}; triangles 1000 J kg^{-1}. Seeds imbedded in aggregates of 0.25–0.50 mm diameter. (b) Total germination of chickpea seeds as a function of time and osmotic soil-water potential: full rhomboids 100 J kg^{-1}; triangles 330 J kg^{-1}; circles 1000 J kg^{-1}; and open rhomboids 2000 J kg^{-1}. Seeds imbedded in osmotic solutions.

diffusivity to water may affect water uptake rates and if there is also decreasing soil–seed interface contact the uptake rates could be even more strongly affected.

Seed-Coat and Seed–Soil Interface Geometrical Configuration

The geometrical configuration of the seed–soil interface zone depends on: the seed-coat surface composition and roughness; the ratio of seed size to sizes of the soil structural units in contact with it; and their spatial arrangement in the contact zone with the seed. The smaller the soil aggregates are, relative to the seed, the greater are the number of contact points and the total contact area. The contact area increases further when those contact points are wetted. Seeds with mucilaginous cover (e.g., like those of some species of the flax family) may achieve complete contact with the soil when they swell while being wetted. Hairy seeds may fail to form any direct seed–soil contact until their hairs either collapse or decay (due to the void found between the seed coat and the soil water by the mat of hairs).

Impedance to Water Flow Across the Seed–Soil Interface

The hydraulic properties of the seed–soil contact zone with respect to water vary during imbibition, because the water content at and next to the contact zone diminishes with water uptake. Since the seed–soil contact zone configuration is difficult to define and its water retention and conduction properties cannot be determined, a compound parameter (impedance to water flow) is used to describe water transfer across the contact zone. The seed–soil contact impedance to flow increases with decreasing water content in the seed–soil contact zone, seed–soil contact area, soil conductivity to water, and coarseness of soil texture and/or structure. The greater the impedance to water flow, the more restricted is the water flow from the soil toward and across the seed–soil contact interface and, consequently, the lower are the imbibition and water uptake

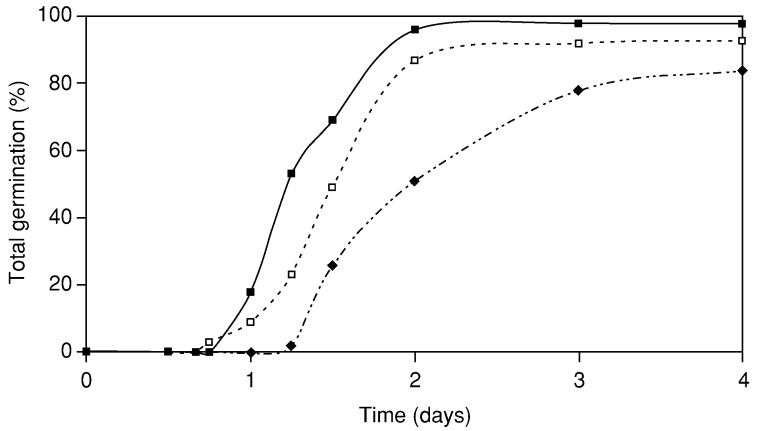


Figure 3 Total germination of chickpea seeds as a function of time and relative wetted seed surface area: filled squares 17%; open squares 6%; and rhomboids 3%. Matric potential range −10 to −33 J kg^{-1}.

rates (Figure 3). Most seeds swell during imbibition, which results in concurrent improvement of the seed–soil contact and increase in the soil mechanical stress on the seed. These changes may affect imbibition and final germination, especially in dense and compacted soils. In saline soils, seed–soil contact impedance effects may be partly obscured by the increase in the soil water osmotic potential caused by accumulation of salts excluded from entering through active membranes of the seed-surface cells during imbibition.

Seedling Emergence and Establishment

The germinated seed extends its radicle or seminal rootlets deeper into the soil to adsorb the water needed for the extension of the rootlets downward, and of the hypocotyl toward the soil surface. Water in ample quantities is needed by the rootlet and hypocotyl, to maintain the pressure required for their elongation within the soil and to overcome the weight of the overlying soil and the tensile strength of the soil surface layer, in order to emerge through it. The emerging seedling starts to transpire and to photosynthesize, and under a favorable soil-water regime, its radicle or roots elongate and start to explore the deeper soil layers in order to draw enough water to sustain the transpirational demands. It is obvious how difficult seed germination, emergence, and establishment become under arid conditions, under which the soil surface is wetted infrequently, the evaporative demands are high, and often the soils tend to slake and form a surface crust. Under such circumstances the germinating seeds and seedlings must cope with scarce and rapidly diminishing supplies of available soil water. Furthermore, soil-water depletion is too rapid either for the seedlings to extend their radicle or rootlets down into deeper, wetter soil layers where water is still available or for the seedling to break through the rapidly hardening soil crust.

Future Research

Timely, fast, and uniform seed germination, emergence, and seedling establishment are as crucial for successful crops as they are for survival of wild plants. Each seed reacts individually to the varying conditions prevailing in its microenvironment. The various aspects of seed germination and the interactions between seeds and the various environmental factors affecting their germination have been discussed above. The information presented, based on many studies carried out to resolve the complex system of seed germination and seedling establishment, is still incomplete and does not furnish a full understanding of the processes and interactions between germinating seeds and their microenvironment.

List of Technical Nomenclature

Critical hydration level ($cm^3\ cm^{-3}$)	The minimal seed volumetric water content below which the seed will not germinate
Critical water potential (kPa)	A given soil-water potential value below which the seed will not germinate
Effective soil mechanical stress (kPa)	The stress transmitted through the soil by intergranular and water menisci in the soil voids
Envelope soil-water pressure (kPa)	The magnitude of change in the total soil-water potential at a given position caused by the mechanical stress imposed by the soil layers above it
Growth	The last phase of seed germination, during which the radicles and rootlets are protruding and elongating into the soil

Imbibition A first phase of seed germination, during which the seed absorbs water and its tissues are rehydrated

Impedance to flow, soil–seed interface impedance to water flow (day^{-1}) A condition which hinders the movement of water through the seed-soil contact zone under the influence of water potential gradient

Matric potential (kPa) The difference between the water potential of source pool of soil solution at a given elevation and air pressure and a pool identical to the source but at the elevation under consideration (above water table)

Mechanical impedance Soil strength conditions which hinder or inhibit the devleopment and growth of plant organs in the soil

Osmotic soil-water potential (kPa) The magnitude of change in soil-water potential due to differences in the chemical composition of the soil solution related to free pure water at the same elevation and air pressure

Seed diffusivity to water ($cm^2\ day^{-1}$) The flux of water per unit gradient of volumetric water content in the seed

Seed germination A general term describing the sequence of processes involved in starting metabolic activities and leading to the initiation of growth in the quiescent embryo in the seed

Seedling emergence A term stating that a growing seedling has broken through the soil surface

Seedling establishment A term describing the stable development and growth of a seedling that has recently emerged

Seed's volumetric water content ($cm^3\ cm^{-3}$) The volume of water per unit bulk volume of the seed

Soil conductivity to water ($cm\ day^{-1}$) The flux of water per unit gradient of matric potential

Soil diffusivity to water ($cm^2\ day^{-1}$) The flux of water per unit of gradient of volumetric water content in the soil

Soil strength A transient, localized soil property which is a combined measure of a soil to resist deformation under external or internal stress imposed on it

Soil structure The spatial arrangement of soil primary particles into secondary particles or units

Soil texture The relative proportions of the various size classed separates in the soil

Soil-water potential (kPa) The amount of work that must be done in order to transport reversibly and isothermally an infinitesimal quantity of pure water from a specified source to a specified destination

Transition The second phase in seed germination

See also: **Crusts:** Structural; **Cultivation and Tillage**; **Nutrient Availability**; **Plant–Water Relations**; **Rhizosphere**; **Root Architecture and Growth**; **Structure**

Further Reading

Arndt W (1965) The nature of the mechanical impedance to seedlings by soil surface seals. *Australian Journal of Soil Research* 3: 45–54.

Benesch-Arnold RL and Sanchez RA (1995) Modeling weed seed germination. In: Kigel J and Galili G (eds) *Seed Germination and Development*, pp. 545–566. New York: Marcel Dekker.

Bewley JD and Black M (1982) *Physiology and Biochemistry of Seeds*. Berlin, Germany: Springer-Verlag.

Bradford KL (1995) Water relations in seed germination. In: Kigel J and Galili J (eds) *Seed Development and Germination*, pp. 351–396. New York: Marcel Dekker.

Corbineau F and Come D (1995) Control of seed germination by dormancy and by the gaseous environment. In: Kigel J and Galili J (eds) *Seed Development and Germination*, pp. 397–424. New York: Marcel Dekker.

Hadas A (1982) Seed soil contact and germination. In: Khan AA (ed.) *The Physiology and Biochemistry of Seed Development, Dormancy and Germination*, pp. 507–527. Amsterdam, the Netherlands: Elsevier.

Hadas A (1997) Soil tilth – the desired soil structural state obtained through proper soil fragmentation and reorientation processes. *Soil Tillage Research* 43: 7–40.

Hadas A and Stibbe E (1977) Soil crusting and emergence of wheat seedlings. *Agronomy Journal* 69: 547–550.

Koller D and Hadas A (1982) Water relations in the germination of seeds. In: Lange OL, Nobel PS, Osmond CB, and Zigler H (eds) *Encyclopedia of Plant Physiology,* vol. 12B, pp. 410–431. Berlin, Germany: Springer–Verlag.

Kutilek M and Nielsen DR (1994) *Soil Hydrology.* Cremlingen-Destedt, Germany: Catena-Verlag.

Marshal TJ, Holmes JW, and Rose CW (1996) *Soil Physics*. Cambridge, UK: Cambridge University Press.

Mayer AM and Poljakoff-Mayber A (1989) *The Germination of Seeds*, 4th edn. London, UK: Pergamon Press.

Van Duin RHA (1956) *On the Influence of Heat, Diffusion of Air and Infiltration of Water in the Soil*, Versl. Landb. K. Onderz. 62, pp. 1–82. Pudoc: Wageningen, the Netherlands.

Global Warming *See* **Carbon Emissions and Sequestration**; **Climate Change Impacts**; **Greenhouse Gas Emissions**

GRASSLAND SOILS

J A Mason, University of Wisconsin–Madison, Madison, WI, USA
C W Zanner, University of Nebraska–Lincoln, Lincoln, NE, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Soil formation in grasslands is strongly influenced by the climatic conditions under which grassland vegetation predominates as well as the distinctive characteristics of grassland ecosystems. In most grasslands, frequent soil-moisture deficits limit the rate of mineral weathering and often lead to secondary carbonate mineral accumulation in lower soil horizons. In grassland ecosystems, both the relative abundance of belowground biomass and active bioturbation lead to thick, dark, organic matter-rich A horizons. In well-drained grassland soils, organic matter content increases with increasing effective moisture and decreasing mean annual soil temperature; depth to secondary carbonates decreases with decreasing effective moisture. Topographically controlled local variation in soil hydrology can produce similar gradients within the length of a single hillslope.

Global Distribution of Grasslands and Grassland Soils

Grassland ecosystems presently (*c.* 2003) occupy between 14% and 26% of the earth's land surface. These estimates do not include large areas of former grasslands recently brought into cultivation, where the present soils formed as part of natural grassland ecosystems. The range in estimated extent in part reflects varying classification of natural plant communities that include both grasses and scattered shrubs and trees. Soil development beneath all of these communities may be strongly influenced by grasses.

From a global perspective, the distribution of grassland is largely a function of climate, particularly climatic factors that influence soil moisture availability. Effective moisture (annual precipitation minus annual potential evapotranspiration) is a useful index of those factors, although it neglects important seasonal variations in moisture supply. At the broad, continental scale, grasslands occupy regions of intermediate effective moisture, between the humid regions that support forest and arid regions dominated by desert shrub communities. The soil-water balance in grasslands varies widely, but, in the most extensive grassland regions such as the steppes of Russia and Central Asia or the North American Great Plains, potential evapotranspiration frequently exceeds precipitation both on an annual basis and during the growing season. Large interannual variability in precipitation is common and droughts can cause drastic reduction in plant productivity as well as changes in species composition. Geologic records also indicate longer-term shifts between wetter and drier climates during the Quaternary (the last 1 700 000 years); thus, most grassland soils have formed under a range of climatic conditions and plant communities.

Wind-blown sand and loess are probably the most common geologic materials underlying grasslands, mainly as a consequence of the climatic setting. Eolian sand and dust were generated locally within grasslands during past periods of dry climate, or are derived from neighboring arid regions. In either case, the wind-blown sediment can subsequently be stabilized by grassland vegetation and preserved over long time periods. Grassland ecosystems also thrive on a wide range of other geologic materials, from weathered bedrock to glacial till and alluvium.

At the local scale, factors other than regional climate become important in determining boundaries between grassland and other vegetation. Fire, ignited by lightning or by humans, is believed to have maintained grassland or savanna in regions where the climate would otherwise have allowed forest development. Recurrent fires generally favor grasses with ground-level or underground growth points over woody plants and shrubs that have most of their biomass aboveground. For example, in the 'prairie peninsula' of the midwestern USA, grasslands formerly extended far eastward into zones receiving 800–1000 mm of annual precipitation, under which forest might be expected. In this region, the natural

forest–grassland boundary was often associated with obstacles to fire such as river valleys or lakes. This effect of fire has resulted in a complex pattern of natural vegetation, with isolated patches of forest surrounded by grassland or vice versa (Figure 1).

Certain soil orders are traditionally thought of as archetypal grassland soils, including Mollisols in the US Department of Agriculture's (USDA) *Soil Taxonomy*, and Chernozems and Kastanozems in the system adopted by the Food and Agriculture Organization of the UN (FAO). It is important to recognize that the global extent of these soil orders is much smaller than the extent of all soils formed under grassland or savanna ecosystems. Classification in the Mollisol order requires a mollic epipedon, one or more dark-colored, near-surface horizons that reach a minimum total thickness (generally 25 cm, though thinner mollic epipedons are allowed in some cases). Base saturation (percentage of the cation exchange capacity that is occupied by Ca^{2+}, Mg^{2+}, K^{+}, and Na^{+}) also must exceed 50% in a mollic epipedon. Subtropical or tropical grassland soils often do not qualify as Mollisols because of low organic matter content, resulting in light-colored upper horizons; some of these low-latitude soils may also be too acidic to qualify. Even at higher latitudes, large portions of most grassland regions are occupied by weakly developed soils that do not meet mollic epipedon thickness requirements (FAO: Regosols and Lithosols; USDA: Entisols and Inceptisols), particularly on eolian sand deposits and on steep slopes in a variety of parent materials. Nonetheless, Mollisols or 'Mollisol-like' soils provide the best examples of pedogenic processes specific to grasslands. Mollisols (and roughly equivalent soil orders in other systems) are found almost entirely within grasslands, or grasslands recently converted to agricultural use, whereas other soil orders that occur in grasslands can also form in a range of other ecosystems.

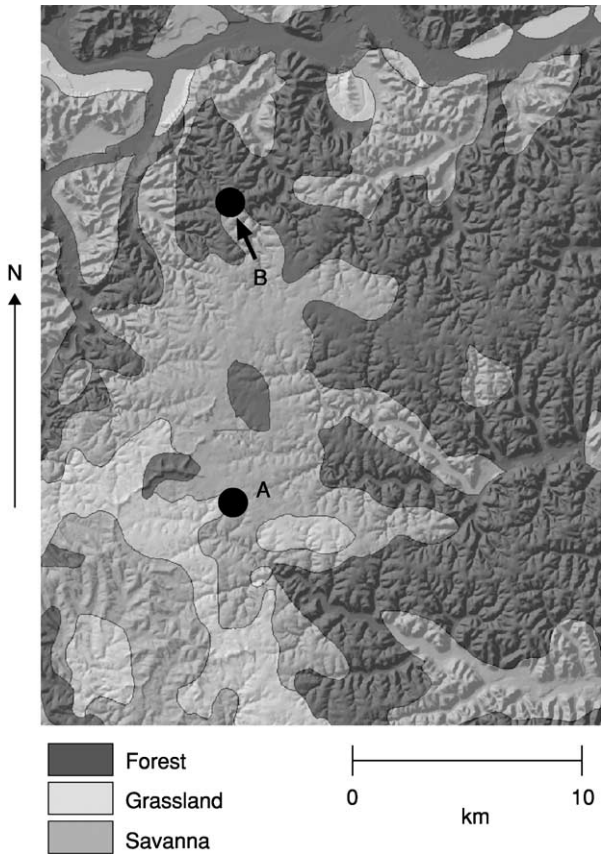


Figure 1 Complex pattern of grassland, savanna, and forest vegetation in a small area of southeastern Minnesota, USA. Vegetation types before conversion to agriculture in the 1850s to 1870s are superimposed on a shaded relief image of topography. Forest is largely confined to dissected areas where deep narrow valleys blocked rapid spread of fire. Grassland occurs on broad, gently sloping ridgetops and in some wide river valleys, where fires spread quickly. Savanna occurred in intermediate landscapes. Locations of grassland (A) and forest (B) soil profiles compared in Figure 5 are indicated. Vegetation cover is based on public land survey records, interpreted by Marschner FJ (1930) *The Original Vegetation of Minnesota* (1:500 000 scale map), US Department of Agriculture, and digitized by the Minnesota Department of Natural Resources. Topographic data are derived from US Geological Survey topographic maps.

Environment of Pedogenesis in Grasslands

The distinctive characteristics of grassland soils in part reflect the influence of the climatic conditions under which grassland vegetation predominates. The frequent soil-moisture deficits in many grassland regions inhibit silicate mineral weathering and organic matter oxidation, and favor precipitation of secondary carbonates within the soil profile.

Ecologic processes also influence grassland soil development. Net primary productivity (biomass production by photosynthesis minus biomass consumed through respiration by the photosynthesizers themselves) is relatively modest in most grasslands, relative to many forest ecosystems. The total standing biomass in grasslands is also small compared with mature forests. A larger proportion of the total biomass produced by photosynthesis is transferred to plant roots in grasslands than in forests, however. In many grasslands, underground plant biomass is greater than aboveground biomass, in some cases by a factor of 5 or more. As a consequence, the fraction of the total organic matter input that is added to the soil belowground is larger in grasslands than in forest, where most input occurs in the form of surface litter.

Consistent with the relative importance of belowground biomass, burrowing activity by animals

appears to be greater in grasslands than in other ecosystems, although the difference is difficult to quantify. There are many observations of remarkably extensive burrowing by earthworms, insects (especially ants and cicadas), and rodents in North American and Eurasian grasslands. Termite-burrowing and mound construction is extensive in tropical and subtropical savannas and grasslands.

Geomorphic processes that interact with pedogenesis may also differ between grasslands and other ecosystems. As noted above, eolian sand and loess deposits are widespread in grassland regions, indicating the potential for past truncation or upbuilding of soil profiles by wind erosion and dust deposition. This potential is quite evident in grass-covered dune fields, where excavations commonly reveal multiple buried soils. The potential for reworking of soil material by slopewash is also higher in most grasslands than in forest.

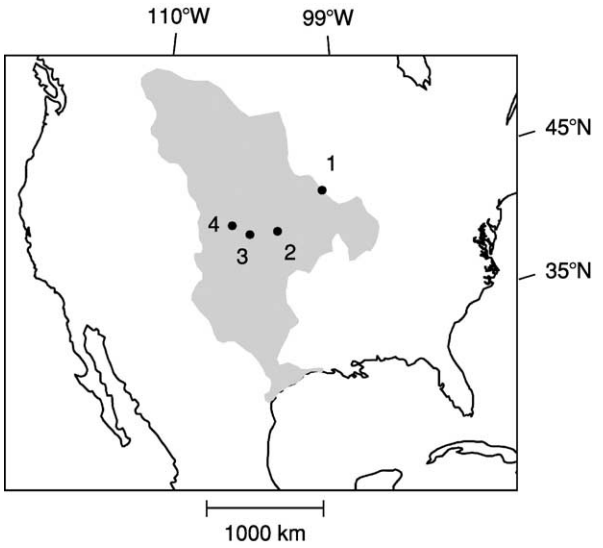


Figure 2 Locations of grassland soil profiles shown in **Figure 3** and **4**, within the grasslands of central North America (shaded). 1, Caledonia, Minnesota; 2, Seward, Nebraska; 3, Beaver City, Nebraska; 4, Brule, Nebraska. Boundaries of grassland region are approximate and encompass numerous outlying patches of forest and savanna, particularly near the eastern margin of grassland.

Organic Matter Accumulation, Bioturbation, and Structure Development

One of the most distinctive aspects of soil formation under grassland in temperate climates was recognized in an early classic of pedology, V.V. Dokuchaev's *Russian Chernozem*. The term 'chernozem,' or 'black earth,' is an apt description of the thick, dark-colored zone of organic matter accumulation, or A horizon, in soils of relatively humid grasslands in Russia and the Ukraine. Moving southward from the chernozem zone in Russia, effective moisture decreases and A horizons become thinner and lighter-colored, reflecting lower organic carbon content. Similar trends in carbon content and A horizon color are observed along east-to-west transects of decreasing effective moisture in the grasslands of central North America (**Figures 2–4**) and are likely to be present in other regions. Local anomalies occur within these broad regional trends in areas where A horizons are being built upward by ongoing eolian or alluvial deposition, and soil carbon is thereby being sequestered in deeper horizons beyond the depth of most active decomposition (**Figure 4b**). A temperature effect independent of soil-water balance is also evident. In North America, soils in the cold grasslands of the northeastern Great Plains store more organic carbon than grassland soils farther south that receive similar effective moisture but formed under warmer temperatures. Organic carbon content also increases as temperature decreases and precipitation increases, from low- to high-elevation grasslands.

These trends with respect to climate can be interpreted using a simple model, in which total soil organic carbon content is in a steady state, with additions equal to losses. The rate of addition is a function of primary productivity, which increases with both increasing effective moisture and increasing temperature. For example, in the central Great Plains of North America, primary productivity increases by two- to threefold in an east-to-west traverse from 250 to 750 mm of annual precipitation. Each year's addition of organic matter is offset by loss through oxidation, largely through microbial activity, of a fraction of the total carbon pool. The fraction oxidized also increases with temperature and effective moisture, but not as fast as primary production. Thus, a much larger pool of soil carbon is required to maintain steady state in cool, humid grasslands than in areas that are warmer and/or drier. This model is simplistic, because the climate in most grasslands is too variable to allow a true steady state to be maintained, and various forms of organic matter are oxidized at rates differing by orders of magnitude. Most recent modeling efforts consider multiple pools of soil carbon, turning over at varying rates, but the results are still consistent with observed trends of grassland soil carbon content in relation to climate.

Soil texture also strongly influences organic matter content. Regression models for prediction of soil organic carbon content often include clay content or some other textural index as a key predictor. This effect is related to the formation of complexes containing both humus and mineral particles, which can

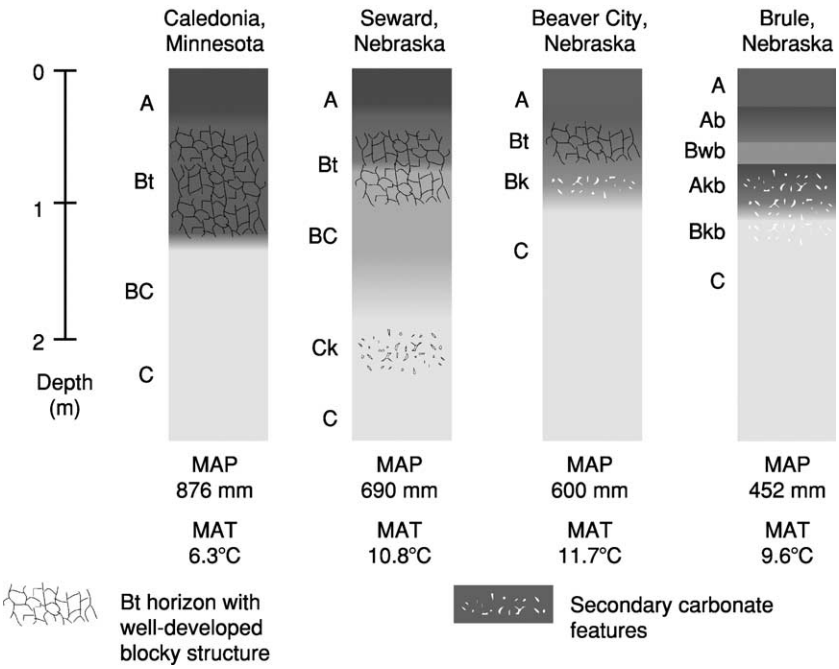


Figure 3 Schematic view of soil profiles from humid to semiarid grasslands in central North America (locations shown in **Figure 2**), showing typical trends associated with climatic gradients. Climatic normals indicated below each column (MAP, mean annual precipitation, 1961–90; MAT, mean annual temperature, 1961–90). The climate at each site has varied significantly over geologic time but the regional climate gradients are likely to have been maintained throughout the period of soil development. Major soil horizons are indicated. Note that 't' indicates clay accumulation, 'k' secondary carbonates, and 'b' buried horizons covered by younger sediment. All profiles are formed in late Pleistocene loess deposited 25 000–12 000 years ago, except for the soil at Brule, Nebraska. The latter is formed in late Pleistocene loess at greater than 75 cm depth and in at least two increments of Holocene loess at less than 75 cm. Thickness of dark-colored upper horizons decreases westward as rainfall decreases from Caledonia to Beaver City. The Brule profile deviates from this trend because renewed loess deposition has buried the upper Ab horizon, and organic matter then accumulated in the newly deposited loess, in effect producing an overthickened A horizon. Thickness of the Bt horizon and depth to secondary carbonates also decrease with decreasing rainfall.

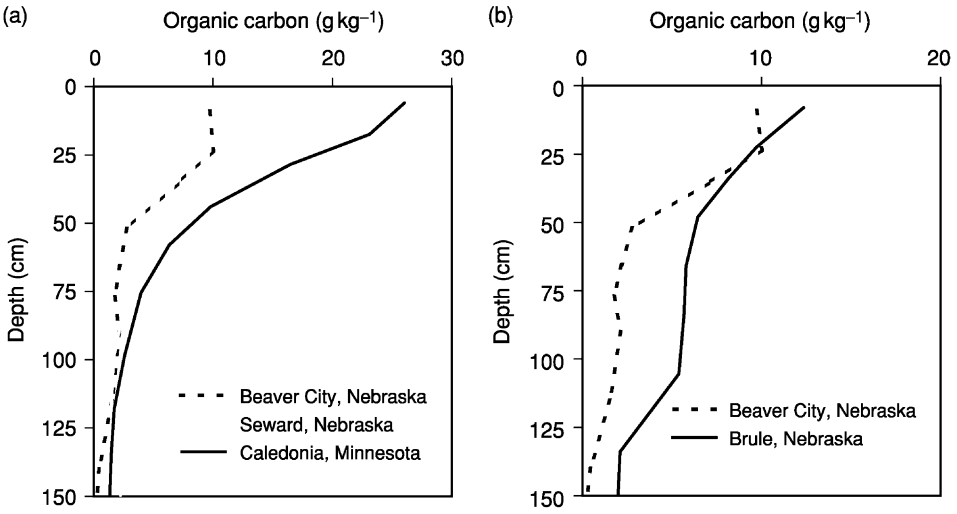


Figure 4 (a) Profiles of organic carbon content in three of the soils shown in **Figure 3**. Total organic carbon storage decreases with decreasing mean annual precipitation. The Caledonia profile also formed in a somewhat colder climate, which further enhances organic carbon accumulation; (b) effect of ongoing loess deposition on carbon storage. The Brule profile, where 75 cm of loess has accumulated in the past 10 000 years contains more organic carbon to a greater depth than the Beaver City profile, where mean annual precipitation is somewhat greater but loess deposition in the same time interval has been less than 40 cm.

enhance resistance to organic matter decomposition. Organic carbon associated with coarse clay and fine silt is more stable than the more rapidly turned over portion associated with fine clays. Strongly aromatic humic acids are associated with the coarser clays. Original resistance, complexation with phenolic polymers, and absorption to clay can maintain humic materials in the soil for thousands of years.

Locally, the vertical distribution of organic matter within the soil profile can change rapidly across the transition from grassland to forest. Scattered trees have limited influence on surface soil characteristics, but, as the forest becomes denser, organic matter addition to the soil is increasingly dominated by litter deposited on the soil surface rather than belowground production. Also, a portion of tree root biomass is comprised of large, suberized and lignified roots that are longer lived and less readily decomposed than fine grass roots and therefore contribute less organic matter to the soil on an annual basis. Addition primarily at the surface and limited bioturbation lead to much thinner A horizons in near-surface horizons under closed forest than under nearby grasslands. Soils with intermediate A-horizon thickness occur in the grassland–forest transition zone. Enhanced moisture in the litter layer in a shaded forest floor environment causes higher decomposition rates than in prairie and increases the production of translocatable organic acids, with the effect of reducing total carbon storage in near-surface horizons under forest (Figure 5).

Both bioturbation and formation of mineral–organic matter complexes play a role in the development of the fine, exceptionally stable structure of many A horizons in temperate grasslands. Macroscopically, soils under undisturbed natural grassland often have moderate to strong granular structure. Microscopically, these soils display a spongy fabric made up of rounded aggregates with abundant interaggregate pore space. Many of the most stable and well-rounded aggregates are in fact excrements of soil fauna. Others are produced through extensive burrowing activity and propagation of the fine root network of grasses. An extreme case is represented by soils that are predominantly made up of earthworm casts (Vermustolls and Vermudolls). Aggregates in undisturbed grassland soils are stabilized by mineral–organic matter complexes (particularly microbial polysaccharides interacting with clays at submicron size), fungal hyphae, and other bonding agents, and are resistant to physical disruption and slaking or dispersion in water.

Agricultural tillage breaks up soil aggregates, distributes oxygen and moisture through the plow layer, and agriculture can also add nitrogen to the system. Conversion of grasslands to agricultural use results in a reduction of the organic carbon content by 20–50% in 20–40 years. Loss of organic carbon reduces the soil's water-holding capacity and structural stability, and may reduce cation exchange capacity.

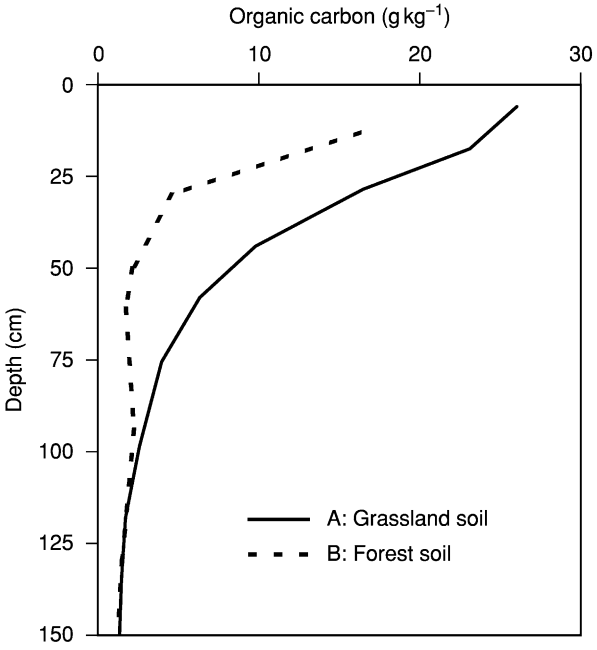


Figure 5 Contrasting organic carbon content profiles of soils in Minnesota, USA, that were under grassland or forest at the time of public land surveys in the nineteenth century. The soils formed under identical parent materials and climatic conditions and are in similar topographic settings. Locations shown in **Figure 1**.

Mineral Weathering, Clay Translocation, and Secondary Carbonate Mineral Formation

The typical climatic setting of grassland soils does not favor rapid silicate mineral weathering, although some studies have demonstrated that measurable chemical weathering does occur over periods of thousands to tens of thousands of years. The soil-water balance in this setting does not allow large volumes of water flux through soil; thus, the products of weathering are not effectively removed from the soil profile. Particularly in temperate grassland soils, the cation exchange complex remains dominated by Ca^{2+} and Mg^{2+}, in contrast to soils with greater throughput of water, in which these cations are depleted over time. The base saturation requirement for mollic epipedons is consistent with this observation. In the subtropical grasslands and savannas of the Sahel region in Africa, exchangeable Ca^{2+} and Mg^{2+} are also replenished by dust influx from the Sahara. A similar process may occur in other grasslands bordering desert dust sources. The predominance of Ca^{2+}, Mg^{2+}, and

other base cations buffers soil pH in the circumneutral range in which silicate mineral weathering is generally slowest. Divalent cations also inhibit clay particle dispersion and translocation. In the most humid grasslands, these inhibitions on weathering and clay movement may be less effective. For example, some Mollisols in the humid Midwestern USA appear to have accumulated abundant clay in excess of the parent material clay content (Figure 6; Caledonia, Minnesota, profile).

Ecosystem characteristics as well as climate limit weathering and clay translocation in grassland soils. This is indicated by rapid changes in soil morphology across the grassland–forest transition. With increasing forest density, more prominent eluvial E horizons develop, with a more abrupt downward transition to clay-rich B horizons. Forest soils are reported to have lower pH in upper horizons (under natural vegetation), higher rates of silicate mineral weathering, and increased depletion of total phosphorus than nearby grassland soils. These contrasts can be explained by contrasting organic matter dynamics, weathering intensity, and clay mobility under the two vegetation types. Grassland organic matter decomposition tends to produce organic compounds that form strong organoclay complexes and inhibit downward translocation of base cations and clays. Under forest cover, organic acids produced as the surface residue decomposes facilitate leaching of base cations and clays. Less total organic matter is available to produce resistant organoclay complexes. These effects combine to facilitate downward translocation of clay under forest. Soils near the transition between present-day forest and prairie often show evidence of both ecosystems. This is because this boundary is not static and will shift in reaction to changes in climate.

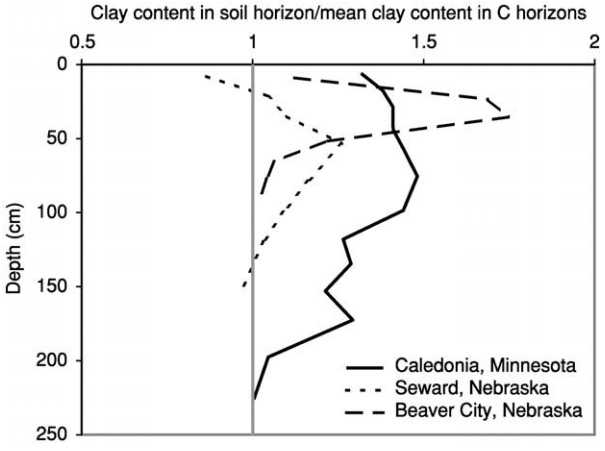


Figure 6 Clay accumulation in three of the soil profiles shown in Figure 3. Clay contents in A, B, and BC horizons are normalized by taking the ratio of actual clay content to average C-horizon clay content. Vertical bar represents a ratio of 1; values greater than 1 represent net gain of clay accumulation relative to parent material. Note prominent but shallow clay accumulation peak at the westernmost and driest site (Beaver City), and thick zone of clay accumulation at the most humid site (Caledonia).

Clay translocation in grassland soils is a controversial subject. Argillic horizons (zones of clay enrichment through translocation) are uncommon in grassland soils of Eurasia, except near the grassland–forest boundary, consistent with limited clay translocation under conditions of circumneutral pH and abundant exchangeable Ca^{2+} and Mg^{2+}. Some supposed argillic horizons in the North American Great Plains are quite prominent (Figure 6), although microscopic studies of soil thin sections provide little or no evidence of translocated clay. The German pedologist Arnt Bronger has proposed that these clay-rich horizons represent paleosols buried by loess with much lower clay content. This explanation has been received with skepticism by some North American researchers, but clear evidence for ongoing Holocene loess deposition at some sites suggests that Bronger's hypothesis is plausible (Figure 3; Brule, Nebraska, profile).

Precipitation of secondary calcite ($CaCO_3$) is nearly ubiquitous in fine-textured grassland soils of temperate regions, except in the most humid areas near the grassland–forest boundary. Secondary carbonates are less common in soils formed in sandy parent materials. In well-drained soils, the processes of secondary carbonate formation are straightforward and are clearly controlled by climate. Soil water contains dissolved Ca^{2+} derived from rainwater, dry dust deposition, or weathering of carbonate or silicate minerals. During major infiltration events, the Ca^{2+} is translocated to deeper soil horizons, but the amount of infiltrating water is insufficient to completely remove Ca^{2+} from the profile. As the soil subsequently dries, the soil water becomes saturated with respect to calcite, which precipitates (the required anion, HCO_3^-, is ubiquitous in soil water). As effective moisture decreases across a grassland region, the depth of frequent wetting decreases and secondary carbonates are concentrated at a shallower depth.

When many soils distributed across a wide range of climatic conditions are considered, a systematic decrease in depth to the upper limit of secondary carbonates with decreasing effective moisture is evident (Figure 3). At more local scales, however, depth to secondary carbonates is surprisingly variable. This variation may reflect differing soil hydraulic properties, or local topographically controlled differences in the partition of precipitation between runoff and infiltration.

Grassland Soil Hydrology and Its Influence on Pedogenesis

Closed depressions are common in many midlatitude grassland regions. These include regions that were glaciated during the most recent glacial period, where there has not been enough time for a connected drainage system to form. Closed depressions of eolian origin are common in dune fields and loess-mantled landscapes. Depressions can serve either as points of focused groundwater recharge or as discharge areas for groundwater flow systems. Some depressions alternate between these two roles on a seasonal basis.

Precipitation and snowmelt in recharge depressions is removed by both percolation and evapotranspiration (unless artificial drainage has been installed). Grassland landscapes containing closed depressions frequently illustrate the concept of a soil catena. A catena is a linked group of soils that have different characteristics because of variations in hillslope position and location within local surface and subsurface hydrologic systems. Well-drained soils on broad summits are drier and have moderately deep, dark brown surface colors reflecting more oxidizing conditions. Carbonates are leached as water infiltrates into the soil more than it runs off. Shoulder positions experience erosion and runoff, so soils are thin, and carbonates are close to the surface. Carbonate leaching depth and solum thickness increase downhill on side slopes as infiltration increases. Redoximorphic features may become more strongly expressed and closer to the surface with distance downslope, reflecting more frequent near-saturated conditions in which reducing conditions can develop. Areas low in the landscape produce more biomass and thus have greater organic inputs to the soil than upland positions. Decomposition rates are also slow in the relatively cool, wet soils in depressions or on lower slopes. The result is higher organic matter content and surface horizon colors that are blacker than those observed in the upland positions. In these wet soils, redoximorphic features can be masked by the darkness of the surface horizons, but gleyed colors (gray colors resulting from dissolution of iron oxides and hydroxides under reducing conditions) may be present at greater depth. Soils in some closed depressions display a distinct, light-colored E horizon, which is generally lacking in other grassland soils.

Grassland soils in areas of groundwater discharge often accumulate soluble salts, gypsum, and/or secondary carbonate minerals. These minerals form from ions transported to the site by groundwater flow. As water is extracted from the soil by root uptake and surface evaporation, the concentration of ions exceeds the solubility of the minerals and they precipitate, often in order of increasing solubility along the upward groundwater flow path. That is, carbonates precipitate first, followed by gypsum and soluble salts. Both the vertical and horizontal distribution of secondary minerals are highly sensitive to the local topographic and hydrologic setting. In more humid grasslands, there are extensive soils in depressions and other areas with shallow water tables that contain secondary calcite in the A horizon (Calciaquolls in USDA *Soil Taxonomy*). In other settings, there is a vertical sequence of secondary mineral formation, from calcite at greater depths through a gypsum-enriched zone to soluble salts in shallow horizons. This sequence can lead to high levels of sodium in near-surface soil horizons, either in the form of sodium salts or as Na^+ ions held on cation exchange sites on clay and organic matter. As long as the soluble salt content is high, soil structure is stable. If the upper part of the soil profile periodically experiences downward flux of infiltrated water (e.g., in a depression that is a site of either discharge or recharge depending on the season), then salts will be removed to greater depth. The Na^+-saturated clay and organic matter particles are then easily dispersed and translocated to form a dense, nearly impermeable B horizon. Such sodium-affected soils often occur in patches associated with subtle microrelief, such as shallow depressions a few meters to tens of meters in diameter.

Classification of Grassland Soils

A long-established scheme for classifying temperate grassland soils on the basis of A horizon organic matter content and color was originally developed in Russia, but was applied in the past or is still used in other parts of Eurasia and North America. In this scheme, dark-colored Chernozems occupy the humid end of the climatic range of grassland ecosystems, while lighter-colored Chestnut soils occur in drier grasslands and shrub steppes. The FAO system maintains this subdivision (Chestnut soils become Kastanozems), though with more rigorous definitions. The USDA *Soil Taxonomy* subdivides most well-drained Mollisols on the basis of climatic setting rather than intrinsic soil properties. Many Mollisols of cold climates, such as those formed under high-altitude grasslands, are included in the Cryoll suborder. In regions with Mediterranean climates (winter wet, summer dry), such as the annual grasslands of California, USA, well-drained Mollisols are classified as Xerolls. Most well-drained Mollisols in other climates are placed in the Udoll suborder in areas of

greater effective moisture, and the Ustoll suborder in drier regions. The Udoll–Ustoll boundary generally does not coincide with the Chernozem–Kastanozem boundary. Many common combinations of geologic parent material and pedogenic processes occur in more than one of these climatically defined suborders. For example, grassland soils formed in loess and affected by ongoing dust deposition include both Ustolls in the Great Plains (**Figure 3**) and Xerolls of the Palouse Prairie in eastern Washington, USA.

Both the FAO and USDA systems recognize intergrade soils formed in the grassland–forest transition (e.g., Alfic Hapludolls, USDA system) and grassland soils with evidence of clay illuviation (Luvic Chernozems, FAO system). The USDA *Soil Taxonomy* includes many wet and sodium-affected grassland soils in the Mollisol order, using suborders or great groups to distinguish them, while these soils are placed in separate orders in other systems.

The emphasis on strictly climatic criteria in the USDA *Soil Taxonomy* may be ill-advised if projections of rapid climatic change in the coming century are valid. The classification of many Mollisols could change much more rapidly than any visible or measurable properties of the soils themselves.

See also: **Range Management**

Further Reading

Aandahl AR (1972) *Soils of the Great Plains. Land Use, Crops, and Grasses*. Lincoln, NE: University of Nebraska Press.

Anderson DW (1987) Pedogenesis in the grassland and adjacent forests of the Great Plains. *Advances in Agronomy* 7: 53–93.

Baxter FP and Hole FD (1967) Ant (*Formica cinerea*) pedoturbation in a prairie soil. *Soil Science Society of America Proceedings* 31: 425–428.

Bronger A (1991) Argillic horizons in modern loess soils in an Ustic soil moisture regime: comparative studies in forest-steppe and steppe areas from eastern Europe and the United States. *Advances in Soil Science* 15: 41–90.

Coupland RT (ed.) (1993) *Ecosystems of the World 8: Natural Grasslands*. Amsterdam, the Netherlands: Elsevier.

Dokuchaev VV (1883) *Russian Chernozem*. Jerusalem, Israel: Israel Program for Scientific Translation.

Fenton TE (1983) Mollisols. In: Wilding LP, Smeck NE, and Hall GF (eds) *Pedogenesis and Soil Taxonomy. II. The Soil Order*, pp. 125–164. New York: Elsevier.

Pawluk S and Bal L (1985) Micromorphology of selected mollic epipedons. In: Douglas LA and Thompson ML (eds) *Soil Micromorphology and Soil Classification*. Soil Science Society of America Special Publication No. 15. Madison, WI: Soil Science Society of America.

Weaver JE (1954) *North American Prairie*. Lincoln, NE: University of Nebraska Press.

Weaver JE and Albertson FW (1936) Effects of the great drought on the prairies of Iowa, Nebraska, and Kansas. *Ecology* 17: 567–639.

Green Manuring *See* **Cover Crops**

GREENHOUSE GAS EMISSIONS

K A Smith, University of Edinburgh, Edinburgh, UK

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

The 'greenhouse effect' is the enhanced warming of the Earth's surface and lower atmosphere as a result of human activity, over and above the warming due to natural processes. Absorption of infrared radiation by natural concentrations of water vapor, carbon dioxide (CO_2), and trace gases in the atmosphere is responsible for the mean surface temperature of the Earth being approximately $+15°C$ rather than $-18°C$. Warming due to additional emissions of greenhouse gases is superimposed on this very large, natural warming process.

The most important individual greenhouse gas is CO_2 (*See* **Carbon Emissions and Sequestration, Carbon Cycle in Soils:** Dynamics and Management), but substantial contributions to global warming are made by other trace gases, including methane (CH_4)

and nitrous oxide (N_2O), both of which come in part from soil emissions. One kilogram of CH_4 has a warming potential 23 times greater than 1 kg of CO_2, over a 100-year period, while the warming potential of 1 kg of N_2O is nearly 300 times greater. A further contribution to global warming comes indirectly from tropospheric ozone, which is produced by light-catalyzed reactions involving nitric oxide (NO) and organic compounds emitted by vegetation. NO is a natural product emitted from soils as well as from combustion processes. Soils also act as a minor sink for methane already in the atmosphere.

Methane Sources and Sinks

The concentration of methane in the atmosphere has more than doubled since the preindustrial era, from approximately 0.7 ppm (0.7×10^{-6} mol mol^{-1}) to greater than 1.7 ppm today. The lifetime of CH_4 in the atmosphere is quite short, *c*.10 years; the main removal pathway is through oxidation by UV-created hydroxyl radicals (OH) in the atmosphere, while 5–10% is oxidized by soil microorganisms. No single source is responsible for the atmospheric increase, but soils contribute approximately 40% of total emissions (Table 1).

Soil Sources

Methane is formed in soils by the microbial breakdown of organic compounds in strictly anaerobic conditions (*See* **Anaerobic Soils**). A very low redox potential is required for this process, and CH_4 production does not begin until reduction of molecular oxygen, nitrate, iron(III), manganese(IV) and sulfate (all of which maintain the potential at higher levels) is complete. Such low-redox conditions are predominantly found in soils where prolonged waterlogging is a normal feature, e.g., natural wetlands and flooded rice fields, and beneath the soil cover of closed landfills, as well as in the sediments at the bottoms of lakes. The original identification of methane as a main component of the gas released as bubbles from natural marshes was made by Allessandro Volta in Italy in 1776. The gas is still known by the name 'marsh gas' given to it at that time.

The organic compounds that provide the carbon for CH_4 formation in marshes, bogs, and flooded soils are released from living roots or come from the degradation of plant remains, e.g., decaying peat material and dead plant stems and leaves, in a bogland plant community, or rice straw ploughed into the soil of a paddy field prior to the establishment of the next crop. The cellulosic material in these residues is made up of hexose units, and these break down to give CH_4 and CO_2 in equimolar quantities, according to the equation:

$$C_6H_{12}O_6 \Rightarrow 3CO_2 + 3CH_4$$

The methane formed in this way can migrate to the surface and be emitted into the atmosphere by one of three different pathways (Figure 1). Diffusion can take place in solution from the point of formation in an anaerobic layer, upward to water layers containing oxygen, where much of the CH_4 is oxidized and only a fraction outgasses to the atmosphere. When sufficient gas is produced to exceed the solubility product, bubbles form in the water layer, and force their way to the surface – the process of ebullition that prompted Volta's original investigation. The speed of this process prevents any significant oxidation. The third route is via the internal structures of vascular plants. Many species that are adapted to life in flooded environments contain continuous air spaces which connect the leaves and stems to the roots. These structures, known as aerenchyma (Figure 2), have evolved to transport oxygen needed for cell division to the root meristems, but serve equally well as channels for the transport of methane from the root environment to the atmosphere.

In natural wetlands in boreal environments, CH_4 emissions have been found to vary considerably with the type of vegetation present, but the greatest differences in flux appear to be brought about by the depth to the water table. Figure 3 shows the logarithmic increase in CH_4 emissions from a range of ecosystem types in boreal North America as the measured depth to the water table diminishes from approximately 40 cm to zero. Although many investigations have been made into emissions from temperate and boreal wetlands, the degree of uncertainty about total emissions is still large. The uncertainty is greater still over emissions from tropical wetlands, because fewer studies have been made, and because there is thought to be a much bigger contribution from seasonally

Table 1 Estimated soil sources of methane emissions to the atmosphere

Source	*Global annual emission (Tg CH_4 year^{-1})*
Natural	
Wetlands	115
Termites	20
Anthropogenic	
Rice paddies	50
Landfills	40
Total from soils/soil fauna	225
Total from all sources	500–600

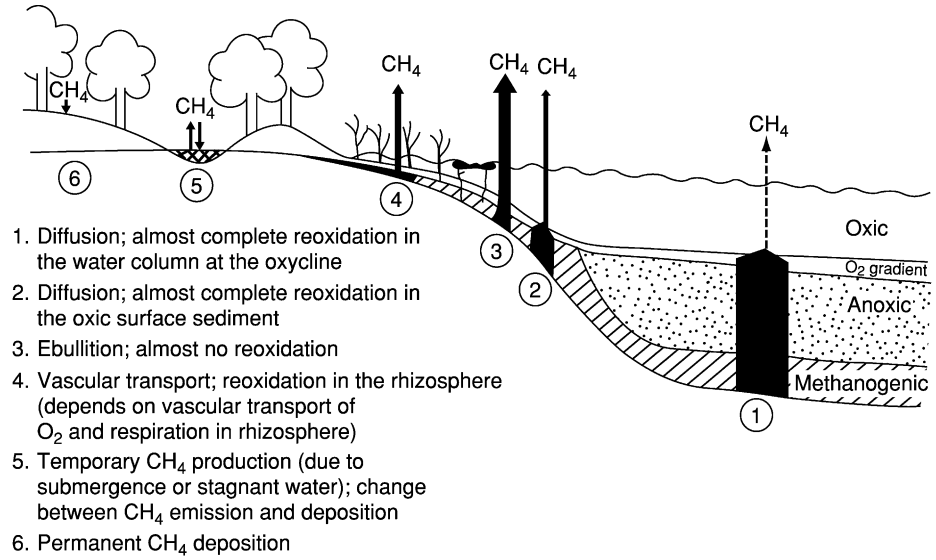


Figure 1 Transport and oxidation processes involved in methane cycling between terrestrial ecosystems and the atmosphere. (Adapted from Conrad R (1989) Control of methane production in terrestrial ecosystems. In: Andreae MO and Schimel DS (eds) *Exchange of Trace Gases between Terrestrial Ecosystems and the Atmosphere*, pp. 39–58. Chichester, UK: John Wiley & Sons, Ltd., with permission.)

flooded environments. Tropical river basins such as those of the Amazon and Parana in South America, and the Congo in Africa are prime examples of such seasonal flood plains, in which there is typically a rapid development of lush vegetation with the advent of the water, followed by decomposition under the floodwater and consequent methane production. However, the flooded areas vary considerably in extent and mean flooding depth from year to year.

In the absence of adequate experimental data, mathematical modeling of regional emissions has made an important contribution to the assessment of wetland emissions. Observed seasonal variations in the concentrations of methane in the atmosphere at different latitudes can now be simulated reasonably well by such models. Generally these models predict that more than 50% of global wetland emissions are of tropical origin.

Rice cultivation is also a major CH_4 source. The most recent estimates compiled by the Intergovernmental Panel on Climate Change have put the annual global emission from rice fields at approximately 50 Tg CH_4 $year^{-1}$ (Table 1). The area devoted to the rice crop has expanded considerably in recent decades, and global production is predicted to rise still further in the future, to meet the needs of rising populations. This trend has undoubtedly contributed to the increase in overall CH_4 emissions observed in the last few decades. An international programme of research has been carried out to investigate the principal factors that determine the magnitude of emissions from rice. This has shown that soil type, crop fertilization, the amount of straw or organic manure incorporated into the soil before planting, and the continuity of the flooding regime, as well as climate, are all important.

In warm climates, CH_4 production also occurs in some aerobic soils, as well as in wetlands, but here the source is the bacteria that live symbiotically in the guts of some species of burrowing and mound-forming termites, and not the free-living soil microflora.

Soil Sink for Methane

The phenomenon of CH_4 oxidation by soil bacteria is general in aerated soils (Figure 1). In well-drained and therefore well-aerated soils, in the absence of termites, any CH_4 present is largely that which has diffused in from the atmosphere, where the concentration is approximately 1.7 ppm as of 2003. This methane is oxidized to CO_2 by bacteria that are adapted to living on this very low concentration of substrate. The overall global sink for atmospheric methane in soils is estimated at approximately 30 Tg of CH_4 per year; this is about 5% of total annual emissions from all sources.

Soil bulk density and water content, and their consequent effects on gas movement and penetration in the soil profile, have a major impact on the rate of oxidation of atmospheric methane in soils. Rates are highest in coarse-textured forest soils with well-developed soil structure and a permeable surface organic layer. Increasing soil acidity reduces oxidation; part of the explanation is that in acid conditions plant litter decomposes less readily and accumulates

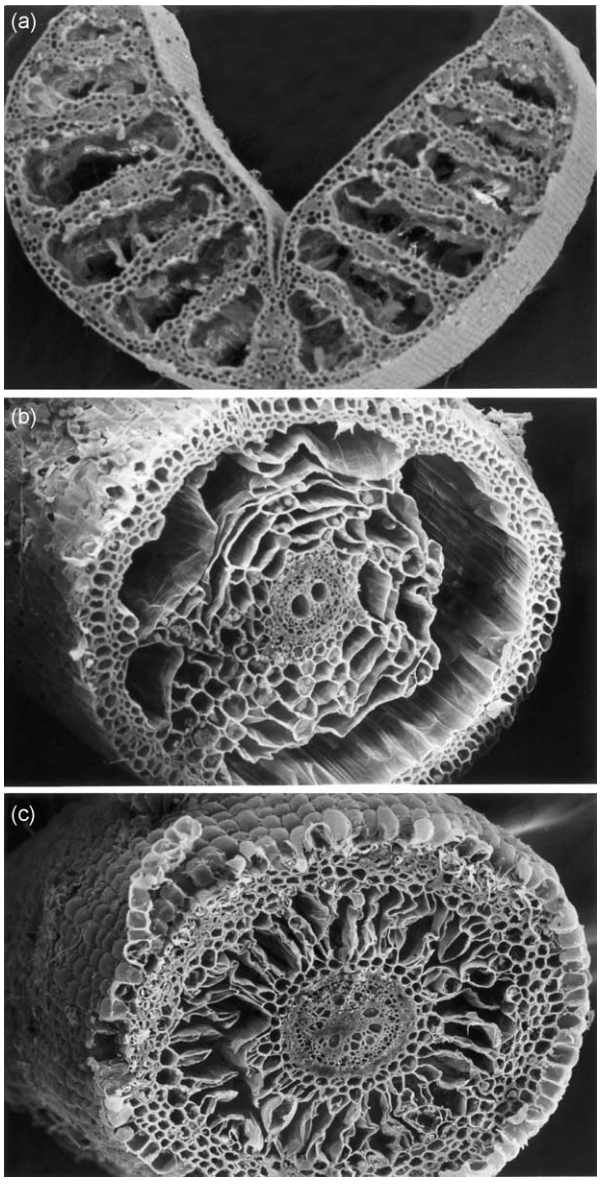


Figure 2 Scanning electron micrographs of transverse sections through the vegetative organs of some vascular wetland plants, showing the aerenchyma through which oxygen and methane can readily diffuse: (a) *Eriophorum angustifolium* leaf; (b) *Carex echinata* root (at a depth of 100–110 mm); and (c) *Molinia caerulea* root (at a depth of *c.* 150 mm). (Reproduced with permission from Thomas KL, Benstead J, Davies KL, and Lloyd D (1996) Role of wetland plants in the diurnal control of CH_4 and CO_2 fluxes in peat. *Soil Biology and Biochemistry* 28: 17–23.)

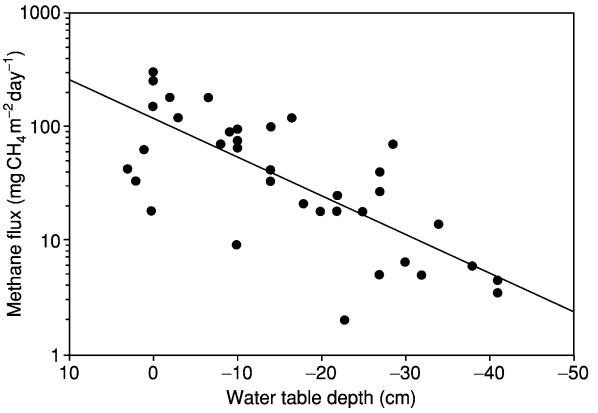


Figure 3 Relationship between methane fluxes from natural wetlands in boreal North America and water table depth (negative values on *x*-axis indicate water tables below the surface). Data compiled from six different studies. (Adapted from Liblik LK, Moore TR, Buber JL, and Robinson SD (1997) Methane emissions from wetlands in the zone of discontinuous permafrost: Fort Simpson, northwest Jemtones, Canada. *Global Biogeochemical Cycles* 11: 485–494.)

at the soil surface. This litter, particularly the leaves of broadleaved deciduous trees, acts as a diffusion barrier to gas exchange between the soil and the atmosphere. The effect of soil temperature on oxidation rate is small, and this is attributed to limitations imposed by the supply of available CH_4 substrate, due to the combined effects of diffusion resistance and low atmospheric concentration.

Conversion of natural soils to agricultural use reduces methane oxidation rates by approximately two-thirds. Additions of nitrogenous fertilizers also decrease oxidation rates. Recovery of oxidation rate from land-use change or fertilization is slow; it may take 100 years or more to return to predisturbance rates. The reasons for this unusually long period of adjustment are unknown.

In addition to the oxidation of atmospheric CH_4 in soils, there is also oxidation of a fraction (perhaps 20–30%) of the CH_4 produced in soils, before it escapes to the atmosphere. The methanotrophic bacteria responsible live in, for example, the thin, aerated sheath of soil that exists around the roots of vascular plants growing in flooded conditions, and in intercellular spaces within plant root systems. They also occur in aerobic hummocks on the tops of bogs, as well as in aerated cover soils above anaerobic refuse layers in landfills. In such environments, average CH_4 concentrations may be very much higher than those resulting from gas diffusion from the atmosphere, and different mixed populations of bacteria, capable of assimilating CH_4 at a faster rate than those adapted to atmospheric concentrations, become established.

Efforts to Reduce Emissions

Attempts are being made to reduce net CH_4 emissions from two major sources: sanitary landfills and flooded rice fields. Paradoxically, the creation of closed landfills for municipal wastes to prevent other problems associated with unconfined piles of refuse has brought about a major increase in the amounts of CH_4 produced. This is because the soil

caps (typically 1–2 m thick) exclude atmospheric oxygen, thus allowing the landfills to become anaerobic. However, the general recognition of the undesirable contribution of these sites to global warming has led to the development of effective engineering solutions for application to large modern landfills. These range from the venting and burning on-site of the CH_4 (converting it to CO_2, which is less than one-twentieth as potent a greenhouse gas), to its collection and use in energy production, thus gaining an additional benefit through reducing the use of fossil fuel for that purpose. For many older and smaller landfills, however, there is little scope for economic installation of CH_4 collection and recovery equipment. For such environments the more effective option is the promotion of CH_4 oxidation in the cover soil. The oxidation is optimal in cover soils of intermediate textures and moderate water contents, which permit a reasonable rate of diffusion through the soil matrix and thus allow oxidation to take place, but avoid the development of cracks through which gas may be vented rapidly to the surface.

There is potential for reducing substantially the average CH_4 emissions from traditionally flooded paddy fields. This involves introducing such variations in agronomic practice as delayed flooding at the start of the season; temporary drainage of the fields in midseason; and avoiding the decomposition under flooded conditions of large amounts of organic additives such as straw and animal manure. Such changes to traditional methods can be made without detriment to crop yields and sometimes bring with them other agronomic or cost benefits – at worst they can usually be introduced without increasing the cost or difficulty of farming operations. However, persuading farmers to use such new methods takes much advisory effort, and the rate of adoption varies greatly among different societies, unless incentives are introduced.

Nitrous Oxide

The natural cycle of N_2O in the environment involves microbial and chemical production in soils and waters, and ultimate decomposition in the stratosphere. Approximately 65% of all atmospheric emissions of N_2O are from soils. Ice-core data show that the concentration of N_2O in the atmosphere in the preindustrial era was approximately 275 ppb (275×10^{-9} mol mol^{-1}), but this had increased to about 311 ppb in the early 1990s. This increase is attributed to increased anthropogenic emissions, through increased production and use of N fertilizers, tropical land conversion from forest to agriculture, increased biomass burning, etc. The extra emissions have not been accompanied by any corresponding increases in the sinks for N_2O. The dominant sink is the reaction with ozone in the stratosphere to produce NO, which was the process that first generated the environmental concern about N_2O because of its potential for increasing UV radiation at the Earth's surface. In addition, a small soil sink has been suggested, but this has not been satisfactorily quantified.

N_2O and NO are produced in soils mainly by two contrasting microbial processes, autotrophic nitrification and heterotrophic denitrification. Nitrification is the microbial oxidation of ammonium (NH_4^+) to nitrite (NO_2^-) and thence to nitrate (NO_3^-). The most studied pathway is that involving autotrophic bacteria of the genera *Nitrosomonas* and *Nitrobacter*. However, heterotrophic microorganisms, mainly fungi, may be the most important nitrifiers in some circumstances, e.g., in acid forest soils.

The nitrification process is fundamentally an aerobic one, for which the presence of molecular oxygen (O_2) is essential. However, when the supply of O_2 is limiting, the nitrifying bacteria are able to use NO_2^- as an electron acceptor and reduce it to NO and N_2O:

$$\begin{array}{c} \text{NO, } N_2O \text{ (emission)} \\ \uparrow \\ NH_4^+ \Rightarrow NO \Rightarrow NO_2^- \Rightarrow NO_3^- \end{array}$$

Rates of NO production by nitrification are much greater than the corresponding rates of N_2O production. The ratio of gaseous products to nitrate produced during nitrification increases as the partial pressure of O_2 decreases, but the efficiency of NH_4^+ oxidation also decreases. This makes it difficult to estimate the effect of changes in soil aeration on the actual amount of N_2O produced. Nonetheless, in conditions in which nitrification is able to proceed, even at a reduced rate, gas diffusion rates within the soil are likely to be sufficient to permit most of the N_2O (and NO) produced to escape to the atmosphere.

The optimum temperature range for nitrification in soils is approximately 25–35°C, although nitrifiers in soils of different climatic regions may have different optima as a result of adaptation.

The other microbial process producing N_2O is denitrification, involving the reduction of NO_3^-, in the absence of oxygen. The necessary conditions for the process occur when respiratory O_2 consumption in the soil by plant roots and soil microorganisms exceeds the rate of replenishment by diffusion from the atmosphere. Anaerobic zones within the soil profile ensue, and within these zones nitrate is the chemical species that most readily acts as an electron

acceptor, and so it becomes reduced by a succession of enzymes, to nitrite, NO, N_2O, and finally molecular nitrogen, N_2:

$$\begin{array}{c} N_2O\ (\text{emission}) \\ \uparrow \\ NO_3^- \Rightarrow NO_2^- \Rightarrow NO \Rightarrow N_2O \Rightarrow N_2 \end{array}$$

The fraction of the total gaseous products of denitrification which is actually emitted to the atmosphere as N_2O depends heavily on soil structure and soil wetness. On the one hand, a very small anaerobic microsite may form in an otherwise well aerated soil profile, caused by a localized region of high respiration, and any N_2O formed in the microsite will have a high probability of escaping before being reduced to N_2. On the other hand, the soil may contain large, virtually saturated (and anaerobic) clods, into which NO_3^- ions may diffuse in solution. Any N_2O produced well within the clod can only escape after diffusing in solution to the surface, and is much more likely to be reduced to N_2 before this occurs.

The denitrification pathway to N_2O has been established as having to go via NO, yet NO is virtually absent from gaseous emissions in wet soils where anaerobic conditions occur. The accepted explanation is that NO has a much greater affinity than N_2O for enzyme binding sites, and consequently, after its formation and release into the soil, NO is much more readily readsorbed by the denitrifying organisms and further reduced. Thus NO emission from denitrifying soils is usually only detected in highly artificial experimental conditions, e.g., when the NO is swept out of a soil by a stream of inert gas (N_2 or helium). Figure 4 shows the general impact of soil wetness on the mean ratio of NO to N_2O emitted from soils. In spite of the characteristic variability associated with observations of this type, there is a clear downward trend in the ratio with increasing soil water-filled pore space (WFPS), with the overall mean falling from near 10 (log (ratio) = 1) at 50% WFPS, to less than unity at greater than 70%.

The only other process of any significance producing N_2O in soils appears to be chemodenitrification. This involves chemical reactions between NO_2^- and organic compounds such as amines under acid conditions, to form N_2, NO, and N_2O, and reduction of NO_3^- and NO_2^- by ions such as Fe^{2+} in subsoils.

The soils of humid tropical forests are the largest natural terrestrial source of N_2O emissions to the atmosphere: estimated at 3 Tg of N_2O-N per year (Table 2). In these soils, nitrogen is more rapidly

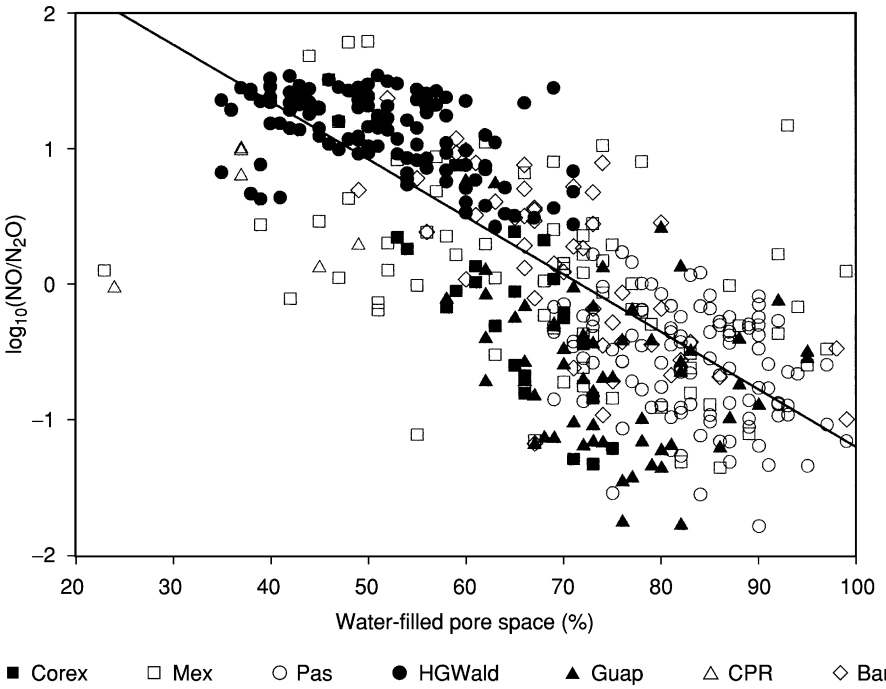


Figure 4 Relationship between soil water-filled pore space and the ratio of NO emitted from soils (g N ha^{-1} day^{-1}) to N_2O emitted (g N ha^{-1} day^{-1}), in seven studies in North and Central America and Germany. Sites: Corex, Pas, GUAP, Ban, all Costa Rica; Mex, Mexico; HGWald, Germany; CPR, Colorado, USA. Drawn from data contained in Davidson EA and Verchot LV Testing the hole-in-the-pipe model of nitric and nitrous oxide emissions from soils using the TRAGNET data base. *Global Biogeochemical Cycles* 14: 1035–1043. (Reproduced with permission from Smith KA, Ball T, Conen F *et al.* (2003) Exchange of greenhouse gases between soil and atmosphere: interactions of soil physical factors and biological processes. *European Journal of Soil Science* 54: 779–791.)

Table 2 Estimated soil sources of nitrous oxide emissions to the atmosphere

Source	*Global annual emission (Tg N_2O-N $year^{-1}$)*
Natural	
Tropical wet forests	3.0
Tropical dry savannas	1.0
Temperate forests	1.0
Temperate grasslands	1.0
Anthropogenic	
Cultivated/fertilized soils	4.2
Total from soils	11.2
Total from all sources	17.7

cycled between organic forms and mineral (NH_4^+ and NO_3^-) forms than in temperate or boreal soils, because of the elevated temperatures and moist conditions. Thus the essential substrates for nitrification and denitrification are in more plentiful supply and the emissions are correspondingly higher. However, emissions from temperate forests, temperate grasslands and tropical dry savannas are not negligible: each is thought to contribute of the order of 1 Tg $year^{-1}$, giving a global total of 6 Tg N_2O-N $year^{-1}$ from all natural soils (Table 2). The uncertainty associated with this estimate is of the order of $\pm 50\%$.

Land-use change from forest to pasture or arable land has been concentrated in recent decades in the tropics. Research indicates that N_2O emissions from such newly created pastures may be several times greater than those from paired forest sites, for the first few years after deforestation, but emissions from pastures more than 20 years old are less than those from forest soil. These emissions from new agricultural land in the tropics constitute only a small fraction of the total emissions from all agricultural land across the globe. The estimated size of the latter has risen gradually, as more and more research has been carried out, and is now put at more than 4 Tg N_2O-N $year^{-1}$ (Table 2), albeit with an uncertainty range bigger than those for the natural ecosystems.

Included within this agricultural total are emissions from land receiving mineral nitrogen fertilizers and from land receiving animal manures, either applied by farmers or directly deposited as the excreta of grazing livestock. The steady increase in consumption of nitrogenous fertilizers and in the size of livestock populations has contributed to a corresponding increase in the size of the agricultural N_2O source to its present-day level. Increases in food production are regarded as essential to meet the needs of a growing population, over the next few decades, and are likely to require increased use of N. This in turn may be expected to increase overall N_2O emissions.

Agricultural Emissions

The emissions of N_2O from agricultural soils are concentrated in those environments where there are substantial quantities of N fertilizer (and/or N in the form of organic manure) applied to the land, and where there are high animal stocking rates. Increases in soil mineral N content as a result of N application are characteristically followed by emission peaks, provided the soil is sufficiently wet (Figure 5). The largest fluxes occur when WFPS values are very high (70–90%), indicating that denitrification is the major process responsible.

On grazed land, the deposition of excreted N is spatially very variable. A small patch of soil surface may receive N as urine at a rate equivalent to some hundreds of kilograms per hectare, whereas land between patches receives none. Also, the treading of the soil surface in wet conditions by animal hooves can create localized compacted zones in which water collects and the soil becomes anaerobic. Such factors lead to the occurrence of 'hotspots' of microbial activity and N_2O emission; overall the average emission rate from grazed grassland may be two to three times that from N-fertilized grass grown as a crop to be cut for winter feed. In western Europe at least, the emissions from cut grassland are, in turn, substantially higher than emission rates from arable land cropped with wheat and barley. This is attributable to a combination of soil physical conditions and the timing of N applications. However, the emissions are very variable between sites and between successive seasons, and for any crop type several years' data are required to obtain a robust estimate of annual N_2O flux.

The response of N_2O emissions to changes in soil temperature can be complex. In conditions in which topsoils are routinely frozen for part of the winter, or are subject to successive freeze–thaw cycles, a substantial part of the total annual emission may occur within a brief period after thawing. This is true of both forest soils and arable land. The principal cause is the development of conditions that stimulate microbial activity, in particular the release of carbon from organic materials in the presence of a high soil-water content. Many research studies have reported nonlinear (exponential) increases in N_2O emission rate with temperature, with values for the Q_{10} for the process (i.e., the ratio of the rate at $T + 10^\circ$ to the rate at T°) ranging up to 10 or more, compared with the general range of 2–3 for most biochemical processes. An example with a Q_{10} of 5.9 is shown

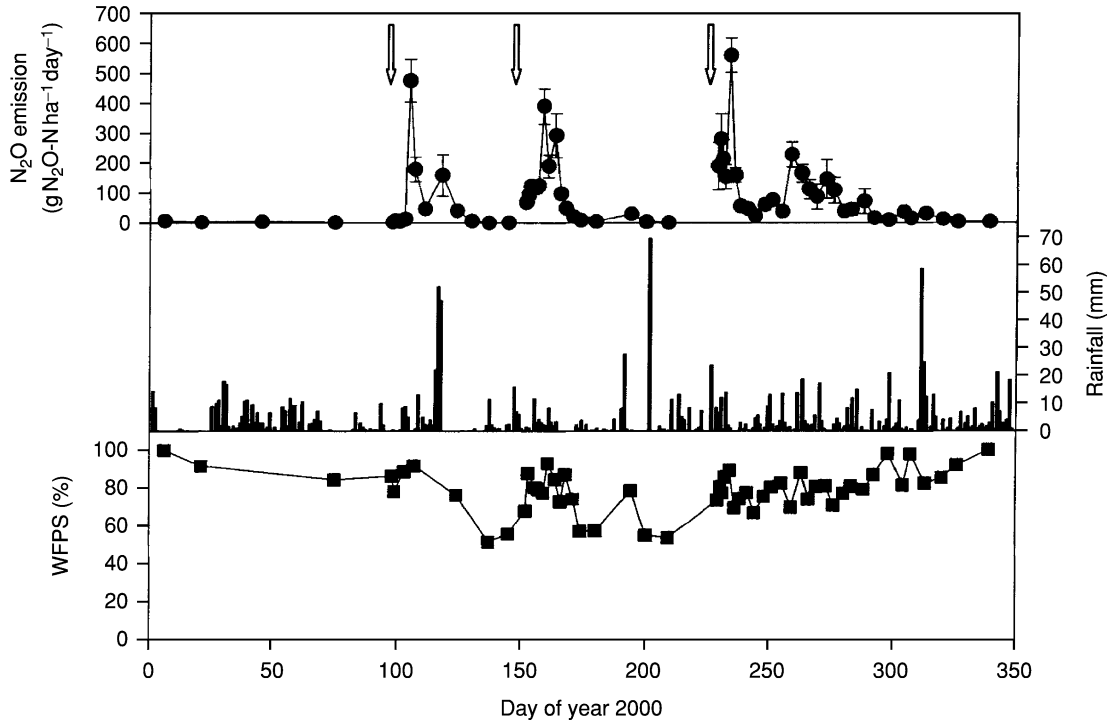


Figure 5 N_2O emissions from a glacial till soil in SE Scotland under grass, fertilized and cut for silage three times per season, together with rainfall and soil water-filled pore space data for the same site. Arrows indicate dates of fertilization (100 kg N as ammonium nitrate on each occasion). (Adapted from Dobbie KE and Smith KA (2003) Nitrous oxide emission factors for agricultural soils in Great Britain: the impact of soil water-filled pore space and other controlling variables. *Global Change Biology* 9: 204–218.)

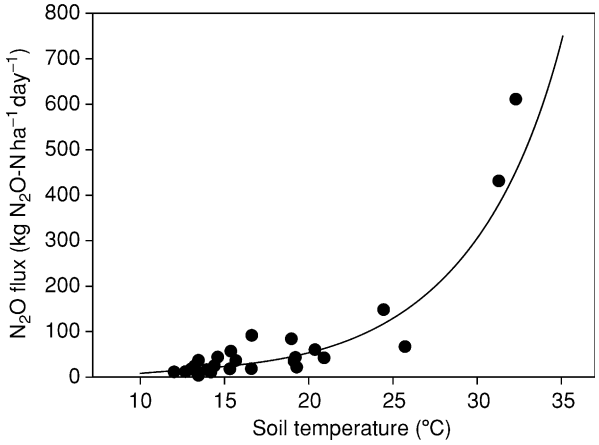


Figure 6 Exponential increase with temperature of the N_2O emission from a grassland soil fertilized with ammonium nitrate. Samples of 40 cm diameter of turf/topsoil enclosed in plastic rings and transferred from the field to the laboratory for incubation at a range of temperatures. Q_{10} for the temperature response = 5.9. (Adapted from Conen F (1994) Influence of temperature on nitrous oxide emissions from soil. MSc Dissertation. University of Edinburgh.)

in Figure 6. An explanation is provided by the theory of anaerobic zone development in soils: all N_2O production by denitrification goes on within such zones, and an increase in temperature leads both to an increase in the size of the zones (because of increased respiration depleting the oxygen supply), and in the rate of denitrification per unit anaerobic volume. Thus the overall change in the rate of N_2O production in a soil mass is the product of these two factors. An increase in soil WFPS, and the consequent decrease in O_2 diffusion rate within the soil, has the same effect, and thus N_2O production/emission increases exponentially with WFPS.

Potential for Reducing N_2O Emissions

The strong link between the amount of N used as a nutrient in agriculture and N_2O emissions means that there is some prospect of reducing emissions by improving the agronomic efficiency with which fertilizer and manure N are used. In general, the procedures recognized as helping in this direction include:

- Using soil/plant testing to determine fertilizer N needs;
- Minimizing fallow periods to limit mineral N accumulation;
- Improving the timing of and/or splitting N applications;
- Using special forms of synthetic fertilizers, such as 'controlled release' materials, which make the

N available for crop uptake over an extended period;
- Adding nitrification inhibitors to urea and ammonium fertilizers;
- Matching the fertilizer type to the soil water regime, e.g., by using an ammonium form in circumstances where wet soil conditions are likely to cause excessive denitrification losses from nitrate forms.

The widespread adoption of these measures could reduce total emissions by 20–25%.

See also: **Anaerobic Soils**; **Carbon Emissions and Sequestration**; **Climate Change Impacts**; **Climate Models, Role of Soil**; **Waste Disposal on Land:** Municipal

Further Reading

Andreae MO and Schimel DS (eds) (1989) *Exchange of Trace Gases Between Terrestrial Ecosystems and the Atmosphere*. Chichester, UK: John Wiley & Sons, Ltd.

Bouwman AF (ed.) (1990) *Soils and the Greenhouse Effect*. Chichester, UK: John Wiley & Sons, Ltd.

Conrad R (1996) Soil microorganisms as controllers of atmospheric trace gases (H_2, CO, CH_4, OCS, N_2O, and NO). *Microbiological Reviews* 60: 609–640.

Firestone MK (1982) Biological denitrification. In: Stevenson FJ (ed.) *Nitrogen in Agricultural Soils*, pp. 289–326. Madison, WI: American Society of Agronomy.

Granli T and Bøckman OC (1994) Nitrous oxide from agriculture. *Norwegian Journal of Agricultural Sciences* (Supplement 12) 1–128.

Houghton JT, Meiro Filho LG, Callander BA *et al.* (eds) (1996) *Climate Change 1995 – The Science of Climate Change*. Cambridge, UK: Cambridge University Press.

Houghton JT, Ding Y, Griggs DJ *et al.* (eds) (2001) *Climate Change 2001: The Scientific Basis*. Cambridge, UK: Cambridge University Press.

Prather M, Derwent R, Ehhalt D *et al.* (1995) Other trace gases and atmospheric chemistry. In: Houghton JT *et al.* (eds) *Climate Change 1994: Radiative Forcing of Climate Change and an Evaluation of the IPCC IS92 Emission Scenarios*, pp. 73–126. Cambridge, UK: Cambridge University Press.

Prinn RG (ed.) (1994) *Global Atmospheric-Biospheric Chemistry*. New York: Plenum Press.

Wassmann R, Lantin RS, and Neue H-U (eds) (2000) *Methane Emissions from Major Rice Ecosystems in Asia*. Dordrecht, the Netherlands: Kluwer.

Watson RT, Zinyowera MC, and Moss RH (eds) (1996) *Climate Change 1995. Impacts, Adaptations and Mitigation of Climate Change: Scientific-Technical Analyses*. Cambridge, UK: Cambridge University Press.

GROUNDWATER AND AQUIFERS

Y Bachmat, Hebrew University, Jerusalem, Israel

© 2005, Elsevier Ltd. All Rights Reserved.

Basic Concepts

While digging a well or drilling a borehole into the ground, two main types of moisture zones might be encountered in the subsurface: An upper, unsaturated zone, in which the void space of the rock contains both liquid water and a gaseous phase (including air, water vapor, and other gases); and a lower, saturated zone in which the entire void space is filled with liquid water only. Water shows up in the well only after it has entered the saturated zone and reached a depth where the water pressure becomes equal to or higher than atmospheric. Such water is termed 'groundwater.'

Groundwater is unconfined if bounded from above by a free surface at atmospheric pressure, the 'water table.' In this case water will appear at the bottom of the well once uncovered. Groundwater is confined if bounded from above by an impervious layer at a pressure higher than atmospheric. In this case the water, once exposed, will rise to a level higher than the bottom of the well.

Although groundwater exists almost everywhere, the focus of attention is usually on groundwater found in geological formations of permeable rock or loose material, which can both contain and transmit significant amounts of water. Such formations are referred to as 'aquifers.' Particularly important among them are aquifers in which the water is renewable, in contrast to aquifers found in desert regions of the world (e.g., the Sahara Desert, the Australian Great Basin, the Sinai Peninsula), which contain fossil groundwater, namely water that has accumulated in the far past and represents a one-time, nonrenewable resource. From here on, only aquifers with renewable groundwater are considered.

Types of Aquifer

Very deep boreholes may encounter two or more layers of aquifers which are separated from each other by impervious formations or semipervious ones, called 'aquitards.' Accordingly, the saturated zone of an aquifer is said to be phreatic or unconfined wherever bounded by a water table, confined wherever bounded at top and bottom by impervious layers, leaky wherever bounded by under- or overlying aquitards, and artesian wherever groundwater is flowing out above the surface of the ground.

Groundwater Basin

Sometimes a group of several aquifers that are interconnected in the subsurface through semipermeable layers or faults may form an independent hydraulic unit. This leads to the concept of a 'groundwater basin,' which resembles a surface watershed. A groundwater basin is defined as a physiographic unit containing several connected aquifers, whose waters are flowing to a common outlet, and which is delimited in the subsurface by impermeable bondaries.

In principle, if water is added to or abstracted from an aquifer at any point, water levels and the patterns of groundwater flow in the entire basin are affected in the course of time. The same applies to the transport and concentration of a pollutant, but on a much longer time scale. The region appropriate for the control of groundwater quantity and quality should, therefore, be a groundwater basin.

A groundwater basin usually crosses the boundaries of overlying administrative areas of jurisdiction in one country and, often, also boundaries between countries. Hence, any analysis of the state of groundwater within the limits of a political or an administrative unit must take into account the effect of subsurface transport of groundwater and its constituents across the boundaries of the unit.

Prominent Features of an Aquifer

Large Areal Extent Compared with Saturated Thickness

Aquifers may stretch over hundreds and thousands of square kilometers underground, whereas the thickness of an aquifer may range from tens to hundreds of meters only. For this reason aquifers are often modeled as large but thin sheets of porous material with properties that are uniform throughout their depth and may vary only on the surface of the sheet. Such a two-dimensional representation, referred to as a hydraulic model of an aquifer, is widely used in the study of groundwater flow on a regional scale whenever water quality variations with depth are practically negligible.

Permanently yet Slowly Renewable Reservoir

Groundwater stored in an aquifer is in a state of permanent motion from areas of natural and artificial recharge to natural and artificial drainage outlets. Renewable groundwater is actually part of the hydrologic cycle, a schematic representation of which is given in Figure 1. It shows a permanent circulation of water from the oceans and land masses to the atmosphere by evaporation and transpiration and then, by condensation and precipitation, on to the surface and back to the oceans via streams or aquifers.

It is worth noting that most of the precipitation on land (in the order of 70%) is lost to the atmosphere by evapotranspiration. But most of the remainder (in the order of 25%) percolates via the unsaturated zone downward to the saturated portion of an aquifer and moves as groundwater from areas of recharge to natural discharge outlets on the surface of the ground, such as springs, marshes, streams, lakes, and oceans (see Figure 2).

It is evident that the recharge of an aquifer from precipitation and/or from other surface sources of water can only occur in areas where the unsaturated zone is permeable and the saturated portion of the aquifer is unconfined. The possibility of direct contact with surface sources of water and pollutants makes the unconfined portions of an aquifer more sensitive to changes in groundwater quality and storage than portions of the aquifer in which the groundwater is confined between impermeable layers.

The movement of groundwater within the saturated zone is usually in the order of meters per day and even per year. As a result, the residence time of a water particle in the saturated zone (i.e., the time it takes a particle to travel from its point of entry into the saturated zone to the point of natural discharge) may range from several weeks to thousands of years. The aforementioned is typical for most aquifers composed of granular material such as sand or gravel. Much faster, yet exceptional, flow velocities may prevail in aquifers of solid rock that is highly fractured or in limestone aquifers, termed 'karst,' where groundwater has dissolved the rock material to form large channels or even caverns.

Large Storage Capacity Compared with Annual Recharge/Discharge

The capacity and volume of groundwater stored in an aquifer are usually large compared with its annual recharge/discharge volume. Under such

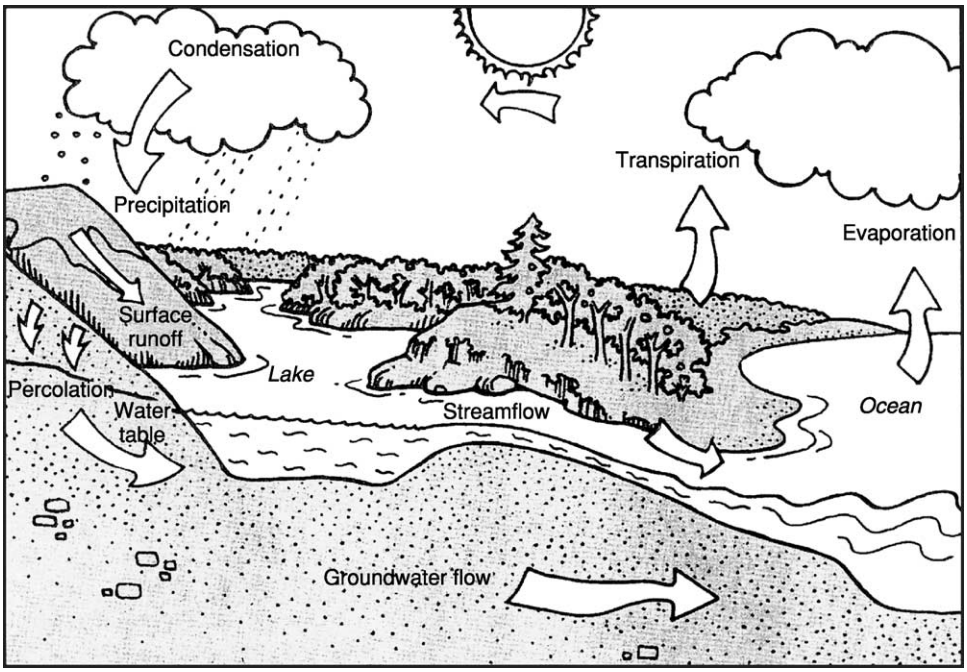


Figure 1 The hydrologic cycle. (Adapted from Bachmat Y (1994) Groundwater as part of a water system. In: Zoller U (ed.) *Groundwater Contamination and Control*. New York: Marcel Dekker, Inc. with permission.)

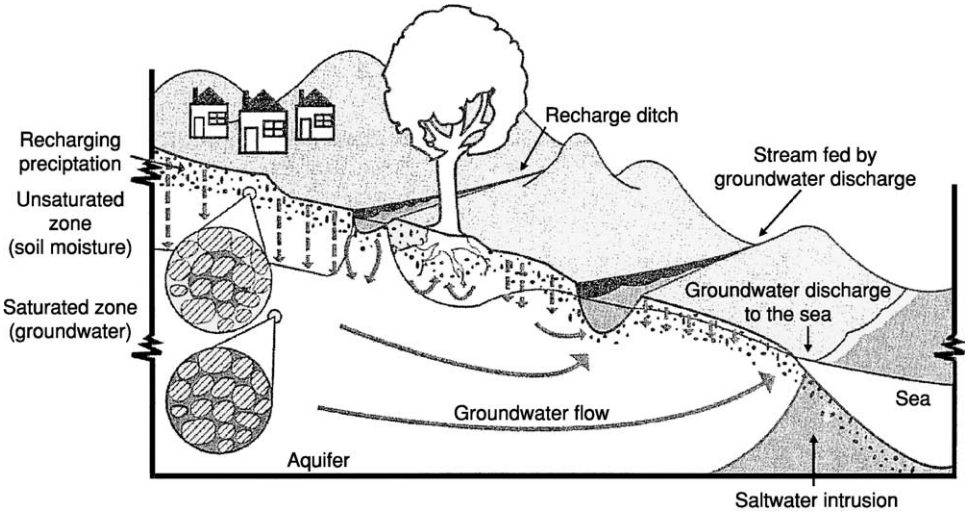


Figure 2 Groundwater flow. (Adapted from Bachmat Y (1994) Groundwater as part of a water system. In: Zoller U (ed.) *Groundwater Contamination and Control*. New York: Marcel Dekker, Inc. with permission.)

circumstances an aquifer can absorb fluctuations in annual recharge without long-term effects on the total volume or quality of the water stored in it.

Hidden Medium

The internal configuration of the void space of an aquifer is not accessible to direct observation. Moreover, samples of various types of rock material provide evidence of its complexity and diversity. Under such conditions, the spatial and temporal pattern of groundwater flow and its response to external effects can only be inferred by the conjunctive use of scientific knowledge and field records of quantities that are both relevant and observable; for example, water level observations in a piezometer.

A piezometer is a small-diameter borehole placed in the ground with an upper end open to the atmosphere and the lower end in contact with groundwater. The water level in a piezometer abve a given reference plane (e.g., the mean sea level) is the 'piezometric head.'

Time series of observations of the piezometric head at different locations and depths in the saturated zone of an aquifer are used by the theory of groundwater flow for assessing the volume and flow rates of groundwater in the aquifer in the course of time.

Functions of an Aquifer

The features of an aquifer listed above can be used to satisfy a multitude of human and environmental needs.

Easily Accessible and Durable Source of Water

The use of groundwater is facilitated by the fact that, within the domain of the areal extent of an aquifer, groundwater is found everywhere and often at small depths, ranging from tens to several hundreds of meters. Therefore, groundwater has always been an important – and in some parts of the world a major or even unique – source of fresh water supply. Groundwater is used primarily for irrigation and as a a source of drinking water in many parts of the world where surface water is nonexistent or extremely costly to develop. However, even in countries where surface water is abundant, the generally good quality of groundwater, its accessibility, and its relatively low cost are factors that stimulate an increasing level of exploitation of this resource. In the USA, for example, groundwater supplies only 25% of the fresh water used for all purposes, but approximately 50% of the population relies on groundwater as its primary source of water for drinking. In rural areas the percentage is even higher. Moreover, the use of groundwater in the USA is increasing more rapidly than the use of surface water. In Canada, 26% of the population rely on groundwater for domestic use. Two-thirds of these users live in rural areas and the rest in small municipalities, where groundwater is the primary source for their supply of water. In Israel, a semiarid country, groundwater provides three-quarters of the fresh water supply.

Subsurface Storage Reservoir

Aquifers can also be used for storing surplus amounts of water from the rainy season to the dry one, as well as from years of exessive precipitation to years of drought. Underground storage of surface water is carried out either by means of artificial recharge wells, which inject the water directly into the saturated zone of the aquifer, or by storing the water in natural depressions or artificial ponds, referred to as 'spreading grounds,' and letting the water drain through the unsaturated zone down to the groundwater. Obviously, this form of indirect recharge is practical in areas where the unsaturated zone is permeable and allows an undisturbed percolation to the underlying unconfined groundwater. In general, preference is given to artificial recharge in unconfined portions of an aquifer where groundwater has free space to accumulate on top of the water table.

An additional advantage of subsurface storage of water is the fact that it saves losses of land and losses by evaporation of water which are associated with storage of water on the surface of the ground. An example of aquifers serving as seasonal and long-term storage reservoirs of water is well illustrated by the water supply system in Israel. Israel's national water supply is based on three major elements: a lake (Lake Kinneret) that drains the water of the Jordan river basin in the northern part of the country; two major aquifers (coastal aquifer, composed of sand and sandstone, and Yarkon–Taninim aquifer, composed of karstified limestone and dolomite) in the central part of the country; and a transportation–distribution system that interconnects the above sources and links them with users. The main conduit of this system is the National Water Carrier, which transports water from the lake to the central and southern semiarid parts of the country. Both the inflow into the lake and the natural replenishment of the aquifers are highly variable, and a storage capacity of 4.5 billion cubic meters is needed for full interannual regulation of the national water supply system. Except for the lake, which has only 10% of this capacity, there are no other favorable surface storage options. Therefore, the main burden of providing the needed volume and facilities for the regulation of water supply falls on the two aquifers. Water brought by the National Water Carrier from the lake and injected into the aquifers, in the rainy winter season and in years of heavy rainfall, is repumped and transported back to the National Water Carrier to meet water needs in the summer season and in years of drought.

Facility for Improving Water Quality

The natural filtering capacity of an aquifer as a porous medium, and in particular processes of adsorption, ion exchange, dilution, and chemical and biochemical reactions that prevail in the unsaturated zone, can reduce nutrients, carbonaceous oxygen demand, certain metals, and bacteriological pathogens. The Dan Region Wastewater Reclamation Plant, located in the sand dunes of the coastal plain of Israel south of Tel Aviv, may serve as an example. This plant handles at present an annual amount of nearly 100 million cubic meters of municipal sewage collected from the greater metropolitan area of Tel Aviv. After some biological and chemical

pretreatment, the partially treated effluent is brought to spreading grounds and left to percolate downwards through the unsaturated zone, which is used as a natural filter. Upon reaching the groundwater table, the water continues to move through the saturated zone, to be intercepted by a chain of pumping wells and transported to the dry south for irrigation of industrial crops.

Aquifer as a Conduit

The large spatial extent of an aquifer permits water that enters the saturated zone by natural or induced replenishment to flow long distances from the region of recharge to points of discharge or artificial withdrawal. The proximity of groundwater to individual consumers spread out over large areas saves the need for large water conveyance and distribution systems overland.

Aquifer Characteristics

Some geometric parameters related to the capacity of an aquifer to store and transmit water are:

- Porosity: volume of voids per unit elementary volume of the aquifer;
- Effective porosity: volume of interconnected voids per unit elementary volume of the aquifer. While porosity is a local measure of the total water storage capacity, effective porosity is a local measure of the aquifer's capacity to store movable water only. Hence the important role of effective porosity in the transport of groundwater and its constituents through the saturated zone of an aquifer;
- Specific storativity: volume of water added to (or released from) storage per unit volume of the aquifer per unit rise (or decline) of the piezometric head.
- Permeability (or intrinsic permeability): a characteristic of the configuration of the void space which affects its ability to transmit a fluid. The permeability is determined by three factors: (1) the effective porosity of the aquifer material; (2) the surface area of the solid matrix of the aquifer which exerts a resistance to flow. The smaller this area per unit volume of the void space, the larger the permeability of the aquifer material; (3) the tortuosity of the actual flow paths. Tortuosity is the ratio of the actual length of a flow path measured between two points to the shortest distance between the two points. The smaller the tortuosity, the larger the permeability.

Another geometric parameter of an aquifer which plays an important role in the transport of a solute is dispersivity. It represents a measure of the capacity of the aquifer to cause spreading of a solute. While permeability depends on the average value of tortuosity in the vicinity of a point in the aquifer, dispersivity depends on the variance of the tortuosity around a point.

Aquifers composed of hard rock material are rigid in the sense that they are not subject to changes of porosity due to changes in external load on the aquifer or changes of water pressure inside the saturated zone of the aquifer. The contrary is true for aquifers containing soft sediments, which are deformable, such as clay. Examples of severe environmental problems associated with the deformability of such aquifers are land subsidence due to compaction because of the drop in water pressure as a result of pumping out large amounts of water from the aquifer, or swelling because of increase in water pressure as a result of injecting large amounts of water in the confined portion of an aquifer.

Aquifers are usually heterogeneous, meaning that the properties listed above may vary from point to point within the aquifer. Aquifers may also exhibit anisotropy, meaning that the permeability at a point within the aquifer may be different in different directions. A typical example of anisotropy is permeability along a deposited layer of sediments being different from the permeability across the layer. Anisotropy is also typical for limestone aquifers where dissolution of rock material has formed large channels with preferred flow directions.

Groundwater State and Behavior

The state of groundwater in an aquifer is described by the values of a set of variables which represent physical, chemical, biological, and other properties of the water. The selection of the set of these variables, referred to as state variables, is not unique. It depends on requirements related to the intended use of the water. Common examples of state variables of groundwater are depth to groundwater, velocity, pressure, concentration of dissolved and suspended substances, and temperature.

The state of groundwater in an aquifer at a given point in time is nonuniform, namely it varies from place to place. Usually, the state of groundwater is also nonsteady, namely it varies in the course of time. The behavior of groundwater in a given period of time is the sequence of states undergone by the groundwater during that period. Knowing the behavior of groundwater in a given aquifer in the past is helpful in recognizing the factors which govern the behavior of the groundwater and in establishing quantitative cause–effect relationships between these factors and the state variables. This knowledge, in turn, can be used for predicting the behavior of the groundwater in the future. The branch of science

which studies the behavior of groundwater is 'groundwater hydrology.'

Levels of Groundwater Study

The behavior of groundwater can be studied at two levels: (1) the microscopic level, which follows the movement and properties of individual water particles within the actual configuration of the void space; (2) the macroscopic level, which views the groundwater and the solid matrix as material continua occupying simultaneously the entire saturated zone. At this level of description, the value of a state variable of the groundwater at each point is an average of local values within an elementary volume around the point, referred to as representative elementary volume (REV).

The behavior of groundwater at the microscopic level is sometimes studied in research laboratories. In the field, however, observations and study of groundwater behavior are possible and of practical interest at the macroscopic level only. Therefore, and because of the ease of mathematical treatment, groundwater hydrology uses the macroscopic level of description. It is worth noting that the parameters of an aquifer presented above make sense and could be defined at the macroscopic level only.

A more simplified level of description of groundwater behavior is the hydraulic level, at which state variables and parameters are averaged over the entire depth of the saturated zone, thereby reducing the description of their variation to two spatial coordinates only. As stated earlier, the hydraulic level of description is applied wherever the variation of the values of parameters and variables of interest along the vertical are negligible in comparison with their averages.

Tracing the Movement of Groundwater

The movement of groundwater in an aquifer is described by a variable termed 'specific discharge.' It is defined at any point as the volume of water which passes per unit time through a unit area perpendicular to the direction of flow at that point. The specific discharge is also termed 'apparent velocity of the groundwater,' since it assumes that the flow takes place through the entire unit area, while the actual flow is through the area of the voids only. The specific discharge is therefore smaller than the actual average velocity by a factor equal to the effective porosity defined earlier.

Since specific discharge cannot be measured directly, attempts have been made to estimate it by indirect observations. In 1856, a French engineer, Henry Darcy, investigated the flow of water in vertical, homogeneous sand columns. He concluded that the discharge per unit cross-sectional area of a column is directly proportional to the difference in water level elevations in the inflow and exit reservoirs of the column, and inversely proportional to the length of the column. The findings of Darcy have been adapted to flow in a porous medium and given the name 'Darcy's law.'

Definitions pertinent to Darcy's law include:

- Hydraulic slope: the drop of the piezometric head per unit distance in a given direction;
- Hydraulic gradient: the vector whose direction and magnitude are those of the largest hydraulic slope. The hydraulic gradient at any point is perpendicular to the surface of constant piezometric head at that point.

Darcy's Law

The specific discharge of a homogeneous fluid (e.g., groundwater of uniform density) at any point of a homogeneous and isotropic porous medium is directly proportional to the hydraulic gradient at that point, states Darcy's law.

The coefficient of proportionality, 'hydraulic conductivity,' is directly proportional to the permeability of the medium, defined earlier. It also depends on the properties of the fluid. The heavier the fluid, the higher its conductivity; the higher the viscosity of the fluid, the lower its conductivity.

Darcy's law, wherever applicable, enables determination of the specific discharge at any point in an aquifer at any time by means of the piezometric head, which is observable. Hence the importance of monitoring and mapping the piezometric head for the assessment of groundwater flow directions and volumes.

An example is the determination of streamlines and discharges in a hydraulic model of an aquifer at any given time. According to Darcy's law, at any point in the groundwater flow domain, a line perpendicular to a contour line (i.e., an isoline) of the piezometric head is a streamline. Hence, observations of the piezometric head in a set of piezometers can be used to draw isolines of the piezometric head and, subsequently, lines perpendicular to them which are streamlines. Then, by knowing the depth of the saturated zone and using Darcy's law, one can compute the discharge (i.e., the volume of groundwater flow per unit time) between any two streamlines.

Darcy's law has also been derived analytically. Indeed, a fluid moving through the void space of a porous material is subjected to the driving forces of gravity and pressure differences on the one hand, and

to the drag force of the solid phase which opposes the movement on the other hand. The piezometric head at any point represents the sum of the gravitational and pressure energy per unit weight of the fluid at that point, and the drop of energy per unit length of a flow path is the sum of the driving forces. Assuming that the drag force is proportional to the specific discharge and the acceleration of the fluid in the neighborhood of a point is negligible, one arrives at Darcy's law.

Darcy's law, wherever applicable, is widely used in groundwater flow models to represent flows across subsurface boundaries within the saturated zone, water exchange between groundwater and surface water bodies (rivers, lakes, the sea, reservoirs) as well as outflow to springs and drainage facilities.

Further extensions of Darcy's law were also made for the cases of an anisotropic aquifer, and a nonhomogeneous fluid. It is worth noting that, in an anisotropic aquifer, the direction of flow and hence also of the specific discharge does not necessarily coincide with the hydraulic gradient, which makes the computations more complicated. In the case of a fluid of variable density, the piezometric head becomes irrelevant and is replaced by the pressure as the relevant state variable.

At this point it is worth becoming familiar with two parameters which are important in formulating a hydraulic, essentially horizontal, groundwater flow model:

- Storativity or storage coefficient: the volume of water an aquifer releases from (or takes into) storage per unit surface area of the aquifer per unit decline (or rise) of the piezometric head. Storativity, at a given point in the horizontal plane, equals the specific storativity, defined earlier, multiplied by the saturated thickness of the aquifer at the given point;
- Transmissivity: the discharge of groundwater through a unit width of the saturated zone as a result of a unit hydraulic gradient. Transmissivity, at a given point in the horizontal plane, equals the hydraulic conductivity, defined earlier, multiplied by the saturated thickness of the aquifer at the given point. In an anisotropic aquifer, the transmissivity at a point depends on both the direction of the normal to the area of the unit width, and the direction of the unit slope. In an isotropic aquifer, the transmissivity is a number independent of direction.

Aquifers and Groundwater as Part of a Water Resources System

At early stages of groundwater exploitation, withdrawals and uses of water in a region are sparse, and their rates constitute a negligible fraction of the total discharge. Under such conditions the natural components of the mass balances of groundwater and its constituents in the region remain essentially undisturbed. However, as populations grow and concentrate on small tracts of land, and as economic development advances, withdrawals and uses of water increase in intensity and variety. As a rule, human settlement and economic activity are also accompanied by uses and releases of substances that, either directly or indirectly, reach natural water bodies, including groundwater. Thereby a new type of a complex water resources system is formed in which natural and anthropogenic components interact with each other and modify the original state of the natural water sources. A schematic diagram of such a system in a region is depicted in **Figure 3**.

The links of the aquifer-groundwater subsystem with other parts of the water resources system can be used in satisfying human needs related to water. Thus, as stated above, groundwater may serve as a reliable and durable source of fresh water for various uses; groundwater levels can be manipulated to regulate springflow as well as the exchange of water and solutes between aquifers and adjacent surface water bodies (streams, lakes, sea). Costly water conveyance and distribution systems can be avoided by utilizing aquifers as conduits for transporting water over large distances and/or to individual users spread out over large areas.

Contrary to the favorable effects listed above, interactions between groundwater in a region and its environment may also produce effects unfavorable to both the groundwater and the environment. The most common of these effects are:

1. Unfavorable effects of the environment on groundwater:
   - Pollution of groundwater by indirect recharge from anthropogenic sources of pollution via the unsaturated zone and/or by release of pollutants that have been accumulated in that zone;
   - Contamination of groundwater by influx of contaminants from natural surface or subsurface water bodies;
   - Contamination of groundwater by direct injection of contaminated water;
   - Accumulation of contaminants in areas of groundwater level depressions caused by concentrated pumpage;
   - Loss of groundwater to undesired outlets, e.g., contaminated surface water bodies or aquifers;
   - Continued decline of groundwater levels by reduction of recharge as a result of land-use alterations and/or by excessive withdrawal;

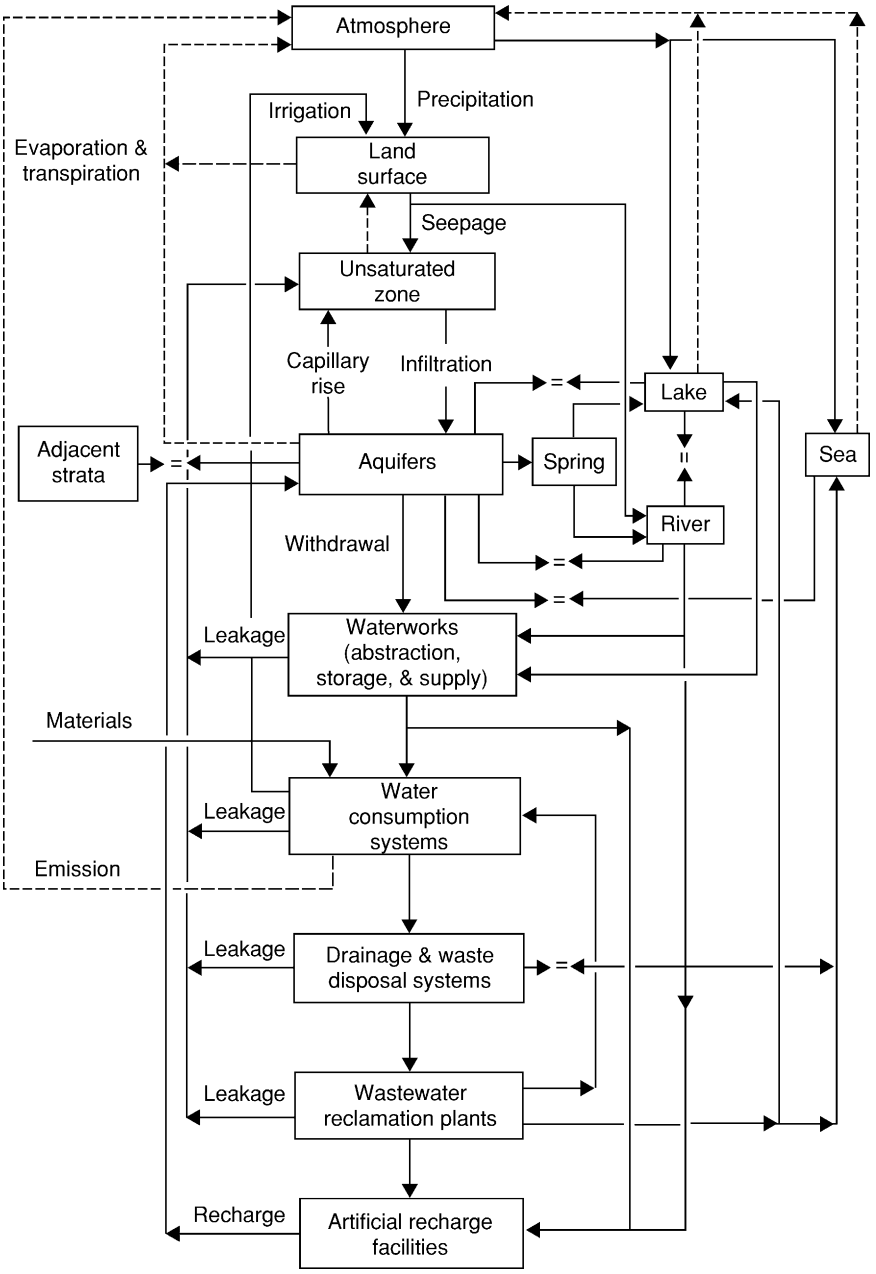


Figure 3 Water resources system. (Adapted from Bachmat Y (1994) Groundwater as part of a water system. In: Zoller U (ed.) *Groundwater Contamination and Control*. New York: Marcel Dekker, Inc. with permission.)

2. Unfavorable effects of groundwater on the environment:
   - Abandonment of pumping wells in areas of contaminated groundwater;
   - Contamination of streams or lakes by contaminated groundwater;
   - Damage to vegetation, crops, soils, structures, and land surface caused by a rise of the groundwater table and/or by deterioration of groundwater quality;
   - Decline of flow and/or water levels in streams and/or lakes as a result of reducing the outflow of groundwater to these water bodies;
   - Reduction or cessation of springflow as a result of declining groundwater levels;
   - Reduction of groundwater outflow to adjacent regions as a result of lowering the groundwater levels in the region;
   - Land subsidence caused by depletion of an underlying compressible confined aquifer.

All of the unfavorable effects listed above may have repercussions on health, material wealth, and other aspects of well-being of the population, not only within the region under consideration, but also beyond it.

Groundwater Contamination Sources and Events

The natural processes and anthropogenic activities that generate groundwater contamination are many and varied, and so are their sources. The sources of groundwater contamination are grouped in two categories: point and nonpoint sources.

Examples of anthropogenic point sources are: open solid-waste dumps; landfills; septic tanks; cesspools; injection wells of inferior-quality water: aboveground and underground storage tanks of hazardous and nonhazardous waste materials; gas stations; poorly constructed production or monitoring wells, which serve as conduits for the vertical transfer of pollutants, either from surface sources to the saturated zone or between geologic layers in the subsurface; and spills from transport facilities. The effect of a point source is, in many cases, local and therefore relatively easy to be monitored and ascribed to the source.

Examples of anthropogenic nonpoint sources are land application of chemicals in agriculture, irrigation with poor-quality waters, percolation of atmospheric pollutants, urban runoff, and patterns of groundwater pumpage and/or groundwater contamination by interaction with surface water bodies. These produce large-scale effects in both time and space, which are often overlapping and difficult to monitor and relate to each source separately.

There is a vast literature of reported groundwater contamination events in all parts of the world. Among the naturally occurring sources of groundwater contamination, whose discharge is created and/or exacerbated by human activity, the best known is seawater intrusion in coastal aquifers. Under undisturbed natural conditions, a state of dynamic equilibrium is maintained whereby flow of fresh groundwater to the sea keeps the landward encroachment of seawater into the aquifer within certain limits. However, as pumpage from the aquifer develops, causing reduction of outflow to the sea and subsequent lowering of the groundwater table (or the piezometrtc surface in a confined aquifer) near the coast, the dynamic balance between the freshwater and seawater is disturbed, permitting seawater to advance inland and contaminate usable parts of the aquifer. As long as the level of pumpage from the aquifer does not exceed its replenishment, a new equilibrium may be reached. Otherwise, seawater intrusion will proceed to contaminate an ever-increasing part of the aquifer. It is worth noting that, although the process of seawater intrusion is slow, it is practically irreversible. Field observations have shown that, once an area has been invaded by seawater, all pumping wells in that area will sooner or later be contaminated and abandoned.

Seawater intrusion has been reported in several parts of the world and is regularly followed up in the Netherlands and in Israel. The best-documented cases in the USA include Long Island, Miami, and many parts of California.

In the USA, contamination of groundwater is usually ascribed to the migration of harmful substances from waste-storage sites, landfills, and farms. According to the Environmental Protection Agency (EPA), industrial wastes are the primary source of groundwater contamination in the USA. Numerous cases of groundwater contamination, particularly from landfills, mine-tailings, and fertilizers, have been reported in Canada. In Israel, the focus of attention is on groundwater contamination by chlorides and nitrates in the coastal aquifer, which is one of the country's major sources of water supply. Originally, when pumpage was concentrated primarily along the coastline, the main source of contamination was considered to be seawater intrusion. As of 2003, when the major well fields are farther inland and concentrations of contaminants in pumped water are nevertheless increasing, the major items of attention are anthropogenic sources on the surface of the ground. An investigation of groundwater contamination by nitrates in a typical part of the coastal aquifer, based on data from 391 wells over an area of 87 km^2, has shown that approximately 61% of the nitrogen in the groundwater originates from fertilizers, 16% from irrigation water, 10% from sewage and refuse, and the remaining 13% is shared almost equally by groundwater inflow through lateral boundaries, urban runoff, and rainfall.

In several European countries with intensive farming (France, Hungary, Russia, the UK, Czechoslovakia, Germany, the Netherlands), a close statistical relationship has been found between the amount of nitrogen fertilizer applied to farmland and the nitrate concentration in groundwater, indicating that under current fertilization practices nitrogen input exceeds its uptake. Contamination of groundwater from domestic, municipal, human, and animal waste is a major problem in many countries in Africa, Asia, and Latin America. Parts of urban and rural environments

in many developing countries, such as India, Kenya, Nigeria, Sudan, Egypt, Iraq, and Brazil, are being contaminated by waste products generated in large quantities as a result of rapid industrial development. Intensive, uncontrolled use of fertilizers and poor farming practices is also reported to be a major cause of groundwater contamination in developed countries.

It is worth noting that, except for particular cases of direct contact between local sources and outlets of water and/or contaminants, the occurrence of any of the above effects is not immediate, namely, there is a time lag between the onset of the cause and the occurrence of its effect. This fact is of particular significance with regard to the contamination of groundwater and/or its impact on the environment. Thus, a pollutant released from a surface source of contamination, within a groundwater recharge area, may reach a certain pumping well after many years, or it may never be observed there. Indeed, the path of such a pollutant and the evolution of its concentration are determined by a multitude of factors, such as the depth of the unsaturated zone, the flow pattern within the saturated zone and, hence, the distance to be traveled by the pollutant within this zone toward the pumping well, the pattern of pumpage (locations and rates), and, finally, the properties of the two zones that affect the rate of transport and transformation of the pollutant.

It may thus be concluded that processes of groundwater contamination may remain undetected for long periods of time, turning short-term benefits from invisible polluting activities, at present or in the past, into visible long-term losses from these activities in the future. Similarly, once contamination of groundwater and/or its impact on the environment has been detected, it may require a very long and costly effort to see the desired result of remedial actions if it is at all practically realizable.

The case of the Aswan Dam in Egypt serves as an example of groundwater pollution problems and environmental damage that can be exacerbated by man-made modifications of the natural water system. Groundwater levels in lower Egypt rose considerably with the building of the Aswan Dam in 1965. The regional rise in water levels of shallow aquifers compounded the problems of pollution spread through subsurface disposal of wastewaters and irrigation water. It was reported that "villages are experiencing considerable difficulties with wastewater disposal due to subsurface saturation, high groundwater, emerging surface pools of septic waters, gross groundwater pollution, deterioration of buildings and structures due to moisture absorption and other related problems."

Groundwater and Aquifers Need to be Managed

The benefits that can be drawn from the links of groundwater and aquifers with their physical environment, on the one hand, and the possible adverse effects of those links, on the other hand, lead to the conclusion that the behavior of groundwater and aquifers needs to be controlled. In order to give it a chance, groundwater within a given administrative region has to be dealt with concurrently as both part of a basin and part of the water resources system in the region. Assuming that the overall objective is to improve the economic well-being and the environmental conditions of the region's population, the task of the water resources system, including its groundwater subsystem within that region, would be to supply water for direct consumption (e.g., drinking, recreation), for the production of (industrial and agricultural) goods and power, and for the provision of services (e.g., navigation). However, the latter two are inevitably accompanied by application and release of contaminants which, as shown earlier, may contaminate and lower the intended economic value of the water and/or harm the health of direct users of the water. Uncontrolled abstraction of groundwater as a source of water supply may also be accompanied by losses from mining groundwater reserves and from influx of contaminants from other natural or anthropogenic sources.

This leads to a second conclusion, namely, that an appropriate balance has to be found between actions aimed at improving the economic well-being of the population in a region and those aimed at securing the quality of the environment, including the natural waters. The term 'appropriate' is used here in the sense that all parties and interests involved are satisfied to a degree that no one's well-being can be further improved without hurting someone else. An adequate framework for accomplishing this task may be an integrated regional water resources management system, whose function is to plan and implement means of control, related directly or indirectly to water, in a way that renders the highest net benefit from the water resources system to the entire population of the region. The activities concerned include withdrawal of water from aquifers, lakes, and streams; artificial recharge of the above; import and export of water; water treatment; reclamation and reuse of wastewater; allocation and mode of supply of water from the various sources to uses and users; land-use activities (e.g., placement, construction, operation, and maintenance of facilities and land application of substances which may contaminate the water sources); and transfer of water

between components within the water system of the region, and between them and the environment.

As groundwater is part of the water resources system in a region, the management of this system should also comprise a groundwater management subsystem. Its particular task should be to plan and implement the control of groundwater quantity and quality so as to maximize the excess of gains from meeting groundwater-related demands, both present and future, over losses incurred by adverse effects of links with the environment within the region and across its boundaries.

Groundwater Management Levels and Models

The groundwater management process is accomplished via a three-level system: (1) the strategic level, which sets long-range objectives, determines decision criteria and constraints, and prescribes policy guidelines; (2) the tactical level, which translates the directives of the strategic level into long- and short-term plans and projects; and (3) the executive or field level, which controls the specification and implementation of the projects.

Each of the above levels of the management process involves the employment of three categories of formal models. First among them are dynamic models, which are used to predict the behavior of the groundwater system in terms of relevant state variables of groundwater quantity and quality under a given pattern of groundwater operation and/or development.

The starting point of building a dynamic model is a hydrogeological investigation aimed at locating the aquifer and the groundwater stored in it, as well as identifying and locating sources and outlets through which water and its constituents can be exchanged between the aquifer and its physical environment. This is followed by performing a sequence of tasks which is summarized in the form of a list of statements and assumptions that constitute the conceptual model of the system. The main items comprising the conceptual model are:

1. Boundaries, subdivision, and interconnection of aquifers and aquitards in the region of interest;
2. Substances of interest (e.g., groundwater, solutes, and/or suspended matter which are pertinent to water quality requirements);
3. State variables which are relevant to the substances (e.g., groundwater density, piezometric head, temperature, concentration of solutes);
4. Sources and outlets of the substances within the domain of interest;
5. Processes of transport and transformation that are relevant to the substances (e.g., groundwater flow, evaporation, mass transport of solutes, chemical reactions within the liquid phase, mass transfer between the solid matrix and the liquid phase) within the domain of interest;
6. The environment of the considered domain and the processes of exchange of mass of the substances and/or heat across the boundaries with the environment;
7. Assumptions about relevant physical, chemical, biological and other properties of the substances.

The next step in the process of building the dynamic model is the translation of the conceptual model into a mathematical statement referred to as the mathematical model of the groundwater-aquifer system. The main items comprising the mathematical model are:

1. Mathematical formulation of the configuration of the boundaries of the system;
2. Formulation of the balance equations of mass (and heat wherever needed) of the substances of interest;
3. Formulation of the transport equation for each substance in terms of relevant state variables;
4. Formulation of source and sink functions of each substance of interest in terms of observable quantities and/or state variables.

   Upon inserting (3) and (4) into (2), the balance equations take on the form of a set of so-called field equations of the state variables of interest:

5. Formulation of initial values of the relevant state variables;
6. Formulation of the boundary conditions which express the transport of each substance across the boundaries of the system in terms of relevant state variables.

The solution of the field equations under the given initial and boundary conditions takes the form of fields of the state variables of interest within prescribed time-and-space domains of the problem at hand. Consider, for example, the construction of a dynamic model of groundwater flow in a given aquifer. The conceptual model of the system states that the substance of interest is groundwater and the variable of interest is the piezometric head. The lateral boundaries of the aquifer form a rectangle; at one side of the rectangle the aquifer is in contact with the sea, whereas all other sides are impervious. The upper boundary of the saturated zone is the water table; the base of the aquifer is horizontal and impervious. The aquifer is

Table 1 Sources and outlets of groundwater in the selected region

Inflows	*Outflows*
Replenishment from rainfall	Pumpage
Return flow from irrigation	Outflow to the sea
Leakage from water supply system on the surface of the ground	
Inflow through artificial recharge wells	

homogeneous, isotropic, and rigid. The groundwater is incompressible, and the density and temperature of the groundwater are uniform and constant.

The sources and outlets of the groundwater are presented in Table 1. The flow of the groundwater is essentially horizontal. Therefore, and owing to the homogeneity of the aquifer and the groundwater, it is sufficient to describe the behavior of the groundwater at the hydraulic level.

The formulation of the mathematical model starts with the volume balance equation of the groundwater (since the density of the groundwater is constant). Table 1 lists the components of the water balance which change the volume of groundwater in the aquifer. The term 'groundwater storage' is often used to denote the total volume of groundwater in the region of interest. The verbal formulation of the global groundwater volume balance reads:

Rate of increment of groundwater storage
= Sum of the rates of all inflows
−Sum of the rates of all outflows

Table 1 shows that only two components of the volume balance are fully controllable: pumpage from wells, and artificial recharge through wells. The rest are not controllable, either fully (replenishment from precipitation) or partly (return flow from irrigation, leakage from surface sources, outflow to the sea). Hence, in order to make the balance equation usable for the purpose of predicting the anticipated change in the state of groundwater (i.e., groundwater level, groundwater flow rates and directions) as a result of pumping or recharge operations, one has to express the uncontrollable terms as functions of relevant measurable quantities (e.g., natural replenishment as a linear function of precipitation depth, return flow from irrigation as a linear function of irrigation depth) and/or relevant state variables (i.e., outflow to the sea and change in groundwater storage as functions of groundwater levels). This step requires utilization of existing scientific knowledge and adaptation of it to the specific conditions in the area of interest. In this example it implies using Darcy's law with the coefficient of transmissivity and the change in storage with the coefficient of storativity defined earlier.

The last step in building the dynamic model is fitting it to the aquifer of interest. This step, referred to as calibration of the model, includes the selection of model parameter values (e.g., the values of transmissivity, storativity, coefficient of natural replenishment and coefficient of return flow from irrigation, in the present example) which ensure a best fit between the historical data of groundwater levels and concentrations with data predicted by the model.

Prediction of future changes of groundwater storage and quality requires data about the configuration, composition, and present state of the groundwater system, data on future inputs of water and contaminants from surficial sources, and also quantitative knowledge about the processes of transport and transformation that may occur within the groundwater subsystem and on its boundaries. Acquisition of this information is hampered by several factors, some of which are specific for a groundwater system. Indeed, any groundwater-aquifer system is hidden and, hence, inaccessible to direct observation. This results in uncertainties concerning its configuration, composition, sources, and outlets. Also, the actual state of the system at a given time is not directly observable. It can only be inferred from assumed relationships with observations of relevant quantities made at discrete locations and times. In addition only a few of the external links of the groundwater subsystem are almost fully controllable. These include direct pumpage and/or recharge. All other links can be grouped into three categories: (1) those that are practically uncontrollable, such as precipitation; (2) those that can only be indirectly controlled by groundwater management, such as flow and mass transfer between the groundwater and external water bodies, whether on the surface of the ground or in the subsurface; and (3) those that are primarily controlled by human activities on the surface of the ground (e.g., irrigation, waste disposal, land-use alterations).

Consequently forecasts of dynamic models are never exact and can at best be expressed in probabilistic terms only. But the error of prediction can still be statistically reduced as more relevant observations are gathered and utilized in the various stages of the modeling process.

The second kind of models are evaluation models, which assign to any course of action and any possible state of the system a set of values of preselected quantitative measures of success ('attributes' or 'criteria') in pursuing the objective(s) of managing the system. The evaluation model can thus be utilized to evaluate the outcome of each alternative course of action, as

predicted by the dynamic model. It should be noted that the evaluation model is also not perfect, for reasons analogous to those of the dynamic model. However, like a dynamic model, its level of reliability can also be raised by means of relevant observations.

The third type of model is the model of choice. It receives the output of the evaluation model for different courses of action and selects according to a chosen method the alternative that is most attractive from the point of view of reaching the objectives of the system. The model of choice may have to cope with two problems: uncertainty in the information received from the evaluation model and multiplicity of noncommensurable attributes such as measures of health, level of contamination by various contaminants, and financial loss.

All three of the above models are sometimes included in a single optimization model. Although formally attractive, the use of optimization models in practice is limited. The reasons for this situation are twofold: The limited capability of an optimization model to represent a complex real system, and the reluctance of decision makers to accept a prescribed course of action without taking part in the process of making a choice.

At the strategic level, one model may represent an entire water sources system. The groundwater subsystem may be represented in this model by several basins, each as a single cell. The same may hold for the surface-water subsystem. Another model at the strategic level may represent the water-supply system, a third, the wastewater disposal and reuse system, and a fourth, the economics of the water resources system. All of the above models are linked to each other. The focus of interest is not so much on the details of each of the subsystems but rather on their interrelation and the resulting long-term consequences.

At the tactical level, each subsystem is modeled separately with an emphasis on the behavior of each of its parts and their interrelations. Obviously, models representing the same system at different levels of aggregation must satisfy certain conditions of compatibility. Of particular interest, in this regard, may be the actual application of a set of such models in Israel in developing a strategy of integrated management of its water resources using the systems approach and multilevel modeling.

To be applicable and useful in practice, all of the above models need data for their formulation, calibration, validation, and operation. This, in turn, requires data-management models, including models for the design of observation programs, models for data quality control, and models for transforming observations to processed data, which serve as inputs to the aforementioned models.

Use and Usefulness of Formal Models in Groundwater Management

Although much progress has been made in the field of development of the above types of models, their actual use and usefulness in the management of groundwater is still wanting. The main reasons for this seem to be:

1. Lack of a mandatory, clearly defined, and duly implemented methodology and procedures of utilizing quantitative models, and information derived therefrom, in the decision-making process;
2. Low credibility of the models, especially those pertaining to groundwater quality.

The situation can be improved by advancement of knowledge about processes of groundwater contamination and their environmental effects, as well as by adoption and improvement of methods of system identification, modeling, analysis, and monitoring. Progress in the above fields and assimilation of the knowledge in practice require an interdisciplinary effort and present a challenge.

Another hindrance to the effective use of formal models in groundwater management may be relatively loose links between modelers and technical staff of model users, as well as between the latter and the managers. The remedy for this is practical training in the fields of interdisciplinary communication and interactive formulation and solution of problems.

Anthropogenic Constraints on Groundwater Management

No management system operates in a vacuum. The same applies to the management of water resources in general and groundwater in particular. Laws and regulations with respect to water, rules prescribed by the socioeconomic and political system of government, along with limited resources and know-how, are constraints that restrict the feasible action space of the water management system.

Legal constraints are often a major obstacle to efficient management of groundwater. For example, despite the fact that groundwater and surface water are interrelated, several states in the USA still legally define the two as being physically disconnected. But the most powerful legal constraints are the property rights on sources of water. Improperly defined property-right systems preclude efficient utilization of groundwater and stimulate overexploitation and contamination.

One type of improperly defined property rights, which, for example, still exists in certain states in

the USA, treats groundwater as a common property resource, i.e., as a resource that is not exclusively controlled by a single authority, but by many individual owners, each of whom is entitled to full control over drilling and extraction of groundwater within the limits of a surface area, which is small in comparison with the geographic extent of the relevant aquifer. Although the actions of individual users affect the state of the groundwater beyond their area, the property right protects them from bearing all of the consequences of their actions. Losses from overexploitation and/or pollution of the aquifer, which are taken into account in the case of exclusive ownership, are an externality to the individual owner, and not a part of their decision-making process. Hence, wherever property rights treat groundwater as a common property resource, it cannot be managed efficiently, because the property rights allow unlimited use of the resource and discourage conservation.

Another type of property right that precludes efficient management of groundwater treats water as a public good, which is fully accessible to all, as if it were an unlimited resource such as air. It is claimed that legal constraints of the aforementioned types are being gradually relaxed as the fundamentals of hydrology become better understood by public officials, legislators, and the courts.

The Water Law in Israel has declared all waters as public property subject to state control and intended to be used for the benefit of its inhabitants and for the development of the country. Although, as in this case, exclusive property rights on water are a necessary condition for efficient management of water resources, they are not yet sufficient. Political and administrative constraints, such as the placement of the water management system within the system of government, and the role of the central government in the management process, may have a major effect on the effectiveness of control over this vital resource.

In Portugal, for example, management of water affairs is under the Secretariat of State for Environment and Natural Resources, which resides within the ministry of Regional Development and Land Use Planning. The placement of water management within an intersectorial ministry, which must integrate different strategies of different sectors in the economy, eliminates the possibility of one sector's interests prevailing and resulting in an allocation of water that is inefficient from the point of view of the public at large. Portugal is also restructuring the actual administration of water affairs. A multilevel, decentralized management system is being introduced, in which a national authority (National Water Institute) representing the government sets the rules and procedures of water management, but the actual management of the system rests with the users through the delegation of power and responsibility to regional and local authorities.

France is another example of a country in which management of water resources is decentralized and practiced on the scale of a river basin. The agency that manages the basin, in terms of water use and water quality control, acts on behalf and under the supervision of the users in the basin, but subject to rules set by the government.

Standards of groundwater quality are often imposed as mandatory constraints. Unlike standards on water quality at the site of use, which are imposed by the respective branches of government and rarely questioned by the public, ambient water quality standards in the aquifer are closely related to the intended use of the water and to its treatment after extraction from the aquifer. But decisions of this kind should be derived on the basis of an analysis of options, outcomes, and consequences. Otherwise, arbitrarily imposed ambient standards may lead to an inefficient allocation of the groundwater resource and heavy financial losses.

Nonphysical Means of Groundwater Management

Apart from physical means of control that can be applied directly to the groundwater subsystem (pumpage and/or recharge) and/or to its physical environment (e.g., technological devices and facilities), there is a variety of nonphysical means of control that can enhance the effectiveness of groundwater management. These may include any combination of the following categories:

- Managerial (e.g., enhancement of public participation in the processes and procedures of decision-making, follow-up, and control of implementation);
- Economic (e.g., prices, charges, taxes, rebates, loans, subsidies);
- Legal (e.g., modification of laws and regulations);
- Administrative (e.g., allotments);
- Political (e.g., treaties);
- Educational (e.g., dissemination of information, provisions of guidance, training);
- Scientific (advancement of knowledge about processes, technologies, and techniques).

Experience shows that economic market measures of control are more flexible in responding and adjusting to changes in the quality of the receiving water body and in the benefits that can be derived from pollution abatement than administrative measures. A switch from regulatory measures of water-pollution control to economic incentives, or an appropriate mix of the two, may therefore be desirable. In the USA, for example, a number of empirical cost-effectiveness studies of water-pollution control have shown that the current EPA uniform effluent standards are much less cost-effective than emission charge or marketable emission-permit systems. Evidently, economic incentive systems have a distinct advantage over bureaucratically defined standards by forcing the sources of pollution continually to search for new opportunities to control emissions at lower cost. In Europe, for example, effluent charges play a prominent role in a number of countries.

It should, however, be emphasized again and again that any set of available means of control can serve as effective management tools if, and only if, their choice and implementation are guided by a rational decision-making process that utilizes information on the present state of the system, on the anticipated effect of any undertaken action on the states of the groundwater subsystem and its environment, as well as on the ensuing benefits and losses, and only then selects that combination of means and those levels of actions that are most effective in reaching the groundwater-management objectives. In doing so, losses associated with uncertainties concerning the future state of the system, and the implementation and impact of the undertaken measures of control, should be taken into account. Since these losses can be reduced as information on the state of the system is being updated, it follows that the management of groundwater, as part of a modern water system, should be based on adaptive control by feedback with information from ongoing monitoring of its actual state.

Finally, even the most rational decisions are useless if not duly implemented. Problems of coordination, adherence to schedules, methods, and procedures, access to data, technology transfer, and delays in allocation of funds are commonplace. Therefore, the physical and managerial parts of the groundwater subsystem should be managed together, maintaining a permanent control over the adoption and actual implementation of decisions by means of an integrated multidisciplinary, multipurpose information service.

Acknowledgment

This article is adapted with permission from Bachmat Y (1994) Groundwater as part of the water system. In: Zoller U (ed.) *Groundwater Contamination and Control*. New York: Marcel Dekker.

See also: **Darcy's Law**; **Drainage, Surface and Subsurface**; **Hydrodynamics in Soils**; **Pollution:** Groundwater; **Vadose Zone:** Hydrologic Processes

Further Reading

Bachmat Y (1989) *Management of Groundwater Observation Programs*. World Meteorological Organization, Operational Hydrology Report No. 31, WMO-No. 705. Geneva, Switzerland: WMO.

Bachmat Y (1994) Groundwater as part of the water system. In: Zoller U (ed.) *Groundwater Contamination and Control*. New York: Marcel Dekker.

Bachmat Y and Chetboun G (1974) *Seawater Encroachment in the Coastal Plain of Israel*. Jerusalem, Israel: Israel Hydrological Service.

Bear J (1979) *Hydraulics of Groundwater*. New York: McGraw-Hill.

Bear J and Bachmat Y (1991) *Introduction to Modeling of Transport Phenomena in Porous Media*. Dordrecht, the Netherlands: Kluwer Academic.

Council on Environmental Quality (1980) *Environmental Quality – 1980*. Washington, DC: Government Printing Office.

Domenico PA and Schwartz FW (1990) *Physical and Chemical Hydrogeology*. Chichester, UK: John Wiley.

Environment Canada (1990) *Water Fact Sheet 4*. Ottawa, Canada: Inland Waters Directorate.

Haimes YY (1977) *Hierarchical Analyses of Water Resources Systems*. New York: McGraw-Hill.

Hall WA and Dracup J (1970) *Water Resources Systems Engineering*. New York: McGraw-Hill.

Mandel S and Shiftan ZL (1981) *Groundwater Resources Investigation and Development*. New York: Academic Press.

Mercado A (1980) Israel as a case study. In: Shural HI (ed.) *Water Quality Management Under Conditions of Scarcity,* pp. 93–146. New York: Academic Press.

Office of Technology Assessment (1984) *Protecting the Nation's Groundwater from Contamination*. OTA-0-233 Washington, DC: Government Printing Office.

Piver WT (ed.) (1989) *EHP Monograph on Groundwater Quality*. Washington, DC: US Department of Health and Human Services.

Preul HS (1985) In: *Proceedings of the 5th World Congress on Water Resources*, pp. 357–366. Belgium: Brussels.

Schwarz J (1980) Israel as a Case Study. In: Shuval HI (ed.) *Water Quality Management Under Conditions of Scarcity,* pp. 51–91. New York: Academic Press.

Stokey E and Zeckhauzer R (1978) *A Primer for Policy Analysis*. New York: WW Norton.

Tahal and Mekorot (1987) *Dan Region Project. Treatment and Groundwater Recharge of Municipal Waste-water, 1986 Report*. Tel Aviv: Tahal.

Tietenberg T (1988) *Environmental and Natural Resource Economics*, 2nd edn. Glenview, IL: Scott, Foresman.

UNESCO World Meteorological Organization (1992) *International Glossary of Hydrology*, 2nd edn. Paris, France: UNESCO, WHO.

van der Heijde P, Bachmat Y, Bredehoeft J, Andrews B, Holtz D, and Sebastian S (1985) *Ground-water Management: The Use of Numerical Models*. American Geophysical Union, Water Resources Monograph 5. Washington, DC: AGU Publications.

Vardi Y (1980) Israel as a case study. In: Shuval HI (ed.) *Water Quality Management Under Conditions of Scarcity*, pp. 37–49. New York: Academic Press.

Willis R and Yeh WG (1987) *Groundwater Systems Planning and Management*. Engelwood Cliffs, NJ: Prentice-Hall.

Groundwater Pollution *See* **Pollution:** Groundwater

H

HEAT AND MOISTURE TRANSPORT

R Horton, Iowa State University, Ames, IA, USA
A Globus, Agrophysical Research Institute, St Petersburg, Russia

© 2005, Elsevier Ltd. All Rights Reserved.

In temperate climates, prompt germination, rapid emergence, and vigorous early growth of crops are critical to achieving profitable yields. Management strategies to deal with cool, wet soil conditions include the timing of tillage and planting operations and decisions regarding residue management, fertilization, and pesticide application. Heat and moisture transfer helps to define the conditions under which each of these management strategies may be most effective. Recent studies have highlighted the effects of temperature and wetting-and-drying cycles on soil C and N dynamics. Microbial dynamics affecting C and N cycling, including the generation of greenhouse gases, are influenced by drying–rewetting cycles and through a temperature–soil-water content interaction. Knowledge of heat and moisture transfer in soil is useful for predicting the soil-water content and temperature conditions that influence microbial dynamics in the C and N cycles. Understanding coupled heat and water flow processes, therefore, will enable significant and wide-ranging improvement of management strategies to optimize production efficiencies and reduce environmental impacts.

Coupled transfer of heat and water in soil is common. By coupled transfer we mean that heat and water transfer processes are intertwined. When water moves in soil, it carries heat with it, and temperature gradients that develop in response to heat transfer cause both liquid water and water vapor to move. Diurnal and seasonal climatic fluctuations cause coupled heat and water transfer to occur. For example, in a typical diurnal cycle in an agricultural field, the soil loses heat during the night. As heat leaves the soil, the surface temperature drops, and an upward thermal gradient develops, resulting in transfer of heat to the surface from the relatively warm soil below. If the surface temperature drops to the dewpoint temperature, water vapor condenses at the cooled soil surface, thus liberating heat. Throughout the night, liquid water flows upward toward the surface, rewetting the surface layers that dried during the previous day and redistributing heat upward. In the morning, solar radiation warms the soil surface. As heat enters the soil, the surface temperature increases, and the thermal gradient reverses direction. Liquid water near the soil surface vaporizes and is transported into the atmosphere, taking large quantities of heat with it. Liquid water and some vapor are also driven by the thermal gradient from the hot surface soil to the cooler soil below. As the sun goes down, this dynamic and intricate cycle begins again. All soil biological, chemical, and physical processes are influenced by the fluctuations of soil-water content and soil temperature that result from coupled heat and water transfer processes.

Although it has been known for many years that coupled heat and water transfer occurs in soil and that the coupled processes have important consequences for near-surface physical, chemical, and biological activities, the coupled processes have not been well quantified. Improved understanding of coupled heat and water transfer is needed to provide the basis for more complete understanding of soil-water and temperature conditions, soil-water evaporation, crop and weed seed germination, pesticide volatilization, surface fluxes of carbon dioxide, and trace gas emissions from soil. Improved understanding of coupled heat and water transfer is also needed to provide the basis for further understanding of the impact of climate change on the near-surface soil environment.

Temperature gradients exist in soil owing to periodicity of insolation, geothermal temperature distribution, functioning of buried cables, heating and cooling pipes, etc. Existence of temperature gradients means violation of thermodynamic equilibrium, which results in fluxes of heat and water in soil. The water flux can induce flux of soluble substances.

Soil moisture transfer induced by temperature gradient is called the thermal moisture transfer (TMT). The flow resulting from interaction of TMT with

moisture flow induced by factors other than temperature gradient is called the nonisothermal moisture transfer (NIMT). TMT can be vapor, liquid, and combined (series-parallel).

Vapor TMT is mainly caused by dependence of water vapor concentration (and/or pressure) on temperature, thus gradients of temperature create gradients of water vapor concentration and hence classic thermal water vapor diffusion in soil, which is directed toward a lower temperature. As with normal gas diffusion, vapor TMT depends upon the liquid-free porosity and on tortuosity of diffusion path, which is influenced by soil texture, structure, and bulk density.

Liquid TMT can be effected by several mechanisms:

1. Expansion and contraction of entrapped air due to temperature change pushes liquid back and forth in nearly saturated soil, for example, in the capillary fringe and near the groundwater table. This is the so-called thermometric effect, which is transient;
2. Due to dependence of surface tension at the liquid–air interface on temperature, a temperature gradient induces a gradient of surface tension and respective gradient of capillary pressure of menisci. This can induce hydrodynamic flow in the same way as the influence of a pressure difference of any other origin. This is 'thermocapillary meniscus' flow;
3. When a thermal gradient exists along the liquid–air interface of a liquid film covering solid particle(s), the induced surface tension gradient produces thermocapillary film flow. The velocity profile of this flow (as a function of distance to solid phase), differs from that of hydraulic flow, because the moving force is applied only to the interface. Both thermocapillary flows are directed toward a lower temperature;
4. When enthalpy of pore liquid differs from that of bulk liquid, and, particularly, when there exists some distribution of enthalpy as a function of distance to the solid phase, thermo-osmotic flow can occur. Direction as well as magnitude of this flow depend upon the actual profile of enthalpy. Since normally the latter diminishes near the solid–water interface, thermo-osmotic flow should be directed toward a higher temperature;
5. In unsaturated soil there exists a special vapor–liquid series-parallel (or combined) TMT, consisting of thermal vapor microdiffusion inside air space of a pore, combined with liquid flow in capillary and film elements of soil matrix. Because of its dual nature, this flow bears features appropriate to both vapor diffusion and to liquid flows. For example, it depends upon ambient gas pressure P as vapor diffusion (diminishing with rise of P), and on phobicity of a medium to liquid flow, vanishing in hydrophobic media. Due to the same dual nature of this mechanism, dense and friable soil layers are almost equally well permeable for the combined TMT, supposedly because the increasing role of the faster liquid component in the denser layer compensates for the reduction of water-free space for vapor diffusion. This property of combined TMT can be used for regulating the root-zone water regime, since a friable interlayer, transparent for downward TMT, is less permeable for upward, 'isothermal' liquid flow, thus creating a trap for the moisture transferred by TMT.
6. A special case of moisture migration in soil under the influence of a temperature gradient is the movement of moisture in soils induced by freezing. This process is often described as TMT, where the driving force is formally represented by the temperature gradient. However, actual water flow in this case is caused by the gradient of chemical potential of water, which arises due to locally reduced liquid water content at and behind the freezing front after freezing of some water. Differences in unfrozen water content create differences in matric potentials. So in this case it would be proper to substitute temperature gradient by hydrostatic pressure gradient with the help of the appropriate form of the Clausius–Clapeyron equation. Just like freezing itself, this type of 'TMT' can be suppressed by salinization and/or overburden pressure, which increase the chemical potential of water and reduce the freezing point of pore solution.

Phase composition of TMT governs its contribution to heat transfer, since pure vapor and combined flows participate in transfer of latent heat (approximately $2500\,\mathrm{kJ\,kg^{-1}}$ of water); meanwhile, liquid TMT carries only $4.2\,\mathrm{kJ\,kg^{-1}}$ of water per unit temperature difference. The ratio of vapor and liquid contributions to total TMT depends on the soil moisture content.

There exist three interrelated types of mathematical description of TMT: Fick's formalism, Darcy's formalism, and formalism of irreversible thermodynamics: $q_{wt}/\rho_w = -D_\mathrm{T}\nabla T$; $q_{wt}/\rho_w = -K_\mathrm{T}\nabla T$; $J_\mathrm{wt} = L_\mathrm{WT}X$, respectively, where: q_wt is time rate of TMT, ρ_w is density of water, D_T is thermal moisture diffusivity, K_T is thermal hydraulic conductivity, J_wt is the same as q_wt, L_WT is the phenomenological coefficient, X is thermodynamic force (the product $L_\mathrm{WT}X$ must have dimensions of internal entropy production time rate, so, for example, X can be set equal to $-\Delta T/T^2$, where ΔT is temperature difference on the system boundaries (one-dimensional case), T is mean absolute temperature). Physically, D_T and K_T are the same, though routinely the symbol 'D_T' is in use. D_T can be split into D_Tv, thermal vapor moisture diffusivity (usually also accounting for combined TMT), and

D_{Tl}, thermal liquid moisture diffusivity. The same can be done with K_T and L_{WT}.

Additionally, $q_{WTf}/\rho_w = -K_{Tfz}\nabla T_f$, where q_{wtf} is time rate of water flow in freezing soil, K_{Tfz} is respective 'thermal' hydraulic conductivity, T_f is temperature in the range of water–ice phase transition temperatures.

Classic thermal vapor diffusivity $D_{Tcv} = D_0(\Pi - \theta)\,\gamma\,\varphi\partial c_0/\partial T$, where D_0 is coefficient of water vapor diffusion in air, Π is total porosity, θ is volumetric water content, γ is tortuosity factor (dimensionless), c_0 is concentration of saturated water vapor in air, φ is equilibrium relative humidity of air in soil. The order of magnitude of D_{Tcv} is $10^{-9}\,m^2\,s^{-1}\,K^{-1}$.

Liquid (thermocapillary) thermal water diffusivity D_{Tcp} is often expressed as $K(\psi/\sigma)\partial\sigma/\partial T$, where K is usual unsaturated hydraulic conductivity, ψ is equivalent hydrostatic head, and σ is surface tension of water on interface with air. This form of D_{Tcp} results from the assumption that the contact angle λ in capillarity law $P = C\sigma\,\cos\lambda/r$ (c is constant, r is radius of equivalent capillary) does not depend on T. If temperature dependence does take place, an additional term will appear in D_{Tcp} in the form of $(\psi/\lambda)\partial\cos\lambda/\partial T$.

D_{TP} for series-parallel transfer can be represented as ξD_{Tcv}, where ξ is enhancement factor, $1 < \xi < 10$, which accounts for increased temperature gradient inside air space of pores and interaction of vapor and liquid flow, which makes the space occupied by the liquid phase 'transparent' for this mechanism of TMT. This enhanced TMT is appropriate not only to water but also to such substances as ethanol, benzene, and carbon tetrachloride. However, the larger the $D_0\partial c_0/\partial T$, the smaller are the respective ratios D_{TP}/D_{Tcv} of these substances. All thermal moisture diffusivities increase exponentially with increasing T: vapor diffusivities owing to increases in D_0 and $\partial c_0/\partial T$ with rises in T; and liquid diffusivities owing to reducing viscosity of a liquid with rise of T, which in the range of normal temperatures is relatively greater than diminishing σ. D_{Tcv} and D_{TP} decrease inversely with increase in ambient gas pressure P. Electrolytes in soil solution depress the moisture flux provided by combined TMT. For sodium and calcium chlorides, this effect can be described by the equation

$$(D_{TP})_N/(D_{TP})_{N=0} = 1 - A\sqrt[3]{N}$$

where A is a constant ≈ 0.5 and N is normality of electrolyte solution.

The order of magnitude of K_{Tfz} is $10^{-7}\,m\,s^{-1}$, so TMT under freezing proceeds much faster than above 0°C, though dehydration of soil next to an ice lens puts an end to the process until the freezing front moves deeper over the freeze-dried zone. Due to moisture transfer toward the freezing front, as well as the freezing of part of the water, the zone before a freezing front accumulates soluble substances.

No matter how large (or small) are the respective values of D_T or their sum, within the full range of possible soil moisture contents there exists a range where TMT is comparable with the flux, induced by matric pressure gradient. This range is the domain of effective NIMT, i.e., soil moisture-content diapason where TMT can result in net changes of moisture-content distribution. Experiments with closed systems showed that this range increases with increase in clay content (or specific surface area S), reduction of bulk density, and increase in aggregate percentage. It also depends on temperature-gradient magnitude and on absolute temperature. Inside this range (expressed in terms of saturation degree concept) there often exists a particular 'critical' degree of saturation DS_c, at which NIMT has its maximum, depending on the same factors as the range itself. For a temperature gradient of the order $100°C\,m^{-1}$ (often occurring in shallow soil) and ambient temperature, $DS_c = 0.085 + 0.024\sqrt{S}$, where S is in meters squared per gram. In closed or quasiclosed systems, steady-state moisture distribution is achieved owing to a sort of circulating system, when TMT carries moisture to the cold boundary of a system, thus creating a gradient of moisture content and matric potential, causing moisture to flow back to the hot boundary predominantly as a liquid. In summer days a quasiclosed system in terms of NIMT occurs in the upper layer of soil, since temperature gradients, maximal at the surface, gradually vanish with depth. Moisture driven mainly by combined TMT into a cooler layer flows back to the surface as a liquid, carrying with it soluble substances. At night TMT with condensation can partly replenish the moisture supply of the uppermost soil layer. The moisture circulation due to the NIMT can enhance salinization of the uppermost soil layer with accumulation of soluble substances (useful as well as harmful ones) in it.

Examples of Coupled Heat and Water Flow in Soils

Three examples of coupled heat and water flow in soil are presented. One example represents field infiltration of water under nonisothermal conditions, and the other two examples are for closed laboratory soil columns that experience nonisothermal conditions. The field example shows heat and water transfer under wet soil conditions in response to diurnal heating and cooling. The laboratory examples

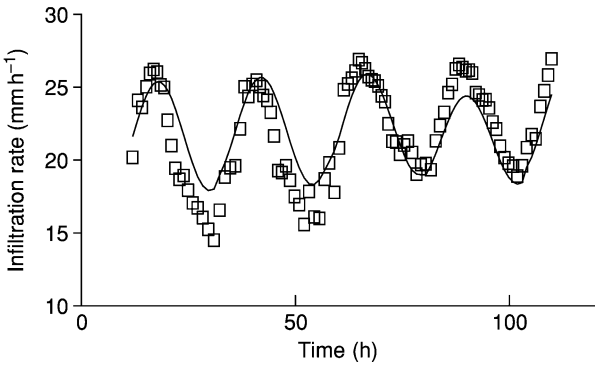


Figure 1 Soil-water infiltration rate as a function of time. Open squares denote measured data, and the solid curve indicates predicted infiltration rates. The model used to predict infiltration rate includes the effect of temperature on water viscosity.

demonstrate the effect of temperature on coupled heat and water transfer and the effect of salinity on coupled heat and water transfer in unsaturated soil.

Figure 1 presents field observations of ponded water infiltration rate as a function of time. The experiment lasted for 5 days. Diurnal fluctuations in water infiltration rate are delineated, and the solid curve represents a simple model of the data. The only parameter included in the model is water viscosity, which is temperature-dependent and thus affects soil hydraulic conductivity. Diurnal soil surface temperature was used to estimate diurnal fluctuations in water viscosity, which in turn was used to estimate diurnal fluctuations in soil-water infiltration rate. As shown in Figure 1, this very simple model captures the trend of the infiltration data and very nearly captures the amplitudes of the diurnal fluctuations. Clearly, heat and water flow occurs in natural environments.

Figure 2 presents laboratory data collected from a series of closed soil columns. Each soil column started at the same uniform initial water content (denoted by the dashed line in Figure 2) and had the same constant thermal gradient imposed on it. The curves in Figure 2 show the water content distributions after a thermal gradient has been applied to each soil column. Although the same thermal gradient was used on all soil columns, the mean temperature varied among the soil columns. For example, mean temperatures were approximately 10°C, 20°C, 30°C, and 40°C. There is a clear effect of mean temperature on the amount of water redistribution in the columns. The mean temperature affects vapor pressure as well as vapor and liquid diffusivities. For large mean temperatures, water transfer can be expected to be large. This has implications for field heat and water transfer as a function of season.

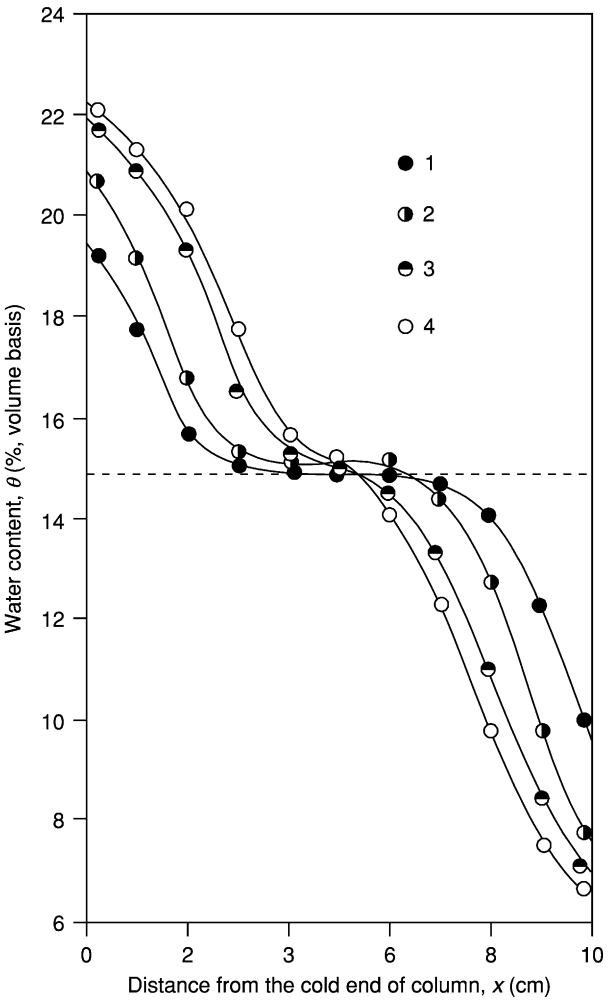


Figure 2 Water-content distributions after soil columns experience the same thermal gradient imposed at different mean temperatures: 1–10°C, 2–20°C, 3–30°C, and 4–40°C. Reproduced with permission from Globus AM (1983) Physics of non-isothermal soil moisture transfer [in Russian]. Leningrad, Russia: Gidrometeoizdat.

Figure 3 presents two water-distribution curves for soil columns that have been exposed to the same thermal gradient. Both soil columns had the same initial water content; however, one soil column contained almost pure water, while the other column had soil solution that was salinized. The curves in Figure 3 indicate that greater water redistribution in response to the thermal gradient occurred in the soil column that was not salinized. In the salinized soil column, much less water redistributed in response to the thermal gradient. These data demonstrate the tempering effect of soil salinity on water movement in response to thermal gradients, which is described by the formula given above. Since most natural soils contain some dissolved salts, it is important to account for soil solution concentrations when estimating coupled heat and water transfer.

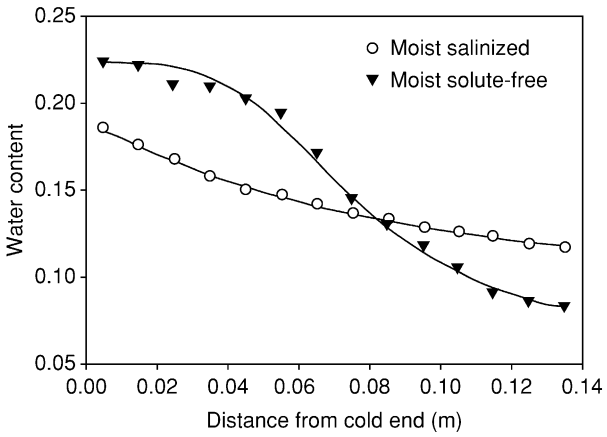


Figure 3 Observed steady-state water-content distributions within moist salinized and solute-free soil columns.

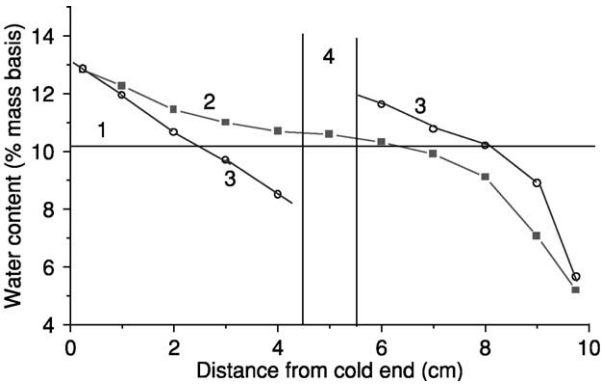


Figure 4 Influence of hydrophobic soil layer on thermal water transfer and moisture distribution in loamy soil. 1, initial moisture content *W*; 2, moisture profile in uniform soil column; 3, the same loamy soil in a column with a hydrophobic layer (4).

Limitations in Our Understanding of Coupled Heat and Water Flow

Although several studies have addressed coupled heat and water transfer in unsaturated soil, fundamental deficiencies in understanding persist. For example, field observations of water vapor movement in soils can be greater than predicted by the widely used diffusion-based theory of Philip and de Vries. Recent measurements highlight our current lack of understanding of the contributions of convection and diffusion to coupled heat and water transfer in soil. Researchers have reported that existing diffusion-based theory is inadequate to describe their field measurements of coupled heat and water transfer. Existing theory cannot explain the large vapor fluxes and evaporation rates measured in some field studies. As a result of these findings, it has been proposed that diurnal heating of the soil causes convective transport of soil air due to temperature-driven expansion of the soil air (which is akin to the thermometric effect, mentioned above). It has also been proposed that this convective transport (soil 'breathing') may explain the large fluxes of water vapor observed in field studies.

Other recent measurements highlight the current lack of understanding of the influence of soil hydrophobicity on coupled heat and water transfer. Current theory of coupled heat and water transfer assumes that soil is perfectly hydrophilic or 'wettable' (i.e., that the contact angle of soil water is 0°). However, recent research suggests that under certain conditions all soils may display water repellency to some degree. Water repellency has been observed in sand, loam, clay, peat, and volcanic ash soils all over the world. Researchers have measured isothermal and nonisothermal (columns heated from below) evaporation from columns containing wettable soil and columns containing hydrophobic soil. Evaporation from the hydrophobic soil was approximately 75% and 50% of the evaporation from the wettable soil under isothermal and nonisothermal conditions, respectively. The available published data of soil wettability effects on coupled heat and water transfer in field soil are limited; however, in laboratory experiments it has been demonstrated (Figure 4) that a porous hydrophobic layer reduces TMT drastically in wet soil. Figure 4 compares NIMT in a uniform column of loamy soil with NIMT in a column of the same soil with a hydrophobic soil layer in the middle section. The data show clearly that soil wettability can have significant impacts on heat and mass transfer in soil. No existing models can describe explicitly these results, although the Philip–de Vries theory assumes implicitly that in hydrophobic soil the series-parallel mechanism does not work.

There are limitations in the models used to predict coupled heat and water transfer in soil. Current models all assume that the state variables can be represented by continuous functions and that volume-averaged transport parameters adequately represent the properties of the soil volume under consideration. These continuum models with volume-averaged parameters have significant limitations for describing coupled heat and water transfer. For example, the widely used de Vries model for soil thermal conductivity requires assumptions about the pore-scale arrangement of soil solids, water, and air. These assumptions are represented in the model by 'shape factors.' The choice of shape factors determines how thermal conductivity varies with soil water content. The lack of a model to describe the pore-scale arrangement of the solids, water, and gas forces us to rely on these arbitrarily chosen shape factors. Hence

our ability to predict thermal conductivity is limited, as is our ability to predict coupled heat and water transfer, which is strongly dependent on thermal conductivity. As a second example, Philip and de Vries have made an effort to improve the limitation of their continuum model by introducing a thermal vapor enhancement factor to account for pore-scale features (e.g., nonequal thermal gradients in the various phases and nongas-phase diffusive pathways in the pores). Researchers have used Philip and de Vries' theory to describe nonisothermal evaporation from columns containing wettable soil and columns containing hydrophobic soil. In the most wettable soil column(s), the thermal vapor enhancement factor is approximately 3.5 (meaning that thermal vapor transport is approximately 3.5 times greater than predicted by classic thermal vapor diffusion theory), but in the most hydrophobic soil column the thermal vapor enhancement factor was approximately 0.35.

In other words, soil columns that differ only in wettability show very different water evaporation rates and water redistribution. No existing model can describe explicitly the pore-scale features that lead to these greatly different thermal vapor enhancement factors. The development of discrete models such as pore-scale network models will aid in the understanding and description of coupled heat and water transfer in soil.

Coupled heat and water transfer is very important in soils; it is also very complex. Clearly, it is important for scientists to continue to seek to understand these important, complex processes.

List of Technical Nomenclature

γ	Tortuosity factor
θ	Volumetric moisture content
λ	Contact angle
ξ	Enhancement coefficient
Π	Total porosity
ρ_w	Specific density of water
σ	Surface tension of liquid on the interface with air
ϕ	Relative humidity
ψ	Equivalent (matric) hydraulic head
A, C	Constants
c_0	Concentration of saturated vapor
D_0	Vapor diffusivity in free air
D_T	Thermal moisture diffusivity
D_{Tcp}	Thermocapillary liquid moisture diffusivity
D_{Tcv}	Classical thermal vapor moisture diffusivity for soil
D_{Tl}	Thermal liquid moisture diffusivity
D_{TP}	Series-parallel liquid-vapor thermal moisture diffusivity
D_{Tv}	Thermal vapor moisture diffusivity
DS_c	Critical degree of saturation
J_{WT}	Thermal moisture flux
K	Unsaturated hydraulic conductivity
K_T	Thermal hydraulic conductivity
K_{Tfz}	Thermal hydraulic conductivity due to freezing
L_{WT}	Phenomenological coefficient
N	Normality of electrolyte solution in soil pores
P	Ambient air pressure
q_{WT}	Time rate of thermal moisture transfer
q_{WTf}	Time rate of thermal moisture transfer to freezing front
r	Radius of equivalent capillary
S	Specific surface area
T	Absolute temperature
T_{fz}	Absolute temperature in the range of water–ice phase transition in soil
X	Thermodynamic force

Further Reading

Cary JW (1963) Onsager's relation and the nonisothermal diffusion of water vapor. *Journal of Physical Chemistry* 67: 126–129.

Cary JW (1965) Water flux in moist soil: thermal versus suction gradients. *Soil Science* 100: 168–175.

Cary JW (1966) Soil moisture transport due to thermal gradients. Practical aspects. *Soil Science Society of America Proceedings* 30(4): 428–433.

Cass A, Campbell GS, and Jones TL (1984) Enhancement of thermal water vapor diffusion in soil. *Soil Science Society of America Journal* 48(1): 25–32.

Globus AM (1983) Physics of non-isothermal soil moisture transfer [in Russian]. Leningrad, Russia: Gidrometeoizdat.

Milly PCD (1980) A simulation analysis of thermal effect on evaporation from soil. *Water Resources Research* 7: 1216–1225.

Nassar IN and Horton R (1992) Simultaneous transfer of heat, water, and solute in porous media. I. Theoretical development. *Soil Science Society of America Journal* 56: 1350–1356.

Nassar IN and Horton R (1997) Heat, water, and solute transfer in unsaturated porous media. I. Theory development and transport coefficient evaluation. *Transport in Porous Media* 27: 17–38.

Nassar IN, Horton R, and Globus AM (1992) Simultaneous transfer of heat, water, and solute in porous media. II. Experiment and analysis. *Soil Science Society of America Journal* 56: 1357–1365.

Philip JR and de Vries DA (1957) Moisture movement in porous materials under temperature gradients. *Transactions of the American Geophysics Union* 38: 222–232.

Heat Capacity *See* **Thermal Properties and Processes**

Heat Flow *See* **Thermal Properties and Processes**

HEAVY METALS

D C Adriano, University of Georgia, Aiken, SC, USA
N S Bolan, Massey University, Palmerston North, New Zealand
J Vangronsveld, Limburgs Universitair Centrum, Diepenbeek, Belgium
W W Wenzel, Universität für Bodenkultur, Wien, Austria

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Certain metals are essential for the normal growth of plants, animals, and humans. In plant nutrition, these are called micronutrients, represented by iron (Fe), copper (Cu), zinc (Zn), manganese (Mn), and molybdenum (Mo). In addition to these micronutrients, arsenic (As), cobalt (Co), chromium (Cr), nickel (Ni), selenium (Se), tin (Sn), and vanadium (V) are essential for animal nutrition. Similarly in human nutrition, the above nutrients are essential, except As and V. They are collectively called micronutrients or trace elements because they are required in only trace amounts (unlike the major nutrients nitrogen, phosphorus, and potassium, required in relatively large amounts) and any excess may result in unfavorable biological response.

Since the advent of the Industrial Age, the environment has been subjected to emission and deposition of anthropogenic chemicals, both organic and inorganic. Consequently, some of these chemicals have accumulated in various ecosystems, both aquatic and terrestrial, to the extent that in numerous instances they have bioaccumulated to exorbitant levels that have the potential to imperil the well-being of biota, including the consumers (i.e., animals and humans). Heavy metals are elements having densities greater than $5\,g\,cm^{-3}$. The term refers to metals and metalloids that are associated with pollution and toxicity, but also elements that are required by biota at low concentrations. While some of the most prevalent xenobiotic organics have only been released to the environment during the last 3–4 decades (e.g., polyaromatic hydrocarbons (PAHs), dioxin, trichloroethylene (TCE)), several heavy metals have been produced and subsequently emitted for hundreds of years. For example, lead (Pb) was initially produced even before the rise of the Roman Republic and Empire (approx. 2000 years ago).

The ever-increasing production and demand by society for metals indicate the mounting probability of their dispersal and contact with the environment. A metal may be dispersed from the time its ore is mined to the time it becomes a finished product. In some cases the ultimate disposal of the finished product may also lead to metal dispersion. The growing world population and the importance of metals in growing economies just ensures more production of metals and the likelihood of more dispersal, especially in developing countries.

Sources of Metals in the Environment

There are two basic sources of metals input in the environment: natural and anthropogenic. Parent rocks and metallic minerals (or ores) represent the main natural sources. Anthropogenic sources range from agriculture (fertilizers, pesticides, livestock manure, etc.), metallurgy (mining, smelting, metal-finishing, etc.), energy production (leaded gasoline, battery manufacture, etc.), and microelectronics, to waste and scrap disposal. Intensive agricultural practices in industrialized countries, especially in North America and Europe, have been increasing the metal burden in soils with heavy and frequent applications of fertilizers, agrochemicals, and soil amendments. For example, the use of certain rock phosphate in manufacturing phosphate fertilizer could contaminate the soil with cadmium (Cd) – already a concern in New Zealand, Australia, and other countries. Swine manure is typically high in Cu content, i.e., it is used as a growth promotant so using this manure as a soil amendment could increase levels of Cu.

In developed countries such as those in Western Europe and North America, atmospheric deposition plays an important role in contaminating arable soils. In England and Wales, atmospheric deposition represents from 33% (for Cu) to 93% of the total input (for mercury, Hg) of heavy metal contamination. Arable soils received approximately 40% Ni and Zn, and more than 50% Cd, As, and Pb from atmospheric deposition. Cu and Zn are contributed mostly by application of animal manure, Cr by application of industrial by-product waste; Cd and, to a much lesser extent, Cr and Ni are added to the soil by fertilizer and lime application. Sewage sludge represents a minor source of metals in arable soils but, with the banning of ocean disposal and incineration of biosolids, it will assume a substantial role in adding metals to the soil, since disposal of biosolids will be shifted on to the land. In forest ecosystems, soil contamination by metals is virtually all from atmospheric deposition. This distribution of metals in the soils of England and Wales is thought to be typical for most developed countries.

Atmospheric deposition is a transboundary issue that ensures contamination of even pristine areas. Certain metals (e.g., selenium (Se), Hg, As, and antimony (Sb)) are fairly volatile and can be transported over long distances in gaseous form or enriched fine particulates; whereas other metals (e.g., Pb, Cu, Zn, and Fe) are transported only in the enriched particulates. Promulgation of 'clean air' policy in many countries may mitigate the impact of atmospheric deposition on the world's ecosystem, but may also switch the burden to the land, owing to disposal and/or landfilling of waste residues collected from air pollution-control devices or accumulation of waste materials that otherwise could have been incinerated or disposed of in the ocean.

While terrestrial systems serve as the main recipient of metal emission, groundwater and freshwater systems can also be impacted. Metals in the anionic form may leach to the groundwater or be transported laterally to nearby freshwater bodies. Metals known to be mobile because they are negatively charged, i.e., anionic, include the arsenites (As(III)), arsenates (As(V)), selenites (Se(IV)), selenates (Se(VI)), and chromates (Cr(VI)). Freshwater systems can be contaminated via runoff, drainage, or direct emission and/or disposal. The main reservoir for metals in aquatic systems is the sediment. Because of the high buffering capacity (i.e., high alkalinity, salinity, etc.) of marine waters, metals are usually not a concern in this system.

Environmental Contamination

Land contamination with metals has been associated with the accumulation of unwanted waste by-products, either aerially or by application and/or disposal, from various human activities. Chemical degradation of soils is a consequence of the post-Industrial Revolution. The following biological and chemical processes may promote the degradation of multifunctionality of soils and the eventual deterioration of soil quality: weathering, pollution, acidification, organic matter buildup or depletion, nutrient buildup or depletion, and salinization. Other chemical processes in soils, such as ion exchange, complexation, oxidation – reduction, sorption – desorption, and precipitation – dissolution may also influence the extent of its degradation. A chemically degraded soil may offer only limited land use rather than one having a multifunctional role. For example, a site contaminated by Pb from battery storage may not be fit as a playground for children unless remediated.

There are practices that have contributed to chemical degradation of soils: agrochemical application, waste disposal, landfilling and storage of waste, combustion of fossil fuels, mining and smelting of metal ores, application of biosolids and compost, use and storage of petroleum products, deforestation, and dredging. Due to urbanization, industrialization, and intensive agriculture, many types of contamination ('pollution' may be a more appropriate term in cases of severe contamination that may result in negative biological response) are recognized as inflicted by gaseous, liquid, and solid forms of metals. Contamination varies from country to country, but generally is more severe in developed countries with high population density. Also, in a given country, the

urban areas and their suburbs tend to have more severe contamination.

The importance of metals as probably the most important group of contaminants has been well established. For example, of the 11 most common and dominant contaminants at the US National Priority List sites (the so-called Superfund sites), six (Pb, As, Cr, Cd, Ni, and Zn) are metals. This reflects the extensive production and use of these metals in the USA; similarly, these metals are also the most dominant at hazardous waste sites in affluent countries in Europe and elsewhere. Fortunately, none of these sites include a highly productive arable land; rather, these sites were impacted primarily by the mining and industrial sectors.

Because of enactment of environmental policy limiting the impact and total burden of metals in soil, the US Environmental Protection Agency (EPA) has issued (in 1993) metal limits relative to soil application of municipal sewage sludge on arable land. In addition to maximum cumulative loading, there are also restrictions in ceiling concentrations in sludge and maximum annual loading. For example, in the case of Cd, the permissible maximum annual loading is $1.9\,kg\,ha^{-1}\,year^{-1}$. The maximum annual loading given for Cr is $150\,kg\,ha^{-1}\,year^{-1}$. The discrepancy between these two metals (also among the other metals) is due to their different biogeochemical behavior and eventual plant uptake, their difference in phytotoxicity threshold level, and their differing sensitivities to edaphic and environmental factors. The European Commission has also adopted a similar regulatory policy on metal input in arable soils, although generally in a more restrictive manner in terms of loading amounts. There are also differences in limits among countries in the European Union. These transcontinental and interregional variations in limits can be attributed to differences in scientific as well as sociopolitical, economic, and international trade orientation.

Retention and Dynamics of Metals in Soils

In terrestrial ecosystems, the soil is the main repository of metals. Likewise, in aquatic systems, the sediment serves as their ultimate sink. As such, it is very important to understand the various dynamic biogeochemical processes in the soil controlling the transformation, fluxes, and fate of these chemicals. Most importantly, it is paramount to evaluate the extent of the bioavailability (or mobility), more or less on a quantitative basis, to microbiota, plants, animals, and humans as influenced by these various processes. Relevant to this phenomenon is the role of the source term (i.e., form) and chemical speciation of the chemicals of concern. And central to this phenomenon is the role of natural attenuation (NA) in chemical dynamics in soil systems. ('Natural attenuation' refers to the state of equilibrium approaching that of the background, i.e., having negligible risk, induced by indigenous physical, biological, and chemical processes in the soil.)

NA is well established as a remedial strategy for a few organic chemicals, primarily BTEX (benzene, toluene, ethylene, xylene). However, these processes cannot destroy metals, although in some cases they can immobilize them. Natural attenuation can be managed for both organics and inorganics. The kinetics of NA can vary widely between the organics and metals. The important parameters to gauge NA also vary between these two groups of contaminants. For organics (e.g., BTEX), the commonly used parameter is the total concentration (or mass) in the affected plume, and NA, even if accelerated, may take months to years, depending on site and contaminant properties. Also certain organics may degrade into daughter by-products that are more toxic and recalcitrant (e.g., TCE degrades to vinyl chloride); whereas with metals they are generally persistent and immutable, so the 'bioavailable' fraction in the soil should be the most relevant parameter for NA. Unlike common organics, the kinetics of NA for metals, as measured by a decrease in bioavailability, can be relatively much shorter, i.e., in a matter of days.

Metal ions can be retained in soil by the following basic biogeochemical processes: (ad)sorption, precipitation, and complexation reactions. Sorption is defined as the accumulation of matter at the interface between the solid adsorbent and the aqueous phase. This can include ion exchange, formation of surface solid complexes, precipitation, and diffusion into the solid. The lower the metal solution concentration and the more sites available for adsorption, the more likely that adsorption process will determine the soil solution concentration. Once deposited in the soil, metals interact with the soil minerals and organic constituents. However, the fate of metals in the soil environment is dependent on both soil properties and environmental factors.

Adsorption

Charged solute species (ions) are attracted to the charged soil surface by electrostatic attraction and/or through the formation of specific bonds. Retention of charged solutes by charged surface is broadly grouped into specific and nonspecific retention. In general terms, nonspecific adsorption is a process in which the charge on the ions balances the charge

on the soil particles through electrostatic attraction; whereas specific adsorption involves chemical bond formation between the ions and the sorption sites on the soil surface. If a nonspecific adsorption process solely controls metal adsorption, then the adsorption capacity of the soil is dictated by its cation exchange capacity (CEC). However, in many soils the amount of metal sorbed exceeds the CEC of the soil. This infers that, in addition to nonspecific adsorption, other processes such as specific adsorption, precipitation, and complex formation also contribute to the overall retention.

Both soil properties and soil solution composition determine the equilibrium of metals between solution and solid phases. The concentration of metals in soil solution is largely influenced by the pH and the nature of both organic and inorganic anions. The effect of pH values greater than 6 in lowering free metal ion activities in soils can be attributed to the increase in pH-dependent surface charge on oxides of Fe, Al, and Mn, chelation by organic matter, or precipitation of metal hydroxides.

There are three reasons for the effect of inorganic anions on the adsorption of metals: (1) inorganic anions form ion-pair complexes with metals, thereby reducing their adsorption on to soil particles; (2) the specific adsorption of ligand anions is likely to increase the negative charge on soil particles, thereby increasing the adsorption of metal cations; and (3) specifically sorbed anions (e.g., phosphate) strongly compete with metal anions such as arsenate and selenates, resulting in their desorption.

Complexation

Metals form both inorganic and organic complexes with a range of solutes in soils. For illustration, the effect of inorganic anionic complex formation on the adsorption of Cd^{2+} by soils is discussed. Generally, chloride can be expected to form a soluble complex with Cd^{2+} as $CdCl^{+}$, thereby decreasing the adsorption of Cd^{2+} on to soil particles. In contrast to inorganic ligand ions, Cd^{2+} adsorption by kaolinite, a variable-charge mineral, could be enhanced by the presence of organic matter via the formation of an adsorbed organic layer on the clay surface. As might be expected, the organic component of soil constituents has a high affinity for metal cations because of the presence of ligands or groups that can chelate metals. With increasing pH, the carboxyl, phenolic, alcoholic, and carbonyl functional groups in soil organic matter dissociate, thereby increasing the affinity of ligand ions for metal cations. The general order of affinity for metal cations complexed by organic matter is as follows:

$$Cu^{2+} > Cd^{2+} > Fe^{2+} > Pb^{2+} > Ni^{2+} > Co^{2+} > Mn^{2+} > Zn^{2+}$$

Precipitation

Precipitation appears to be the predominant process of metal immobilization in alkaline soils in the presence of anions such as carbonate, hydroxide, and phosphate, especially when the concentration of metal ion is high. Coprecipitation of metals, especially in the presence of iron oxyhydroxide, can also be expected, and often such interactions lead to significant changes in the surface chemical properties of the substrate. Precipitation as metal phosphates is considered to be one of the mechanisms for the phosphate-induced immobilization of heavy metals, especially in substrates containing high concentrations of metals.

Liming typically enhances the retention of metals; for example, increased retention of Cr(III), Cd, and Pb with a lime-induced increase in pH can be anticipated. The pH of the lime-treated soil, which ranges from 7 to 8, coincides with the effective precipitation range for a number of metals as their respective metal hydroxides (for example, Cr(III) as $Cr(OH)_3$). An increase in pH due to liming also increases the negative charge of variable-charge soils, enhancing metal adsorption.

Bioavailability–Natural Attenuation Interactions

Since most metal retention in soils occurs in the upper layer of the soil profile (especially those deposited aerially), usually involving a large tract of land, a possible landscape scenario depicts a vegetated system, unless the metal levels in soils are phytotoxic. Otherwise, for soil conservation and environmental protection purposes, highly metal-contaminated soils need to be amended as an integral part of a remediation scheme. Thus plants play a central role in the bioavailability of the metals in question and their NA. We hypothesize that plants, directly or indirectly, heavily influence the degree and kinetics of NA and modify the influence of certain soil amendments when employed to accelerate NA.

A more generic definition of 'bioavailability' is the potential for living organisms to take up chemicals from food (i.e., orally) or from the ambient environment (i.e., externally) to the extent that the chemicals may become involved in the metabolism of the organism. More specifically, 'bioavailability' refers to the biologically available chemical fraction (or pool) that can be taken up by an organism and can react with its metabolic machinery; or it refers to the fraction of the

total chemical that can interact with a biological target, for example a plant. In order to be bioavailable, metals have to come in contact with the plant (i.e., chemical accessibility) to be able to enter a plant root. In essence, for a metal to be bioavailable, it has to be mobile, transported to the plant, and be in an accessible form. Hence, the bioavailability (more specifically, 'phytoavailability') of a metal in soil may be defined by the amount of metal absorbed by growing plants or by concentration in the harvested plant tissue. Plants typically absorb much less than 1% of the metals present in the soil. This expression does not indicate the extent to which metals present in soils are bioavailable nor the duration of bioavailability. Therefore, bioavailability of metals must account for not only the plant uptake in one growing season, but also the total amounts potentially available over time. The concepts of NA and bioavailability are being employed increasingly in remediation strategies.

The bioavailable fraction can be particularly important for metals, despite its dynamic nature and sensitivity to edaphic and environmental factors. (The 'bioavailable' fraction of metals in soil can be estimated by an extraction procedure using weak electrolytes; such a procedure has been used successfully by soil chemists for decades for micronutrient metal (Fe, Cu, Zn, and Mn) requirement by crops.) While the total content of metals is still being widely used for regulatory purposes, there has been a significant move, especially in Western Europe, to adopt the bioavailable fraction in risk assessment.

It is important to emphasize the critical role of source term (or form) and chemical speciation of metals in bioavailability. The 'source term' refers to the physicochemical form of the metal after it is released and deposited in soil. For example, Pb bound in metal ores in mining areas has very low bioavailability to humans compared with Pb bound to dust in urban areas. Thus, greater exposure of children to Pb in big cities can be expected. Speciation refers to the ionic or complexed (i.e., with an ion pair) condition of a particular metal within a given oxidation (i.e., valency) state. There is overwhelming evidence that certain chemical species (rather than total concentration) of metals in the soil solution are more bioavailable to plant roots than others. In microbiology, the free-ion activity model has been proposed as a general hypothesis, claiming that metal uptake by microbes is largely related to the activity of free ions in solution. While this model may apply to microbial uptake, it may not be so with higher plants. Indeed, metal–chelate (ethylenediaminetetraacetic acid, EDTA, or diethylenetriaminepentaacetate, DTPA) complexes have been shown to facilitate phytoextraction of metals, in this form, by *Brassica juncea* (Indian mustard). Furthermore, certain inorganic ligands may facilitate uptake. For example, Cd–chloro complexes are potentially more bioaccumulative, as demonstrated with the field potato in Australia. Thus it is evident that free ions or complexed ions should be favored over the total metal concentration in soil solution as a predictive tool for bioavailability (or, conversely, NA).

Chemical extraction of soil is another tool widely used to assess bioavailability of metals. Selective extractants using dilute acids (e.g., HCl, NH_4OAc) have been extensively used in soil-testing as a diagnostic tool for micronutrient deficiency. On the other hand, sequential extraction provides more information, in addition to the bioavailable pool, on the redistribution of the metal amongst organic and mineral soil constituents. When amended, it can be expected that the majority of the metals in the soil would be sequestered in less labile pools represented by the organic matter, Fe and Mn oxides, and silicate clays.

Quite often plant performance (yield, chlorosis and/or necrosis, root morphology), plant quality (total concentration of metals, nutrients, etc.), plant diversity, and metal-induced biochemical stress responses in plant tissue are used as direct indicators of remediation efficacy. Green, lush vegetation in a highly contaminated soil is a good sign of healthy soil, indicating favorable soil–microbe–plant interrelations. Other assessment tools include bacterial biosensors that detect the bioavailable fraction of a certain metal or suite of metals. For effective monitoring it is often necessary to deploy a mix of these tools.

Biological Response to Metals

When metals exceed certain threshold levels in soils, the organism being impacted may exhibit a negative biological response. The target organisms involved range from microorganisms, small invertebrate animals and other microfauna, and microflora to higher plants, including crops. Either one of these organisms or a combination of them may be used to assess their behavior, which has become a pivotal part of ecologic risk assessment. When their behavior becomes detrimentally affected, impacting on their normal physiological and/or biochemical functions, the metal can be suspected to pose some risk. These alterations in their usual biological functions could lead to disruption in community and/or population structure (i.e., less biodiversity), eventually leading to death. Movement of high concentration of metals in solute affecting groundwater quality is also a form, although not biological, of risk.

Likewise, humans could be at some risk when unnecessarily exposed to metals. The risk depends on

the exposure route and the dose. The total dose represents the sum of specific doses resulting from each exposure pathway:

$$\text{Dose}_{\text{total}} = \text{Dose}_{\text{soil}} + \text{Dose}_{\text{food}} + \text{Dose}_{\text{water}} + \text{Dose}_{\text{air}} + \text{Dose}_{\text{dermal}}$$

with units in milligrams of metal per kilogram of body weight per day.

The first three categories of dose represent ingestion of soil and/or dust (especially in children), edible food, and water, the fourth one representing inhalation of polluted air and the last representing skin absorption from polluted soil, water, or air. The dose relationship above indicates that young children, because of their low body mass, are the most sensitive segment of the population. Industrial hygiene is necessary to limit exposure of workers to metals via inhalation (e.g., vapor and/or fine dust) and/or skin contact (e.g., Ni). The relationship above also indicates that the most common exposure pathway for the general population is via ingestion of food and drinking water. The main concern in the former is the consumption of vegetables high in Cd and seafood high in Hg. Drinking well water high in As is of concern in some parts of the world, particularly in Bangladesh and West Bengal, India, where millions of people are at risk.

Just like in ecological risk, metals can affect humans in a variety of ways – depending on the type, form, and chemical speciation of the metals, age and body weight, dietary and/or nutritional status, exposure route and dose, etc. Because metals typically exist as metallothioneins, i.e., metals binding strongly to proteins, when in excess, they can disrupt the functioning of metalloenzymes. Such dysfunctions can eventually be manifested by disorders, for example in the central nervous system (by Hg and Pb), kidney (by Cd), liver (by As), blood (by Pb), and skin and extremities (by As). Under acute exposure, multiorgans can be affected, which can eventually lead to death.

Soil Remediation

The use of conventional, engineering-type techniques to clean up metal-contaminated soils is typically invasive and expensive. For example, in soil-washing, the soil has to be excavated first and usually treated off-site. The treated soil may be returned to the site or landfilled. The washing not only irreversibly destroys the integrity of the soil in a general sense, but also produces a secondary waste – the washing effluent, which is usually a chelate solution now laden with metals.

The most logical, ecologically friendly and economically viable method is by *in situ* immobilization of metals using abundant, inexpensive soil amendments. This technique is particularly suitable for the treatment of big tracts of land where the soil is mostly surficially contaminated. Soil amendments that have been successfully tested in North America, Europe, and Australia include liming materials, rock phosphate (hydroxyapatite), alkaline biosolids and/or compost, Fe-rich by-products (steel shot, red mud, 'Fe-rich' from aluminum processing), and coal residues (alkaline coal fly ash, beringite).

Traditionally, the use of unenhanced (or noninvasive) natural processes as part of a site-remediation strategy is called natural attenuation. Thus the US-EPA coined the term 'monitored natural attenuation' (MNA) when NA is employed within the context of a carefully controlled and monitored site cleanup strategy to be able to achieve site-specific remediation objectives within a time frame that is more reasonable than that offered by other, more invasive methods. A new twist to the definition has been proposed, that natural processes can be accelerated by adding to the soil inexpensive amendments that are cost-effective and minimally invasive. These materials can be incorporated into contaminated soils to speed up and optimize, in the case of metals, their immobilization as mediated by key processes such as (ad)sorption, precipitation, complexation, and redox reactions. The extent of immobilization can be measured by K_d (partition coefficient), which also serves as an index for bioavailability. A high K_d means very strong binding to the soil solid phase, rendering very low concentrations of the metal in the solution phase. Thus, a high K_d translates to low bioavailability induced by effective immobilization.

The 'driver' processes indicate that, when they are dominantly operative in the soil system, the K_d is high and bioavailability is low (or mobility and leaching potential are also low). In order to sustain the immobility (or high K_d) of metals, the 'counter' processes (e.g., desorption, dissolution, high oxidation potential) should be minimized or prevented, otherwise they can impede metal immobilization processes. For example, precipitation as metal phosphate has been shown to be a major mechanism for the immobilization of metals, such as Pb and Zn by phosphate compounds. These new compounds have extremely low solubility over a wide pH range, rendering phosphate application an attractive technology for managing metal-contaminated soils. The formation of the new solid phase (i.e., precipitates) occurs when the ionic product in the solution exceeds the solubility product of that phase. Recent studies using X-ray absorption fine structure (XAFS) indicate that formation of

surface precipitates may occur even when solution concentration is undersaturated with respect to homogeneous precipitation of the pure metal-precipitate phase. In normal soils, precipitation of metals is unlikely, but in highly metal-contaminated soils, this process can play a major role in immobilizing metals.

The ability of apatite to immobilize dissolved Pb^{2+} as precipitates is more commonly manifested as hydroxypyromorphite or as chloropyromorphite. Two processes for the reaction of dissolved Pb^{2+} with apatite are proposed: first, Pb can react with apatite through hydroxyapatite (HA, $Ca_{10}(PO_4)_6(OH)_2$); second, Pb^{2+} can substitute for Ca^{2+} in the apatite (eqn 3). Thus, (Ca, Pb) apatite could be potentially formed by adsorption of Pb or by dissolution of HA, followed by coprecipitation of mixed apatites. Dissolution of apatite (eqn 1) is an important initial step in the immobilization of Pb as pyromorphite.

$$Ca_{10}(PO_4)_6(OH)_2(s) + 14H^+(aq) \rightarrow 10Ca^{2+}(aq) + 6H_2PO_4^-(aq) + 2H_2O \quad [1]$$

$$10Pb^{2+}(aq) + 6H_2PO_4^-(aq) + 2H_2O \rightarrow Pb_{10}(PO_4)_6(OH)_2(s) \quad [2]$$

$$Ca_{10}(PO_4)_6(OH)_2(s) + xPb^{2+} \rightarrow \left(Ca_{(10-x)}Pb_x\right)(PO_4)_6(OH)_2(s) + xCa^{2+} \quad [3]$$

The immobilization of As, Cd, Pb, and Zn in smelter-contaminated soils using diammonium phosphate (DAP) is also possible. Application of high levels of DAP at a rate of 2300 mg P kg^{-1} could be very effective for immobilizing Cd, Pb, and Zn in the contaminated soil. Activity–ratio diagrams indicate that DAP decreases solution concentration of these metals by forming metal-phosphate precipitates with low solubility products.

Similarly, since bioavailability of metals to plants is typically greater in acidic than alkaline soils, neutralizing agents in the form of lime are usually added to acidic soils. Although the primary incentive in liming acidic arable soils is the suppression of toxic bioavailable Al and Mn to plants, liming is increasingly being practiced as a management tool to immobilize metals in soils, biosolids, and mine tailings, thereby reducing their bioavailability for plant uptake and transport to groundwater. In fact, alkalinization of biosolids is commonly done to enhance their immobilization potential for metals. Several reasons have been attributed to the lime-induced immobilization of metals: increases in negative charge (CEC) in variable-charge soils; formation of strongly bound hydroxyl metal species; precipitation of metals as hydroxides; and sequestration due to enhanced microbial activity. A range of liming materials are available, which vary in their ability to neutralize the acidity and immobilize metals: calcite ($CaCO_3$), burnt lime (CaO), slaked lime ($Ca(OH)_2$), dolomite ($CaMg(CO_3)_2$), and slag ($CaSiO_3$).

Since the main risk in highly contaminated soils arises from the mobility (transport) and bioavailability (uptake), and subsequent accumulation (biotoxicity) in biota, any attempt to remediate such sites should effectively control these two most important parameters. Accordingly, these two processes should be monitored by a battery of tests including, but not limited to, chemical speciation, selective and sequential extraction, microbial tests, plant performance, and quality.

Acknowledgment

Award number DE-FC09-96SR18546 from the US Department of Energy to the University of Georgia Research Foundation is gratefully acknowledged.

Further Reading

Adriano DC (2001) *Trace Elements in Terrestrial Environments: Biogeochemistry, Bioavailability and Risks of Metals*, 2nd edn. New York: Springer.

Bolan NS, Adriano DC, and Naidu R (2003) Role of phosphorus on (im)mobilization and bioavailability of heavy metals in soil-plant system. *Review of Environmental Contamination and Toxicology* 177: 1–44.

Bolan NS, Adriano DC, and Curtin D (2003) Soil acidification and liming interactions with nutrient and heavy metal transformation and bioavailability. *Advances in Agronomy* 72: 215–272.

Deutsches Institut für Normung (1995) *Soil Quality Extraction of Trace Elements with Ammonium Nitrate Solution*. DIN 19730. Berlin, Germany: Beuth-Verlag.

Lombi E, Wenzel WW, and Sletten RS (1999) Arsenic adsorption by soils and iron-oxide-coated sand: kinetics and reversibility. *Journal of Plant Nutrition and Soil Science* 162: 451–456.

Ross SM (ed.) (1994) *Toxic Metals in Soil–Plant Systems*. New York: John Wiley.

Sarret G, Vangronsveld J, Manceau A *et al.* (2001) Accumulation forms of Zn and Pb in *Phaseolus vulgaris* in the presence and absence of EDTA. *Environmental Science and Technology* 35: 2854–2859.

Vangronsveld J and Cunningham SD (1998) *Metal-Contaminated Soils: In Situ Inactivation and Phytorestoration*, pp. 265. Berlin, Germany: Springer-Verlag.

Vangronsveld J, Ruttens A, Mench M *et al.* (2002) *In situ* inactivation and phytoremediation of metal/metalloid contaminated soils: field experiments. In: Wise DL, Trantolo DJ, Cichon EJ, Inyang HI, and Stottmeister U

(eds) *Bioremediation of Contaminated Soils*, 2nd edn, pp. 859–884. New York: Marcel Dekker.

Wenzel WW, Lombi E, and Adriano DC (1999) Biogeochemical processes in the rhizosphere: role in phytoremediation of metal-polluted soils. In: Prasad NMV and Hagemeyer J (eds) *Heavy Metal Stress in Plants – From Molecules to Ecosystems*, pp. 273–303. Heidelberg, Germany: Springer-Verlag.

HILGARD, EUGENE WOLDEMAR

R Amundson, University of California–Berkeley, Berkeley, CA, USA

© 2005, Elsevier Ltd. All Rights Reserved.

The field of soil science emerged during the great scientific advances of the nineteenth century. During this formative period, it is likely that no one in the USA cast as large an intellectual and personal shadow as Eugene Hilgard. Hilgard's initial intellectual contribution to pedology occurred in his report on the geology and agriculture of the state of Mississippi, a report which presented an erudite and novel synopsis of soil formation and the soil geography of Mississippi. Hilgard's interests gradually shifted from his initial geological efforts toward soils and agriculture, and this growing interest in soils followed his academic transfer to Michigan and then ultimately to Berkeley, where the second half of his long career was dedicated to a great expansion of soil science education and research at both a state and a national level. His accomplishments and experiences seem to surpass the 83 years in which he lived, and they continue to serve as a standard with which modern achievements and accomplishments in the natural and agricultural sciences must be compared.

Hilgard was born in Zweibrücken, Germany, on 5 January 1833. His father, Theodore Erasmus Hilgard, was a prominent lawyer and served as Chief Justice of the Court of Appeals in his province. Hilgard's mother, Margaretha Pauli, was the descendant of generations of clergy, many of great importance to royalty and universities in France and Germany. Hilgard's parents possessed liberal views, and his father resigned his court appointment rather than adopt a revised and less progressive legal code. This resignation and the changing political landscape in Germany led to the Hilgard family's decision to emigrate to America. In 1836, the Hilgard family arrived in Belleville, Illinois, a settlement already home to an enclave of cultured and educated German families that had emigrated to the USA for political reasons.

Life in the Hilgard household of Belleville probably bore little resemblance to that of average pioneer families of the time. Eugene was the youngest of nine children, and with local schools being considered unsatisfactory by the family, Hilgard's father undertook the education of the children himself. Emphasis on mathematics and language was particularly strong, and, being educated in largely French institutions, Hilgard's father imposed two 'French only' days at the home each week (forcing the young boys to retreat outdoors to speak German). The children were encouraged to study the local botany and natural history, and this effort was greatly aided by the family's rich library collection. The breadth and rigor of this early education are made even more impressive by the fact that, in subsequent studies at universities or technical institutes, Hilgard was commonly far ahead of his peer group in academic preparation and abilities.

Although Belleville offered many opportunities and experiences, mosquitoes bore typhoid and malaria, the latter of which is believed to have caused fevers and sickness which plagued him for much of the remainder of his life. Malaria is also believed to have contributed to a weakening of his eyes, which also posed a lifelong handicap. At the age of 15, suffering from malarial maladies and greatly missing his older brother Theodore, who had left to study medicine in Europe, Hilgard's father let him travel with his brother Julius to Washington, DC, where Julius was an assistant in the US Coast Survey. Hilgard's time in Washington, and a later visit to Philadelphia, provided the young Hilgard with a variety of intellectual and scientific opportunities. Hilgard's growing interest in chemistry, and the few opportunities to study this emerging field in America, led Hilgard to travel to Heidelberg to study, where his brother Theodore was already a medical student.

At Heidelberg, Hilgard was an eager student, engaging in laboratory experiments and attending lectures. However, many of the courses were found to be, in Hilgard's view, uninspiring, and the social code was found to be unacceptable and disappointing. These deficiencies, combined with political unrest in Germany, led Hilgard and his brother to travel to Zurich. In Zurich, Hilgard spent three

semesters studying botany and geology, and, due to the rigor of his preparatory education, was soon selected to be an assistant in chemistry both in lectures and in the laboratory. Following this brief stay in Zurich, Hilgard moved to the Mining Academy at Freiburg for advanced study in mining and metallurgy. The bleak physical surroundings, unsocial conditions, and physical hazards at Freiburg ultimately forced Hilgard's departure. The final blow came when Hilgard was overcome by mercury vapor during a distillation process, causing Hilgard's physician brother to declare that Eugene should follow the outdoor life of a geologist or botanist, medical advice he provided consistently on multiple occasions during later years.

Leaving Freiburg, and after engaging in recuperative exercise in the Swiss countryside, Hilgard returned to Heidelberg, this time to more advantageous academic prospects. He was soon drawn to the recently hired Professor Robert Bunsen (of Bunsen burner fame), whose research interests spanned geology to chemistry, much as Hilgard's interests did. Bunsen encouraged Hilgard to choose, as his doctoral thesis, an enquiry into the chemistry of the four components of a burner flame. Hilgard produced a definitive dissertation on the chemistry of the gases in the flame components and, on 3 October 1853, received a degree that was bestowed with the university's highest honor, *summa cum laude*. Hilgard was then only 20 years old and, amazingly, had not received any formal degrees of any type prior to that point.

Following his dissertation examination, Hilgard suffered a relapse in his health, and was advised to travel to a suitable climate, this time Spain. While there, he engaged himself by learning the language and customs, and his love of music led to his introduction to the talented daughter of a retired Colonel, Jesusa Bello, who in 1860 became his wife. In the spring of 1855, his health improved, Hilgard obtained passage to New York and then to Washington, where, for lack of other suitable employment, he continued his chemical research at the Smithsonian Institution. Soon, Hilgard received an offer to be the assistant to the State Geologist of Mississippi, a position in which he was adequately forewarned not only of the difficult director then in charge, but also of the lack of scientific challenges in the Paleozoic formations of the state. But Hilgard accepted the position, setting into motion challenges which served to test his intellect, creativity, fortitude, and political talents – an arsenal of skills which served him well for nearly 50 more years.

In retrospect, one cannot envision a more unlikely launching site for a such an impressive career, or for many of the initial conceptual and quantitative developments in soil science. As Fredrick Slate wrote, in the National Academy's biographic memoir of Hilgard, "In all these relations, it seems probable that Hilgard's training, because it was foreign, must have helped him to a certain detachment that would favor freshness in attack and in conception of aims." Indeed, Hilgard's supervisor in Mississippi proved to be ineffective, if not incompetent, and soon left the position in 1857. Hilgard used this intervening period to his advantage, meeting and consulting with the geologists of the neighboring states of Kentucky, Arkansas, and Alabama. Notably, Dr. David Dale Owen, of Indiana, provided Hilgard with the advice to place emphasis on the soils and agriculture of the state, advice that Hilgard clearly seized. Following a trip to Washington in 1857, to escape the intolerable conditions under the former Head Geologist, Hilgard returned to Mississippi and was himself appointed State Geologist.

Only 25 years old, and for the most part intellectually isolated, Hilgard set into motion his plan for the geological investigation of the state. Legislative support and approval of his work were initially almost withdrawn, but political skirmishing succeeded in reestablishing the survey and providing Hilgard the time to complete his report. The completed document was published in 1860, but was not released until 1866 due to the onset of the Civil War.

The 1860 report is remarkable in that it, and subsequent investigations, established Hilgard as a preeminent scholar in the already established field of geology, later earning him the praise of being the 'father of the American Tertiary.' Hilgard's files of correspondence contain letters from, among others, Louis Agassiz and the father of geology himself, Sir Charles Lyell. Hilgard's work in Mississippi earned him the later opportunities to explore and clarify the geology of Louisiana, and the 'Mississippi Embayment,' and many of his early publications center on a variety of geological and paleontological problems of the then 'Southwest.' This eminent standing in the geological community was later important in Hilgard's efforts to establish a federal program of soil survey.

However, even more innovative was the second half of the 1860 report, which focused on the principles of agriculture and the agricultural features of the state. The first portion, 'Principles of Rational Agriculture,' is truly remarkable. It begins, most interestingly, with a question: 'What is a soil?' Hilgard answers it, and addresses soil's formation and relationship to plant production, in a most rigorous and enlightening discussion. It is important that Hilgard posed the question using the article 'a,' which

explicitly indicates that he recognized that soils are naturally varying entities that, in a given location, possess characteristic features. In Hilgard's mind, at the time, soil (in the present sense) contained 'soil,' or what we might now call the A horizon, and 'subsoil,' the material lying below the 'soil.' Though not expressing it in these terms, Hilgard therefore recognized the 'profile' and 'horizon' concept. More importantly, he presented a revealing discussion of soil biogeochemistry, including a discourse on the oxidation and reduction of iron compounds, and included a discussion on the formation and role of humus in agricultural practices.

Published in 1860, the report sat in storage until the end of the war. During this period, Hilgard retained his position as State Geologist, and (as a sign of the regard in which he was held by the Legislature) he continued to receive a salary of $1250 per year and an account to buy chemicals and carry out analyses. He also served the Confederacy by exploring for nitrates, and he attempted to install calcium lights on the bluffs above Vicksburg to illuminate federal gunboats – a plan that ultimately failed. He was also placed in charge of university property during the war, maintaining the integrity of the campus through the hostilities and during periods when it served as a hospital for the wounded.

In 1866, due to political insecurities, Hilgard joined the university as a professor of chemistry, the beginning of his long and varied academic career, one which marked a continuous shift from a primarily geological focus (though he maintained such an interest to the end) to one in soils and agriculture. In 1871, his professorial title was changed to Professor of Experimental and Agricultural Chemistry. This same year, Hilgard was the delegate from his university to the first convention of agricultural colleges established under the Morrill Act of 1862. While there, Hilgard first met D.C. Gilman, who at that time was at Yale. A few years later, it would be Gilman, then President of the fledgling University of California, who would induce Hilgard to move to Berkeley.

However, in the meantime, Mississippi moved to expand academic programs and instruction in agriculture. A small student enrollment, and political opposition and threats to separate the Agricultural College from the remainder of the campus, ultimately seemed to drain Hilgard and, in 1873, he accepted an offer, that had twice before been extended, to become Professor of Geology and Natural History at the University of Michigan in Ann Arbor.

While Ann Arbor, reportedly the premier public institution at that time, offered the cultured Hilgard family (which by that time included three children) a stimulating atmosphere, a number of issues forced Hilgard to look far beyond Michigan for his future. First, he was obligated to teach large classes, leaving him little time for original research. Second, the agricultural components of the Michigan educational system were located in Lansing, a situation which Hilgard unsuccessfully lobbied against. And finally, to a family of southerners (and a wife from Spain), the Michigan winters proved to be unbearable. In 1874, while attending a meeting of the American Association in Hartford, Connecticut, Hilgard again met Gilman, who was now President of the University of California, and who was also looking for a suitable professor of agriculture (after recently dismissing the first occupant of this position). It is reported that Gilman immediately offered the Chair to Hilgard, but Hilgard demurred. Gilman then succeeded in inducing Hilgard to travel with him back to California and to take a leave of absence from Michigan. This request was, as might be expected, met with strong objections by President Angell of Michigan, who "warned strenuously against going to such a hornet's nest as the university at Berkeley was reputed to be." None the less, Hilgard took leave, met Gilman in Chicago, and crossed the continent in the company of the president of the new University of California. Upon his arrival in Berkeley, Hilgard delivered lectures over a 6-week period on 'The Origin, Properties, and Functions of Soils,' which was attended by a relatively large group of students, many of them graduates.

While all is reported to have gone well during the visit, Hilgard returned to Ann Arbor "without a lingering thought of returning to California." However, yet another brutal winter and a medical relapse caused Hilgard to change his mind and, in 1875, he accepted Berkeley's offer. As would be expected, the Michigan faculty did their best to prevent the the loss of Hilgard (who was elected to the National Academy of Sciences in 1873). President Angell even promised to keep Hilgard's chair open for 6 months should he eventually change his mind.

Upon his arrival in Berkeley, it certainly must have entered Hilgard's mind to return directly to Michigan. His recollection of that period, and his response to it, summarizes the monumental struggle, and cultural sacrifices, that Hilgard subjected himself and his family to in what was now the second half of his career:

> When, after an uneventful overland voyage, we landed in Oakland, I found to my intense disappointment that Dr. Gilman was on the point of leaving the University of California for the presidency of Johns Hopkins University, at Baltimore. Had I known this, I might have never left Ann Arbor, and I thought for a moment of turning round immediately. But Dr. Gilman protested that

> I ought at least to try and study the situation; and, although he was leaving, I would have the strong help of the Board of Regents, and a new and fruitful field of investigation instead of the mere routine of teaching in Ann Arbor. And he gave me as his farewell a numerously attended reception at which he introduced me to a large number of influential people with warm recommendations for aid in my task. So I did not return to Ann Arbor, but resolved to fight it out as I had done in Mississippi before. The fight turned out to be quite as hard as anticipated, for a number of years. But I won in the end; and the climatic advantages proved for myself and my family an ample offset in the prolongation of life and health to the social pleasantness we had relinquished in making the change.

The cultural and academic contrasts between Berkeley and Ann Arbor at the time must have been severe. Photographs of the Berkeley campus at the time show a few isolated university structures in the midst of what appears to be pastureland. Streets were consistently reported to be rutted and muddy, and cattle and coyotes were reported to roam across the campus. To a man whose command of language (speaking German, Spanish, English, French; reading Sanskrit, Italian, Greek, Latin, and Portuguese) was second only to the university's Professor of German, this had to have been a dismal turn of events.

The Berkeley phase of Hilgard's career, beginning in 1875 as Professor of Agriculture and Botany, began with numerous challenges to his and the fledgling College of Agriculture's future. Internally, it is reported that suspicion by his colleagues over the role of agriculture in the university posed a challenge, one that was diffused by Hilgard's command of languages and literature. More significantly, opposition by the state's agricultural constituency, many of whom promoted the concept of a more practical, as opposed to scientific, approach to agricultural education, was a serious issue. Hilgard diffused this opposition head on, through a series of public lectures and discussions with the agriculturalists of the state. His efforts on public outreach and education seem extraordinary to the modern academician and certainly give a sense of the provincial nature of the state at that time and to the tenuous foothold that agricultural education held. Hilgard never ceased his outreach and educational efforts, writing and lobbying in newspapers and the general press for the remainder of his career.

However, to the scientific and international community, it was Hilgard's reinvigorated research, stimulated by new circumstances and climates, that is the hallmark of this period. It is challenging, and ultimately a disservice, to try to summarize his accomplishments in such a brief fashion. Hilgard's bibliography lists 229 publications: books (including his famous *Soils* of 1906), multivolume reports, newspaper and magazine articles, an abundance of research articles in the premier journals such as *Science, American Journal of Science*, etc. – on topics ranging from geology, soils, and plant pathology to education and language. In addition, due to his poor eyesight, Hilgard possessed a typewriter as early as 1876. As a result of his scientific notoriety, his records include carbon copies of more than 20 000 letters to people ranging from local farmers to Sir Charles Lyell. The range of his intellectual prowess, and his sheer fortitude to maintain scholarly output and contact, is staggering even to those now accustomed to various means of computer- and electronic-aided correspondence and publishing.

Hilgard's biographers have consistently emphasized several key achievements during his Berkeley era, which will be highlighted here. First, Hilgard accepted the commission by the federal government to conduct and publish the Cotton Census for the southern states and California (1884). The commission provided him with $25 000 to hire staff, collect samples, conduct analyses of cotton and, most importantly, soils – putting Hilgard's developing College of Agriculture on the national stage (and, of additional importance, at no cost to the state of California). On the surface this may appear an unlikely undertaking for someone of Hilgard's background, but a perusal of the California report reveals Hilgard's intent. The report is primarily the first synthesis of the soil geography of the state, complete with regional discussions and chemical analyses. Included also is the first 'Agricultural Map' of the state (outlining the geographical/soil provinces in a way still used today).

The second contribution mentioned here is Hilgard's quantitative attack on the origin and remediation of soil salinity and sodicity. California, like many arid and semiarid regions, contains soils in certain geographical settings that naturally possess high levels of salt. These are also areas which, due to irrigation, are susceptible to human-induced expansions of salinity due to poor irrigation management. Remarkably, Hilgard (using his nineteenth-century chemical technology) was able to decipher the general framework of the related, but differing, roles of various salts and sodium on crop production and water infiltration. He was able to determine that the application of gypsum would prove to be an effective means of remediating sodic soils, and thereby improving structure and hydraulic characteristics. He also, as part of this work, perceptively wrote that irrigation systems without adequate drainage are doomed to waterlogging and salinization, a

principle that has yet to be embraced by the state of California's government and its citizenry. The work became widely disseminated around the world, in regions, arid or not, where salt problems hindered agricultural productivity.

Hilgard's displacement to the semiarid California, and his summer expeditions in the arid Pacific Northwest as a consultant for his railroad 'king' cousin Henry Villard (on the Northern Transcontinental Railroad Survey), provided him with insights into arid soils. These trips provided a stark contrast to his soil experiences in Mississippi. Thus, he developed a clearly conceived comparison of soils and soil-forming processes in the arid versus humid regions of the country, focusing in particular on carbonate and plant essential nutrients.

Among Hilgard's great efforts at Berkeley is an equally important 'near miss' in an activity of far-ranging importance and intrigue. Hilgard's experiences in Mississippi and during the Cotton Census effort sharpened his view that the nation needed an organized program of soil (or, using his term, 'agricultural') survey. His reading of the charter of the newly formed US Geological Survey (USGS; one of the first, if not the first, federally funded scientific programs) revealed that one of its missions was the 'classification of public lands,' which Hilgard viewed as including soils. Hilgard's attention on the USGS coincided with the selection of John Wesley Powell as its director (1881). Powell, like Hilgard, possessed broad-ranging interests, and additionally shared Hilgard's concern over the distribution and use of land. Powell's tenure as USGS director was marked by great highs and lows and, while agricultural survey was but a small part of his overall concern, he responded to Hilgard's (sometimes almost annoying) lobbying by offering him a position in the USGS to undertake the agricultural survey program. Hilgard, now fully entrenched at Berkeley and in the agricultural arena, declined this overture, but continued to lobby his eastern colleagues (many with powerful political connections) to push Powell to engage in agricultural pursuits.

In 1888, the Agriculture Department was elevated to a cabinet-level department, and Hilgard was nominated (and approved by Congress) to be the Assistant Secretary in charge of scientific investigations. This would have given Hilgard the opportunity personally to formulate a federal survey program. Additionally, the new US Department of Agriculture (USDA) appealed to Powell as a more stable funding base for the USGS, which then as now was part of the Department of Interior. To the shock and dismay of his eastern colleagues, Hilgard turned down the USDA position. The reasons that Hilgard cited for declining included issues which by then were well known to those familiar with his writings: the unfavorable climate of Washington and the lack of financial compensation (which he struggled to maintain for both himself and his research programs throughout his career). Additionally, one senses, in reviewing his correspondence, that he was hesitant to relinquish the fame and stature he enjoyed in Berkeley and in California (he was well acquainted with senators, governors, and business leaders in the region).

This decision proved a disaster for Hilgard, and produced ramifications for soil survey and the earth sciences in the USA that continue today. First, his friends and colleagues were embarrassed by his decision (since the congress and president had already been convinced on Hilgard). Second, Powell, lacking personal support within the USDA, never succeeded in transferring the USGS to the USDA and he himself was then subjected to a series of continuing political attacks which eventually forced his retirement. Third, while soil survey did indeed develop in the USDA (1899), it was headed by a young and ambitious scientist named Milton Whitney. Early correspondence between the two men reveals that Hilgard was amused by, and somewhat paternal to, Whitney. However, this early bemusement changed to raging opposition with Whitney's development of the Soil Survey. This controversy hinged on Whitney's inexplicable dismissal of the importance of soil chemical analyses to agriculture, and his belief that soil physical properties were the primary way in which soils affected vegetation. Hans Jenny, in his biography of Hilgard, paints a detailed portrait of this conflict, one in which Hilgard, by his own choice isolated in California, could only agitate against but could not change. The new soil survey, and scientific investigations associated with it, centered squarely on soil physical properties and analyses. Hilgard rose to the defense of his colleague, F.H. King, who had been Head of the Soil Management Investigations of the soil survey under Whitney. King was forced out of this position by Whitney, and three of his manscripts were censored by Whitney, forcing King to publish them privately. Hilgard rallied to King's defense in *Science*, and positive support for his defense came from such prominent scientists as T.C. Chamberlin.

Whitney's ascent corresponded with Hilgard's waning but, in typical fashion, highly productive years on the Berkeley faculty. Hilgard was obligated to retire in 1906, the first 'victim' of the university's emeritus policy. He continued to write. His book *Soils*

was published 1906; and he coauthored a book on *Agriculture for Schools of the Pacific Slope*, and published numerous articles and comments.

The late nineteenth and early twentieth century was also a period for Hilgard to receive the awards and accolades that his work warranted. He was only the second graduate of Heidelberg to receive the 'Golden Degree,' commemorating the fiftieth anniversary of his dissertation. He received honorary doctorate degrees from Columbia, Michigan, Mississippi, and Berkeley. In regards to the latter award at Berkeley, received at a relatively advanced age in 1914, Hilgard remarked that he lived to see this award because "my persistence in living beyond the scriptural years carried the day." When Hilgard died, in early 1916, of pneumonia, the memorial accolades from the University of California and the National Academy of Sciences give the modern reader a keen appreciation of his stature, and the regard and esteem in which he was held.

Historians of science frequently note the fleeting nature of scientific stature and prominence in a human enterprise that unavoidably honors and discusses only the most recent publications, concepts, and trends. Even on the Berkeley campus, and from within Hilgard Hall where I write this biography, it is rare to find modern scholars who are familiar with Hilgard's life and contributions. Yet, in soil science, and in pedology in particular, his ground-breaking work and remarkable concepts resonate in much of what we do and think, in most cases without citation. Even courses still taught on the Berkeley campus originate from those first taught by Hilgard in the late nineteenth century, again without recognition. So, even at the beginning of the twenty-first century, his intellectual influence still persists, albeit in muted tones.

But Hilgard's contemporaries, possibly aware of the transient nature of academic fame, also chose to honor his contributions in ways of great permanence. To this day, Hilgard Avenues are found in Berkeley and Los Angeles, *Hilgardia* exists as a publication of the University of California, and Mount Hilgard, a towering massif of 13 361 ft, is found in a 'hall' of Sierra Nevada peaks named in honor of the great natural scientists of the nineteenth century. It is enlightening to read of the naming of Mount Hilgard as described by the surveyor who explored that region: "Mount Hilgard from the west is a stiking mass . . . It was named at the suggestion of an admiring former pupil of Professor Hilgard, Mr. Ernest C. Bonner, who accompanied me on one of my outings." It is instructive that a natural monument of such magnitude be chosen by a member of the generation that followed and was inspired by Hilgard. His work will surprise and stimulate those of the present generation who wish to step back in time and walk again with one of the giants of the natural sciences.

Further Reading

Amundson R and Yaalon D (1995) Hilgard EW and John Wesley Powell: efforts for a joint agricultural and geological survey. *Soil Science Society of America Journal* 59: 4–13.

Hilgard EW (1860) *Report on the Geology and Agriculture of the State of Mississippi*. Jackson, MS: E Barlesdale.

Hilgard EW (1884) General discussion of the cotton production of the United States: embracing the cotton-oil industry, methods and utility of soil investigation, and tables of cotton fiber measurements. In: Hilgard EW (ed.) (special agent in charge) *Report on the Cotton Production in the United States; Also Embracing Agricultural and Physico-geographic Descriptions of the Several Cotton States and of California*, vol. 5, pp. 7–93. Washington, DC: US Census Office.

Hilgard EW (1892) A report on the relations of soil to climate. *US Department of Agriculture Weather Bureau Bulletin* 3: 1–59.

Jenny H (1961) *E.W. Hilgard and the Birth of Modern Soil Science*. Pisa, Italy: Collana della Rivista agrochimica.

Slate F (1919) Biographical memoir of Eugene Woldemar Hilgard. *National Academy of Sciences Biographical Memoirs* 9: 95–155.

Wheeler BI, Wickson EJ, Loughridge RH, and Smith EA (1916) In memoriam: Eugene Woldemar Hilgard. *University of California Chronicle* 18: 1–50.

HOOGHOUDT, SYMEN BAREND

P A C Raats, Wageningen University, Wageningen, The Netherlands
R R van der Ploeg, University of Hannover, Hannover, Germany

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

In the absence of dikes against sea and river water, more than half of The Netherlands would be flooded. Moreover, The Netherlands has a humid climate with, on average, precipitation exceeding evapotranspiration during autumn, winter, and early spring. Historically, crop yield records show a strong correlation between wet autumn and winter seasons and later low crop yields. However, from about 1850, the ability to get rid of excess water was the determining factor in crop yield. In the first half of the nineteenth century, much effort was put into modernizing the design, maintenance, and management of waterways. Replacement of windmills by steam-driven pumping stations started in 1845 and made removal of water independent of the whims of the weather. At the field scale, originally the removal of water was by a ridge-and-furrow system or a dense network of trenches. Already early in the nineteenth century, trenches filled with brushwood covered with soil were used incidentally, but such drains were found to decay rapidly. Earthenware tiles were introduced in England around 1845, and the first experiments in The Netherlands with these 'drying pipes' were done in 1851. Gradually, ditches were replaced by tiles of 0.05 m inside diameter, with a slope of 0.01–0.02 m m^{-1}, and draining into the remaining main ditches spaced 100–300 m apart, the larger spacings always being used in new polders. The choice of drain depth and distance depended on the nature of the soil and the drainage requirement. Leading the introduction of tile drainage, especially on arable land, was the province of Groningen, in the northeast, followed by the province of Zeeland, in the southwest, and the polders being reclaimed from the IJsselmeer. When, around 1960, tube drains succeeded tile drains, the practice also spread to other parts of the country. The drains not only serve to remove water, but in late spring and summer may also supply water from the main ditches to the fields by subirrigation.

In the 1850s, drainage experiments were conducted by the French engineer S.C. Delacroix. In the 1860s this work was noted by the Danish engineer D. Hannemann, who in turn inspired Ludvig August Colding (1815–88) to derive the ellipse equation. This Dupuit-type equation describes the shape of the water table for steady flow Q, resulting from uniform input s at the water table, to tile drains spaced l apart and resting on an impermeable base (Figure 1):

$$Q = ls = \frac{4kH_0^2}{l} \qquad [1]$$

with k representing the hydraulic conductivity of the soil and H_0 the height of the water table above the impermeable base at the midpoint between the drains. The ellipse equation was widely used in Denmark and also became known in other European countries. In The Netherlands it was used only sporadically, but nevertheless became the point of departure for S.B. Hooghoudt's drainage research in the early 1930s.

From Physical Chemistry to First Contacts with Drainage Theory and Practice

Symen Barend Hooghoudt (Figure 2) was born on August 31, 1901 at Slochteren, in the province of Groningen, in the Netherlands. He studied physical chemistry at the State University of Groningen, receiving a BSc degree in 1922, an MSc degree in 1926, and a doctoral degree in 1928. His doctoral thesis, supervised by Frans Maurits Jaeger (1877–1945), dealt with a method for the accurate measurement of the Becquerel effect. This photovoltaic effect is the basic physical process through which a solar cell converts sunlight into electricity.

Hooghoudt joined the Soil Science Institute at Groningen on March 16, 1929 and remained with it till his death on August 30, 1953. In the late 1920s, the Institute's Director, David Jacobus Hissink

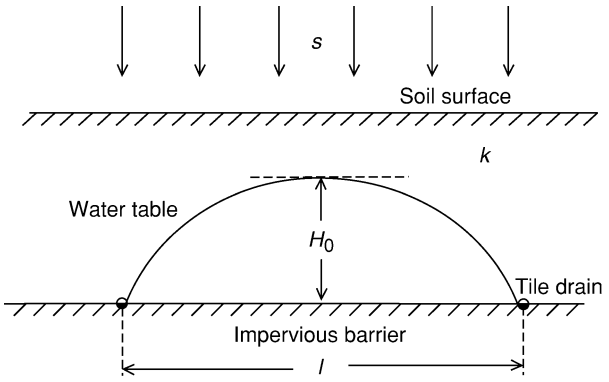


Figure 1 Flow to tile drains, as considered by L.A. Colding in 1872.

Figure 2 Symen Barend Hooghoudt (1901–53). (Reproduced with permission from Bruin P and Visser WC (1953) Het werk van Dr S.B. Hooghoudt in dienst van het landbouwkundig onderzoek. *Landbouwkundig Tijdschrift* 65: 650–657, © KLV Royal Netherlands Society for Agricultural Sciences.)

(1874–1956), was heavily involved in discussions within the International Society of Soil Science (ISSS) on the measurement of soil pH and its use as a diagnostic tool. The chinhydron electrode was the best method available, but unfortunately it could not be used in soils high in manganese and pH >8.5. Hooghoudt's first assignment was to study the antimony electrode as an alternative for soils with pH >8.5. Soon after this study was completed successfully, the glass electrode was introduced, allowing simple and rapid measurement up to pH 12–13, and as a result the antimony electrode was widely considered obsolete almost immediately.

In 1930–31 Hooghoudt was involved with Hissink in a first study of some physical properties in a wide range of Dutch soils, namely the permeability for water, the dry bulk density, and the air capacity, using methods devised by J. Kopecky early in the twentieth century. The prime motivation was the need for guidance in developing drainage systems for the planned IJsselmeerpolders, in particular the pilot polder Andijk and the first full-scale polder Wieringermeer. For reference, subsidence and prognosis of temporal changes in soil properties were studied in a series of polders reclaimed in 1550, 1844, 1924, and 1925. Several fields devoted to studies of drainage or subirrigation and an area intended to be used for a complex of sports and playing fields were also included in the study. The practical context exemplified by these field sites remained a hallmark of Hooghoudt's research.

The soil permeability was expressed in terms of the *D*-value, defined as the amount of water in meters per day flowing under a ponding depth of 0.04 m vertically through a sample of 0.07 m freely into the air. Despite its operational nature, the *D*-values enabled Hissink and Hooghoudt to delineate important aspects of the flow of water in Dutch soils. It was shown that for sandy soils without structure the particle-size distribution can be used to estimate the *D*-value, but that large structural pores dominate the conduction of water in all other soils. In soils with such structural pores, the *D*-values measured with ring samples of 3850 mm^2 cross section were found to be extremely variable, even at very small distances.

In December 1931, Hooghoudt went on a study tour to Breslau, Prague, and Zurich. At Breslau, F. Zunker studied the dependence of various physical properties upon the specific surface derived from the particle-size distribution. In his trip report, Hooghoudt emphasized that Zunker's methods could only be used for sandy and very light loamy soils. At Prague, Hooghoudt learned about many practical aspects of soil and water management. But he found little of direct use, mainly because the structure of the soils differed too much from the generally rather young Dutch soils with comparable distributions of particle sizes, and also because some of the concepts being used were somewhat esoteric. The visit strengthened his belief that Dutch soils reclaimed from the sea have unique properties and that field methods were needed to determine soil properties relevant to the drainage process. The visit to E. Diserens at Zurich was by far the most fruitful. It brought Hooghoudt in contact with an analysis of drainage problems based on what he called the law of Dupuit–Darcy, and the associated laboratory and field methods for determining the required soil physical properties.

In July 1932, the 6th Commission of the ISSS held a meeting at Groningen on water management, with emphasis on drainage and reclamation of land from the sea. Together, the ISSS contacts, the impressions from the European study tour, the first study of soil physical properties with Hissink, and the practical needs of water management in The Netherlands shaped Hooghoudt's research plans for the next 10 years.

Physical Characterization of Soils in the Laboratory

From the very beginning, attempts were made to relate the physical properties of soils to the composition of the solid phase and the volume fractions of the solid, aqueous, and gaseous phases. The composition of the solid phase was given in terms of the contents of humus and calcium carbonate, and the content and size distribution of particles other than humus and calcium carbonate. At Groningen, the size distribution of particles was originally determined using either the Atterberg sedimentation cylinder method for the range 2–30 μm or the Kopecky multiple-tube elutriation method for the range 16–100 μm, complemented with sieving of the coarse fraction to 2000 μm. Hooghoudt made very detailed evaluations of these and other methods. The 323-page report of this methodological study includes a 59-page review of validity, limitations, and extensions of the Stokes law for the rate of fall of solid particles in a fluid. This report was submitted in August 1944. As for many scientists, for Hooghoudt scholarly work of this kind was a way to cope with the stagnation caused by the war-time conditions. In 1946 this study resulted in a switch to the combined sieve–pipette method for routine determinations, the pipette method being used for the fraction less than 35 μm and the sieve method for the fraction greater than 35 μm.

Following F. Zunker, Hooghoudt introduced the specific surface, U, which is defined as the ratio of the surface area of a given mass of soil divided by the surface area of this same mass of soil in the form of spheres having a reference diameter $\sigma_r = 1$ cm. He showed that if $w = f(\sigma)$, where w is the weight fraction of the total amount of particles smaller than a specific particle diameter σ, then the specific surface U is given by:

$$U = \sigma_r \int_{w=0}^{w=1} \sigma^{-1} \mathrm{d}f = \sigma_\mathrm{r} \int_{\sigma=\sigma_1}^{\sigma=\sigma_2} \sigma^{-1} (\mathrm{d}f/\mathrm{d}\sigma)\mathrm{d}\sigma \qquad [2]$$

A hypothetical soil with uniform particle diameter $\sigma_{\mathrm{eff}} = \sigma_\mathrm{r}/U$ has the same specific surface U as the actual soil. Hooghoudt refers to σ_{eff} as the effective particle diameter of the actual soil. The specific surface area U and effective particle size σ_{eff} derived from it are particularly relevant in the range 16–2000 μm. Clearly, U and σ_{eff} are proper measures of the coarseness or fineness of a soil and σ_{eff} can be regarded as a characteristic microscopic length scale of a soil, comparable with that used in the Miller–Miller scaling theory of modern soil physics. Consistent with this, in expressions used by Hooghoudt, the hydraulic conductivity is inversely proportional to U^2 and the height of capillary rise is proportional to U.

From the very beginning, Hooghoudt distinguished two broad classes of soils: a class of soils without structure, and a class of soils with structure arising from aggregation, cracking, and perforation. For the class of soils without structure, the physical properties can be determined in the laboratory on disturbed samples and the results can be extrapolated to the field situation, using formulas involving the porosity, volume of entrapped air, and temperature. For the class of soils with structure in their natural condition, meaningful measurements of the hydraulic conductivity cannot be made in the laboratory, but must necessarily be made in the field. In fields without drainage systems, this can be done by the auger-hole method. In fields with drainage systems, the hydraulic conductivity can also be inferred from the relationship between the groundwater level and the flux into or out of drains and ditches.

The phase distribution of soils was characterized by the total porosity, considered as the sum of the drainable porosity and the water-filled porosity after drainage, and the air-filled porosity below the water table. Related to the dynamics of the water, Hooghoudt determined two physical parameters in the laboratory: the height of capillary rise and the hydraulic conductivity below the water table. To interpret data on height of capillary rise and hydraulic conductivity, he made use of rational formulas, showing the dependence upon the specific surface U and the porosity. By modern standards, the concepts used by Hooghoudt may seem simplistic, but they were perfectly suited for his goals.

Hooghoudt also paid some attention to rheological properties of soils, resulting in a 200-page publication on hardness and on rupture by various means of soils in the dried state. The aim was to get a basic understanding of the mechanical properties of dry clods and crusts. He used methods suitable for dried pastes borrowed from research related to the pottery and porcelain industry. He focused on the influence of particle-size distribution and the composition of the adsorption complex on the mechanical properties.

In the context of a large-scale study of irreversible drying of peat soils, Hooghoudt studied the causes of the phenomenon and developed a method to express the degree of irreversible drying by a single number.

Auger-Hole Method for Determining Hydraulic Conductivity in the Field

Observations of the rate of rise of the water level in an auger-hole following lowering of the water table by pumping can be used to calculate the hydraulic

conductivity. E. Diserens and J. Donat pioneered this method. Following his discussions with Diserens in December 1931, Hooghoudt greatly improved the underlying theory, arriving for homogeneous soils at a formula expressing the rate of rise of the water level as the sum of a term due to horizontal flow through the wall of the hole and a term due to vertical flow through the bottom of the hole. He also derived corresponding formulas allowing for soil heterogeneity, either in the form of layering or continuous variation with depth. The formulas were verified by laboratory tests in a 62.5-cm-diameter drum with a perforated copper tube placed at the center. In the experiments the diameter of the tube, the depth of the bottom of the tube below the undisturbed water table, the initial lowering of the water in the tube, and the location of the impermeable layer were varied. Both homogeneous and layered soils were used in the tests. To study the possible influence of the small diameter of the drum, tests were also performed in a soil bin 10 m long, 2 m wide, and 2 m deep.

By necessity the tests of the theory in the small soil drum and large soil bin were done with sandy soils. However, Hooghoudt also developed the equipment needed for measurements in the field, allowing measurements to be made not only in sandy soils, but also in structured soils. With this equipment, measurements were made in the Wieringermeerpolder in a drained field, allowing comparison with hydraulic conductivity inferred from the relationship between groundwater levels and drain fluxes. The agreement between the hydraulic conductivities determined with the two methods was quite satisfactory, and Hooghoudt proceeded to use the auger-hole method widely and used the results to give advice with regard to design of drainage systems. In his extensive 1936 report on the auger-hole method, he gave as examples data for a proposed airfield near Leeuwarden and for the Rietwijkeroorderpolder near Amsterdam, intended to be developed as a recreational park.

The theory underlying the auger-hole method was developed further after World War II. At Groningen, L.F. Ernst in 1950 used J.J. van Deemter's numerical relaxation method to obtain a new formula to calculate the hydraulic conductivity from auger-hole measurements. In the USA, Don Kirkham and his associates made important contributions from 1948 onward. Based on their very general analytical theory for auger-hole seepage, C. Boast and D. Kirkham in 1971 concluded:

> Hooghoudt's formula should be used only for a restricted range of geometries. Ernst's formula or graph, or table we give should be used otherwise.

Flow of Water in Shallow Soils Without or With Structure in Their Natural Condition

In a 1937 publication, Hooghoudt developed a theory for flow to ditches and drains in shallow soils. Building upon the hydraulic approach of Colding and his followers, he assumed that the vertical distribution of the pressure is hydrostatic at any point in the field, and that as a consequence the slope of the water table can be regarded as the driving force for the flow in the horizontal direction. For structured soils, he ignored the contribution from horizontal flow in a capillary fringe. For a steady flow Q resulting from uniform input s at the water table, to ditches spaced l apart and reaching to the impermeable base, and with a water table everywhere below the soil surface, Hooghoudt derived the following expression for the flux Q (Figure 3):

$$Q = ls = \frac{4k\left(H_0^2 - h_0^2\right)}{l} = k\frac{8h_0 + 4m_0}{l}m_0 \qquad [3]$$

where H_0 is the height of the water table above the impermeable base at the midpoint between the drains, h_0 is water level in the ditch, and $m_0 = H_0 h_0$. If the hydraulic conductivity k is known, say from measurements with the auger-hole method, then Eqn [3] can be used to calculate the desired drainage spacing l for given s, H_0, and h_0. Alternatively, if the hydraulic conductivity k is not known, it can be calculated if l, s, H_0, and h_0 for a drainage system are known.

To verify Eqn [3], Hooghoudt used a soil bin 10 m long, 2 m wide, and 2 m deep, filled with river sand. He found that the estimates of hydraulic conductivity on the basis of Eqn [3] agreed very well with laboratory data, provided the effect of flow in the capillary fringe of thickness h_c was included by replacing H_0, and h_0 by $H_0 + h_c$ and $h_0 + h_c$.

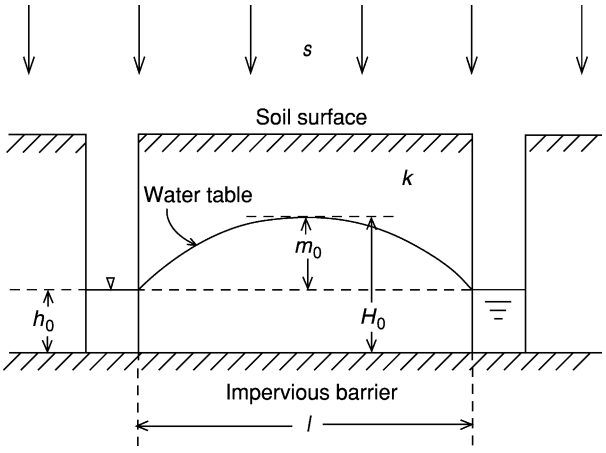


Figure 3 Flow to ditches, as considered by Hooghoudt in 1937.

Still for homogeneous soils, Hooghoudt considered various complicating factors:

- Situations with the input s exceeding the value for which the water table at the midpoint between the ditches or drains reaches the soil surface;
- Situations with the drain or the bottom of the ditch being some distance above the impermeable base. Tentatively these were handled by increasing l in the denominator of Eqn [3] by $r(r+b)/(2h_0)$ where b is the width of the bottom of the ditch ($b=0$ for a drain) and r is the depth of the impermeable layer below the bottom of the ditch or drain;
- Situations with flow not only to parallel drains and small ditches but also toward a larger ditch that intercepts the flow from the parallel drains or small ditches; for this case he also used the large soil bin to test the results.

Hooghoudt also extended the hydraulic approach to heterogeneous soils, considering in detail cases with jump discontinuities of the hydraulic conductivity at interfaces between horizontal, homogeneous layers, and cases with gradual changes of the hydraulic conductivity with vertical position. In effect he replaced, at any point between the drains, the hydraulic conductivity k by an apparent hydraulic conductivity k_{av}, being defined as the vertically averaged conductivity between the impermeable base and the location of the water table. Clearly, even with the heterogeneity being restricted to the vertical direction, k_{av} is inherently dependent on the location of the water table. Nevertheless, the results of the analysis can still be written in the form of Eqn [3], provided k is replaced by an effective conductivity k_{eff}, which is a function not only of the vertical distribution of the real conductivity, but also of H_0 and h_0. For example if the real conductivity k increases linearly with height y according to $k=k_0+ay$, where a is a constant, then the effective conductivity k_{eff} in an expression of the form of Eqn [3] will be given by:

$$k_{eff} = k_0 + \frac{\frac{1}{3}a\left(H_0^3 - h_0^3\right)}{H_0^2 - h_0^2}. \qquad [4]$$

Note that if $a=0$ then k_{eff} reduces to k_0.

The Hooghoudt Drainage Equation

In a 1940 publication, Hooghoudt removed the restriction of the impermeable base being relatively close to the drain or the bottom of the ditch. To this end he introduced the device of partitioning the flow region in two parts, one part away from the drain or ditch in which the flow is approximately horizontal and another part close to the drain or the ditch in which the flow is radial. The result was cast in a form similar to Eqn [3], with the radial resistance being accounted for by replacing h_0 by an equivalent thickness h_{eq} of the water-conducting layer below the drain, and allowing for the hydraulic conductivity in the layers above and below the drain having different values of, respectively, k_1 and k_2 (**Figure 4**):

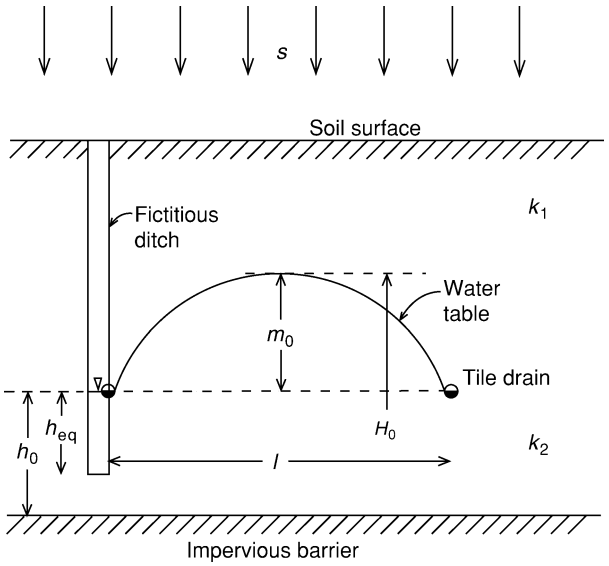


Figure 4 Flow to tile drains, as considered in the Hooghoudt drainage equation of 1940.

$$Q = ls = \frac{8k_2h_{eq} + 4k_1m_0}{l}m_0 \qquad [5]$$

with the equivalent thickness h_{eq} being a known function of drain spacing l, the real thickness of the water-conducting layer below the drain h_0, and the diameter of the drain. Eqn [5] is usually referred to as the Hooghoudt drainage equation.

Applications of Drainage Theory

The planning of reclamation of the IJsselmeerpolders stimulated a rapid development of the study of drainage of soils for agricultural purposes, with theory and practice going hand in hand. D.J. Hissink and his staff started activities in the pilot polder Andijk in the late 1920s and made extensive studies in the Wieringermeerpolder in the early 1930s. Hooghoudt was responsible for the scientific basis of the design of the drainage systems. In the same period, smaller polders were also studied. For example, for the Linthorst Homanpolder in the north of the province of Groningen a drainage plan was developed, and for

the Nijkspolder in the southeast of the province of Friesland a study was made of the expected increase in seepage into the polder as a result of an increase in the size of this polder and a lowering of the water levels in the open water within the polder. On an even smaller scale, numerous drainage recommendations were given for individual farmers' fields.

Three reports concerning sports fields were submitted in 1938. In the post World War II period 1947–51, 15 more reports followed. Recommendations included design of drainage systems and alterations of the soil profile to improve playability by changes in layering and by supplementing clay soils with sand. A comparison of the 1939 and 1953 guidelines for the design of soccer fields illustrates the progress.

Hooghoudt and his colleagues J.D. de Jong and A.J. Wiggers assisted the Laboratory of Soil Mechanics at Delft in a government assignment to make recommendations regarding suitability and possible improvements of sites for civilian and military airfields. Two reports concerning the military airfield Valkenburg in the province of South Holland were presented in 1939. In the post World War II period 1945–48, about 20 more reports were presented, concerning nearly as many locations.

The Committee Forestry Plan Amsterdam prepared an ambitious plan for a recreational park at the southwestern edge of the city of Amsterdam. Hooghoudt was asked to design a drainage plan. In the period 1936–39, Hooghoudt submitted eight reports concerning the soils in the area, including measurements of the hydraulic conductivity by the auger-hole method, and recommendations for spacing and depth of drains. Based on these, 300 km of tiles were laid.

Hooghoudt was also responsible for the first regional hydrologic study. In the Veenkoloniën, the thickness and the hydraulic conductivity of the water-conducting layer were studied to a depth of 20 m. On the basis of these data, the functioning of canals and ditches was analyzed. It was the first study at this scale, allowing a prognosis of the consequences of a variety of measures such as eliminating or reducing the size of canals originally dug for the purpose of transport of peat being mined for fuel, eliminating superfluous ditches, and the breaking of poorly conducting layers inside and underneath remaining peat layers.

Water Management of the Plant Root Zone

All along, it was of course realized that generally the purpose of drainage and subirrigation is to promote a favorable environment for the growth and functioning of plant roots. In February 1936, Otto de Vries (1881–1948), son of the botanist Hugo de Vries (1848–1935), gave a lecture on 'Soil, water, and plant' at the first meeting of the Dutch section of the ISSS. Following his 22 years in agricultural chemistry in the Dutch Indies, in 1930 Otto de Vries became director of the Agricultural Experiment Station and Soil Testing Laboratory at Groningen and, in 1939, following the retirement of Hissink, director of the Soil Science Institute. In 1943 the Departments for Agricultural Research and for Agricultural Industries of the Central National Council for Applied Scientific Research in the Netherlands (TNO) were founded and Otto de Vries was appointed chairman of both. He left Groningen in 1944 and lived in The Hague until his death in 1948.

In 1942 a field experiment for studying the influence of controlled water table depth upon the growth and yield of crops was realized at Nieuw Beerta in the eastern part of the province of Groningen. The experiment was continued till autumn 1955. The detailed description and interpretation of the results in eight publications illustrate Hooghoudt's evolving interest in the agronomical effects of drainage. Of particular interest is the early demonstration that crop response to drainage is complicated, being mediated by the influence of the water table depth on soil structure, plant root development, and the fate of soil nitrogen.

In 1943 a national Working Committee for Evaporation Research was formed, and in 1946 the more broadly based Committee for Hydrological Research TNO. Under the leadership of Hooghoudt, plans were made for a study of the evaporation from a polder, comparing three methods: (1) determining the water budget of the entire polder, (2) direct measurement of evaporation on the basis of a calculation of vertical vapor transport from measured vertical humidity and wind velocity profiles by the Royal Dutch Meteorological Institute, and (3) determining the soil-profile water balance using a lysimeter and soil-moisture sampling. For the latter purpose, the use of tensiometers and thermal conductivity probes was initially considered, but later abandoned for various reasons. The spatial variability of the soil moisture was recognized as a major obstacle. Hooghoudt played a leading role in the choice of the location and the detailed planning of this experiment. Perhaps not surprising, from a large number of possibilities the Rottegatspolder near Groningen was selected. Data collection started August 1, 1947 and terminated February 1, 1971, thus long after the agrohydrologic research had moved in 1956 from Groningen to Wageningen.

Mentor of Young Talents: The Theorists Van Deemter and Ernst

After World War II, Hooghoudt concentrated on the application of drainage theory and the exploration of its agronomical implications. Eventually he coached a large team working throughout the country. Further development and experimental verification of the theory were the responsibility of, respectively, J.J. van Deemter and L.F. Ernst.

Jan Josef van Deemter, after earning in 1945 an MSc degree in physics from the University of Groningen, was employed in the Agrohydrology group from June 1946 till some time in 1947, when he joined the Royal/Dutch Shell Laboratory at Amsterdam. On the basis of the results of his research started in the brief period at Groningen, in 1950 he was awarded a doctoral degree in applied mathematics at the University of Amsterdam. In his thesis, he adapted the hodograph method to give exact solutions of the problems of drainage and infiltration for systems of parallel, equidistant drains or ditches in soils with homogeneous hydraulic conductivity to infinite depth, including the effects of rainfall, evaporation, and upward or downward flow. He also used the so-called relaxation method of Southwell, which can be applied to all two-dimensional steady-state problems, as well as for soils with heterogeneous hydraulic conductivity.

After van Deemter left Groningen, he was succeeded by Lodewijk Ferdinand Ernst, who improved the auger-hole method and generalized the drainage formulas. Ernst became involved in several large-scale field studies and continued research in the style of Hooghoudt after he moved to Wageningen in 1956. In 1962, Ernst was awarded a doctoral degree in applied mathematics at the University of Utrecht on the basis of a thesis about groundwater flow in the saturated zone in the presence of parallel open conduits.

Hooghoudt's Legacy

In the autumn of 1952, there was a proposal to move the Agricultural Experiment Station and Institute for Soil Research TNO partly or entirely to a more central location at Wageningen. Hooghoudt was strongly opposed to such a move and pointed out that in his experience it was quite possible to lead joint research projects throughout the country from Groningen. At a staff meeting on December 9, 1952, he illustrated this with a description of recent and current agrohydrologic research by him and his team, emphasizing the great advantage of having agronomic and hydrologic expertise under one roof. Theory, experiment, and practical application were perfectly balanced in the research program of Hooghoudt.

Hooghoudt died at Groningen on August 30, 1953. At a memorial meeting on October 1, 1953, the director, P. Bruin, gave a detailed survey of his work. Three years later the agrohydrogic program was moved to Wageningen to become the core of the newly formed Institute for Land and Water Management Research (ICW). At the ICW, Wageningen University, and the International Institute for Land Reclamation and Improvement (ILRI; founded in 1955), Hooghoudt's ideas flourished and were promoted worldwide. The ILRI International Course on Land Drainage has, since its first session in 1962, drawn more than 1000 drainage professionals from all over the world. Dutch-born PhD students at Iowa State University and Cornell University have facilitated the export of some of Hooghoudt's ideas published mainly in Dutch to the USA and the English-speaking world in general. This was consolidated by the 1957 monograph *Drainage of Agricultural Lands*, published by the American Society of Agronomy, as well as in the later, updated editions of 1974 and 1999.

See also: **Drainage, Surface and Subsurface**

Further Reading

Bear FE (ed.) (1952) *Soil Science* 74: 1–101.

Bruin P and Visser WC (1953) Het werk van Dr S.B. Hooghoudt in dienst van het landbouwkundig onderzoek. *Landbouwkundig Tijdschrift* 65: 650–657.

Childs EC (1969) *An Introduction to the Physical Basis of Soil Water Phenomena,* Chapt. 16. London, UK: John Wiley.

De Vries JJ (1982) *Anderhalve Eeuw Hydrologisch Onderzoek in Nederland*. Overzicht van de ontwikkeling van de wetenschappelijke kennis van het water in Nederland tussen 1830 en 1980, beschouwd vanuit een geohydrologische gezichtshoek. Amsterdam, The Netherlands: Rodopi.

Harmsen K (1990) *Het Instituut voor Bodemvruchtbaarheid 1890–1990*. Haren (Gr), The Netherlands: Knoop.

Hissink DJ and Hooghoudt SB (1931) Bijdrage tot de kennis van eenige natuurkundige grootheden van den grond. Eerste mededeeling [Contribution to the knowledge of some physical parameters of soils. First commnunication]. *Reports of Agricultural Research* 37B: 101–190. The Hague, The Netherlands: General Government Printing and Publishing Office.

Hooghoudt SB (1934–1950) Bijdragen tot de kennis van van eenige natuurkundige grootheden van den grond, 2–10 [Contribution to the knowledge of some physical parameters of soils, 2–10]. *Reports of Agricultural Research* 40B: 215–345 (1934); 41(1)B: 589–615 (1935); 42(13)B: 449–541 (1936); 43(1)B: 1–11 (1937);

43(13)B: 461–676 (1937); 46(14)B: 515–707 (1940); 49(16)A: 721–913 (1943); 50(13)A: 671–993 (1945); 56(3): 1–209 (1950). The Hague, The Netherlands: General Government Printing and Publishing Office.

Luthin JJ (1957) *Drainage of Agricultural Lands*. Madison, WI: American Society of Agronomy.

Van Deemter JJ (1950) Bijdragen tot de kennis van enige natuurkundige grootheden van de grond, no. 11. Theoretical and numerical treatment of flow problems connected with drainage and irrigation (in Dutch). *Reports of Agricultural Research* 56.7: I–VIII, 1–67. The Hague, The Netherlands: General Government Printing and Publishing Office.

Van der Loeff JA (1954) *Publications and reports of Dr. S.B. Hooghoudt*†, *Principal Scientific Officer with the Agricultural Research Station and Soil Science Institute T.N.O. Groningen* (in Dutch). Literatuurlijst nr. 12. Centrum voor Landbouwdocumentatie. The Hague, The Netherlands: Ministerie van Landbouw, Visserij en Voedselvoorziening.

Van der Ploeg RR, Kirkham MB, and Marquardt M (1999) The Colding equation for soil drainage: its origin, evolution, and use. *Soil Science Society of America Journal* 63: 33–39.

Van der Ploeg RR, Marquardt M, and Kirkham D (1999) On the history of the ellipse equation for soil drainage. *Soil Science Society of America Journal* 61: 1604–1606.

Van Schilfgaarde J, Kirkham D, and Frevert RK (1956) Physical and mathematical theories of tile and ditch drainage and their usefulness in design. Research Bulletin 436. Ames, IA: Iowa Agricultural Experimental Station.

Zunker F (1930) Das Verhalten des Bodens zum Wasser. In: Blanck E (ed.) *Handbuch der Bodenlehre*, vol. 6, pp. 66–220, *Die physikalische Beschaffenheit des Bodens*. Berlin, Germany: Springer-Verlag.

HUMIFICATION

T C Balser, University of Wisconsin–Madison, Madison, WI, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Soil organic matter represents the largest dynamic, terrestrial carbon pool. While many aspects of the processes controlling its dynamics are understood, the mechanisms leading to the accumulation and storage of carbon in soil have not been substantially elucidated. The quantity of carbon in soil is greater than that in vegetation and the atmosphere combined (Figure 1), and a large part of that carbon is in a stable form, either physically protected or chemically recalcitrant (humified). Physically protected carbon is that associated with soil mineral surfaces. The process and degree of physical protection are topics that have been widely studied in order to quantify and predict carbon sequestration in soil. Soils with differing mineralogy will protect and sequester carbon to varying degrees (Figure 2). Humification (the alteration of biologically derived carbon to chemically complex forms) also represents a critical process driving carbon sequestration. Because humic substances (HS) in soil are highly recalcitrant, with turnover times ranging to thousands of years, humified carbon is sequestered carbon. While carbon sequestration has been the subject of attention recently due to changes in global climate and atmospheric chemistry, it has been known for more than a hundred years that soil carbon (particularly HS) is essential to soil fertility and productivity. HS are important regulators of trace metal concentrations in soil solution and may act as plant growth promoters. In addition, HS are critical determinants of soil properties such as water-holding capacity, structure and aggregation, and nutrient retention and exchange.

Soil microorganisms are dominantly responsible for the formation and turnover of HS in soil. Yet models

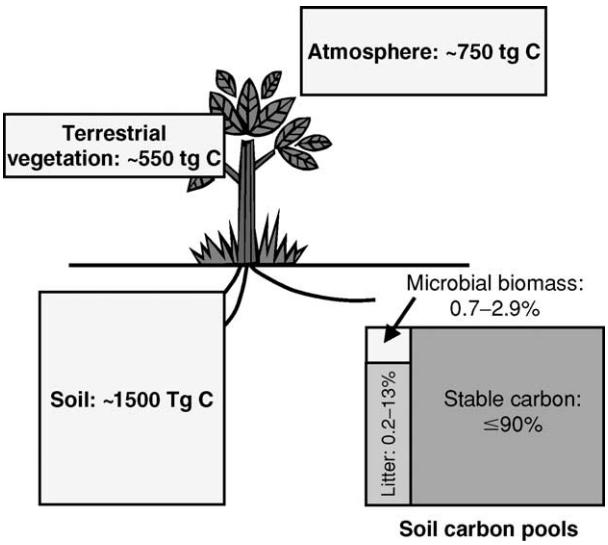


Figure 1 Global carbon pools. Soil carbon content is greater than that in terrestrial vegetation and the atmosphere combined. Up to 90% of soil carbon is in stable, humified form. Tg, teragrams.

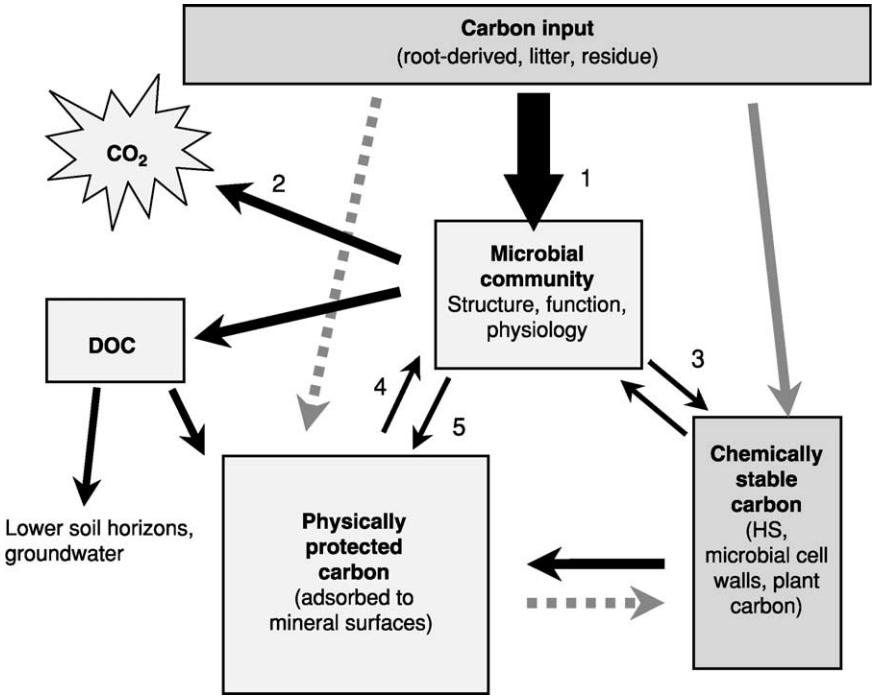


Figure 2 Soil carbon storage showing microbial community control. Community physiology and adaptation to incoming carbon quality determines partitioning between loss as CO_2 or dissolved organic carbon (DOC), and chemically stable or physically protected forms (1–3). Gross scale community composition determines contribution of cell walls to stable C as senescent biomass. Soil mineralogy can alter all fluxes, especially between the physically protected pool and microbial community (4–5). HS, humic substances.

and field studies of carbon storage have tended to concentrate on the physical factors controlling the size of carbon pools, rather than on microbiological controls over carbon transformation in soil. Intuition suggests microorganisms play a critical role in determining the loss or storage of soil carbon, yet gaps remain in our understanding. The microbial role in promoting carbon loss from soil is obvious: heterotrophic metabolism results in production of CO_2. Even so, we do not fully understand the controls on heterotrophic metabolism in the soil. Microbial physiology can tell us how different chemical constituents might be utilized for biomass and energy, but will not elucidate the long-term implications of metabolism and physiology for carbon storage and humification. Similarly, typical studies of humification show changes in chemical structure, but we can only infer process from such studies. Studies of organic-matter chemistry cannot elucidate the mechanistic relationships that determine carbon partitioning between microbial biomass, CO_2, and HS.

There is a difference in perspective about the role of microorganisms in humification that arises from chemical versus microbiological studies of HS genesis and characterization. Chemical studies are designed primarily to elucidate structural details of soil carbon 'pools' and infer the process of HS genesis. This provides critical information on the behavior of HS in soil, and their role in soil fertility, water-holding capacity, and structure. In contrast, a microbiological perspective addresses the question of why humification happens, as well as where, and how rapidly. A dynamic, microbiologically based understanding of humification will be necessary in the context of current and future research in soil carbon cycling.

Historical Perspective

It has long been recognized that HS play a critical role in soil functioning. From the late 1700s, soil HS have been studied in the context of agriculture and soil fertility. The 1700s up to the 1970s saw major advances in our understanding of the nature and properties of soil HS, as wet chemical techniques gave way to increasingly sophisticated analytical equipment. In the 1970s, advances in organic geochemistry analytical techniques, coupled with interest from the petroleum industry, fueled research in humification and humus genesis and characterization beyond the agronomic realm. The 1980s saw growing interest in the role HS play in environmental and ecological arenas. HS are involved in disinfection by product formation

in drinking water (e.g., organohalide formation), and in toxic-metal binding and bioavailability. The interaction between HS and soil, sediment, and water contaminants such as pesticides and polyaromatic hydrocarbons has received increasing attention. In addition, the role of humification as a mechanism of soil carbon sequestration has become an area of importance in light of current, global-scale changes in atmospheric carbon concentration.

With increasing emphasis on humification and carbon storage has come a renewed interest in understanding and quantifying the microbial role in carbon cycling. The first recognition that microorganisms play an important role in HS genesis came in the early 1900s, but the nature and importance of that role continue to be debated even today. Because of the long time scale necessary for humus genesis and its complex chemical nature, HS research has traditionally focused on isolation and characterization rather than process dynamics. The discipline has been dominated by organic chemists and geochemists rather than microbiologists, resulting in long-standing operational definitions of HS and an entire language used to describe the properties and nature of HS. Terms such as 'humic acid,' 'fulvic acid,' and 'humin' refer to classic soil-extraction procedures based on solubility in aqueous media, as a function of pH. While past chemical emphasis on the genesis and behavior of humic substances has contributed greatly to our understanding, there is a high degree of uncertainty regarding the microbial role played in humification. Yet, with the development of recent analytical techniques, there is ever-increasing evidence that microorganisms play an important role in the genesis and composition of HS in soil.

Chemical Perspective on HS Genesis

HS are by nature highly complex mixtures. Understanding how HS form, until recently, has required us in effect to blow up HS molecules, characterize the fragments, and then attempt to understand how they were put together in the first place (Figure 3). This is a bit like wrapping a watch in a handkerchief, smashing it with a sledge hammer, then handing the bits to someone and asking them not only to reconstruct the watch, but also to explain how the watch was formed in the first place! While this is the domain in which soil chemists and geochemists excel, this type of 'bottom-up' analysis is a challenge, to say the least. Ideally, an understanding of HS genesis could be gained from a 'top-down' approach, observing formation from biogenic materials. However, HS genesis takes place over a time scale of hundreds to thousands of years, rendering direct observation impractical. Thus, an understanding of HS genesis really does require an understanding of the chemical nature and properties of extant HS extracted from soil. Methods developed over the past several decades have led to ever-more sophisticated means of investigating HS chemistry, allowing more detail in structural inference. As methods have been developed and the knowledge base expanded, so too has the postulated role of microorganisms been reassessed (Table 1).

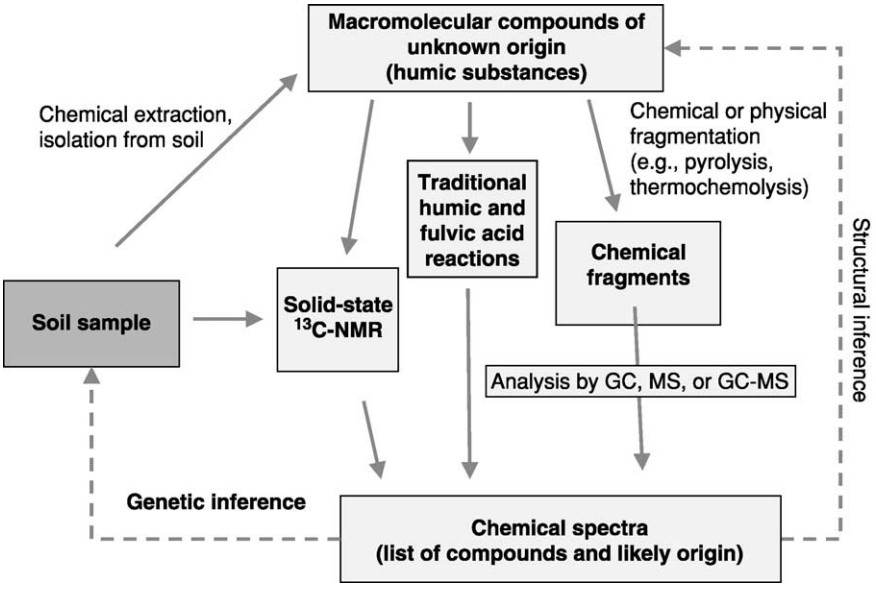


Figure 3 Flow diagram of deductive analytical techniques for investigating humic substances (HS) genesis and composition. GC, gas chromatography; MS, mass spectrometry; NMR, nuclear magnetic resonance.

Table 1 Comparison of traditional and newer methods for humic substances (HS) analysis

	Traditional degradative methods	*Nondegradative methods and combinations of newer techniques*
Method description	HS separated on the basis of chemical solubility (operational definitions rather than structural or compositional)	HS studied in unaltered form or in chemical or physical fractions
Structural inferences drawn	HS are macromolecular, highly aromatic	HS are macromolecular, highly aliphatic
Postulated microbial role	Metabolic action in early stages results in selective preservation of plant structural carbon, and abiotic condensation of smaller molecules	Large contribution of microbial cell and metabolic components into HS in addition to importance of metabolic role

Traditional (Degradative) Methods of HS Characterization

Common traditional methods for HS characterization have typically been 'degradative' approaches that involve the physical or chemical breakdown of macromolecular compounds into various subunits which are subsequently isolated and identified (Figure 3). Structural inferences are drawn based on the identities and yields of degradation products and the reaction processes responsible for their generation. This class of methods includes a range of extraction techniques (e.g., alkaline, organic, or neutral salt extractants) that are designed to solubilize and isolate chemical fractions from HS. For example, the humic acid fraction is not soluble in water under acid conditions, the fulvic acid fraction is soluble in aqueous solution under all pH values, and humin is neither soluble nor extractable in aqueous media. Traditional isolation methods begin by treating the sample with a strongly basic solution. Both humic and fulvic acids are coextracted in basic solution. The unextracted residue contains the humin. Upon acidification (with a strong acid such as HCl) of the alkaline extract, humic acids precipitate. Organic material remaining in solution is considered to be the fulvic acid fraction.

A drawback to this type of extraction is that it is extremely difficult to separate all nonhumic material from HS. Thus, not all the extracted material is HS. For example, long-chain aliphatic compounds may be closely associated with extracted HS. Also, the specific conditions chosen for extraction can produce chemical artifacts that may lead to false structural inferences. Strong alkaline extraction, in particular, can dissolve organic components from fresh plant material, incorporating them into humified organic matter. In addition, condensation reactions (e.g., between amino acids and reducing sugars or quinones) can occur in alkaline solution, resulting in humic materials created in the laboratory during extraction. These types of artifacts have led to the belief that humic substances are dominantly aromatic in structure (Table 1).

Nondegradative and Newer Degradative Techniques

Newer methods avoid the artifacts of traditional degradative techniques and indicate different structures for HS from those determined by traditional extraction techniques (Table 1). Nondegradative techniques are among the methods that may indicate the role of microorganisms in HS formation more reliably than past methodologies. Nondegradative approaches allow analysis of HS directly, usually without further chemical alteration. They include solid-state nuclear magnetic resonance (NMR), UV-visible and infrared spectrophotometry, electron spin resonance (ESR) spectrometry, X-ray analysis, electron microscopy, and electron diffraction.

A method that has received considerable attention is solid-state NMR spectroscopy. Liquid-state NMR has been used in chemical studies for a long time, but until relatively recently the complexity of HS and the presence of paramagnetic minerals (such as iron oxides) in soil has prevented the effective use of solid-state techniques. With the introduction of methodological refinements such as Fourier transform analysis and cross-polarization magic angle spinning (CP-MAS), researchers have been able to look at unaltered chemical mixtures in soil directly. A primary advantage of solid-state over liquid-state NMR is that most soil organic matter is insoluble and can only be analyzed in solid state. In addition, solid-state NMR is nondestructive – the sample is unaltered by solvent effects. Using NMR, it has been clearly shown that HS are not as aromatic in nature as previously thought. Studies of the composition of soil organic matter in relatively undisturbed soil profiles (forest floor samples, in particular) indicate a shift from plant-derived carbon near the top to carbon that has a microbial origin lower in the profile. Microbial components and metabolic products accumulate in HS as soil carbon becomes humified, indicating a significant microbial role in humification.

While nondegradative techniques have provided some of the first indications of the extent of microbial contribution to humification, there are still drawbacks to their exclusive use. For example, ^{13}C and ^{15}N-NMR require the mitigation of issues associated with solid samples, such as C-H and N-H dipolar interactions, chemical shift anisotropy, and interference by paramagnetic minerals (e.g., iron oxides) that can cause loss of sensitivity and line-broadening in the resulting spectra.

Advances in degradative techniques, and combinations of nondegradative and degradative techniques, have proven effective in allowing further understanding of HS chemistry and indicating the role of microorganisms in HS genesis. While new degradative techniques still involve the extraction of HS from soil, degradation and/or fragmentation, and analysis of the resultant product, changes in reaction conditions and method combinations reduce the incidence of chemical artifacts.

Thermal decomposition (pyrolysis) produces fragments of the main structural elements of HS. Pyrolysis with gas chromatography and mass spectrometry (GC-MS) has provided a significant new body of information about HS chemistry. A major advantage of pyrolysis is that it does not require solubility. During pyrolysis, the sample is heated rapidly in an inert environment. Thermal energy causes fragmentation of the HS macromolecules. The resulting volatile products are chromatographically separated (GC) and identified (MS). Alternatively, the volatile products can be ionized and identified using soft field-ionization mass spectrometry (FIMS). Pyrolysis GC-MS or FIMS yields data on the molecular composition of HS from the proportions of several classes of compounds in summed and averaged mass spectra. Furthermore, thermograms of volatilization give information on the strengths of chemical bonds within the organic macromolecules, or between organic materials and minerals. The thermal energy required to volatilize a compound indicates bond strength; and because bond strength is a proxy for resistance to microbial attack, pyrolysis-MS thermograms can indicate degree of humification. Drawbacks to pyrolytic techniques are that only a fraction of the sample is pyrolyzed, the high temperature can alter the structural 'building blocks' being studied, and secondary reactions between volatile products (such as rearrangement or polymerization) can occur. Structural inference must be drawn with caution. Despite these drawbacks, pyrolysis provides molecular-level detail about the original HS that is at least as representative of the original material as any other technique and requires very little sample. The molecular-level information provided complements the broader scale of information from NMR, and studies such as these have proven useful in indicating the role of microorganisms in humification.

The use of capillary GC (e.g., in pyrolysis GC-MS) is a powerful tool for resolution of compounds in complex mixtures, but the compounds must be of low molecular weight and low polarity. The majority of HS are too large and too polar to be analyzed effectively. Thermal and chemical degradation coupled with derivitization can be highly efficient in releasing smaller volatile compounds. Derivitization protects the fragments from reaction and reduces hydrogen bonding (reducing polarity of the molecule). Tetramethylammonium hydroxide thermochemolysis–gas chromatography/mass spectrometry (TMAH-thermochemolysis GC-MS) is a relatively new technique that uses derivitization. Because of the reaction conditions and derivitization step, this technique samples a unique set of components from HS, which are complementary to those found using other degradative techniques. TMAH chemolytically hydrolyzes and methylates esters and ether linkages, releasing fatty acid methyl esters of plant and microbial origin. Thus, TMAH coupled with pyrolysis gas chromatography–mass spectrometry (Py-GC/MS) and ^{13}C-NMR has been used to show the contribution of microbially derived fatty acids to HS.

Chemical Perspective on Microbial Role in Humification

Based on the information obtained from newer chemical methods, it is now widely believed that humic materials are less aromatic in nature than previously thought and are dominated by aliphatic components. The source of the aliphatic components is thought to be plant-derived polymers such as cutin, cutan, or suberin. However, it is now also thought that microbial biomass components such as lipids or lipopolysaccharides may play a part in HS genesis. Aliphatic components can be due to incorporation of microbial materials as well as microbial alteration of primary carbon structures. Thus newer methods indicate a greater role for microbes and microbial activity than has been previously assumed (Table 1).

A top-down approach to genetic inference suggests that microbes are actively involved in HS genesis at several levels. From a strictly chemical (in contrast to an ecologic or biological) viewpoint, there are several microbially mediated 'pathways' by which HS substances may form (Figure 4):

1. Degradation and abiotic condensation: Microbial degradation products and metabolites (amino sugars, semiquinones, or phenolic acids) undergo spontaneous condensation reactions to form complex,

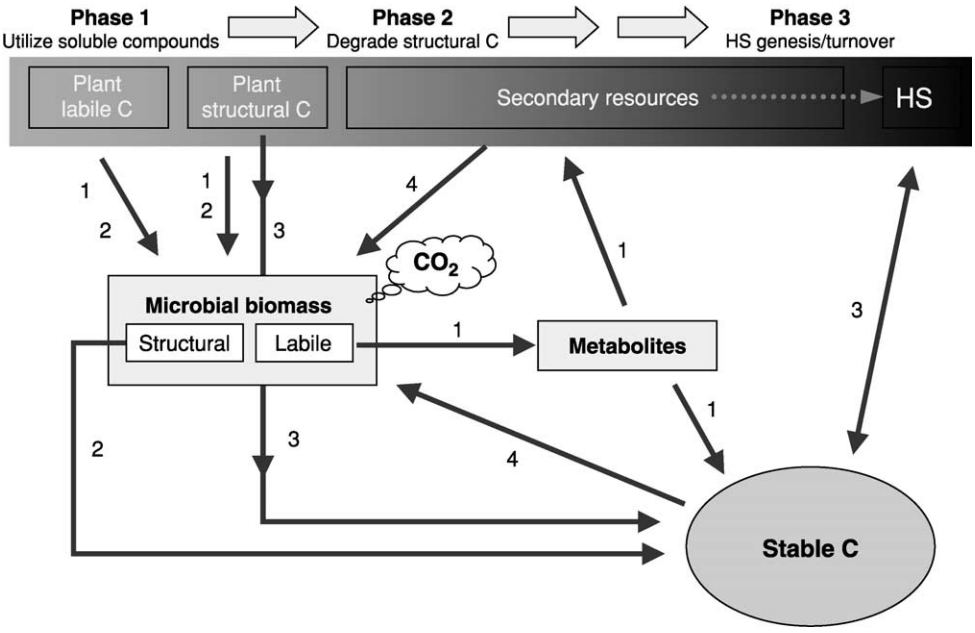


Figure 4 Pathways of microbial involvement in humification. From a strictly genetic standpoint, microorganisms participate in humification via four possible avenues: (1) degradation and abiotic condensation of metabolites and biomass components; (2) degradation of primary resources and resynthesis into refractory components; (3) selective preservation; (4) direct participation in HS production and degradation.

random, macromolecular humic materials. This pathway has long been recognized as one of the primary ways that HS form, particularly in aquatic environments;

2. Degradation of primary resources and resynthesis into recalcitrant components: Microbes consume plant litter, incorporating the carbon into their biomass. Upon death, their cell wall and membrane components are nascent humic substances. This mechanism of HS formation has been suggested fairly recently, with the rise in nondegradative techniques for studying HS. In particular, fungal melanins and bacterial lipid materials may be highly recalcitrant;

3. Selective preservation: Microbes degrade readily accessible plant carbon (e.g., small compounds and side chains), leaving only large, resistant structures such as the 'backbone' from lignin or tannin compounds, alkyl carbon from plant cuticles. The selective preservation of carbon from plant lignins and tannins has also long been acknowledged as a primary microbially mediated process in humification;

4. Direct participation in HS production and degradation: Microbes participate directly in humus genesis and turnover via enzymatic activities. This is a critical yet often overlooked role for microorganisms in humification. Through enzymatic activities, microorganisms can directly polymerize subunits into larger, humic molecules. In addition, microbial activity is a dominant means for HS turnover and loss from soil.

These pathways represent the current chemical view of the role microorganisms play in humification. However, the occurrence of a given microbial pathway and the rate of carbon movement along it will depend on microbial activity and community composition.

Microbiological Perspective

Chemical studies of HS have provided a large array of structural information, and, while structure can indicate the properties and reactivity of HS in soil, it cannot indicate the rate and reasons for HS formation. Ultimately, humification is a consequence of microbial growth and activity. The organisms performing humification are surrounded by carbon at all stages along a decay continuum (Figure 5). Their requirements for energy and survival lead to degradation and processing of soil carbon. The result is humification.

Microorganisms in the soil are influenced by, and interact with, their entire chemical and physical environment rather than with specific chemical compounds in an isolated HS fraction. Thus an understanding of humus genesis and rate of formation requires an understanding of the reasons why humification occurs. This falls within the realm of microbial ecology. From a microbiological standpoint, humification occurs in three phases: (1) rapid initial decomposition of primary plant residues; (2) slow decomposition of plant structural

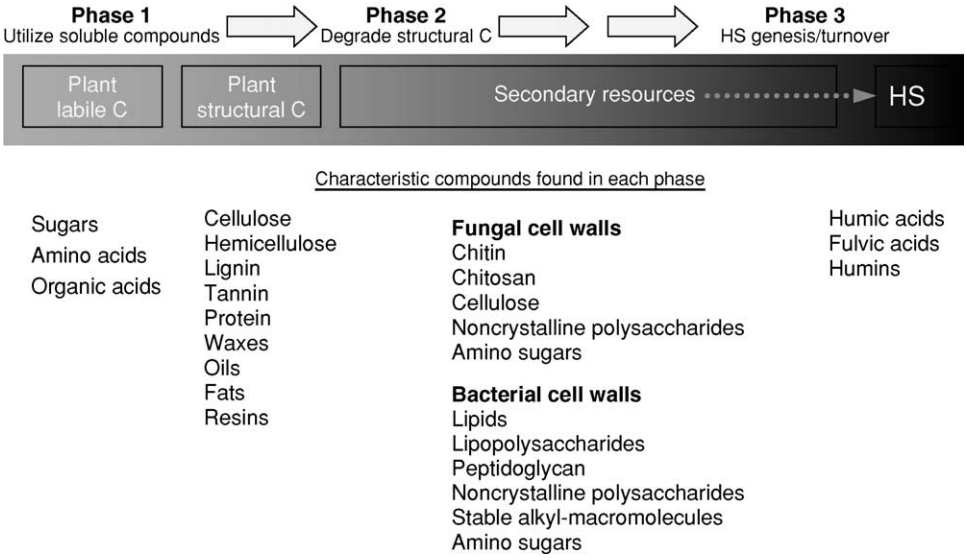


Figure 5 Microorganisms exist surrounded by a continuum of organic matter at all stages of decay and humification. These include primary resources (plant-derived), secondary resources (microbial biomass components) and partially or fully humified organic materials. From a microbial perspective, humification can be thought to occur in three phases that move carbon along the decay continuum: phase 1 is utilization of soluble plant compounds (e.g., amino acids and sugars); phase 2 is degradation of plant structural carbon (e.g., polymeric compounds such as cellulose or hemicellulose); phase 3 is the direct genesis or turnover of humic materials.

components; and (3) alteration of soil organic carbon and HS genesis.

Phase 1: Initial Decomposition of Primary Resources

The first phase in humification is decomposition of labile, plant-derived carbon (litter, or 'primary resources'). Decomposition of terrestrial plant litter is the mechanism by which carbon and nutrients are returned to either the atmosphere or to a plant-available state. Microorganisms utilize soluble, low-molecular-weight compounds ('labile' substrates such as sugars, amino acids, and organic acids) for energy generation and biomass production. Due to their small size, these substrates are readily taken into the cell and metabolized. Because labile substrates (such as glucose) are often low in available nutrients, degrading organisms immobilize nutrients such as nitrogen, phosphorus, or calcium in their biomass during this first phase of decomposition. The microbial community able to utilize these substrates is highly diverse; nearly all organisms will use labile substrates in preference to larger, more complex forms of carbon.

Phase 2: Slow Decomposition, Energy Limited

Degradation rate slows once labile sources of carbon have been utilized, and primarily lignin-encrusted cellulose and hemicelluloses remain. These plant structural compounds are significantly larger in size than water-soluble or labile components from phase 1 and cannot be taken into the cell for degradation. Instead, they must be depolymerized extracellularly prior to uptake and metabolism (Figure 6). Plant structural components surround proteins and other nitrogen-containing resources, making them physically inaccessible. To access nitrogen, organisms must attack the structural components. This requires systems of multiple enzymes, and often multiple organisms (a microbial 'consortium'). In comparison with utilization of labile carbon, there are fewer organisms capable of carrying out depolymerization reactions. Many fungi and eubacteria, such as actinomycetes, are important during this phase of humification. They are capable of producing a variety of cellulolytic and lignolytic enzymes.

Phase 3: Direct Genesis and Degradation

The final phase of humification is direct genesis or degradation of HS. In contrast to phases 1 and 2, microbial contribution to phase 3 is not only through metabolic activity, but also through incorporation of metabolites and biomass components into HS.

The macromolecular random structure of HS makes them highly resistant to microbial enzymatic attack. Microbial enzymes are generally designed for regular repeating carbon units and/or binding sites. Thus, depolymerization and oxidation of HS in soil generally occur incidentally (usually by 'cometabolism'). Cometabolic degradation is oxidation of HS via enzymes produced for another purpose, such as decomposition of plant structural carbon. It yields no

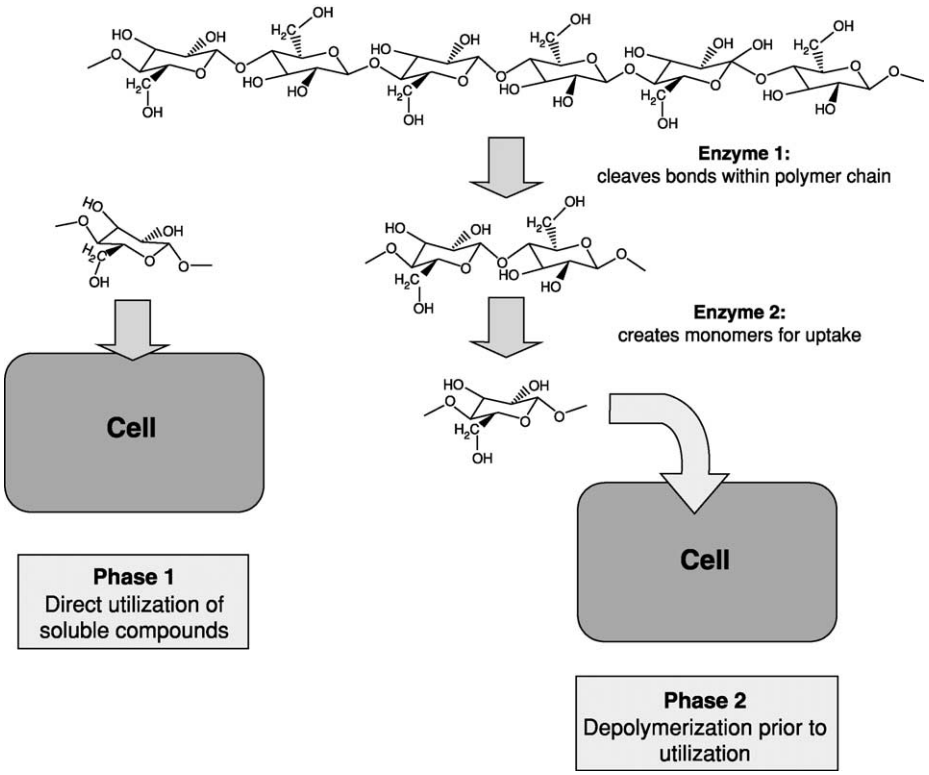


Figure 6 In phase 1 of humification, simple monomeric compounds are taken directly into the cell and metabolized. There are a large array of organisms capable of utilizing labile carbon monomers. In phase 2, plant structural components must be depolymerized prior to utilization. This requires a suite of enzymes and often more than one organism working together in a 'consortium.' Fewer microorganisms are capable of utilizing polymeric compounds.

energetic benefit for the microorganism; rather, an alternate energy source is required (e.g., simple sugars leaked from plant roots). A second source of 'accidental' HS degradation is via enzymes stabilized on mineral surfaces. These can remain active for years, oxidizing and altering HS. Soil mineralogical composition can greatly affect the stability and longevity of this cell-free enzyme pool.

In some circumstances, microorganisms can directly generate and polymerize aromatic compounds to form HS. For example, under low-oxygen conditions, it has been found that microorganisms can make polyaromatic hydrocarbons such as phenanthrene, fluoranthrene, pyrene, and chrysene. It is not known whether this is incidental or serves a biological purpose.

Finally, some of the most important recent studies of humification focus on the incorporation of microbial components into humic materials. First, microbial metabolites and cell components can condense abiotically to form HS. For example, phenolic compounds that are released into the soil from litter or microbial activity are highly reactive and spontaneously undergo nonenzymatic chemical reactions to form more complex molecular structures. In addition, techniques such as TMAH, NMR, and pyrolysis on soil chemical fractions have shown that microbial metabolites and cellular material regularly contribute to HS. For example, the contribution of microbial fatty acids increases with degree of soil organic matter humification and decreasing particle size, suggesting selective preservation of these molecules. In addition, fungal and actinomycete pigments or melanins can be very similar in structure to older HS. Recent ^{15}N-NMR studies also indicate that the majority of organic nitrogen is present in the form of amide and amino N, potentially from preserved microbial proteins. It is well known that microbially derived enzymes can become irreversibly bound to HS in soil. Part of the refractory amide nitrogen may also derive from the bacterial cell wall materials such as peptidoglycan and fungal chitin (Table 2, Figure 7). Lastly, studies of amino sugars have also indicated a critical role for microbial components in HS structures. Muramic acid uniquely originates from soil bacteria, while the compound glucosamine is found in both bacterial and fungal cell walls. Using these marker compounds, it has been possible to show that microbial products accumulate in HS with age and depth in the soil.

Table 2 Some important microbial cell wall components

Cell wall component [a]
N-Acetylglucosamine (amino sugar monomer: glucosamine)
N-Acetylmuramic acid (amino sugar monomer: muramic acid)
Peptidoglycan structure
Chitin
Chitosan

[a]Structures shown in **Figure 7.**

Regulation of Humic Substances Genesis and Turnover

Humification, like all microbially mediated processes, is controlled by factors that affect microbial activity and access to carbon and energy sources. These factors include environmental constraints (temperature, water, pH), nutrient availability, adsorption to surfaces of soil minerals or encryption within a clay mineral coating, and occlusion within micropores in soil structure (Figure 8). In addition, factors such as the resource 'quality' available for microbial growth (e.g., plant litter) act as controlling variables.

Climate and Soil Type

Typical studies of carbon dynamics in soil focus on physical and environmental conditions. Environmental conditions that prevent microbial utilization of C lead to high levels of stored organic matter – as seen in the tundra (where temperature limits utilization) and in peat bogs (where oxygen is limiting). In these situations, microbial enzyme activity limits carbon degradation and humification. For example, monooxygenase enzymes are required for degradation of aromatic structures such as phenolic compounds. These enzymes are inactive in anoxic environments. The carbon in these environments is not humified, but is nevertheless unavailable for microbial utilization. When the environment changes, for example, if the peatlands dry or the climate warms, then the stored carbon may be released rapidly to the atmosphere.

Soil type (e.g., mineralogy and texture) determines physical protection of soil carbon via interaction with soil minerals or within aggregates. Both past and recent work has demonstrated that soils with higher silt and/or clay contents or better aggregate structure tend to have higher levels of carbon. Soils with poorly crystalline minerals such as allophane also maintain high carbon levels (Figure 2). Adsorption to mineral surfaces represents an important process in preventing microbial degradation of stable carbon compounds.

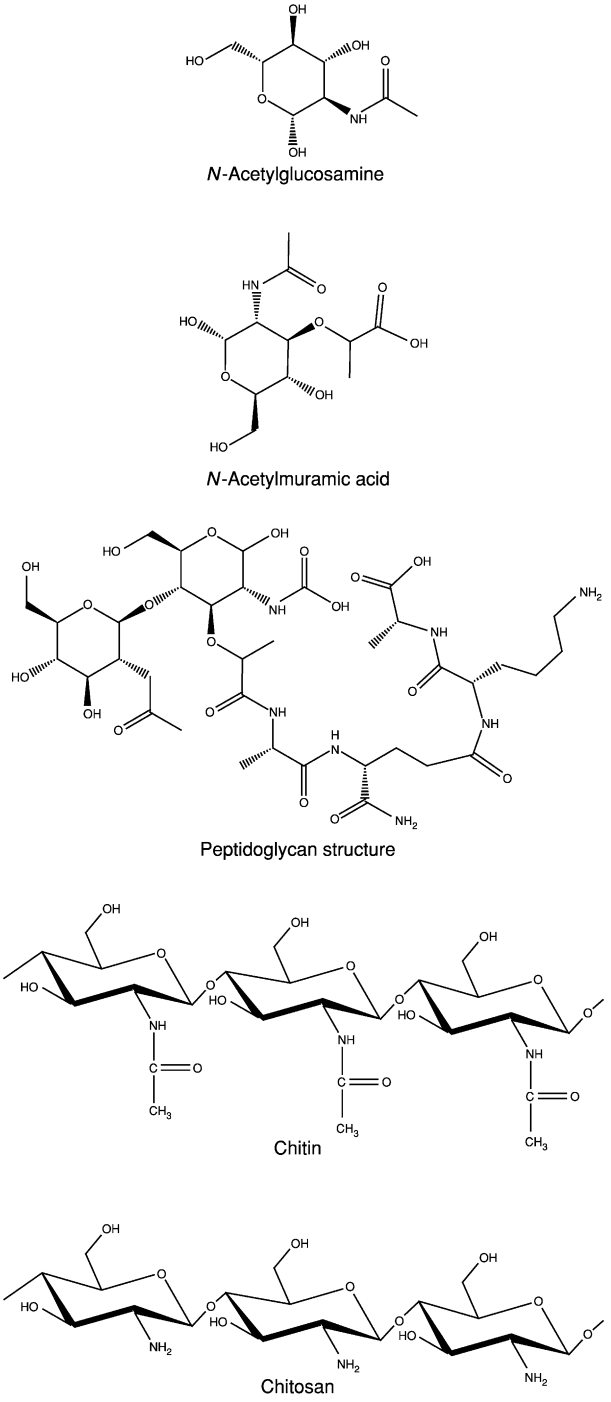


Figure 7 Chemical structures of some important microbial cell wall components. (See **Table 2**.)

Studies employing bulk dissolved organic carbon (DOC) compounds from soils indicate that the extent of sorptive preservation is related to mineral surface area, but, more importantly, to mineral surface properties. More strongly adsorbed molecules, including intrinsically labile compounds, appear to accumulate in soils. In effect, organic compounds bound to clay

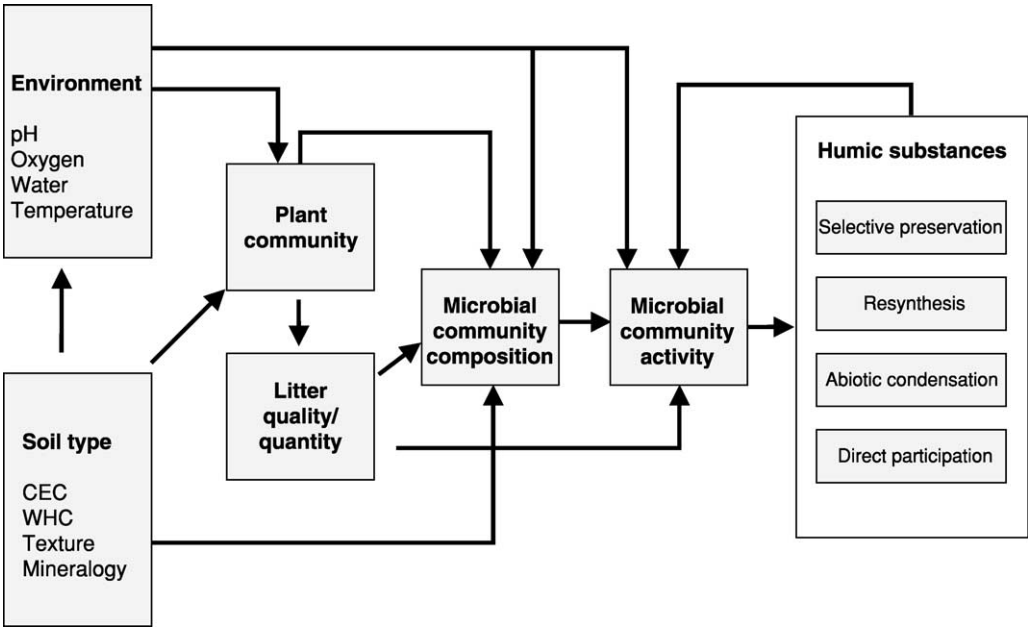


Figure 8 Points of regulation for humification and carbon storage. Physical environment, soil type, and litter quality and quantity all affect microbial community activity and composition, which in turn influence the rate and occurrence of humification. CEC, cation exchange capacity; WHC, water-holding capacity.

minerals are not readily accessible to microbial attack. This slows the rate of humification considerably.

Amount and Chemical Composition of Starting Resources

A second level of control over carbon storage is the importance of litter quality and quantity. Fresh plant litter is degraded at a rate that is proportional to its initial C:N ratio, or perhaps more accurately, proportional to its initial lignin:N ratio. The ratio of C:N or lignin:N is a proxy for substrate quality. Substrate quality controls carbon dynamics via potential energetic gain. Potential energetic gain is determined by oxidation–reduction status, and molecular length and heterogeneity. Highly reduced compounds yield more energy upon oxidation by microorganisms. These types of substrates are utilized preferentially. Molecular length and heterogeneity will also determine whether or not a substrate is preferred. Monomeric compounds (such as glucose) can be taken directly into the cell and require only 'standard' enzymatic machinery for degradation. Polymeric compounds (e.g., cellulose) have the same energetic yield per unit carbon, but require the concerted action of several enzymes. The 'cost' to produce the additional enzymes makes degradation more difficult and therefore less likely. Polymers having irregular or random structures (such as lignin or HS) are also more difficult to degrade than those possessing regular structure. The energetic yield from a random macromolecular compound is low; thus microorganisms will preferentially utilize other carbon sources.

Role of Microbial Community Composition

This topic deserves special mention, because past humification studies have tended to ignore the composition and diversity of the organisms involved. However, the organisms present in the soil can determine the occurrence and rate of humification. A wide range and diversity of organisms participate in phase 1 (simple decomposition). However, phases 2 and 3 require organisms capable of producing somewhat specialized extracellular enzyme complexes. Thus microbial community composition can potentially be a determining factor in the quality and quantity of HS production and turnover in soil (Figure 9).

Microbes, and microbial community composition, impact humification in two ways: (1) via cellular components, and (2) through metabolic activity. Microbes have the ability to take labile carbon and turn it into resistant and/or reactive compounds, but not all microorganisms are equal in this ability. For example, microbial cell wall composition (and recalcitrance) varies widely between Gram-negative bacteria, Gram-positive bacteria, and fungi. If cell walls become nascent humic substances, then the relative abundance of a given wall type should affect the amount of humus produced. Or, since decomposition

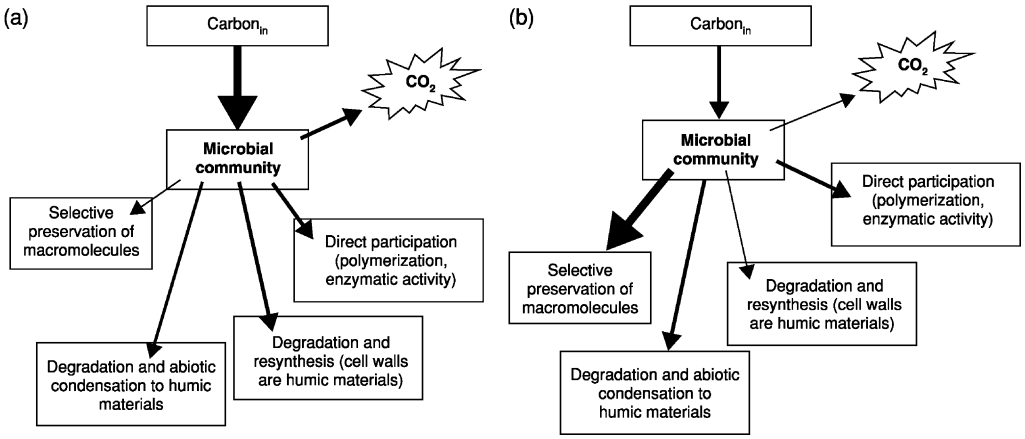


Figure 9 Hypothetical carbon cycle differences for (a) fungally versus (b) bacterially dominated microbial communities. The flow and fate of carbon inputs will vary: fungi degrade incoming plant litter more completely than bacteria, releasing more carbon as CO_2. Because fungal cell walls are highly recalcitrant, humification by degradation and resynthesis may also be high. In contrast, bacteria may contribute more to humification via selective preservation of macromolecular plant litter compounds.

of complex molecules requires a suite of enzymes that not all microbes possess in entirety, the relative importance of selective preservation versus degradation will depend on the genetic makeup of the microbial community. Thus the humification pathway of a given carbon molecule in soil is determined: (1) by the genotype/phenotype of the organism that the C encounters (does it have the ability to use the C?); (2) by the organism's metabolic status (is it starving, or growing, or dormant?); and (3) by the organism's proximity to others (if the initial organism cannot use it, is there one nearby that can?). An excellent example is seen in the forests of the Pacific Northwest USA; the flow and fate of carbon from fallen logs in these wet temperate forests can be entirely determined by microbial community composition. Colonization of logs by wood decay fungi of differing functional types dramatically impacts the resulting carbon cycle. White rot fungi are able to degrade lignin completely to CO_2 and water, and typically leave cellulose untouched. Brown rots are specialists on cellulose, and leave lignin mostly unaltered. White and brown rots do not typically coexist on a single log; the initial colonizer will actively prevent colonization by other fungal species. A log colonized by white rot is reduced to nearly pure cellulose, and one colonized by brown rot is left as a nearly pure source of lignin. Thus at a gross level, the flow and fate of carbon in the system are determined by which fungus colonizes a given log first.

Another example of when microbial community composition might impact humification and carbon storage is in locations within an ecosystem or soil that vary in microbial community composition. For example, forest soils differ from managed and grassland soils; they have highly complex plant litter, and the microbial community is often dominated by filamentous organisms (fungi). Because fungi are generally considered to be more metabolically 'talented' than bacteria, and also have cell walls that are highly resistant to degradation, humification in a forest soil is likely to be a balance between fungal degradation of and contribution to humic materials. The flow of carbon from litter to microbial biomass and into HS will differ from that in a nonfungal-dominated soil (Figure 9). In contrast, the microbial community in a heavily managed soil is likely to be dominated by bacteria, coupled with differing carbon inputs. The process and rate of humification are likely to differ from that in a forest soil.

Finally, even within a given soil, the amount of carbon moving from pool to pool (Figure 2) and the dominant humification pathway (Figure 8) can vary widely. The rhizosphere is generally dominated by Gram-negative bacteria and mycorrhizal fungi, which differ considerably in their ecology (e.g., different metabolic requirements and preferences) from Gram-positive bacteria or saprotrophic fungi. Thus the flow and fate of carbon in the rhizosphere will differ from that in bulk soil (which is dominated by Gram-positive bacteria and saprotrophic fungi).

Summary

Because the generation and turnover of humic substances in the soil is largely a by-product of microbial growth and activity, an understanding of humus genesis must be predicated upon an understanding of microbial community ecology. Factors that control microbial activities and community composition will also control the rate and quality of humification. Ultimately, this understanding of why, where, and

how rapidly humification occurs will be critical in our understanding of global carbon sequestration, as well as our general understanding of the nature and properties of the world's soils.

List of Technical Nomenclature

Carbon stabilization The process of converting carbon readily utilized as a microbial energy source to recalcitrant compounds having a long turnover time. This can be done via chemical alteration (humification) or via environmental or physical restraints on degradation

Degradative approach to study Macromolecular HS are chemically or physically broken down into various subunits which are subsequently isolated and identified. Structural inferences are drawn based on the identities and yields of degradation products and the reaction processes responsible for their generation

Humic substances (HS) Conventionally defined as yellow to black-colored, high-molecular-weight, heterogeneous, refractory, naturally occurring organic compounds. They tend to be a range of organic materials that are lumped together almost by default; they cannot be easily classified into discrete categories such as proteins, polysaccharides, or nucleotides. They are often further operationally defined in terms of the methods used to isolate them from soils. Humic acids, fulvic acids, and humin are the three primary fractions, defined on the basis of their solubility in aqueous media, important in soil carbon sequestration due to long turnover times

Humification Alteration of biotically derived carbon to macromolecular, structurally amorphous forms (e.g., humic and fulvic acids)

Microorganisms or microbes Organisms not identifiable with the naked eye, including bacteria, fungi, algae, and archaea

Nondegradative approach Isolated humic substances are analyzed directly, usually without further chemical alteration, utilizing techniques that allow structural inferences to be made

Primary resources Identifiable plant residues

Rhizosphere Zone of soil influenced by plant roots. The microbial community in the rhizosphere is generally dominated by Gram-negative bacteria and mycorrhizal fungi

Secondary resources Residues of soil animals and microorganisms, as well as partially humified compunds

Soil organic matter Primary and secondary resources, plus dissolved organic carbon, root exudates, and HS. Totality of biologically derived carbon in soil

See also: **Bacteria:** Soil; **Carbon Cycle in Soils:** Dynamics and Management; **Carbon Emissions and Sequestration**; **Fungi**; **Microbial Processes:** Environmental Factors; **Organic Matter:** Principles and Processes; Genesis and Formation; **Organic Residues, Decomposition**; **Pollutants:** Biodegradation

Further Reading

Almendros G and Dorado J (1999) Molecular characteristics related to the biodegradability of humic acid preparations. *European Journal of Soil Science* 50: 227–236.

Chefetz B, Chen Y, Clapp E, and Hatcher PG (2000) Characterization of organic matter in soils by thermochemolysis using tetramethylammonium hydroxide (TMAH). *Soil Science Society of America Journal* 64: 583–589.

Chefetz B, Tarchhitzky J, Desmukh AP, Hatcher PG, and Chen Y (2002) Structural characterization of soil organic matter and humic acids in particle-size fractions of an agricultural soil. *Soil Science Society of America Journal* 66: 129–141.

Gramss G, Ziegenhagen D, and Sorge S (1999) Degradation of soil humic extract by wood- and soil-associated fungi, bacteria and commercial enzymes. *Microbial Ecology* 37: 140–151.

Hatcher P (2001) Modern analytical studies of humic substances. *Soil Science* 166(11): 770–794.

Kaiser K and Guggenberger G (2000) The role of DOM sorption to mineral surfaces in the preservation of organic matter in soils. *Organic Geochemistry* 31: 711–725.

Kogel-Knabner I (2002) The macromolecular organic composition of plant and microbial residues as inputs to soil organic matter. *Soil Biology and Biochemistry* 34: 139–162.

Linhares AA, Linhares LF, and Coelho RRR (1998) Neutral sugars in melanins synthesized by actinomycetes from Brazilian soils. *Biology and Fertility of Soils* 27: 162–167.

Martin JP and Haider K (1971) Microbial activity in relation to soil humus formation. *Soil Science* 111(1): 54–63.

McCarthy P (2001) The principles of humic substances. *Soil Science* 166(11): 738–751.

Oades JM (1989) An introduction to organic matter in mineral soils. In: Dixon JB and Weed SB (eds) *Minerals in Soil Environments*, pp. 89–159. Madison, WI: Soil Science Society of America.

Paul E and Clark F (1996) *Soil Microbiology and Biochemistry.* New York: Academic Press.

Schulten HR and Leinweber P (2000) New insights into organo-mineral particles: composition, properties, and models of molecular structure. *Biology and Fertility of Soils* 30: 399–432.

Stevenson FJ (1994) *Humus Chemistry: Genesis, Composition, Reactions*. New York: John Wiley.

Swift MJ, Heal OW, and Anderson JM (1979) *Decomposition in Terrestrial Ecosystems*. Berkeley/Los Angeles, CA: University of California Press.

Wershaw RL (1993) Model for Humus in Soils and Sediments. *Environmental Science and Technology* 27(5): 814–816.

Zech W, Senesi N, Guggenberger G, Kaiser K, Lehmann J, Miano TM, Miltner A, and Schroth G (1997) Factors controlling humification and mineralization of soil organic matter in the humid tropics. *Geoderma* 79: 117–161.

HYDRAULIC PROPERTIES, TEMPERATURE EFFECTS

S A Grant, US Army Engineer Research and Development Center, Hanover, NH, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

The three soil hydraulic properties upon which temperature has its most pronounced effects are hydraulic conductivity, matric potential, and water-vapor contents. Well-developed theories exist for the effect of temperature on hydraulic conductivity and water-vapor contents. Despite intensive research in the last half of the twentieth century, the mechanism responsible for the effect of temperature on matric potential has yet to be elucidated. Hydraulic conductivity is inversely related to the viscosity of the fluid. The viscosities of fluids, including water, decrease proportionally to the exponent of the reciprocal of temperature – so that hydraulic conductivity increases with increasing temperature. The absolute value of the matric potential decreases linearly with temperature. The mean decrease for the soils studied thus far is $-0.8\%\,K^{-1}$. In dry soils, water transport in the gas phase can be comparable with that in the liquid phase. The partial pressure of water vapor and its diffusivity increases exponentially with temperature.

Soil-Water Characteristics

Water in unsaturated soils is maintained under negative matric potential. The Young–Laplace equation is generally believed to capture the physics of the phenomenon:

$$\psi = -2\gamma^{lg}\cos\Theta V_{m,H_2O(l)}/(rgM_{H_2O}) \qquad [1]$$

where ψ is matric potential (meters of water); γ^{lg}, surface tension of water (newtons per meter); Θ, the contact angle where the wetting–nonwetting interface meets the soil matrix (radians); $V_{m,H_2O(l)}$, molar volume of water (cubic meters per mole); g, acceleration due to gravity (meters per second squared); r, the effective pore radius (meters); and M_{H_2O}, molecular weight of water (kilograms per mole). When applied to porous media, the Young–Laplace equation states that at a given matric potential all the pores with effective radii greater than r are empty of the wetting liquid; that is, matric potential increases as the degree of saturation by the wetting liquid decreases. In the hydrology literature, the term $\cos\Theta$ is used to express the wettability of the porous matrix. The value of $\cos\Theta$ can vary from 1, signifying completely water-wet, to $+1$, which signifies completely nonwetting-phase wet.

Empirical relations such as the van Genuchten equation relate soil water content to matric potential:

$$S_e = \left\{\frac{1}{[\alpha\psi]^n + 1}\right\}^{\frac{n-1}{n}} \qquad [2]$$

where S_e is the water saturation defined by:

$$S_e = \frac{\theta - \theta_r}{\theta_s - \theta_r} \qquad [3]$$

where θ is volumetric water content (cubic meters per cubic meter); θ_s, saturated water content (cubic meters per cubic meter); θ_r, residual water content (cubic meters per cubic meter); and α (per meter) and n, fitted parameters. Figure 1 presents an exemplary isothermal soil-water characteristic, which is the relationship between matric potential and water content.

Relative Permeability

Once parameterized, the van Genuchten equation can be used to calculate relative permeabilities of the porous medium as a function of degree of saturation:

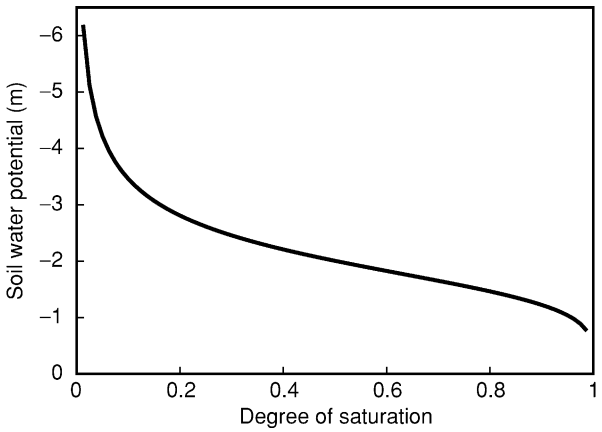


Figure 1 Soil-water characteristic for a loam soil.

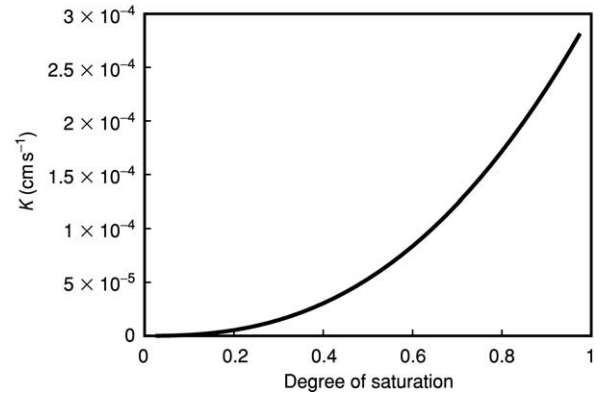


Figure 2 Predicted hydraulic conductivity as a function of degree of saturation predicted by the van Genuchten equation for a loam soil.

$$k_r = K/K_s = S_e^{1/2}\left[\frac{\int_0^{S_e}\frac{dx}{\psi(x)}}{\int_0^{1}\frac{dx}{\psi(x)}}\right]^2 \qquad [4]$$

where k_r is the relative hydraulic conductivity (dimension 1); K, unsaturated hydraulic conductivity (meters per second); and K_s, saturated hydraulic conductivity (meters per second). Figure 2 presents the hydraulic conductivity estimated with the van Genuchten equation for the loam soil presented in Figure 1 with an assumed saturated hydraulic conductivity at 25°C of 0.0003 cm s^{-1}.

Effect of Temperature on Soil-Water Characteristics

At room temperature, the matric potential in an 'average' soil decreases 0.8% for every 1 K increase in temperature. The surface tension of water also decreases linearly with temperature. By inspection of eqn [1], soil physicists early in the twentieth century proposed that:

$$\frac{1}{\psi}\frac{d\psi}{dT} \stackrel{?}{=} \frac{1}{\gamma^{ls}}\frac{d\gamma^{ls}}{dT} \qquad [5]$$

Subsequent experiments have consistently demonstrated that the conjecture presented in eqn [5] is incorrect. The relation underestimates the effect of temperature on matric potential by a factor of approximately 4. Why this is so is not known. Three general mechanisms have been proposed: expansion of trapped gas, solute effects on soil-water surface tension, and temperature-sensitive contact angles. None of these explanations can be reconciled with the available data.

The effect of temperature on matric potential appears to be well described by the relation:

$$\psi(T) = \psi(T_r)\frac{\beta_0 + T}{\beta_0 + T_r} \qquad [6]$$

where T_r is a reference temperature and β_0 is a constant (K), which for most soils has a value between −350 and −450 K. The available evidence indicates that β_0 is unaffected by soil water content. The matric potential at a reference temperature in eqn [6] can be inserted into the van Genuchten equation to yield:

$$S_e = \left\{\frac{1}{\left[\alpha\psi\left(\frac{\beta_0 + T_r}{\beta_0 + T}\right)\right]^n + 1}\right\}^{\frac{n-1}{n}} \qquad [7]$$

Figure 3 presents an example of the matric potential–saturation–temperature relationship predicted by eqn [7] for a loam soil. The effect of increasing temperature is to decrease the matric potential gradients within the soil due to differences in water content.

Effect of Temperature on Hydraulic Conductivity

Hydraulic conductivity has been found to be directly proportional to liquid density, the reciprocal of liquid viscosity, and the square of the mean grain diameter. This result leads to the assumption that hydraulic conductivity is separable into distinct contributions owing to the structure of the porous matrix, and the density and viscosity of the fluid. Symbolically, this yields:

$$K = \frac{k\rho g}{\eta} \qquad [8]$$

where ρ and η are the density and viscosity of the liquid (kilograms per cubic meter and pascal-seconds, respectively), g is the gravitational constant (meters per second squared), and k is the intrinsic permeability of the porous matrix (meters squared). The density

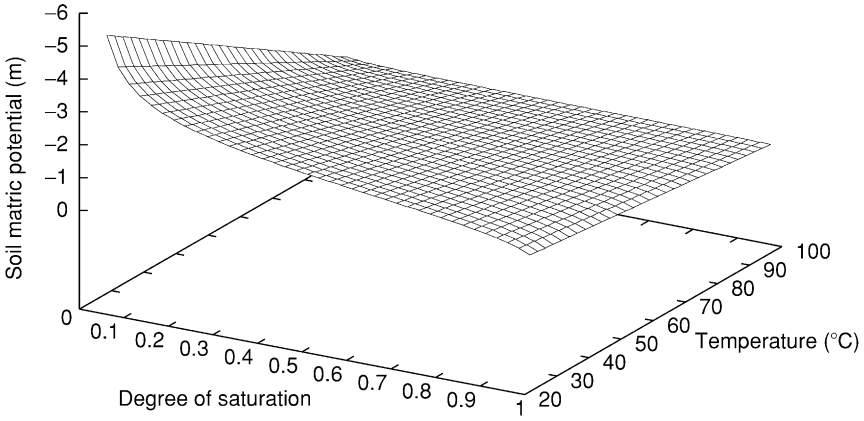


Figure 3 Matric potential as a function of both degree of saturation and temperature for a loam soil.

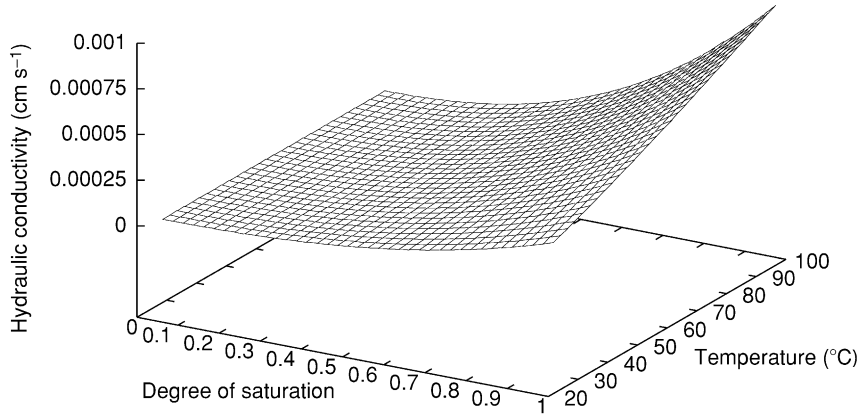


Figure 4 Hydraulic conductivity as a function of both degree of saturation and temperature for a loam soil assuming that the saturated hydraulic conductivity at 25°C is 0.003 cm s^{-1}.

of water changes little between its conventional melting and boiling temperatures, and the approximate value of 1000 kg m^{-3} can be used. The viscosity of water can be fitted to the Vogel–Tamman–Fulcher (VTF) equation:

$$\eta_l = \eta_0 \exp\left(\frac{B}{T + T_0}\right) \qquad [9]$$

yielding the following parameter estimates: $\eta_0 = 31.8903\ (\pm 1.3059)\ \mu$Pa s, $B = 479.9\ (\pm 11.9737)$ K, $T_0 = -154.05\ (\pm 1.7891)$ K. This equation has a maximum error of −1.1% at 100°C. Figure 4 presents the estimated hydraulic conductivity of the loam soil again assuming that the saturated hydraulic conductivity at 25°C is equal to 0.0003 cm s^{-1}.

Water and Energy Transport in a Nonisothermal Soil

There is a general model of liquid and vapor water and energy transport in soil (*See* **Thermal Properties and Processes**):

$$\frac{\partial \theta}{\partial t} = \nabla \bullet (D_T \nabla T) + \nabla \bullet (D_W \nabla \theta) - \frac{\partial K}{\partial z} \qquad [10]$$

$$C_v \frac{\partial T}{\partial t} = \nabla \bullet (\kappa \nabla T) - L \nabla \bullet (D_W \nabla \theta) \qquad [11]$$

where t is time (seconds); D_T, thermal water diffusivity (meters squared per second per Kelvin), D_W, water content-based water diffusivity (meters squared per second); z, depth (meters); C_v, volumetric heat capacity (Joules per Kelvin per cubic meters); κ, apparent thermal conductivity of the soil (watts per meter per Kelvin); and L, latent enthalpy of vaporization (Joules per cubic meter). The total volumetric water content is the sum of the liquid (θ_l, in cubic meters per cubic meter) and gas water contents (θ_v, cubic meters of liquid-water equivalent per cubic meter of soil).

The thermal water diffusivity can be disaggregated into its liquid- and gas-phase contributions to:

$$D_T = D_{lT} + D_{vT} \qquad [12]$$

where

$$D_{\mathrm{lT}} = -\frac{K_s k_r \rho}{\eta}\frac{\partial \psi}{\partial T} \quad [13]$$

and

$$D_{\mathrm{vT}} = -(\phi - \theta_l) D \frac{\partial \theta_v}{\partial T} \quad [14]$$

where D is the gas diffusion coefficient (meters squared per second).

By eqn [5], the temperature derivative of the matric potential is:

$$\frac{\partial \psi}{\partial T} = \frac{\psi}{\beta_0 + T} \quad [15]$$

The vapor water content of the soil is described by:

$$\theta_v = (\phi - \theta_l)\frac{p^{l+g} M_{H_2O}}{RT\rho}\exp[\psi g M_{H_2O}/(RT)] \quad [16]$$

where ϕ is porosity (cubic meters per meters cubed), p^{l+g} is the saturated vapor pressure of water (pascals); R, gas constant (joules per Kelvin per mole). The saturated vapor pressure at a particular temperature can be estimated accurately with the relation:

$$\ln\left(p^{l+g}/p_c\right) = (T_c/T)\left(a_1\tau + a_2\tau^{1.5} + a_3\tau^3 + a_4\tau^{3.5} + a_5\tau^4 + a_6\tau^{7.5}\right) \quad [17]$$

where T_c and p_c are the critical temperature and pressure for water, having values of 647.096 K and 22 064 000 Pa, respectively. The reduced temperature τ is equal to $1 - T/T_c$. The coefficients a_1, a_2, a_3, a_4, a_5, and a_6 have values of −7.8595 17, 1.8440 8259, −11.7866 497, 22.6807 411, −15.9618 719, and 1.8012 2502, respectively. The temperature derivative of vapor water content can be obtained by taking the derivative of eqn [15]:

$$\frac{\partial \theta_v}{\partial T} = \Big(\exp[\psi g M_{H_2O}/(RT)] M_{H_2O}(\phi - \theta_l)\Big\{RT^2\left(\partial p^{l+g}/\partial T\right) - p^{l+g}[gM_{H_2O}\psi + T(R - gM_{H_2O}(\partial\psi/\partial T))]\Big\}\Big)/(R^2T^3\rho) \quad [18]$$

As in eqn [4], the water-content diffusivity can be disaggregated into its gas- and liquid-phase elements:

$$D_W = D_{lW} + D_{vW} \quad [19]$$

where

$$D_{lW} = -\frac{K_s k_r \rho}{\eta}\frac{\partial \psi}{\partial \theta} \quad [20]$$

$$D_{vW} = -(\phi - \theta_l) D \frac{\partial \theta_v}{\partial \psi}\frac{\partial \psi}{\partial \theta} \quad [21]$$

To evaluate eqn [18], the constitutive equations describing ψ and k as functions of liquid water content and temperature are needed. In all but the driest soils, a negligible error is introduced into eqns [1], [5], and [7] by assuming that $\nabla\theta = \nabla\theta_l$, and $\partial\psi/\partial\theta = \partial\psi/\partial\theta_l$. Equations [19], [20], and [21] can be evaluated exploiting the van Genuchten equation and noting that the derivative of water-vapor content with respect to matric potential is:

$$\frac{\partial \theta_v}{\partial \psi} = (\phi - \theta_l)\frac{p^{l+g} M_{H_2O}^2}{R^2T^2\rho}\exp[\psi g M_{H_2O}/(RT)] \quad [22]$$

Summary

The degree to which the temperature effect on a given soil hydraulic property is understood depends on the property in question. The theory of the effects of temperature on hydraulic conductivity and gas-phase water concentration are well known and yield precise estimates of these effects. The effect of temperature on matric potential can be described well only for those soils for which the matric potentials have been measured at more than one temperature.

List of Technical Nomenclature

α	Fitted parameter in van Genuchten equation (m^{-1})
β_0	Constant (K)
γ^{lg}	Surface tension of water ($N\,m^{-1}$)
η	Viscosity ($Pa\,s^{-1}$)
η_0	Fitted parameter ($Pa\,s^{-1}$)
Θ	Contact angle (rad)
θ	Volumetric water content ($m^3\,m^{-3}$)
θ_l	Liquid water content ($m^3\,m^{-3}$)
θ_r	Residual water content ($m^3\,m^{-3}$)
θ_s	Saturated water content ($m^3\,m^{-3}$)
θ_v	Gas water content (m^3 of liquid-water equivalent per m^3 of soil)
κ	Apparent thermal conductivity of the soil ($W\,m^{-1}\,K^{-1}$)
ρ	Density ($kg\,m^{-3}$)

τ	Reduced temperature
ϕ	Porosity ($m^3\,m^{-3}$)
ψ	Matric potential (m of water)
a_1	Fitted parameter
a_2	Fitted parameter
a_3	Fitted parameter
a_4	Fitted parameter
a_5	Fitted parameter
B	Fitted parameter (K)
C_v	Volumetric heat capacity ($J\,K^{-1}\,m^{-3}$)
D	Gas diffusion coefficient ($m^2\,s^{-1}$)
D_T	Thermal water diffusivity ($m^2\,s^{-1}\,K^{-1}$)
D_W	Water-content-based water diffusivity ($m^2\,s^{-1}$)
g	Acceleration due to gravity ($m\,s^{-2}$)
K	Unsaturated hydraulic conductivity ($m\,s^{-1}$)
K_s	Saturated hydraulic conductivity ($m\,s^{-1}$)
k	Intrinsic permeability (m^2)
k_r	Relative hydraulic conductivity
L	Latent enthalpy of vaporization ($J\,m^{-3}$)
M_{H_2O}	Molecular weight of water ($kg\,mol^{-1}$)
n	Fitted parameter in van Genuchten equation
p_c	Critical pressure (Pa)
p^{l+g}	Saturated vapor pressure of water (Pa)
R	Gas constant ($J\,K^{-1}\,mol^{-1}$)
r	Effective pore radius (m)
S_e	Degree of water saturation ($m^3\,m^{-3}$)
T_c	Critical temperature (K)
T_r	Reference temperature (K)
T_0	Fitted parameter (K)
t	Time (s)
$V_{m,\,H_2O(l)}$	Molar volume of liquid water ($m^3\,mol^{-1}$)
z	Depth (m)

See also: **Thermal Properties and Processes**

Further Reading

Bouyoucos GT (1915) Effect of temperature on the movement of water vapor and capillary moisture in soils. *Journal of Agricultural Research* 5: 141–172.

Constantz J (1991) Comparison of isothermal and isobaric water retention paths in non-swelling porous materials. *Water Resources Research* 27: 3165–3170.

Dullien FAL (1992) *Porous Media, Fluid Transport and Pore Structure*. New York: Academic Press.

Hopmans JW and Dane JH (1986) Temperature dependence of soil hydraulic properties. *Soil Science Society of America Journal* 50: 4–9.

Kakaç S, Kilkis B, Kulacki FA, and Arinç F (eds) (1991) *Convective Heat and Mass Transfer in Porous Media*. Norwell, MA: Kluwer Academic.

Nassar IN and Horton R (1992) Simultaneous transfer of heat, water, and solute in porous media. I. Theoretical development. *Soil Science Society of America Journal* 56: 1350–1356.

Philip JR and de Vries DA (1957) Moisture movement in porous materials under temperature gradients. *Transactions of the American Geophysics Union* 38: 222–232.

Raats PAC, Smiles DE, and Warrick AW (eds) (2002) *Environmental Mechanics, Water, Mass and Energy Transfer in the Biosphere*. Geophysical Monograph Series vol. 129. Washington, DC: American Geophysical Society.

Sageev A, Gobran BD, Brigham WE, and Ramey HJ Jr. (1990) The effect of temperature on absolute permeability to distilled water of unconsolidated sand cores. *Proceedings of the 6th Workshop on Geothermal Reservoir Engineering*, pp. 297–300. SGP-TR-50. Stanford, CA: Stanford University Press.

Wilkinson GE and Klute A (1962) The temperature effect on the equilibrium energy status of water held by porous media. *Soil Science Society of America Proceedings* 26: 326–329.

HYDRIC SOILS

G W Hurt, University of Florida, Gainesville, FL, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

The concept of hydric soils is relatively new to soil science, the term 'hydric soil' having been coined in the late 1970s to designate soils that have formed in wetlands. Only since the 1980s, in response to increased concern over the status of wetlands, have these seasonally wet soils received enough attention to merit a special designation of their own. Hydric soils are soils that support or are capable of supporting wetland ecosystems, i.e., soil modifications are not needed to maintain or restore a wetland. In the case of drained hydric soils, only removal of hydrologic modifications are needed to restore wetlands. Conversely, nonhydic soils do not support, nor are they capable of supporting wetland ecosystems; soil modifications are needed to create a wetland on nonhydric soils.

Hydric soils are most often defined as soils that have formed under conditions of saturation, flooding, or ponding long enough during the growing season to develop anaerobic conditions in the upper part. Nearly all hydric soils exhibit characteristic morphologies that result from repeated periods of near-surface saturation and/or inundation for more than a few days. Saturation or inundation when combined with microbial activity in the soil causes a depletion of oxygen. This anaerobiosis promotes biochemical processes such as the accumulation of organic matter and the reduction, translocation, and/or accumulation of iron and other reducible elements. These processes result in characteristic morphologies that persist in the soil during both wet and dry periods, making them particularly useful for identifying hydric soils. Hydric soil indicators are formed predominantly by the accumulation or loss of iron, manganese, sulfur, or carbon compounds.

Hydrology of Hydric Soils

Hydric soils form where water low in oxygen accumulates in the near surface of a soil or on the surface because water cannot leave the soil as fast as it enters. The difference between water inputs and outputs is storage, where the inputs are precipitation (P), overland flow onto the soil (F), and groundwater discharge (D); outflows are evapotranspiration (ET), runoff from the soil (R), and infiltration into lower soil horizons (I); and storage (S) is the water remaining in the soil or on the site:

$$(P + F + D) = (ET + R + I) + (S) \qquad [1]$$

If storage is great enough for a long-enough duration, a hydric soil forms.

Hydric soils form in highly complex wetland ecosystems under a variety of soil properties such as texture, hydrologic conductivity, and bulk density, and a wide variety of hydrologic and other climatic conditions and geomorphic settings; and eqn [1] is difficult if not impossible to quantify for all these conditions. A hydric soil can form in any soil material. It has therefore become convenient to determine the hydric status of a soil based on its morphology. The morphology of hydric soils is most often termed 'hydric soil indicators,' which were developed after observing and describing soil morphology of literally thousands of soils that supported wetland ecosystems. Some of the publications that explain morphologies of hydric soil field indicators are listed in Table 1.

Table 1 Publications of field indicators of hydric soils

Date	*Publication*
1990	Hurt GW, Henderson WG Jr, and Puckett WE. Field identification of hydric soils. In: Carlisle VW (ed.) *Hydric Soils of Florida Handbook*. Florida Association of Professional Soil Classifiers and Florida Cooperative Extension Service. Gainesville, FL: University of Florida Press
1992	Hurt GW and Puckett WE. Proposed hydric soil criteria (indicators) and their field identification. In: Kimble JM (ed.) *Proceedings of the Eight International Soil Correlation Meeting (VII ISCOM): Characteristics, Classification, and Utilization of Wet Soils*. Lincoln, NE: USDA/SCS
1995	Hurt GW, Watts FC, and Henderson WG Jr. Soil and water relationships of Florida's ecological communities and hydric soil indicators. In: Carlisle VW (ed.) *Hydric Soils of Florida Handbook*, 2nd edn. Gainesville, FL: Florida Association of Environmental Soil Scientists
1996	Hurt GW, Whited PM, and Pringle RF (eds) *Field Indicators of Hydric Soils in the United States* (Version 3.2). Fort Worth, TX: USDA–NRCS
1997	Mid-Atlantic Hydric Soil Committee. *Field Indicators of Hydric Soils in the Mid-Atlantic United States*. EPA Region III. Philadelphia, PA
1998	New England Hydric Soil Technical Committee. *Field Indicators for Identifying Hydric Soils in New England*, Version 2.0. New England Interstate Water Pollution Control Commission. Wilmington, MA
2002	Hurt GW, Whited PM, and Pringle RF (eds) *Field Indicators of Hydric Soils in the United States* (Version 5.0), Fort Worth, TX: USDA–NRCS

Important Chemical Reactions

The morphologies used to describe field indicators of hydric soils (Table 1) are created by oxidation–reduction (redox) chemical reactions that occur when a soil is anaerobic and chemically reduced. Most redox reactions produce soil colors and odors indicating that they have occurred. A reduced soil is one in which redox reactions have caused reduced forms of O, N, Mn, Fe, or S to be present in soil solution. 'Reduced' is a general term that implies that some element(s) or compound(s) in addition to O are present in reduced form. Common reduced forms found in hydric soils include: H_2O, N_2, Mn^{2+}, Fe^{2+}, and H_2S; their oxidized forms are O_2, NO_3^-, Mn^{4+}, Fe^{3+}, and SO_4^{2-}, respectively.

To understand how morphological features of reduction form, it is best to simplify oxidation and reduction processes and consider them to be separate reactions even though the two types of reactions occur simultaneously. Redox reactions are the chemical reactions in soils that allow us to assess the wetness of a soil. Although lithology and the amounts and types of vegetation have a major role in determining soil color and amounts of organic matter content at a regional scale, at a site-specific site, these reactions control soil color, organic matter content, the amount of oxygen in soil water, and the amounts and kinds of N, Mn, Fe, and S in soil water. Understanding redox reactions allow us to construe the morphology of a soil and determine the probable wetness (location in the soil).

Oxidation is the production of electrons. When a soil is aerobic (soil pores mostly filled with air), although there are many chemical reactions occurring, the one of interest to us is oxidation. Oxidation occurs when bacteria and other micro- and macroorganisms (collectively called microbes) decompose organic matter. This decomposition of organic matter produces substances that donate electrons and hydrogen ions (protons) which are then available for reduction. Organic matter decomposed by microbes produce electrons (e^-) and protons (H^+). The chemical reaction (oxidation) explained above is the reaction that occurs most of the time in soils. Oxidation is the norm; however, when a soil starts to become wet and eventually saturated (most all soil pores filled with water), reduction reactions begin at microsites and, if the soil is wet enough for a long enough duration, spread through the soil. Reduction is the reverse of oxidation: oxidation produces electrons; reduction consumes electrons. In most soils oxidation and reduction reactions are occurring simultaneously at different microsites.

Reduction of Oxygen

Even after a soil becomes saturated, O is still in the water that is in soil pores. As long as air is present in soil water, all electrons produced by organic matter decomposition are consumed by O to make water:

$$24e^- + 6O_2 + 24H^+ = 12H_2O \qquad [2]$$

This is an aerobic reduction reaction and is the first reducing reaction. When most of the O has been reduced, a soil is said to be anaerobic; As long as there is O in soil water, the reaction shown in eqn [2] continues. Once all the O is gone and can no longer enter the soil because it is waterlogged (saturated), the soil is anaerobic.

After the loss of O (O_2) to form water (H_2O), anaerobic microorganisms that are able to use substances other than O dominate the microbial community. Anaerobiosis results in the reduction of nitrates (NO_3^-) to form N (N_2) followed by Mn from manganic (Mn^{4+}) to manganous (Mn^{2+}), Fe from ferric (Fe^{3+}) to ferrous (Fe^{2+}), sulfates (SO_4^-) to form hydrogen sulfide (H_2S), and carbon dioxide (CO_2^-) to form methane (CH_4).

Reduction of Nitrogen

N reduction may explain the increased interest in wetland retention that occurred in the late 1960s and early 1970s. Nitrates in well water do little damage to most humans. The gastric acid in most stomachs generally kills the bacteria that might otherwise reduce the nitrate to nitrite. For infants less than approximately 2 months old, the nitrate is reduced to nitrite and binds with the infant's hemoglobin. The result is methemoglobinemia, commonly called blue baby syndrome, in which red blood cells carry less oxygen than normal and arrested brain development or death result. Hydric soils in wetlands reduce the nitrate to N_2. This is one of the denitrification reactions:

$$10e^- + 12H^+ + 2NO_3^- = N_2 + 6H_2O \qquad [3]$$

There are similar denitrification reactions that produce nitrite and ammonium (NH_4^+). Excess nitrate in wetlands can lead to eutrophication.

Iron and Manganese Reduction

MnO_2 (usually black in color) and $Fe(OH)_3$ (usually reddish in color) are solids which coat soil particles or other soil surfaces (peds, pores, etc.). Mn^{2+} and Fe^{2+} are colorless and in solution; they are dissolved Mn and Fe ions that move through soil water. Soils that are oxidized most all of the time are usually red in color, and soils that are reduced most all the time are usually grayish in color. Soils with intermittent

wetness can be a mixture of reds and grays; these are the soils that are the most interesting, because they occur near hydric soil boundaries:

$$2e^- + 4H^+ + MnO_2 = Mn^{2+} + 2H_2O \quad [4]$$

$$4e^- + 6H^+ + Fe(OH)_3 = Fe^{2+} + 3H_2O \quad [5]$$

Sulfur Reduction and Methane Production

When H_2S is produced the soil has a smell similar to rotten eggs. Once all the Mn and Fe has been reduced (Table 2), E_h (redox potential) will continue to fall, and when it reaches −150 mv, SO_4^{2-} anions may be reduced to H_2S gas; and if the E_h continues to drop, CO_2 will be reduced to methane gas (CH_4). These reactions occur only in very wet conditions:

$$8e^- + 10H^+ + SO_4^{2-} = H_2S + 4H_2O \quad [6]$$

$$8e^- + 8H^+ + CO_2^- = CH_4 + 2H_2O \quad [7]$$

Redox Potential

Redox potential is a voltage differential as measured between a platinum electrode and a reference electrode both in contact with the soil solution. The voltage is related to the chemical composition of the soil solution (reduced Fe, oxidized Fe, O content, etc.). The voltage reading must be converted to E_h by adding a correction of approximately 200 mV. The correction factor varies with temperature and the type of reference electrode and ranges from approximately 197 to 260). The various reduction reactions (explained above) occur when the soil solution is at various redox potentials. These reactions also vary with pH and mineral species present. One suggested correlation between the reduction reactions and E_h is presented in Table 2.

Figure 1 shows specific E_h/pH lines for O and Fe with the E_h values at pH 7 of +350 and +120, respectively, adjusted with a slope of −60 for each pH value less than pH 7. If E_h values measured at specific pHs plot above the upper line O is present; if below, Oxygen is reduced. If E_h values measured at specific pHs plot above the lower line, Fe is oxidized; if below, Fe is reduced.

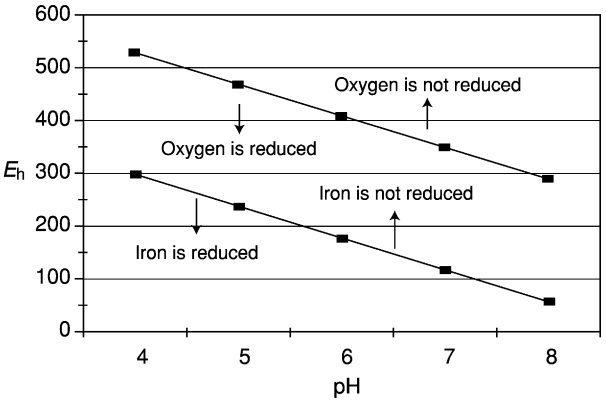


Figure 1 Determination of oxidation–reduction status for oxygen and iron. E_h, redox potential.

Table 2 Reduction reactions and E_h at pH 7

Reduction reaction	E_h *(mV)*
$24e^- + 6O_2 + 24H^+ = 12H_2O$	+350
$10e^- + 12H^+ + 2NO_3^- = N_2 + 6H_2O$	+220
$2e^- + 4H^+ + MnO_2 = Mn^{2+} + 2H_2O$	+200
$e^- + 3H^+ + Fe(OH)_3 = Fe^{2+} + 3H_2O$	+120
$8e^- + 10H^+ + SO_4^{2-} = H_2S + 4H_2O$	−150
$8e^- + 8H^+ + CO_2^- = CH_4 + 2H_2O$	−250

Factors that Control Redox Potential

The factors that control reduction in soils and redox potentials are: (1) organic matter must be present (source of electrons). Good sources of organic matter include dead roots, microbial biomass, plant debris such as leaves, and pieces of roots that come from live roots. Roots even excrete a substance similar to sap upon which bacteria and other organisms thrive. Dissolved organic carbon (material previously excreted by an organism) is utilized by other organisms; (2) air must not enter the soil (soil is saturated); (3) organisms must be decomposing organic matter. This activity is termed 'microbial active.' For most of the world, wet soils are microbial active when air temperature is above approximately 5°C. Reduction, which occurs only when microbes are active, is believed to occur in tundra and Antarctica when the air temperature is below 5°C because of the presence of ethylene glycol (antifreeze) in microbes; (4) dissolved O in the soil water must be removed.

Reduction occurs where water moves through soil slow enough that O is removed and occurs best where water does not move through the soil and most all pores are filled with water. Reduction does not occur where water flows too quickly to allow O to be removed. O must be removed before reduction of other elements can occur. Soil may be saturated but not reduced if: (1) organic matter content is low, (2) microbial activity is lacking, and/or (3) water moves too quickly (O is not removed).

Hydric Soil Indicators

Redox reactions leave signs in the soil that they have occurred as shown in Table 3.

Table 3 Examples of features formed by reducing conditions

Reduced element	*Morphological feature formed*	
	Feature type	*Examples*
Oxygen	Organic (C)-based	Organic soil material Mucky mineral soil material Some black A horizons
Nitrogen	None	None
Manganese	Mn-based	Mn concentrations (mostly black) in a dominantly depleted matrix
Iron	Fe-based	Fe concentrations (mostly red) in a dominantly depleted matrix Reduced matrix
Sulfur	S-based	Rotten egg odor
Carbon	None	None

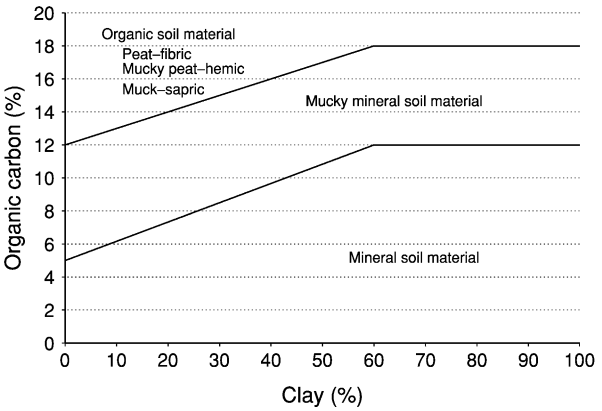


Figure 2 Percentage of organic carbon required for organic soil material, mucky mineral soil material, and mineral soil material as it relates to percentage of clay content.

Carbon Features

Carbon-based features occur when decomposition (of carbon) is slower than production and occur because oxygen is absent. Carbon features are widespread and must be used to identify reduction in soils that contain little or no iron/manganese. Organic-based features consist of one of three kinds of materials: organic soil material and mucky mineral soil material (Figure 2), or mineral soil material with a black color. An exact amount of carbon is usually required (Figure 2). Color and carbon content are the results of relatively high amounts of organic material that accumulate under reduced conditions.

N-based features do not exist. There are no known soil morphological features that are the results of the redox reactions of nitrogen.

Iron and Manganese Features

The reduction, translocation, and oxidation of Fe/Mn oxides form Fe/Mn-based features. The types of features are redox concentrations and redox depletions collectively known as redoximorphic features and are formed by the reduction, movement, and oxidation of Fe and Mn compounds. There are three basic kinds of redoximorphic features: redox concentrations, redox depletions, and the reduced matrix. Both Fe and Mn are used, because the two are very similar in soils. In the absence of additional information, the guidelines in Table 4 can be used to identify the oxidized state of Fe and Mn oxides.

Redox concentrations are zones (usually reddish in color) where Fe/Mn is concentrated. There are three types: (1) nodules and concretions, which are extremely firm (hard when dry), irregular-shaped bodies; (2) masses, which are soft bodies within the soil matrix; and (3) pore linings, which are coatings or impregnations on a natural soil surface (root pore, air pore, ped face, crack face). Pore linings are the most reliable evidence of contemporary wetness, and nodules and concretions are the least reliable evidence of contemporary wetness, because most often they reflect prior wetness. Masses may be the results of reduced iron being reoxidized or evidence that that volume of the soil has never been reduced.

Table 4 Color and composition of iron and manganese oxides

Chroma	*Value*	*Dominant composition*
≤ 2	≤ 2	Manganese
3	3	Iron and Manganese
≥ 4	≥ 4	Iron

Redox depletions are zones (usually grayish in color) where Fe/Mn oxides have been removed. Usually, when a soil has areas of redox depletions, it also has areas of redox concentrations. If something is removed and translocated, it usually shows up somewhere else. A specialized case where this in not true is the 'reduced matrix.' A reduced matrix is gray when freshly exposed and turns redder when exposed to air, because the Fe has been reduced and is thereby colorless but still present in soil water. The Fe oxidizes when exposed to air.

If microbes are active, if Fe/Mn oxides are present in the soil, if the soil is saturated, if the water is stagnant, if there is a source of usable organic matter in the soil, and if O and N have been reduced then the electrons produced by microbes decomposing the organic matter will reduce and translocate the Fe/Mn present and redox features will result. Usually the greater the length of the reduced period the greater will be the percentage of the soil mass redox depletions.

Except for a few exceptions that occur on specific landforms, hydric soils with redox features (features formed by Fe/Mn redox reactions) have more redox depletions than redox concentrations. Fe/Mn features are widespread but occur only in soils with Fe/Mn. Soils that have formed in parent material low in Fe/Mn may have features which appear as depletions. In these soils, carbon-base indicators are used to determine their hydric status.

The reduced matrix is a soil matrix that has a low-chroma color *in situ* because of the presence of Fe^{2+}, but whose color changes in hue or chroma when exposed to air as the Fe^{2+} is oxidized to Fe^{3+}. This phenomenon occurs because, although reduction occurs, the reduced iron lacks an outlet (lateral or vertical) and oxidizes when exposed to air.

Sulfur Features

The only sulfur-based feature is the production of H_2S gas; no particular morphology is associated with H_2S production. The presence of H_2S gas is a very positive indicator of reduction. It is rarely identified in the field because it occurs in soils so wet that few question or look for the presence of hydric soils. Mercaptans are often misidentified as hydrogen sulfide. Mercaptans are s-containing organic compounds with the general formula RSH where R is any radical, especially ethyl mercaptan, C_2H_5SH. Swamp gas (methane or natural gas) features do not exist. There are no known soil morphological features that are the results of the redox reactions that result in the formation and dissolution of methane.

Marl Features

One additional morphology not associated with oxidation–reduction reactions but associated with hydric soils is the production of marl at the soil surface. Marl is a limnic material deposited in water by precipitation of calcium carbonate ($CaCO_3$) by algae. Mineralogical analyses have indicated that almost 100% of the soil minerals are calcite that had precipitated from water saturated with calcium bicarbonate. Marl is formed when the land is flooded for several months each year in the summer (hydroperiod) followed by a number of dry months during the winter. During the annual wet season when these lands are flooded, several hundred species of microalgae (periphyton) grow on the surface water (eqn [8]):

$$Ca^{2+} + 2HCO_3 = CaCO_3 = H_2CO_3 \qquad [8]$$

Periphyton growth produces marl. Flooding creates anaerobic conditions and increases the accumulation of organic matter. During the dry period, this organic matter is oxidized. The relative lengths of the dry and wet periods are approximately the same and therefore marl normally contains little organic matter and is mostly white in color. The rate of calcite mud (marl) production in present coastal areas, as measured by radiocarbon dating, has averaged 1.2 cm per 100 years from 1000 years ago to the present. Generally marl soils form a layer 5–120 cm thick above limestone bedrock.

Summary

Oxidation–reduction reactions are mostly responsible for gray colors in soils. These reactions need organic matter, bacteria, and no O to be present in order to occur. The amount of low-chroma color in soil is usually related to how long it has been reduced, not how long it has been saturated; however, some sandy soils are naturally low in Fe and have low-chroma colors not related to reduction. Redoximorphic features form by Fe/Mn reduction, translocation, and oxidation, and the types of redoximorphic features are concentrations and depletions. Soil morphology that results from redox reactions is used to identify hydric soils in the field.

See also: **Factors of Soil Formation:** Climate; **Wetlands, Naturally Occurring**

Further Reading

Carlisle VW and Hurt GW (eds) (2000) *Hydric Soils of Florida Handbook*, 3rd edn. Gainesville, FL: Florida Association of Environmental Soil Scientists.

Environmental Laboratory (1987) *Corps of Engineers Wetland Delineation Manual*. Technical Report Y-87-1. Vicksburg, MS: US Army Engineers Waterways Experiment Station.

Hurt GW and Richardson JL (eds) (1998) *Hydric Soils of the United States: A Guide to Their Recognition,* (CD-ROM). Fort Worth, TX: USDA–NRCS.

Hurt GW, Whited PM, and Pringle RF (eds) (2002) *Field Indicators of Hydric Soils in the United States, Version 5.0*. Fort Worth, TX: USDA–NRCS.

Mausbach MJ and Richardson JL (1994) Biogeochemical Processes in Hydric Soils. *Current Topics in Wetland Biogeochemistry* 1: 68–127.

Munsell® Soil Color Charts (2000). New Windsor, NY: Gretag Macbeth.

Richardson JL and Vepraskas MJ (eds) (2001) *Wetland Soils: Genesis, Hydrology, Landscapes, and Classification*. Boca Raton, FL: Lewis Publishers.

Schoeneberger PJ, Wysocki DA, Benham EC, and Broderson WD (1998) *Field Book for Describing and Sampling Soils*. National Soil Survey Center. Lincoln, NE: USDA–NRCS.

Soil Survey Staff (1999) *Soil Taxonomy: A Basic System of Soil Classification for Making and Interpreting Soil Surveys*. USDA Agricultural Handbook 436. Washington, DC: US Government Printing Office.

Vepraskas MJ (1994) *Redoximorphic Features for Identifying Aquic Conditions*. Technical Bulletin 301, North Carolina Agricultural Research Service. Raleigh, NC: North Carolina State University Press.

HYDROCARBONS

P Kostecki, University of Massachusetts, Amherst, MA, USA
R Morrison, DPRA, Inc., San Marcos, CA, USA
J Dragun, Dragun Corporation, Farmington Hills, MI, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

This article presents a scientific overview of the basic principles involved with remediating sites containing soils contaminated with petroleum hydrocarbons. The first section discusses the physical and chemical composition and properties of petroleum hydrocarbons. The subsequent sections discuss factors affecting the transport and transformations of petroleum hydrocarbons in soil systems, as well as the primary engineering technologies utilized to remediate soils contaminated with petroleum hydrocarbons.

Chemistry of Petroleum Hydrocarbons

Petroleum hydrocarbons and their refined products originate from a highly complex mixture, although they are comprised mainly of only two elements, hydrogen and carbon, which can be arranged in an almost infinite number of structures. There are some distinct rules that govern the arrangement of these atoms. The carbon atoms can be attached or bonded to up to four different atoms, carbon or hydrogen, while the hydrogen atoms are bonded to only one carbon atom. The carbon atoms are bonded to each other and form a structure called the carbon backbone. Using the same number of atoms, many different arrangements of the atoms can be made. Each arrangement constitutes a unique organic compound.

Although each compound has its own unique chemical properties, the vast number of compounds present in crude oil makes it necessary to group the compounds whenever possible. Rather than knowing the chemical properties of all compounds, the general properties of a group or a class of compounds are used to estimate or approximate the expected property for an individual compound. Where more detailed information is needed, the actual properties for a specific compound can be determined.

One method employed to group a hydrocarbon is by the number of carbon atoms that are present in the compound. For example, all of the hydrocarbons that have four carbon atoms can be designated as C_4. The C_4 compounds commonly found in petroleum products include butane, isobutene, 1-butene, 2-butene, and isobutylene. In this example, C_4 can refer to as many as five different compounds, while C_{10} can refer to more than 470 individual compounds.

Compounds from C_1 to C_4 are gases at ordinary temperature. They are used as fuels and in the manufacturing of specialty chemicals, including plastics or polymeric materials. Compounds from C_5 to C_{12} are liquids that will easily turn into gases when heated. These compounds are used as fuels and as cleaning solvents. Compounds from C_{13} to C_{17} are generally liquids and are used as fuels and lubricants. Those compounds greater than C_{17} are usually solids and are used as lubricants, heavy fuels, and as coatings such as tars and asphalt. The solid hydrocarbons usually have the consistency of wax. These solids will often pour or flow when heated or dissolved in a solvent.

A second major grouping is based on the chemical reactivity of the hydrocarbon compounds. The reactivity of these compounds is based on the molecular structure or the actual arrangement of the hydrogen and carbon atoms. The nature of the carbon backbone and the number of hydrogen atoms that are attached to it determine the chemical properties of these compounds. There are three major classifications of hydrocarbons that are found in petroleum products: alkanes, alkenes, and aromatics.

Alkanes

Alkanes are the major constituents of crude oil and the major constituents of most petroleum products. Alkanes are also called saturates or saturated hydrocarbons. The carbon atoms that make up the carbon backbone are linked together to form a chain (linear or branched alkanes), a circle (cyclic alkenes), or a mixture of both. The alkanes are distinguished from the other hydrocarbons by the fact that they are completely

saturated with hydrogen. This means that no additional hydrogen atoms can be added to these hydrocarbons without destroying the carbon backbone.

Linear and branched alkanes are sometimes called paraffins. These compounds have an empirical formula of C_nH_{2n+2}. A linear or branched alkane that has four carbon atoms will also have 10 and only 10 hydrogen atoms. Linear alkanes, i.e., alkanes whose carbon atoms are arranged in a straight chain or row, are called normal paraffins, normal alkanes, or *n*-alkanes. The branched alkanes are often called isoparaffins or isoalkanes. The carbon atoms of the carbon backbone of the branched alkanes might look like a T, Y, or X.

The simplest alkane is methane. Methane contains only one carbon atom and four hydrogen atoms. The other alkanes can be made by substituting carbon atoms for hydrogens.

The cycloalkanes are similar to the linear and branched alkanes, except that the carbon backbone has the carbon atoms arranged in a circle. These compounds are sometimes called cycloparaffins and have an empirical formula of C_nH_{2n}. These compounds are also saturated. The carbon atoms that make up the backbone are arranged in circles containing usually five, six, or seven carbon atoms.

Hydrocarbons also exist that are a combination of linear alkanes and cycloalkanes. These compounds have the empirical formula C_nH_{2n}. The fact that they have a cyclic groupin dominates their chemistry and they are usually called cycloalkanes or alkyl cycloalkanes. If these hydrocarbons contain a five- or a six-membered ring, they are called napthenes (not to be confused with naphthalenes.)

Alkenes

The second major group of petroleum hydrocarbon is the alkenes, also known as olefins. These compounds are usually not found in crude oil but are the by-product of a refining process. The alkenes are distinguished from the alkanes by the fact that they contain two less hydrogen atoms than the alkane with the same carbon backbone. Because the alkenes have two fewer hydrogen atoms, they contain a carbon–carbon double bond. This carbon–carbon double bond imparts special chemistry to these compounds different from that shown by the alkanes. The chemistry of the alkenes is usually of little importance to petroleum products since most are simply used as a fuel and burned. The alkenes, especially ethylene and isobutylene, are very important starting materials in the manufacture of specialty chemicals.

The alkenes are considered to be unsaturated because hydrogen can be added (to form the corresponding alkane) without breaking the carbon backbone. Addition of two hydrogen atoms to isobutylene yields isobutene. Alkenes have the empirical formula C_nH_{2n}.

Aromatics

The third basic type of hydrocarbon found in petroleum products is the aromatics. Aromatic hydrocarbons typically contain a six-membered carbon ring and they have an empirical formula of C_nH_{2n-6}. The six-membered ring has three carbon–carbon double bonds. This arrangement of three double bonds within a six-membered ring imparts very special chemistry to these compounds. They can behave in ways that are very different from the other hydrocarbons.

Benzene is the simplest aromatic hydrocarbon. Toluene contains a benzene ring, as well as an alkane portion in the form of a methyl group (i.e., CH_3.) Toluene can also be called a C_1-benzene. Ethyl benzene contains a benzene ring with an attached ethyl (CH_2CH_3) group. There are three different xylenes; each contains two methyl groups attached to a benzene ring. The xylenes are distinguished from each other by the relative location of the two methyl groups on the benzene ring. Ethyl benzene and the three xylenes can be referred to as C_2-benzenes. The higher aromatics such as naphthalene and benzo (*a*) pyrene contain more than one benzene ring. The benzene rings are fused where each ring shares two carbon atoms with its neighboring ring. These latter two compounds are part of a class of hydrocarbons called polynuclear aromatic hydrocarbons (PAH or PNA).

There are two very useful terms that can be used to describe similar yet different hydrocarbons. These terms make it easier to express the complexity of the hydrocarbons that are found in petroleum and in petroleum products: 'homologous series' and 'isomers.'

Homologous Series

A homologous series is a series of compounds that differ from each other by the removal or addition of a common unit or group of atoms. An example of a homologous series of *n*-alkyl benzenes would begin with benzene (C_6H_6) and would include toluene ($C_6H_5CH_3$), ethyl benzene ($C_6H_5CH_2CH_3$), *n*-propyl benzene ($C_6H_5CH_2CH_2CH_3$), *n*-butyl benzene ($C_6H_5CH_2CH_2CH_2CH_3$), and so on. Except for benzene, these compounds all contain a normal alkane attached to a benzene ring. Each succeeding compound contains a normal alkane containing one additional CH_2 unit.

A second homologous series based on benzene could be benzene, C_1-benzene (toluene), C_2-benzenes

(ethyl benzenes, *o*-xylene, *m*-xylene, *p*-xylene), C_3-benzenes, and so on. This group differs from the previous in that it includes isoparaffins, as well as more than one alkyl group attached to a benzene ring. Similar series can be made for cyclohexane and naphthalene.

An important homologous series for petroleum products are the *n*-alkanes. The series includes: *n*-C_1 (methane), *n*-C_2 (ethane), *n*-C_3 (propane), *n*-C_4 (butane), *n*-C_5 (pentane), *n*-C_6 (hexane), *n*-C_7 (heptane), *n*-C_8 (octane), *n*-C_9 (nonane), *n*-C_{10} (decane), *n*-C_{11}, *n*-C_{12}, and so on. This latter homologous series is particularly important in the characterization of middle-range petroleum distillates and residual petroleum products such as diesel, fuel oil, jet fuel, kerosene, and bunker oil.

Isomers

Isomers are compounds that contain exactly the same number of atoms, i.e., they have exactly the same empirical formula, but differ from each other by the way in which the atoms are arranged. Examples of isomers with the formula C_8H_{10} are ethyl benzene, *m*-xylene, *p*-xylene, and *o*-xylene. This particular group of isomers is C_2-benzenes, since they all have two carbon atoms attached to a benzene ring.

There can be several isomers for each empirical formula. The total number of possible isomers containing 40 carbon atoms and 82 hydrogen atoms (C_{40} alkanes) is 62 491 178 805 132. Crude oil and its products probably do not contain all possible isomers; however, some highly refined products may contain a significant fraction of the total number of isomers that are theoretically possible. These numbers will provide some insight into the complexity and wide variety of compounds that can be present in petroleum.

Transport of Petroleum Hydrocarbons in Soil

The fundamental principles governing advective (mass) transport of water in soil generally apply to those for hydrocarbon transport (i.e., gravitational and capillarity). Hydrocarbons move through the soil under a three-phase flow condition, displacing air and water. Variations in soil permeability result in a deviation from the gravitationally dominated vertical flow; as the hydrocarbon encounters layers of slightly less permeable materials, or if materials with smaller pores are encountered, it will tend to flow mostly in the horizontal direction until it encounters a path of less resistance.

This conceptual model is more complex, however, because other transport and transformation processes occur. For example, simultaneous processes may occur upon the release of a hydrocarbon at the ground surface:

1. Volatilization of the lighter components, such as benzene, into the atmosphere;
2. Runoff over the soil surface, driven by gravity and affected by the surface configuration and possibly by rainfall, with consequent contamination of nearby surface waters or concentration of the materials in surface depressions;
3. Adherence of the heavier components to the soil surface, possibly resulting in clogging and hydrophobization of the soil;
4. Soil infiltration;
5. Downward and lateral flow within the unsaturated zone in the soil profile;
6. Retention of the hydrocarbon at soil textural interfaces, such as clay, within the soil profile;
7. Retention in soil pores and attachment to soil grains as well as soil organic matter;
8. Volatilization and vapor diffusion within the soil and out of the surface, or possible movement of a vapor cloud vertically into the saturated zone or soil pore water, and possible redissolution into the groundwater at the water table;
9. Chromatographic separation of components within the profile, resulting in selective migration of lighter and less-viscous components;
10. Partial dissolution of soluble or emulsified components within the water phase of the soil;
11. Degradation resulting from both nonbiological (i.e., chemical) and biological processes;
12. Internal drainage or leaching from the soil either within or alongside the water phase toward the water table;
13. Mounding over the water table and divergence of the petroleum hydrocarbon over it, possibly including convergence in cones of depression or drawdown regions at wells or streams;
14. Penetration into the aquifer of the soluble and high specific density compounds;
15. Dispersion and additional migration within the aquifer and eventual appearance in the water supplies.

An example illustrating these possible interactions is useful. If a large volume of gasoline or diesel is released at or near the surface, it will initially tend to infiltrate vertically through the soil. If the volume of release is sufficient to exceed the residual soil-retention capacity, the migration of the hydrocarbon will continue until the fluid reaches the capillary fringe, where it accumulates. As the hydrocarbon is transported through the soil, it may encounter a less-permeable layer, which may cause lateral spreading

until a more permeable zone is encountered for the hydrocarbon to move vertically.

On a smaller scale, the distribution of the hydrocarbons in the soil is dependent on the pore size between the soil grains and the pore pressures of the air, water, and hydrocarbons occupying these pore spaces. The permeability of a soil for a hydrocarbon is different from the permeability for water, owing to differences in viscosity. Also, gasoline components will preferentially dissolve from the hydrocarbon. Furthermore, these components may migrate at different velocities through the soil; this phenomenon is called chromatographic separation. Chromatographic separation is often observed in the separation of more mobile compounds such as methyl *tert*-butyl ether (MTBE) moving ahead of the center of mass of a hydrocarbon in groundwater.

Residual Saturation

'Residual saturation' is defined as the fraction of the total soil space that is filled with a liquid due to capillary forces. Residual saturation plays an important role in hydrocarbon transport through the subsurface. As hydrocarbons migrate through soil, a small amount of the total hydrocarbon mass remains attached to soil particles via sorptive forces. The hydrocarbon that is retained by the soil is also known as the immobile saturation. It is dependent on soil-moisture content, soil porosity, and soil texture. Ranges for gasoline and diesel residual saturations for different soils are listed in Table 1.

The residual saturation of a hydrocarbon is also affected by its viscosity. As the viscosity of the hydrocarbon decreases, the residual saturation concentration also decreases. The residual saturation for petroleum products, expressed as concentration (grams per kilogram), in soils is shown in Table 2.

Residual saturation is significant because it remains as a source of contamination for water infiltrating through the soil column that contacts the residual hydrocarbons. Infiltrating water can 'flush' entrapped residual hydrocarbons from the larger pores in a soil.

Residual-saturation contamination by hydrocarbons can occur downgradient from a spill that may be on an adjacent property. This situation occurs when hydrocarbon floating above a water table comes into contact with otherwise clean soils above it due to a rising water table. Subsequently when the water table drops, the previously clean soil retains a certain percentage of residual saturation and thus becomes a source for further contamination. This phenomenon is important to consider when examining soil gas survey data as an indicator of the source of contamination, especially with shallow groundwater. Soils with residual hydrocarbon downgradient of the original spill may volatilize, resulting in the appearance of a release and/or masking of the original source of the contamination.

Table 1 Ranges for gasoline and diesel fuel residual saturation for different soils

Soil texture	*Gasoline (%)*	*Diesel fuel (%)*
Coarse gravel	0.5–2.0	1–4
Coarse sand	1.0–3.0	3–8
Medium sand	2.0–4.0	4–10
Fine sand	3.0–5.0	6–20
Silt	5.0–9.0	10–30

Table 2 Residual saturation for petroleum products in soils (grams/kilogram soil)

Soil type	*Gasoline*	*No. 2 fuel oil*	*Lube oil*	*No. 6 fuel oil*
Coarse gravel	–	0.8	1.6	–
Gravel to coarse sand	–	1.6	3.2	–
Coarse to medium sand	–	2.8	5.6	–
Medium to fine sand	2.0	4.8	9.6	60.0
Fine sand to silt	–	8.0	16.0	–

Estimation of the Distribution of Hydrocarbons in Soil

The soil volume required to immobilize hydrocarbons depends on soil porosity, the physical properties of the hydrocarbons, and the volume of released hydrocarbon.

The American Petroleum Institute (API) has approximated the soil volume required to immobilize a volume of hydrocarbons as residual saturation by:

$$V_s = (\alpha V_C / P)(\mathrm{RS})$$

where V_C is the volume of discharged hydrocarbons, V_s is volume of soil required to attain residual saturation, P is soil porosity, RS is residual saturation capacity, and α is coefficient, depending on the units used (whether metric or other units).

The maximum depth of hydrocarbon penetration in the subsurface by hydrocarbons can be estimated:

$$D = KV_{HC}/A$$

where V_{HC} is volume of discharged hydrocarbons, A is area of infiltration, and K is a constant that is dependent on the soil's sorptive capacity and hydrocarbon viscosity. Table 3 lists representative K values for three hydrocarbons in soils of various textures.

An approximation to estimate the area over which a spill has occurred is described by:

Table 3 Typical values of K for various soil textures

Soil texture	*K (gasoline)*	*K (kerosene)*	*K (light fuel oil)*
Stone and coarse gravel	400	200	100
Gravel and coarse sand	250	125	62
Coarse and medium sand	130	66	33
Medium and fine sand	80	40	20
Fine sand and silt	50	25	12

$$A_s = 53.5(V_C)^{0.89}$$

where A_s is spill area in square meters, and V_C is volume of discharged hydrocarbon in barrels (1 barrel $\approx$ 160 l). The actual area may vary from the calculated area, due to climatic, topographic, soil, and hydrocarbon properties.

For a steady-state release of a hydrocarbon (a leaking underground storage tank, piping, etc.) into a soil overlying a relatively impermeable layer under isothermal conditions, the radius (R) of the mass of the bulk hydrocarbon is estimated as:

$$R = (Q/3.142h)^{1/2}t^{1/2}$$

where Q is the rate of hydrocarbon addition to the soil, h is height of the oil film, and t is time after the initial hydrocarbon addition to the soil.

Hydrocarbon Interactions at the Capillary Fringe

Vertical migration ceases when the hydrocarbon encounters water-saturated pores (i.e., the capillary fringe or a perched layer). As the volume of floating product increases, saturation at the capillary fringe becomes sufficient for lateral spreading to occur. A mound above the water table occurs along with a corresponding depression in the water table. Spreading takes place along the groundwater surface as a result of the natural groundwater gradient and the gradient caused by the hydrocarbon accumulating above the groundwater surface. The hydrocarbons will eventually move laterally over the water table in a layer roughly as thick as the capillary fringe.

Some of the hydrocarbon residing in the capillary fringe will slowly volatilize into the vadose zone. The rates of volatilization are controlled by the flow of air. Hydrocarbons resting on the water table tend to compress the capillary zone and lower the water table. The compression depends on the hydrocarbon density. As the specific gravity of gasoline and light hydrocarbons are approximately 0.7–0.8, approximately 70%–80% of the hydrocarbons will lie below the depth of the original water table.

Floating free-phase hydrocarbons that are less dense than water (low-density nonaquaeous phase liquids, LNAPLs) may accumulate as a mound and then travel radially in all directions outward from the mound center. This hydrocarbon lens will spread laterally until equilibrium is achieved between the lens thickness and the capillary pressure. As expected, this balance of forces results in a thick layer in fine-grained sediments and a thinner layer in coarse sediments. The center of hydrocarbon mass, however, will move downgradient with respect to groundwater. Volatile components within the LNAPL may also evaporate into the air and diffuse upward, while the higher molecular weight compounds remain in the LNAPL pool. Compounds with low oil–water partition coefficients will tend to diffuse outward and transfer into the aqueous phase, leaving the heavier-weight chemicals within the mound.

Upon entering the capillary fringe, sufficient hydraulic pressure may exist to cause petroleum hydrocarbons to move hydraulically up gradient. This upgradient distance is minimal when compared with the downgradient distance. As the hydrocarbon lens moves laterally, water in the capillary fringe impedes its movement. In the upper section of the capillary fringe, where small amounts of water are present, the hydrocarbons tend to migrate laterally. In the lower section of the capillary fringe, where relatively large amounts of water are present, the bulk hydrocarbons also migrate laterally but more slowly than in the upper portions of the capillary fringe.

The maximum spread of the hydrocarbon lens over the groundwater table can be approximated with the following equation:

$$S = (1000/F)[V - (Ad/K)]$$

where S is maximum spread of the bulk hydrocarbons in meters squared, F is thickness of the hydrocarbons in millimeters, V is volume of bulk hydrocarbon infiltration in meters cubed, A is area of infiltration (meters squared), d is depth to groundwater in meters, and K is a constant dependent on the soil's sorptive capacity for a given hydrocarbon and the viscosity of the hydrocarbon. The vertical thickness of this lens shadows fluctuations of the groundwater table, with hydrocarbon sorption and flushing occurring.

Solubilization of Hydrocarbons into Groundwater

The solubilization of products from a petroleum product results in a chemical composition for the dissolved component in the groundwater that is different than

the chemical composition of the initial product. For nonoxygenated gasoline, the most soluble components are benzene, toluene, ethyl benzene, and xylene (BTEX). While these compounds comprise 23–55% of the bulk product in a gasoline, they will comprise from 42 to 74% of the hydrocarbons in the groundwater. Benzene concentrations, for example, are enriched up to 10 times in the groundwater relative to the gasoline. Solubilization experiments have shown that while an unleaded gasoline contains 49% of the aromatic compounds, water in contact with the gasoline contains 95% of these compounds after 30 min of mixing.

For crude oils, the solubility is partially dependent on the API gravity of the crude oil. API gravity is described by the following expression:

$$\text{API} = (141.5 \times \text{density of water at } 60^\circ\text{F} / \text{density of petroleum at } 60^\circ\text{F}) - 131.5$$

Crude oil with a high API is termed 'light' (less dense than water), while a low-API crude is 'heavy.' The solubility of the crude oil in contact with groundwater increases with higher API gravities, because high-API crude oils tend to be rich in low molecular weight compounds with high solubilities. The solubility of crude oil with various API gravities as measured in parts per million (milligrams per liter) total dissolved hydrocarbons is given in Table 4.

The mass of soluble hydrocarbons that actually dissolve into the groundwater is poorly defined. One expression for approximating the amount of dissolved substance originating from free-phase hydrocarbons is to quantify the 'source strength' represented by the free-phase hydrocarbons. The term 'source strength' is used to describe the intensity with which dissolved chemicals are released from a hydrocarbon into water. Source strength can be expressed as a mass per time per unit contact area:

$$S = K_m A$$

where S is source strength (mass per time per meter squared), K_m is mass exchange coefficient (mg per meter squared per second), A is contact area (meter squared).

Table 4 Solubility of crude oil with various API gravities (milligrams/kilogram soil)

	Distilled water		
API gravity	*25°C*	*35°C*	*45°C*
11	3.5	4.5	6.0
15	4.0	5.0	7.0
20	10.0	14.0	18.0
28	65.0	75.0	90.0

The variable A is the interface across which mass exchange occurs over a given period of time. Due to the variety of hydrocarbon distribution patterns in the soil, estimating A is usually qualitatively defined. Maximum mass-exchange coefficients estimated by the US Geological Survey in 1984 were reported to be: 1.0 mg m^{-2} s for gasoline and tar oil; 0.01 for fuel oil, diesel, and kerosene; and 0.001 for lubricating oils and heavy fuel oil.

Petroleum Hydrocarbon Transport in Groundwater

Hydrocarbon transport in groundwater includes both the movement of the free phase in contact with the capillary fringe or within the groundwater, and the dissolved phase within the saturated zone. The movement of both plumes is controlled not only by the physical characteristics of the hydrocarbon as it enters the groundwater, but also by the media through which it flows.

Free-Phase Groundwater Plume

The free-phase component of the contaminant plume normally resides in greatest thickness at the point at which the hydrocarbon enters the groundwater. The lateral movement of the free-phase plume is controlled by those soil properties that control the movement of fluids and by the interaction of the free-phase plume with the groundwater gradient.

A situation can arise in which a free-phase hydrocarbon with a density less than water (an NAPL such as benzene) is found below the water table. One mechanism that accounts for this situation is the result of a water table increase and the entrainment of the benzene below or within a low-permeability sediment.

A dissolved phase can subsequently originate from the trapped free-phase liquid under the low-permeability layer. The rate at which the dissolved phase is created and its concentration can be estimated from solubility data and from information on the chemical concentration of the compound.

Dissolved-Phase Groundwater Plume

Most hydrocarbon fractions are immiscible with water. Some components in hydrocarbons, however, are soluble (e.g., BTEX), as are additives such as MTBE. This property provides a useful means for characterizing hydrocarbons and for understanding whether BTEX concentrations represent near-saturation conditions. Once the soluble fraction of the hydrocarbon has entered the groundwater, its movement is controlled by the same advective and dispersive processes used for estimating the mass transport of groundwater.

Length of Petroleum Hydrocarbon Plumes

Recent examination of hydrocarbon plumes has resulted in the finding that plumes rarely exceed 80 m in length. These observations in turn provide evidence for consideration and/or adoption of natural attenuation strategies as a remedial approach in documents such as the American Society for Testing and Materials *Standard Guide for Remediation by Natural Attenuation at Petroleum Release Sites*.

Lawrence Livermore National Laboratory conducted a study of hydrocarbon plumes in groundwater for the California State Water Resources Control Board to support revision of the Leaking Underground Fuel Tank corrective action process. As part of the data interpretation, the lengths of the hydrocarbon plumes were summarized; these results are similar to the findings of a study by Chevron of 119 sites in California. The results of both studies and the number of sites examined are summarized in Table 5.

These studies also classified hydrocarbon plumes in terms of the change in plume geometry for benzene. The plumes were described as 'shrinking,' 'no significant trend,' and 'increasing.' The results of this classification as percentages of the total sites are shown in Table 6.

Analytical solutions for groundwater transport and attenuation of hydrocarbons used to estimate the length of hydrocarbon plumes have been found to be similar to the Lawrence Livermore and Chevron experimental findings. The similarities between modeled and observed plume lengths include the finding that BTEX plumes may be relatively short (<90 m). Longer plumes might be expected when groundwater velocity is high ($>3\,\mathrm{cm\,day^{-1}}$) and when attenuation rates are relatively low. Modeling results of hydrocarbon plumes also suggests that BTEX plumes may reach stable or steady state conditions within 3–7 years.

Table 5 Results of two studies on lengths of hydrocarbon plumes in groundwater

Plume length (L)	*Lawrence Livermore (271)*	*Chevron (62)*
$L < 30$ m	50%	47%
30 m $< L <$ 80 m	40%	40%
$L > 80$ m	10%	13%

Table 6 Results of two studies on changes in hydrocarbon plume geometry

	Lawrence Livermore (271)	*Chevron (62)*
Shrinking	59%	51%
No significant trend	33%	40%
Increasing	8%	9%

Transport in Fractures

Fractures such as joints, faults, or bedding-plane separations can occur in various geologic strata. The significance of a release of hydrocarbons into a fracture system is (a) the velocities can be greater than for an unconsolidated media, and (b) the distribution of the hydrocarbons within the fracture system becomes difficult to define.

The flow of hydrocarbons in a fracture network is controlled by a number of factors, including the fracture aperture (i.e., the distance between the fracture walls), fracture density (i.e., spacing between the fractures), and fracture length and orientation. The inability to easily measure fracture apertures is primarily due to their size (10–100 μm) and the interferences caused from drilling or excavation stress-relief. As a result, aperture values are usually estimated using the 'cubic law.' The cubic law is based on the relationship between hydraulic gradient and the discharge for a viscous fluid flowing between two parallel smooth surfaces. In reality, apertures can vary greatly in size and wall roughness; these are not easily measured.

The transport of a dissolved-phase hydrocarbon in a fractured porous medium occurs by convection through the fractures, advection into the fracture wall matrix, and diffusion into the fracture walls (assuming that the material is porous). In most cases, advection into the wall is negligible; diffusion controls solute transport from a rapidly moving liquid through the fractures into the relatively immobile pore water of the wall matrix.

The process of matrix diffusion can theoretically retard the overall migration of a hydrocarbon by several orders of magnitude relative to transport within the fracture system. For example, modeling simulations show that, in a scenario, that included matrix diffusion, chloride migrating through a 25-μm aperture fracture in a glacial clay would take about 50 years to reach the underlying aquifer present at a depth of 10 m. For the same modeling scenario, but with no matrix diffusion, the solute would reach the aquifer within about half a day.

For NAPLs moving vertically through a fracture system, the transport is rapid after the release; diffusion has little impact as the aperture size and hydrocarbon properties (density, interfacial tension, and viscosity) impacts the rate of flow. Once the hydrocarbon becomes immobilized within the fracture system, diffusion into the porous matrix occurs. The diffusion of the hydrocarbon into the porous matrix significantly complicates the remediation of the hydrocarbons present within the porous media.

Remediation of Petroleum Hydrocarbon-Impacted Soil and Groundwater

Physical Separation

There are three significant physical separation processes that are utilized for petroleum hydrocarbon-impacted soil. Each of these remedial technologies uses equilibrium driving forces as the primary separation mechanism.

Soil vapor extraction Soil vapor extraction (SVE) is a vacuum extraction process. It utilizes the volatility of the petroleum hydrocarbons as the primary method of remediation. SVE works by utilizing vapor pressure differences between petroleum and water or petroleum and soil. The process is generally utilized to remove the more volatile petroleum products such as gasoline and aviation fuel. It can also be used to reduce the mobility and hazard of diesel, kerosene, and some heating oils by removing the more volatile fractions from the soil. SVE is usually limited to chemicals with a boiling point less than 250°C and high vapor pressure (greater than 0.5 mmHg). Chemicals with a Henry Law constant greater than 100 atmospheres tend to be removable from water using SVE.

During the SVE process, slotted wells are placed in the vadose zone in or around the contaminated soil. The well is sealed at the surface and connected to a vacuum pump. A vacuum is applied to the slotted wells, which causes a subsurface airflow to pass over the adsorbed soil–petroleum matrix. Utilizing vapor-phase equilibrium as the driving force, volatile components migrate from the liquid phase into the vapor phase and then into the slotted wells. Once these components reach the slotted wells, they are removed via the vacuum extraction system.

If the soil is initially wet, it is usually desiccated before there is significant recovery of the petroleum constituents. Low soil permeability and high soil-moisture content inhibit airflow and therefore inhibit petroleum constituent recovery. In general, the efficacy of SVE is enhanced whenever the temperature of the constituents or soil can be increased. Techniques that provide this increased temperature include microwaves, heated and injected air, and injection of steam.

Soil-washing (solvent extraction) Soil-washing is a solvent-extraction process. It uses the difference in the solubility of petroleum and the soil in water as the driving force. First, contaminated soils are usually excavated (although *in situ* soil flushing systems do exist) and the soil screened to remove large rocks and materials. If there are large amounts of clay or other very fine soil particles, soil-washing can be limited by the ability of the soil particle to settle after the washing/extraction has been accomplished.

Second, a surfactant is typically added to the petroleum-contaminated soil. Early surfactants were simple detergents. These detergents work by attaching a hydrophobic part of the detergent molecule to the petroleum while the hydrophilic part of the detergent molecule associates with water molecules. This association makes the petroleum-contaminated soil particle soluble in water.

When this principle is applied to soil-washing, the use of detergent-type surfactants creates an oil–water emulsion that must be treated, and the suspended soil must be separated from the solution. A major impediment to using soil-washing as a remediation technique has been the inability of this process to remediate fine-particle substrates and to treat the wash water in a cost-effective way.

Newer surfactants have been developed that form a separate phase that is easily separated from the soil-water phase. If there are many fine (e.g., clay) particles in the soil, mechanical separation of the soil from the liquid phase may be instituted in order to remove the water from the soil.

Some soil-washing units utilize extracting solvents such as triethylamine, carbon dioxide, or propane. For heavy oil recovery, water with an added surfactant is the most common processing solvent. Some new surfactants also contain many of the base nutrients necessary to support microbiological activity.

Thermal desorption Thermal desorption is a process that heats contaminated soils with an external heat source such as heated air flow or with an internal heat source such as an oil-heated screw feeder. The major difference between thermal desorption and incineration is that thermal desorption either captures or combusts volatilized chemicals emanating from the soil.

Thermal desorbers utilize several different designs. Rotary drum driers utilize either co-current or countercurrent heated airflow. Thermal screw designs use hollow screw feeders that usually contain hot oil as the heat source; the thermal screw design has one inherent advantage, that it can maintain a low-oxygen atmosphere in the chamber. Conveyor furnaces move the soil through a heated chamber on a metal belt. It is similar to the rotary drier in principle.

Thermal desorption systems usually operate at 200°–300°C, but some are operated at temperatures as high as 600°C. In order to determine the temperature at which a thermal desorber is to be operated, the boiling point range of the treated petroleum mixtures

should be considered. Gasoline volatilizes at 40°–225°C, kerosene at 180°–300°C, diesel at 200°–340°C; lubricating oil and crude oil have constituents that volatilize at temperatures greater than 400°C. However, the amount of energy that must be added to the desorber to volatilize the hydrocarbons in the soils is also dependent in many systems on soil particle size and moisture content.

Biochemical Reactions

Several remediation technologies are built around biodegradation processes in which, ideally, the final products are CO_2 and water. The following subsections discuss common remediation techniques that utilize biochemical reactions. It is important to note that these techniques, like every successfully engineered bioremediation process, must create an environment where there are sufficient oxygen and nutrients and where the buildup of waste products such as carbon dioxide, water, and dead cells can be controlled.

Natural attenuation and biopiles Natural attenuation utilizes the natural soil environment to decontaminate a petroleum hydrocarbon-impacted soil. It is popular, because it is one of the most cost-efficient technologies available. However, it may not be the most time-effective approach available.

'Biopiles' refers to the controlled biodegradation of the organic fraction of a solid waste pile and is very similar to compost piles and sludge stabilization piles. Piles of petroleum hydrocarbon-impacted soils are amended with nutrients and water to stimulate microbial activity and the subsequent decomposition of hydrocarbons. This approach has been utilized for centuries in Asia and is used in developed countries to degrade many different types of waste.

Monitored natural attenuation (MNA) and biopiles seem to have lost appeal since 1995 as a technology for remediating contaminated soils. Several barriers to biopile use include a lack of confidence in the technology, costs, lack of understanding by states, and extended cleanup time.

Land farming This treatment technology is commonly known as land treatment, land cultivation, land spreading, land disposal, land application, refuse farming, and sludge farming. The primary objective of land farming is the degradation of a chemical or waste residing in topsoil by soil microorganisms. During land farming, hydrocarbon-impacted soil amended with nutrients and/or microorganisms is spread in thin layers on a field. Over time, microbial processes degrade the hydrocarbons.

Landfills Despite pollution often associated with unlined landfills, those which are lined and well managed are basically huge biopiles supporting the same biochemical processes that occur in small-scale biopiles and land farming.

Physicochemical Reactions

Some remediation techniques utilize a combination of physical as well as chemical reactions to remediate soils, thereby earning the name 'physicochemical reaction processes.' Some of the more common techniques are discussed below.

Solidification During the solidification remediation technique, a solidifying agent such as Portland cement or a calcium-rich pozzolanic material such as coal fly ash is mixed with a contaminated soil. A chemical process creates a new, hardened solid encapsulating the soil and rendering petroleum hydrocarbons immobile. Solidification is usually chosen when it is necessary to decrease the leachability (i.e., mobility) of the contaminants on the soil.

Asphalt incorporation Asphalt incorporation has been used for many years to recycle petroleum-contaminated soils. Contaminated soil is transported to a hot-batch asphalt plant. The volatile chemicals are vaporized in the initial combustion chamber. The lighter ends are volatilized in the asphalt dryer (200°–250°C) and are subsequently combusted. The heavy ends remain in the asphalt. The soil contaminated with the heavy ends makes a very good aggregate replacement for clean fine aggregate if the soil particle size is appropriate.

The production of asphalt from the contaminated soil requires a complete investigation of the properties of the petroleum-contaminated soil, the percentage of petroleum in the mix, and the quality of the petroleum. A special treatment and testing program may be needed to produce and verify a quality product meeting acceptable standards and specifications, such as ASTM specifications.

The production of asphalt concrete with petroleum-contaminated soil is considered a practical and economical procedure that produces a useful and marketable product inexpensively, and it cleans the environment. The final product is used as a construction material for secondary roads, roadbeds, road subbase, impermeable layers for landfill and containment facilities, or stabilizers for steep embankments.

Incineration Contaminated soil is burned in a suitable incinerator at temperatures of 600°–1000°C. Air emissions are generally required to be captured and

treated if the soil contains significant quantities of chlorine or volatile metals such as mercury.

Vitrification Vitrification uses large quantities of electrical energy to physically melt the contaminated soil and to oxidize any petroleum hydrocarbons in the soil. The resulting melt vitrifies into a glass-like mineral that encapsulates any metals and other chemicals contained in the melt for hundreds or thousands of years.

Summary

Contamination of soils and of groundwater by hydrocarbons appears to be a more widespread and complex problem than previously recognized. The presence of various additives in petroleum products (including lead, Pb, as well as certain persistent organic compounds) further exacerbates the problem. Better techniques are needed for identifying, delineating, and remediating the problem in sites where it already exists (as a consequence of past practices or of mishaps such as oil spills or leaky fuel tanks), and – even more importantly – for prevention of the problem from occuring at new sites in the future.

See also: **Pollutants:** Biodegradation; Effects on Microorganisms; Persistent Organic (POPs); **Remediation of Polluted Soils**

Further Reading

American Petroleum Institute (1972) The migration of petroleum products in soil and groundwater. *Principles and Counter Measures*. Washington, DC: API.

American Society for Testing and Materials (1996) *Standard Guide for Remediation by Natural Attenuation at Petroleum Release Sites*. (Draft.) ASTM.

Bruce L and Schmidt G (1994) Hydrocarbon fingerprinting for application in forensic geology: review with case studies. *American Association of Petroleum Geologists Bulletin* 78(11): 1692–1710.

Calabrese EJ and Kostecki PT (1992) *Risk Assessment and Environmental Fate Methodologies*. Chelsea, MI: Lewis Publishers.

Coleman W, Munch J, Stricher R, Ringhand H, and Kopfler F (1984) The identification and measurement of components in gasoline, kerosene, and No. 2 fuel oil that partition into the aqueous phase after mixing. *Archives of Environmental Contamination and Toxicology* 13: 171–178.

Domenico P (1987) An analytical model for multidimensional transport of a decaying contaminant species. *Journal of Hydrology* 91: 49–58.

Dragun J (1991) Geochemistry and soil chemistry reactions occurring during *in situ* vitrification. *Journal of Hazardous Materials* 26: 343–364.

Dragun J (1998) *The Soil Chemistry of Hazardous Materials*, 2nd edn. Amherst, MA: Amherst Scientific Publishers.

Hillel D (1988) Chapter 7 – Movement and Retention of Organics in Soil: A Review and a Critique of Modeling. *Petroleum Contaminated Soils. vol. I. Remediation Techniques, Environmental Fates, Risk Assessment*. In: Kostecki P and Calabrese E (Eds) Chelsea, MI: Lewis Publishers.

Johnson P, Kemblowski M, and Colthart J (1990) Quantitative analysis for the cleanup of hydrocarbon-contaminated soils by *in-situ* soil venting. *Ground Water* 28: 413–429.

Kaplan I, Galperin Y, Lu S, and Lee R (1997) Forensic environmental geochemistry: differentiation of fuel-types, their sources and release time. *Organic Geochemistry* 27(5/6): 289–317.

Kueper B and McWhorter D (1991) The behavior of dense, non-aqueous phase liquids in fractured clay and rock. *Ground Water* 29: 716–728.

Kung S (1990) Preferential flow in a sandy vadose zone: field observation. *Geoderma* 46: 51–58.

Kung S (1990) Preferential flow in a sandy vadose zone. 1. mechanism and implications. *Geoderma* 46: 59–71.

Morrison R (2000) Application of environmental forensic techniques for age dating and source identification of a contaminant release. *International Journal of Environmental Forensics* 1(2): 131–153.

Murphy B and Morrison R (2002) *Introduction to Environmental Forensics*. Oxford, UK: Academic Press/Elsevier Science.

Ryan R and Dhir V (1993) The effect of soil-particle size on hydrocarbon entrapment near a dynamic water table. *Journal of Soil Contamination* 2(1): 59–92.

Hydrodynamic Dispersion *See* Solute Transport

HYDRODYNAMICS IN SOILS

T P A Ferré and A W Warrick, University of Arizona, Tucson, AZ, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

'Hydrodynamics' means 'the movement of water.' In contrast to hydrostatics, hydrodynamics considers the response of a fluid to an imposed energy gradient. The subfield of soil-water dynamics can be further subdivided into flow through a saturated or unsaturated medium, and steady-state or transient processes. Consideration of a saturated medium applies to groundwater flow problems. Flow through the vadose zone requires study of saturated and unsaturated flow through variably saturated media. Both saturated and unsaturated flow problems may be steady-state or transient. Soil-water dynamics is so central to topics involving the movement of water, heat, and solutes through the subsurface that it is fundamental to most, if not all, soil physical processes and groundwater flow problems.

Steady-State Flow Through a Saturated Medium

The Darcy equation describes the response of a fluid in a porous medium to an imposed energy gradient and forms the foundation of discussions of hydrodynamics. The one-dimensional form of the Darcy equation can be written as:

$$Q = -A\frac{k\rho g}{\mu}\frac{\Delta(z+\psi)}{\Delta L} \qquad [1]$$

where Q is the flow (L^3/T), A is the cross-sectional area perpendicular to flow (L^2), k is the intrinsic permeability (L^2), ρ is the fluid density (M/L^3), g is the acceleration due to gravity (L/T^2), μ is the dynamic viscosity of the fluid (M/LT), ψ is the pressure head (L), z is the elevation (L), and L is the distance in the direction of flow (L). The negative sign is necessary due to the definition of a gradient to define flow as occurring from a region of higher energy to lower energy. This can be rewritten in terms of the flux density, $q(L/T)$ as:

$$q = \frac{k\rho g}{\mu}\frac{\Delta(z+\psi)}{\Delta L} \qquad [2]$$

Flow occurs in the direction of decreasing energy of the fluid, and the rate of flow depends directly upon the ability of the medium to permit flow (permeability), the fluid density, and the acceleration due to gravity, and inversely upon the resistance of the fluid to flow (dynamic viscosity). Introducing the hydraulic conductivity, K (L/T) allows simplification to:

$$q = -K\frac{\Delta(z+\psi)}{\Delta L} \qquad [3]$$

For flow in the horizontal, x-direction, the flux density equals:

$$q_x = -K\frac{\Delta\psi}{\Delta x} \qquad [4]$$

The flux density for flow in the vertical, z-direction, is:

$$q_z = -K\left(1+\frac{\Delta\psi}{\Delta z}\right) \qquad [5]$$

Furthermore, if the medium is homogeneous and flow is steady-state, then the gradients of the hydraulic and pressure heads must be constant during one-dimensional flow. That is, the pressure head and hydraulic head will vary linearly with distance in the direction of flow. This is shown for a 200-cm-long, homogeneous horizontal column for which the left boundary is maintained at a constant pressure head of 70 cm and the right boundary is held at a constant pressure head of 0 cm (Figure 1).

The two-dimensional form of the Darcy equation follows directly from the one-dimensional form. For instance, two-dimensional flow in the x-z plane is described as:

$$q_i = -K_{i,j}\frac{\Delta(z+\psi)}{\Delta x_j} \qquad [6]$$

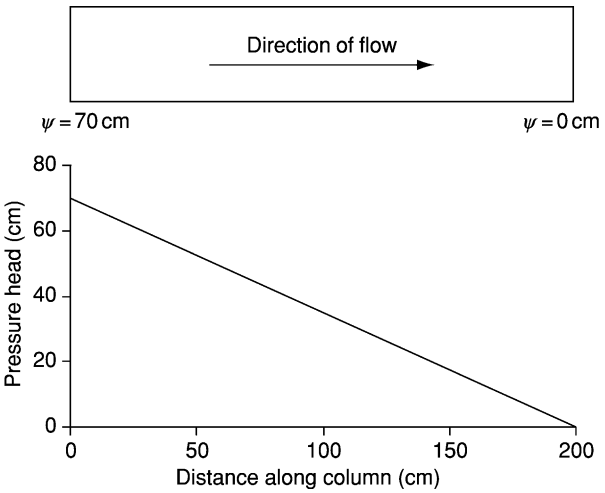


Figure 1 Pressure-head profiles along a horizontal column during steady-state saturated flow.

The subscripts i and j represent either x or z in this case. The hydraulic conductivity matrix, K_{ij}, allows for flow in the x-direction due to gradients in the x-direction ($K_{x,x}$) and gradients in the z-direction ($K_{x,z}$). Furthermore, the convention is followed that a sum is taken over the index that is included twice on the right-hand side of the equation. For example the flux density in the x-direction is:

$$q_x = -K_{x,x}\frac{\Delta(z+\psi)}{\Delta x} - K_{x,z}\frac{\Delta(z+\psi)}{\Delta z} \qquad [7]$$

The flux expression can be written in terms of partial derivatives as:

$$q_i = -K_{i,j}\frac{\partial(z+\psi)}{\partial x_j} \qquad [8]$$

While the modifications of the equations to describe multidimensional flow are straightforward, the results are not trivial. Specifically, for one-dimensional geometries, flow is constrained to a single direction. For multidimensional cases, water will tend to flow preferentially in the direction of highest permeability. This phenomenon is referred to as 'anisotropic flow' if it arises due to a directional dependence of the hydraulic conductivity (e.g., along or across bedding planes).

Transient Flow Through a Saturated Medium

If the energy gradient applied to the fluid in a medium is constant in time, the flow through the medium will achieve a steady-state condition given enough time. When this state is reached, the total volume of water entering any region of the porous medium in a period of time will be exactly equal to the total volume of water leaving that region. The region can be defined as a point for one-dimensional flow or as an area or a volume for two- and three-dimensional flow, respectively. Before steady state is reached, the medium responds in a transient manner to the changes in the imposed energy gradient. The mechanism by which the medium responds is through a change in the volume of water stored within the medium.

Consider the three-dimensional region shown in Figure 2. Each edge has a constant length, ΔL. Flow is two-dimensional in the x and z directions. At steady state the sum of q_x in and q_z in will be equal and opposite to the sum of q_x out and q_z out. However, under transient flow conditions, these sums will not be equal. The net difference in the inflow and the outflow will equal the change in the volume of water stored, ΔV_w, per unit time elapsed, Δt, per unit area of the region. This is written as a mass balance equation:

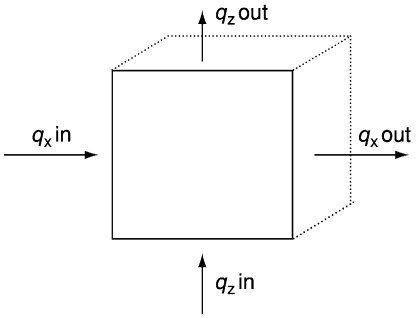


Figure 2 Three-dimensional unit cell for consideration of conservation of mass. The flux density into (q in) and out of (q out) the cell is shown in the x and z directions.

$$(q_x\,\text{in} + q_z\,\text{in})(\Delta L\Delta L) - (q_x\,\text{out} + q_z\,\text{out})(\Delta L\Delta L) = \frac{\Delta V_w}{\Delta t} \qquad [9]$$

For the fluxes shown in Figure 2, the net flux in the positive x direction can be defined as $\Delta q_x = q_x\,\text{out} - q_x\,\text{in}$. Substituting for the net fluxes and dividing by the areas of the sides gives:

$$-\Delta q_x - \Delta q_z = \frac{\Delta V_w}{\Delta t \Delta L^2} \qquad [10]$$

Dividing both sides by $-\Delta L$ gives:

$$\frac{\Delta q_x}{\Delta L} + \frac{\Delta q_z}{\Delta L} = -\frac{\Delta V_w}{\Delta t \Delta L^3} \qquad [11]$$

By considering infinitesimally small regions and infinitesimally small time steps, ΔL and Δt approach zero and the mass balance equation can be written in partial differential form:

$$\frac{\partial q_i}{\partial x_i} = -\frac{\partial}{\partial t}\left(\frac{\Delta V_w}{\Delta L^3}\right) \qquad [12]$$

where x_i is x and z. Combined with the Darcy equation, this gives:

$$\frac{\partial}{\partial x_i}\left(K_{i,j}\frac{\partial(z+\psi)}{\partial x_j}\right) = \frac{\partial}{\partial t}\left(\frac{\Delta V_w}{\Delta L^3}\right) \qquad [13]$$

This can be extended to three dimensions by allowing i to correspond to x, y, and z.

Saturated water flow is caused, in part, by changes in the pressure head of the fluid. Therefore, it is convenient to describe the changes in the volume of stored water per unit area in terms of the response of the medium to changes in pressure head as well. The specific storage, S_s, is defined as the volume of water released from storage per unit volume of medium per unit decrease in the pressure head of the water:

$$S_s = \left(\frac{\Delta V_w}{\Delta\psi \Delta L^3}\right) \quad [14]$$

The specific storage is a lumped term that describes the results of two physical phenomena: medium compaction and water expansion. Water that resides in a medium and is under a positive pressure head acts to expand the matrix of the medium. If the water pressure is reduced, the effective stress on the medium will increase and the medium will compact. By compacting, the porosity of the medium is reduced as water is released from storage. Simultaneously, the decreased pressure of the water allows the water itself to expand, thereby occupying more volume per unit mass of water. This is also accompanied by a release of water from storage. Of the two mechanisms, the first is dominant due to the very low compressibility of water.

Substituting the specific storage into the mass balance equation gives:

$$\frac{\partial}{\partial x_i}\left(K_{i,j}\frac{\partial\psi}{\partial x_j}\right) = S_s\frac{\partial\psi}{\partial t} \quad [15]$$

This form of the mass balance equation conveniently describes the transient response of a porous medium to changes in the applied pressure heads. Physically, this demonstrates that a change in pressure at some location within a medium will take a finite time to propagate to a distant point in the medium. This delay is caused by changes in water storage (either increases in response to an increased pressure head or decreases due to a decreased pressure head) that are caused at each point past which the pressure change is transmitted. The time that it will take for the pressure pulse to be transmitted increases with an increase in the value of the specific storage of the medium. This time also increases with a decrease in the hydraulic conductivity of the medium, because this leads to slower flow of water through the medium in response to an imposed energy gradient.

Steady-State Flow Through an Unsaturated Medium

The Darcy equation is equally applicable to steady-state flow through both saturated and unsaturated media. However, there is a critical difference from the preceding analysis when this relationship is applied to unsaturated flow. Specifically, the hydraulic conductivity (or intrinsic permeability) of the medium is no longer independent of the pressure head of the water in the medium. As the pressure head of the water decreases below the air entry pressure of the medium, the volumetric water content of the medium (θ), defined as the volume of water per unit volume of medium, also decreases. As a result, some of the pore spaces are no longer water-filled and, as such, are not available for water flow. This reduction in the water-filled porosity of the medium reduces the ability of the medium to transmit water under an implied energy gradient. This can be described as a decrease in the hydraulic conductivity with decreasing water content. The equation describing unsaturated flow is deceptively simple:

$$q_i = -K_{i,j}(\theta)\frac{\partial(z+\psi)}{\partial x_j} \quad [16]$$

The ramifications of the dependence of the hydraulic conductivity on the volumetric water content can be quite dramatic. For example, consider the case of steady-state, one-dimensional, horizontal flow through a homogeneous medium. The pressure-head profile is shown in Figure 3 for a 200-cm-long, sand-filled column with the left boundary held at a constant pressure head of 0 cm and the right boundary held at a constant pressure head of −70 cm. Notice that, unlike the case of steady-state one-dimensional flow through a saturated medium, the pressure head varies nonlinearly in the direction of flow. This occurs because the hydraulic conductivity decreases steadily with decreasing pressure head from the saturated value at the left boundary to a much lower value at the right boundary. To maintain steady-state flow throughout the column, the magnitude of the gradient must increase to compensate exactly for this decreasing hydraulic conductivity. Therein lies the difficulty of calculating the pressure-head distribution during unsaturated flow. Even under the apparently simple conditions of steady-state, one-dimensional flow, the pressure-head distribution depends on the change in the hydraulic-head

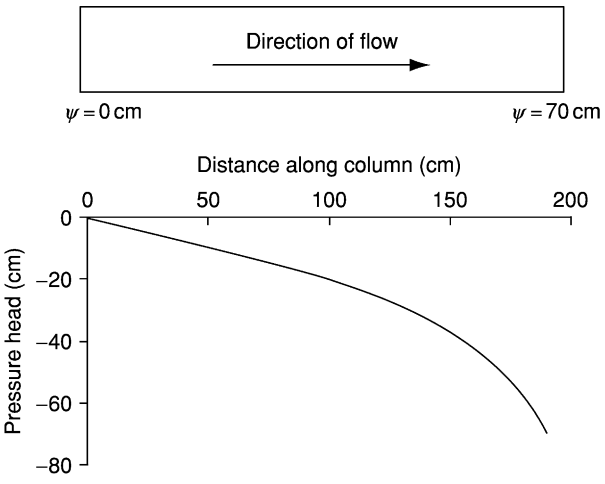


Figure 3 Pressure-head profiles along a horizontal column during steady-state unsaturated flow.

gradient through the system. The gradient depends upon the hydraulic conductivity distribution. However, the hydraulic conductivity distribution depends upon the water content distribution, which, in turn, depends upon the pressure-head distribution. Equations that describe such a circular dependence of the hydraulic parameters on the energy state and of the energy state on the hydraulic parameters are nonlinear.

An alternative approach to describing unsaturated flow in the absence of gravity is to consider the matric flux potential, ϕ, which is defined as:

$$\phi = \int_{-\infty}^{\psi} K(h)\mathrm{d}h \qquad [17]$$

The derivative of the matric flux potential is related to the flux as:

$$-\frac{\mathrm{d}\phi}{\mathrm{d}x} = q_x \qquad [18]$$

Given that the flux is constant throughout the column under steady-state flow, the derivative of the matric flux potential will also be constant with x. That is, ϕ will vary linearly with x. The value of ϕ at any position will depend on the values of pressure head at the ends of the column and on the conductivity function $K(\psi)$.

Transient Flow Through an Unsaturated Medium

Transient unsaturated flow is fundamentally different than saturated flow and steady-state unsaturated flow. During transient unsaturated flow, water enters pores that were previously occupied by another fluid. Commonly this fluid is air; more commonly, it is assumed that the displacement of the resident air does not impede the advance of water into a pore. This is the underlying assumption of the Richards equation, a special expression of mass conservation that describes the movement of water through an unsaturated porous medium. The Richards equation can be written in many forms. One form that allows ready comparison with the mass balance equation developed for saturated flow is written:

$$\frac{\partial}{\partial x_i}\left(K_{i,j}(\theta)\frac{\partial(z+\psi)}{\partial x_j}\right) = \frac{\partial\theta}{\partial t} \qquad [19]$$

Direct comparison with the mass balance equation developed for transient, saturated flow shows two differences. As for steady-state unsaturated flow, the hydraulic conductivity is now a function of the water content. This leads to the same nonlinear behavior seen above. In addition, the storage term is now written as the change in the unitless volumetric water content, θ, per unit time. The use of the change in water content alone to describe storage changes suggests that the mechanisms that allow for changes in water storage in a saturated medium do not apply in an unsaturated medium. In fact, medium compaction and water expansion do occur under conditions of negative pressure head in the water phase. However, the amount of storage change that can occur owing to changes in water content are far larger than those that can occur owing to both of these mechanisms combined. As a result, they can commonly be ignored for unsaturated conditions to allow for more convenient solutions.

Transient unsaturated flow differs from transient saturated flow as a result of the nonlinearity of the Richards equation, through the dependence of the hydraulic conductivity on the water content and of the large potential for water storage changes through changes in the volumetric water content. The first of these effects gives rise to relatively slow movement of water under a given energy gradient in an unsaturated medium compared with the response to the same gradient under saturated conditions. A secondary result is the formation of sharp wetting fronts. (*See* **Infiltration**.) The second of these effects results in even slower propagation of pressure changes through an unsaturated medium due to the large capacity for changes in water storage in response to pressure changes. The validity of the Richards equation and of pressure–saturation relationships determined under static conditions for describing transient flow was still a topic of research at the beginning of the twenty-first century.

See also: **Darcy's Law**; **Diffusion**; **Heat and Moisture Transport**; **Infiltration**

Further Reading

Darcy H (1856) *Les Fontaines Publique de la Ville de Dijon*. Paris, France: Dalmont.

Gardner WR (1958) Some steady state solutions of the unsaturated moisture flow equation with application to evaporation from a water table. *Soil Science* 85(4): 228–232.

Richards LA (1931) Capillary conduction of liquid through porous media. *Physics* 1: 318–333.

HYSTERESIS

J H Dane, Auburn University, Auburn, AL, USA
R J Lenhard, INEEL, Idaho Falls, ID, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

When studying fluid flow in porous media such as soils, we can analyze fluid behavior at the molecular, the microscopic, or the macroscopic level. At the microscopic level, we assume that only one phase is present at a given physical point P and at a given time, i.e., at P only the solid phase or one of the fluids is present. This is in contrast to the macroscopic level, where we assume that all phases can be present at the same physical point and at the same time (principle of superposition). At both the microscopic and the macroscopic level, all variables are considered to be continuous in space and time, so that they are differentiable. Fluids present in the soil are wetting fluids, nonwetting fluids, or of intermediate nature. Although this discussion can be extended to any set of fluids, for the sake of simplicity only the presence of water (wetting fluid) and/or air (nonwetting fluid) is considered here. When only water occupies the pore spaces, the soil is said to be completely saturated. When only air is present, the soil is said to be completely dry. In most cases, both air and water are present, and a relationship exists between the energy status of the soil water and the volumetric water content. The governing potential that determines the water content is the matric potential, ϕ_m (energy per unit volume, with units: joules per cubic meter, or newtons per meter squared, or pascals; energy per unit mass, with units joules per kilograms: or energy per unit weight, with units joules per newton or meters). It consists of three components that bind the water to the solid phase:

$$\phi_m = \phi_p + \phi_a + \phi_\omega \qquad [1]$$

where ϕ_p, ϕ_a, and ϕ_ω refer to the pressure potential, the adsorptive potential, and the osmotic potential, respectively, all at the microscopic level. The osmotic forces are due to concentration differences between the bulk solution and the diffuse double layer. The adsorptive forces are forces of attraction between the water molecules and the solid phase and operate at very short distances only. Assuming the osmotic forces to be part of the adsorptive forces, eqn [1] simplifies to:

$$\phi_m = \phi_p + \phi_a \qquad [2]$$

When the water is at static equilibrium, the total mechanical potential, ϕ, is:

$$\phi = \phi_g + \phi_m \qquad [3]$$

where the gravitational potential, ϕ_g, is constant when considering a horizontal cross-sectional area through a pore located at P. By assuming $\phi = 0$ at some reference level, it follows that $\phi = 0$ everywhere. Hence, the matric potential is constant through this horizontal cross section, i.e.,

$$\phi_m = -\phi_g \qquad [4]$$

When sufficient water is present, the value for the matric potential is completely determined by the value of the pressure potential at the water–air interface, i.e., the adsorptive and osmotic forces in the liquid at the interface are negligible. This will be the case as long as water is present so that a concave air–water interface exists. The qualitative potential distributions in a horizontal cross section of a pore are shown in Figure 1. The solid phase is at $x = 0$, while the concave air–water interface is at the dashed line. The question thus arises as to what determines

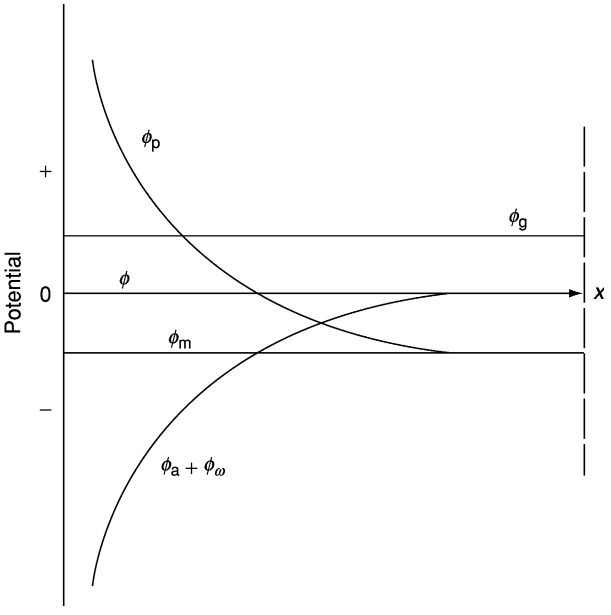


Figure 1 The qualitative potential distributions in a horizontal cross section of a pore inside a soil column with its water at static equilibrium. The pore water has an assumed concave meniscus. The matric potential $\phi_m = \phi_p + \phi_a + \phi_\omega$, where ϕ_p refers to the pressure potential, ϕ_a the adsorption potential, and ϕ_ω the osmotic potential, all at the microscopic level. The latter only exists inside the diffuse double layer. The total mechanical potential is given by $\phi = \phi_m + \phi_g$, where ϕ_g is the gravitational potential. ϕ is assumed to be zero at all elevations.

the pressure in the water just across the air–water interface.

The Capillary Pressure Concept

If more than one fluid is present, such as in an air–water system, the fluids are separated by an interface. If this interface is curved, a pressure jump, ΔP (newtons per meter squared), exists across the interface, which is determined by the Young and Laplace equation:

$$\Delta P = \sigma\left(\frac{1}{R_1} + \frac{1}{R_2}\right) \quad [5]$$

where σ (newtons per meter) is the surface or interfacial tension, and R_1 and R_2 (meters) are the radii defining the curvature. For spherically shaped interfaces these radii of curvature are equal, and eqn [5] can be written as

$$\Delta P = \frac{2\sigma}{R} \quad [6]$$

It is necessary to make conventions about the exact definitions of ΔP and R. ΔP can be defined:

$$\Delta P = P_l - P_g \quad [7]$$

where the subscripts l and g refer to the liquid (water) and gas (air) phase, respectively. It should be noted, however, that the pressure jump across the interface can also be defined as the pressure in the gas phase minus the pressure in the liquid phase. Positive values can be assigned to R if R resides in the liquid phase, and negative values if it resides in the gas phase. It can thus easily be seen that $\Delta P > 0$ for a raindrop in air and $\Delta P < 0$ for water in a (hydrophilic) capillary tube or in a soil pore. In the case of a flat interface, $R \to \infty$ and $\Delta P = 0$. Hence no pressure jump exists across a flat interface.

As long as the air–water interface is concave toward the water, the matric potential in the pore water at a given point is determined by the pressure potential in the water at the air–water interface, which is determined by the water pressure at the air–water interface. When expressed on a mass basis, and assuming the air to be at atmospheric pressure ($P_g = 0$), the matric potential is defined:

$$(\phi_m)_{mass} = (\phi_p)_{mass} = \frac{P_l}{d_l} \quad [8]$$

in joules per kilogram; while the matric potential on a volume basis is defined:

$$(\phi_m)_{volume} = (\phi_p)_{volume} = P_l \quad [9]$$

in newtons per meter squared, or pascals; and the matric potential on a weight basis:

$$(\phi_m)_{weight} = (\phi_p)_{weight} = \frac{P_l}{d_l g} \quad [10]$$

in meters, where d_l is the density of water (kilograms per cubic meter) and g is the gravitational field strength (newtons per kilogram). The matric potential on a weight basis is often referred to as the matric head, h_m. It should be noted that the pressure potential increases, while the adsorptive and osmotic potentials decrease, when approaching the solid phase. The water pressure at the air–water interface, P_l, and consequently the matric potential are determined by eqns [6] and [7]. It should be emphasized that, for capillary water (R in the gas phase), values for P_l, and hence for the matric potential, are negative. Using the alternative definition of ΔP (eqn [7]), i.e., $\Delta P = P_g - P_l$, and assuming $P_g = 0$, results in values of the capillary pressure $P_c = -P_l$ (pascals) or the capillary pressure head $h_c = -P_l/d_l g$ (in meters). The use of P_c or h_c is sometimes preferred, to deal with positive rather than negative values.

Water-Retention Curve

The relationship between matric potential and volumetric water content (θ) is referred to by many names, such as 'water-retention curve,' 'soil-water characteristic curve,' 'pressure head–water-content relationship,' 'water content–matric potential curve,' and 'capillary pressure–saturation curve.' Stepwise reduction of the matric potential of an initially completely saturated soil, i.e., starting at a matric potential value of zero, results in the initial drainage curve (IDC; Figure 2). When no more water can be drained from the soil, the residual water content (θ_r) has been reached. Reversing the process at this point, by increasing stepwise the matric potential until it reaches zero, produces the main wetting curve (MWC; Figure 2). The two curves are noticeably different, which is referred to as 'hysteresis.' Reversing the process again yields the main drainage curve (MDC; Figure 2). Reversal points can start at any point during drainage or wetting and will form scanning curves, which are always inside the loop formed by the MWC and MDC.

All variables in the previous sections have been assumed to occur at the microscopic level, or at least not at a level greater than the pore level. Most of the measurements are, however, obtained at the macroscopic level, as is the case with water-retention curves. Transitions can be made from the microscopic level to the macroscopic level, and it can be shown that, under most conditions, expressions for

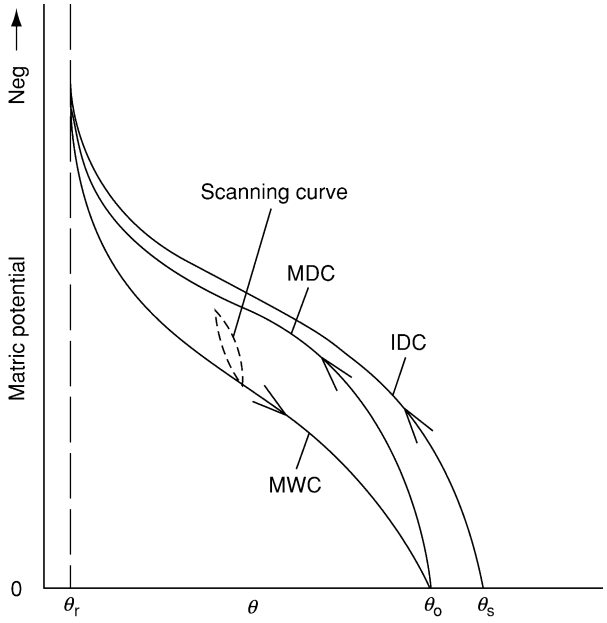


Figure 2 The initial drainage curve (IDC), the main wetting curve (MWC), the main drainage curve (MDC), and a scanning curve. θ refers to the volumetric water content and the subscripts r, o, and s, to residual, satiated, and saturated.

potential have the same form at both levels. Consequently, at the macroscopic level, average pressures within macroscopic volume elements are dealt with.

Causes of Hysteresis

Hysteresis in the water content–matric potential relation is attributed to: (1) the 'ink bottle' effect, (2) the contact angle effect, (3) air entrapment, and (4) deformation of the solid phase.

The ink bottle effect is probably the most important reason for hysteresis to occur and will be explained with the help of the Young and Laplace equation (eqn [6]). Envision a spherically shaped soil pore with two bottlenecks at opposite ends of the pore. The radius at the widest section of the pore is r_w (>0), whereas the radii of the bottlenecks, the smallest parts of the pore, are r_s (>0). One bottleneck is connected to a body of free water in which the pressure can be changed. The pore is small enough so that gravity plays no role in the drainage or imbibition of the pore water. Starting with a fully saturated pore, the water pressure in the body of free water is slowly reduced, and hence in the water in the pore, from zero to some negative value. The pressure in the water at which the pore will drain is determined by:

$$P_{l,d} = \Delta P = -\frac{2\sigma\cos\beta_d}{r_s} \qquad [11]$$

where R in eqn [6] has been replaced by $r_s/\cos\beta_d$ and β_d is the contact angle between the water and the solid phase during drainage. Once the pressure as expressed by eqn [11] has been reached, the pore will abruptly and completely drain. When the pressure in the water is then slowly increased, i.e., from some negative value in the direction of zero, part of the pore will slowly refill. The pore will suddenly completely fill when the pressure in the water has increased to:

$$P_{l,w} = \Delta P = -\frac{2\sigma\cos\beta_w}{r_w} \qquad [12]$$

where R has now been replaced by $r_w/\cos\beta_w$ and β_w is the contact angle during wetting. Assuming the contact angles during wetting and drying not to be too much different, it follows from eqns [11] and [12] that the pressure at which a pore drains is more negative than that at which it refills. Averaging this process over a number of pores, i.e., moving from the microscopic to the macroscopic level, results in data points that show that for the same water content the matric potential will be more negative during drainage than during wetting.

It has been shown that the contact angle between water and the solid phase is greater during imbibition of water than during drainage, i.e., $\beta_w > \beta_d$. According to eqns [11] and [12], this makes the pressure during drainage more negative than during wetting at the same water content, which reinforces the ink bottle effect.

Considerable effort needs to be applied to saturate a soil sample completely. This can be accomplished by first flushing the sample with CO_2, which readily dissolves in water, or by saturating the sample under vacuum. In this manner the IDC (**Figure 2**) can be established. Upon rewetting the soil, however, air cannot always escape and becomes trapped. This causes the water content to be smaller during wetting than during drying at the same value of the matric potential (same radius of curvature at the air–water interface). Because of entrapped air, the soil water content at zero value of the matric potential is often approximately 0.85 times the completely saturated value. The term 'satiated water content' is sometimes used in this respect; and the term 'apparent saturation' is used to account for hysteresis that is due to entrapped air or other nonwetting fluids.

Finally, swelling, shrinking, and aging phenomena may result in different soil structures during wetting and drying. One may argue, however, that the causes for hysteresis should strictly be explained by assuming the porous medium to be rigid. Changes in the soil structure obviously change the water-retention

curve, as demonstrated when solutions of differing composition and concentration are used.

Modeling

It is obvious from Figure 2 that considerable differences may occur in water content values during drying and wetting if compared at the same value of the matric potential. As the ultimate goal is to predict water content values in the field by solving the general flow or the Richards equation for a set of initial and boundary conditions, it is important to account for hysteresis. Although the model of Dane and Wierenga shows this importance, the generated scanning curves do not form closed loops, and as such are not necessarily unique. The best available hysteresis model for air–water relations is probably Lenhard's.

The model by Lenhard *et al.* is the air–water version of the two- and three-fluid model proposed by Parker and Lenhard. The concept of apparent saturation ($\bar{\bar{S}}$) is used to model the effects of hysteresis on subsurface fluid behavior. Apparent saturation applies to fluids that are more wetting than the nonwetting fluid. Hence, for air–water fluid systems in porous media, apparent saturation pertains to water but not to air. The apparent water saturation ($\bar{\bar{S}}_w$) of an air–water system is defined as the summation of the effective water saturation ($\bar{S}_w$), defined:

$$\bar{S}_w = \frac{S_w - S_r}{1 - S_r} \quad [13]$$

and the effective entrapped-air saturation ($\bar{S}_{at}$), defined:

$$\bar{S}_{at} = \frac{S_{at}}{1 - S_r} \quad [14]$$

Consequently:

$$\bar{\bar{S}}_w = \frac{S_w + S_{at} - S_r}{1 - S_r} \quad [15]$$

where $S_w = \theta/\epsilon$ is the water saturation, $S_r = \theta_r/\epsilon$ is the residual saturation, and $S_{at} = \theta_{at}/\epsilon$ is the entrapped air saturation; θ is the volumetric water content, θ_r is the residual water content, θ_{at} is the volumetric entrapped air content, and ϵ is the porosity.

To describe relations between fluid content and capillary pressure that account for hysteresis, two water-rentention curves are parameterized: one being the IDC (Figure 2) and the other being the MWC (Figure 2). It should be noted that these two curves (Figure 2) first need to be converted from θ (P_c) to $\bar{\bar{S}}_w$ (P_c) relations. For the IDC this poses no problem, because no entrapped air exists. Therefore, the apparent saturation is equal to the effective saturation. The IDC is described by the van Genuchten retention function:

$${}^{D}\bar{\bar{S}}_w = [1 + (\alpha_d P_c)^n]^{-m} \quad [16]$$

where ${}^{D}\bar{\bar{S}}_w$ is the apparent water saturation of the IDC, α_d is a curve-fitting parameter for the IDC possessing the same units as $1/P_c$, P_c is the capillary pressure, n is a fitted shape parameter, and $m = 1 - 1/n$. The MWC is similarly described:

$${}^{I}\bar{\bar{S}}_w = [1 + (\alpha_w P_c)^n]^{-m} \quad [17]$$

where ${}^{I}\bar{\bar{S}}_w$ is the apparent water saturation of the MWC (I is 'imbibition' or 'wetting') and α_w is a curve-fitting parameter for the MWC possessing the same units as $1/P_c$. The fitted parameter n is identical for the IDC and the MWC. The starting point for the MWC is the residual water content, while the end point is at a capillary pressure of zero. The conversion from ${}^{I}\bar{\bar{S}}_w$ to θ is accomplished by determining $\bar{S}_{at}$ at the ${}^{I}\bar{\bar{S}}_w$-value of interest. Once $\bar{S}_{at}$ is determined, then θ can be calculated from eqn [15]. Lenhard *et al.* modeled $\bar{S}_{at}$ to vary linearly from $\bar{S}_{at} = 0$ at the saturation-path reversal from the IDC (${}^{\Delta}\bar{S}_w$) to a maximum value at ${}^{I}\bar{\bar{S}}_w = 1$. Because the MWC is under consideration, the reversal point is at θ_r and $\bar{\bar{S}}_w = \bar{S}_w = {}^{\Delta}\bar{S}_w = 0$, i.e., there is no entrapped air at this point. When reversing from the IDC at any point other than θ_r, ${}^{\Delta}\bar{S}_w$ will be different from zero. The maximum value of $\bar{S}_{at}$ is a function of ${}^{\Delta}\bar{S}_w$ and is determined using the method of Land:

$$\bar{S}_{ar} = \frac{1 - {}^{\Delta}\bar{S}_w}{1 + R_{aw}(1 - {}^{\Delta}\bar{S}_w)} \quad [18]$$

in which

$$R_{aw} = \frac{1}{{}^{I}\bar{S}_{ar}} - 1 \quad [19]$$

where $\bar{S}_{ar}$ is the maximum entrapped air saturation for a given wetting path and ${}^{I}\bar{S}_{ar}$ is the maximum entrapped air saturation of the MWC. For an intermediate point along any wetting path, $\bar{S}_{at}$ is determined from:

$$\bar{S}_{at} = \bar{S}_{ar}\left(\frac{\bar{\bar{S}}_w - {}^{\Delta}\bar{S}_w}{1 - {}^{\Delta}\bar{S}_w}\right) \quad [20]$$

Note that $\bar{\bar{S}}_w$ will always be greater than or equal to ${}^{\Delta}\bar{S}_w$. If $\bar{\bar{S}}_w$ is equal to ${}^{\Delta}\bar{S}_w$, then the IDC is being followed. For any intermediate point on the MWC, eqn [20] reduces to:

$$\bar{S}_{at} = {}^{I}\bar{S}_{ar}\,{}^{I}\bar{\bar{S}}_w \quad [21]$$

because ${}^{\Delta}\bar{S}_w = 0$ at the residual water content. Using eqns [18–20], $\bar{S}_{at}$ can be decoupled from $\bar{\bar{S}}_w$, or ${}^{I}\bar{\bar{S}}_w$, to calculate θ.

To obtain drying and wetting scanning curves, the IDC (eqn [16]) and MWC (eqn [17]) are scaled through two pertinent reversal points. For drying scanning paths, the IDC is scaled through the saturation path reversal from the previous wetting path to the current drying path and the preceding drying-to-wetting reversal. The equation for predicting drying scanning paths is:

$$\bar{\bar{S}}_w(P_c) = \left\{ \frac{\left[{}^{D}\bar{\bar{S}}_w(P_c) - {}^{D}\bar{\bar{S}}_w({}^{DI}P_c)\right]\left[{}^{ID}\bar{\bar{S}}_w - {}^{DI}\bar{\bar{S}}_w\right]}{\left[{}^{D}\bar{\bar{S}}_w({}^{ID}P_c) - {}^{D}\bar{\bar{S}}_w({}^{DI}P_c)\right]} \right\} + {}^{DI}\bar{\bar{S}}_w \qquad [22]$$

in which ${}^{D}\bar{\bar{S}}_w({}^{DI}P_c)$ is the apparent water saturation of the IDC at the capillary pressure of the last drying-to-wetting reversal point (${}^{DI}P_c$), ${}^{D}\bar{\bar{S}}_w({}^{ID}P_c)$ is the apparent water saturation of the IDC at the capillary pressure of the last wetting-to-drying reversal point (${}^{ID}P_c$), and ${}^{DI}\bar{\bar{S}}_w$ and ${}^{ID}\bar{\bar{S}}_w$ are the apparent water saturations at previous drying-to-wetting and wetting-to-drying reversal points, respectively.

For wetting scanning paths, the MWC is scaled through the saturation path reversal from the previous drying path to the current wetting path and the preceding wetting-to-drying reversal. The equation for predicting wetting scanning paths is:

$$\bar{\bar{S}}_w(P_c) = \left\{ \frac{\left[{}^{I}\bar{\bar{S}}_w(P_c) - {}^{I}\bar{\bar{S}}_w({}^{ID}P_c)\right]\left[{}^{DI}\bar{\bar{S}}_w - {}^{ID}\bar{\bar{S}}_w\right]}{\left[{}^{I}\bar{\bar{S}}_w({}^{DI}P_c) - {}^{I}\bar{\bar{S}}_w({}^{ID}P_c)\right]} \right\} + {}^{ID}\bar{\bar{S}}_w \qquad [23]$$

in which ${}^{I}\bar{\bar{S}}_w({}^{ID}P_c)$ is the apparent water saturation of the MWC at the capillary pressure of the last wetting-to-drying reversal point, and ${}^{I}\bar{\bar{S}}_w({}^{DI}P_c)$ is the apparent water saturation of the MWC at the capillary pressure of the last drying-to-wetting reversal point. For wetting paths from the IDC, the preceding drying-to-wetting reversal apparent water saturation equals 1.

The scaling assumes that the scanning paths have the same general curvature as either the IDC or the MWC. Furthermore, the scaling enforces that the scanning curves close at the reversal points. Using the model of Lenhard *et al.*, an infinite number of saturation paths can be modeled from the IDC and MWC parameters and the reversal-point saturations. Using eqns [18–20], $\bar{S}_{at}$ can be decoupled from the $\bar{\bar{S}}_w$ calculated with either eqn [22] or [23] to yield θ.

To account for effects of entrapped air on the conduction of water, Lenhard and Parker have modified an equation frequently used by soil scientists to predict the ability of the pore spaces to conduct water. As previously discussed for water retention, they utilize the concept of apparent saturation and make a correction for the pore space containing entrapped air. Their equation to predict water relative permeability, k_{rw}, which is a measure of the ability of the pore space to conduct water relative to when all of the pore spaces are filled with water, is:

$$k_{rw} = \bar{S}_w^{0.5}\left[\frac{\int_0^{\bar{\bar{S}}_w}\frac{d\bar{S}_e}{P_c(\bar{S}_e)} - \int_0^{\bar{S}_{at}}\frac{d\bar{S}_{at}}{P_c(\bar{S}_e)}}{\int_0^{\bar{\bar{S}}_w}\frac{d\bar{S}_e}{P_c(\bar{S}_e)}}\right] \qquad [24]$$

where, for an air–water system, $\bar{S}_e = \bar{\bar{S}}_w$ is the proportion of the effective porosity in which water flow can potentially occur; and $P_c(\bar{\bar{S}}_e)$ is the inverse of the function describing the IDC (eqn [16]). The effect of eqn [24] is to yield a larger k_{rw} for scanning curves than for the IDC at the same water content, because the entrapment of air forces water to move into the larger pores, allowing for a greater potential to conduct water. Typically, soil scientists have ignored hysteresis in relations between k_{rw} and water content because the differences are relatively minor at low-to-moderate water-content values. However, the effects of air entrapment, one of the hysteresis components, may yield significant differences in k_{rw} relations at high-water-content values.

It is important to consider the effects of hysteresis on water behavior in the subsurface. A fluctuating water table experiment has been conducted in which water pressures and saturations were measured. A computer model that either accounts for or ignores hysteresis in the saturation–capillary pressure–relative hydraulic conductivity relations is used to simulate the experiment. Comparisons of simulation and experimental results are shown in **Figures 3 and 4**. In **Figure 3**, water saturation values are plotted as a function of time for an elevation of 60 cm above the lower boundary of the experimental domain, which has a height of 72 cm. The experiment started with the water table at the surface. During a 2-h period, the water table was decreased 5 cm every 10 min to an elevation of 7 cm above the lower boundary of the experimental domain. Starting at $t = 3$ h, the water table was subsequently increased 5 cm every 10 min until an elevation of 42 cm had been reached. This increase was then followed again by a decrease at $t = 5$ h in the water table elevation to 17 cm at the same rate as before. Finally, at $t = 6.67$ h, the water table was

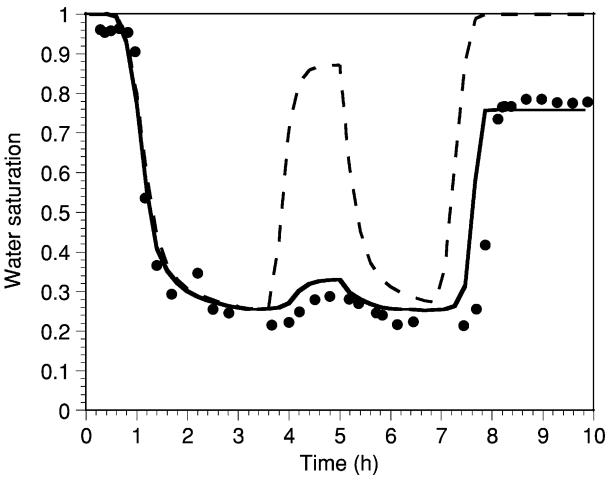


Figure 3 Comparison of measured (solid circles) with simulated fully hysteretic (solid line) and simulated nonhysteretic water content values during water table fluctuations at 60 cm above the bottom boundary of the flow domain.

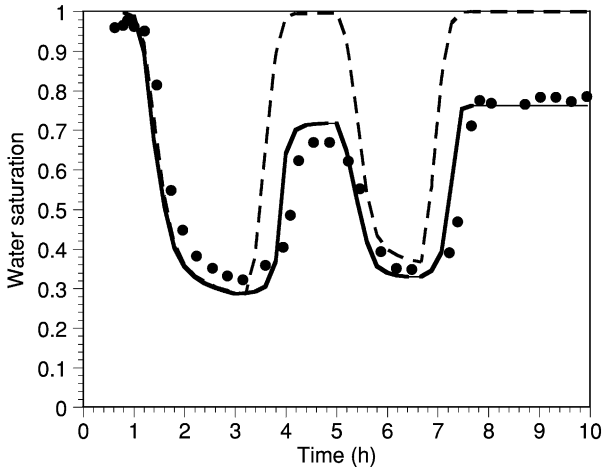


Figure 4 Comparison of measured (solid circles) with simulated fully hysteretic (solid line) and simulated nonhysteretic water content values during water table fluctuations at 70 cm above the bottom boundary of the flow domain.

increased to the surface at the same rate as before. It can be seen that increasing the water table from an elevation of 7 to 42 cm, and subsequently lowering it from 42 to 17 cm had only a small effect on the measured water content values ($3.5 < t < 6$ h). It was not until later, when the water table was raised from 17 cm to the soil surface, that a significant increase in water content occurred. For $8.5 < t < 10$ h, the water table is located above the 60-cm measurement elevation and the water content remains constant with time, as is expected for a strongly water-wet soil. The water saturation does not approach 1 during this time, because air was entrapped in the pore spaces, which constitute slightly more that 20% of the pore volume. Note that, over the short duration of

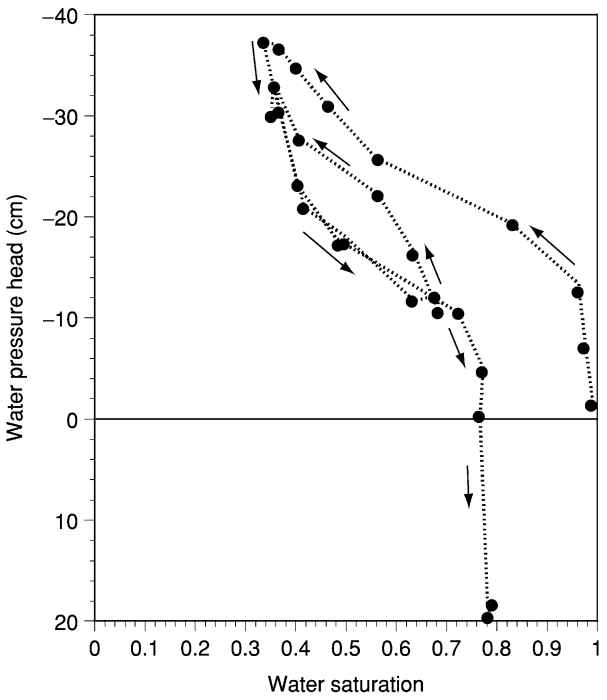


Figure 5 Measured saturation–matric-head (in this case equal to the pressure head) history at 50 cm above the bottom boundary of the flow domain showing a closed internal scanning loop. The solid circles represent the data points, while the arrows indicate the direction in which the water content is changing.

the experiment, the entrapped air does not move or dissolve into the water. In Figure 4, the results are shown for an elevation of 50 cm above the lower boundary. Different water content values from those in Figure 3 were measured because the saturation path history is different than for the data in Figure 3. In fact, all elevations will possess a different saturation path history. The results shown in Figures 3 and 4 demonstrate that the simulation results obtained using an algorithm that accounts for hysteresis are significantly better than those when hysteresis is ignored. Figure 5 shows the measurements of the soil water matric-head versus saturation at the 50-cm elevation. The closed scanning loop was produced from the first raising of the water-table elevation to 42 cm and the subsequent lowering of the water table to 17 cm.

Summary

Hysteresis in water-retention curves causes these curves to be different during wetting than during drying, i.e., at the same water-content values, different values for the soil water matric head or capillary pressure exist or vice versa. Ignoring hysteresis when predicting soil water content profiles by means of simulation models results in erroneous information.

See also: **Capillarity**; **Drainage, Surface and Subsurface**; **Infiltration**; **Leaching Processes**; **Porosity and Pore-Size Distribution**; **Vadose Zone:** Hydrologic Processes; **Water Potential**

Further Reading

Bomba SJ (1968) *Hysteresis and Time-Scale Invariance in a Glass-Bead Medium*. PhD Thesis. Madison, WI: University of Wisconsin Press.

Dane JH and Klute A (1977) Salt effects on the hydraulic properties of a swelling soil. *Soil Science Society of America Journal* 41: 1043–1049.

Dane JH and Wierenga PJ (1975) Effect of hysteresis on the prediction of infiltration, redistribution and drainage of water in a layered soil. *Journal of Hydrology* 25: 229–242.

Haines WB (1930) Studies in the physical properties of soils. V. The hysteresis effect in capillary properties and the modes of moisture distribution associated therewith. *Journal of Agricultural Science* 20: 97–116.

Hillel D (1980) *Fundamentals of Soil Physics*. New York: Academic Press.

Hillel D and Mottes J (1966) Effect of plate impedance, wetting method and aging on soil moisture retention. *Soil Science* 102: 135–140.

Koorevaar P, Menelik G, and Dirksen C (1983) *Element of Soil Physics*. Developments in soil science 13. Elsevier Science.

Land CS (1968) Calculation of imbibition relative permeability for two- and three-phase flow from rock properties. *Society of Petroleum Engineering Journal* 8: 149–156.

Lenhard RJ (1992) Measurement and modeling of three-phase saturation-pressure hysteresis. *Journal of Contaminant Hydrology* 9: 243–269.

Lenhard RJ and Parker JC (1987) A model for hysteretic constitutive relations governing multiphase flow. 2. Permeability–saturation relations. *Water Resources Research* 23: 2197–2206.

Lenhard RJ, Parker JC, and Kaluarachchi JJ (1991) Comparing simulated and experimental hysteretic two-phase transient fluid flow phenomena. *Water Resources Research* 27: 2113–2124.

Miller EE and Miller RD (1955a) Theory of capillary flow. I. Practical implications. *Soil Science Society of America Proceedings* 19: 267–271.

Miller EE and Miller RD (1955b) Theory of capillary flow. II. Experimental information. *Soil Science Society of America Proceedings* 19: 271–275.

Miller EE and Miller RD (1956) Physical theory for capillary flow phenomena. *Journal of Applied Physics* 27: 324–332.

Parker JC and Lenhard RJ (1987) A model for hysteretic constitutive relations governing multiphase flow. 1. Saturation-pressure relations. *Water Resources Research* 23: 2187–2196.

Philip JR (1964) Similarity hypothesis for capillary hysteresis in porous materials. *Journal of Geophysics Research* 69: 1553–1562.

Topp GC (1969) Soil water hysteresis measured in a sandy loam and compared with the hysteresis domain model. *Soil Science Society of America Proceedings* 33: 645–651.

Topp GC and Miller EE (1966) Hysteresis moisture characteristics and hydraulic conductivities for glass-bead media. *Soil Science Society of America Proceedings* 30: 156–162.

van Genuchten MTh (1980) A closed-form equation for predicting the hydraulic conductivity of unsaturated soils. *Soil Science Society of America Journal* 44: 892–898.

IMMISCIBLE FLUIDS

R J Lenhard, Idaho National Engineering Environmental Laboratory, Idaho Falls, ID, USA
J H Dane, Auburn University, Auburn, AL, USA
M Oostrom, Pacific Northwest National Laboratory, Richland, WA, USA

Published by Elsevier Ltd.

Introduction

Soils are located near the surfaces of land on Earth. As such, they contain variable amounts of air and water, which are immiscible fluids. Soils are also subject to contamination by human activities. Many of the contaminants commonly found in soils are liquids that are immiscible with air and water. These immiscible contaminants are principally fuel hydrocarbons and industrial solvents, and are referred to as nonaqueous-phase liquids (NAPLs). Fuel hydrocarbons (petroleum liquids) tend to be NAPLs that are less dense than water (LNAPLs). Industrial solvents, especially chlorinated hydrocarbons, tend to be denser than water (DNAPLs). The increased interest in better understanding NAPL behavior in soils is related to our concern for water quality. As water moves downward through soils, dissolved NAPL components can migrate to potable water supplies. Furthermore, high vapor-pressure components of NAPLs may partition to the soil air and move through the soil, thereby mixing the volatilized components with the soil air. These volatilized components can later partition to potable water supplies. The migration rate of the dissolved and volatile NAPL components and the NAPL movement toward groundwater are affected by the distribution of NAPL in soils.

The distribution and movement of air, water, and NAPLs in soils are governed by differences in fluid pressures among the fluid phases, the gradient of pressure in each fluid phase, and the size of the pores containing the fluids. The partitioning of components among the fluid phases may vary spatially and temporally. Overall, the behavior of NAPLs in soils is a consequence of complex physical, chemical, and biological processes. In the following sections, we outline some of the processes.

Fluid Distribution in the Pore Spaces

The distribution of air, water, and NAPL in pore spaces is a function of their pressures and the molecular forces of attraction for each other and for the surfaces of the soil particles. 'Wettability' describes the propensity of fluids for solid surfaces and other fluids. It is a complex function of fluid and solid properties. One measure of wettability is the contact angle that a fluid makes with a solid surface in the presence of other fluids. The wettability will determine whether fluids will occupy larger or smaller pores in soils. A 'wetting fluid' means that the fluid will spread over solid surfaces and a 'nonwetting fluid' means that the fluid is not in contact with solids. In soils, water is generally accepted to be the wetting fluid, and NAPL and air are nonwetting fluids. It is also generally accepted that NAPL will 'wet' (spread over) water surfaces. Therefore, water typically occupies the smallest pore spaces in soils and air occupies the largest pore spaces. NAPL occupies pore spaces intermediate to those occupied by water and air. However, there are several exceptions to these conditions. One exception occurs when organic compounds coat soil particles. Another exception is when the NAPL has a complex structure in which one end of the compound has a hydrophilic characteristic and the opposite end has a hydrophobic characteristic. For the former exception, the NAPL will probably wet the organic-coated soil particles instead of water. For the latter exception, a condition of mixed wettability may develop that can result in water occupying pores smaller and larger than those occupied by NAPL during water imbibition.

A measure of the molecular forces of attraction between immiscible fluids is the interfacial tension, which is a macroscopic and statistical property of the interface separating two immiscible fluids. The interfacial tension is a measure of the free energy per unit area of the interface. The SI units for interfacial

tension are force per unit length (newtons per meter). When one of the fluids is a gas and the other a liquid, then the term 'surface tension' is commonly used instead of 'interfacial tension.' Surface tension and interfacial tension are different words for the same phenomenon. In soils, the distribution of immiscible fluids is different for identical pore geometries and fluid pressures if the interfacial tensions are different.

A basic equation of capillarity is the Young–Laplace equation, which was proposed in 1805. The equation relates the difference in fluid pressures across the interface separating immiscible fluids to the interfacial tension and the radii of curvature of the interface at mechanical equilibrium. The Young–Laplace equation is:

$$\Delta P = \sigma\left(\frac{1}{R_1} + \frac{1}{R_2}\right) \qquad [1]$$

in which ΔP (pascals) is the difference in pressures of the fluids across the interface, σ (newtons per meter) is the interfacial tension of the fluid pair, and R_1 and R_2 (meters) are radii of curvature of the interface in orthogonal directions. In soils, ΔP is commonly called the capillary pressure, which is defined by:

$$P_c = P_{nw} - P_w \qquad [2]$$

in which P_c (pascals) is the capillary pressure, P_{nw} (pascals) is the nonwetting-fluid pressure, and P_w (pascals) is the wetting-fluid pressure. The interface separating the nonwetting and wetting fluids is curved convex toward the nonwetting fluid, which means that the pressure of the nonwetting fluid is greater than the wetting fluid. Consequently, the capillary pressure is always positive using the definition given by eqn [2]. The radii of curvature of the fluid–fluid interface in eqn [1] can be converted to radii of soil pores by:

$$\cos\alpha = \frac{r}{R} \qquad [3]$$

in which α is the contact angle, r is a radius of the soil pore, and R is a radius of the fluid–fluid interface curvature.

In air–NAPL–water systems in soils, there are two capillary pressures. When water is the wetting fluid and NAPL wets water surfaces, there are interfaces separating the water from the NAPL and the NAPL from the air. The pressure difference across the NAPL–water interfaces is called the NAPL–water capillary pressure. The pressure difference across the air–NAPL interfaces is called the air–NAPL capillary pressure. For a given pore geometry in soils, there is a relationship between capillary pressures and fluid contents. One measure of the fluid content is the ratio of the volume of fluid (cubic meters) per unit volume of soil (cubic meters), i.e.:

$$\theta_i = \frac{V_i}{V_t} \qquad [4]$$

in which θ_i is the volumetric fluid content of fluid i and V_i is the volume of fluid i in the total volume of soil (V_t), including solid and void spaces. Another measure of fluid content is the ratio of the volume of fluid (cubic meters) per unit of void volume in a soil sample (cubic meters) i.e.:

$$S_i = \frac{V_i}{V_v} \qquad [5]$$

in which S_i is the saturation of fluid i and V_v is the total void volume in the soil sample containing the volume of fluid i. The volumetric fluid content and the fluid saturation are related to the porosity of the soil through:

$$\phi = \frac{\theta_i}{S_i} = \frac{V_v}{V_t} \qquad [6]$$

in which ϕ is the porosity of the soil. Because the measures of fluid content and porosity are expressed as ratios of volumes, they are dimensionless.

The air–water capillary pressure (P_{aw}) and the water, or aqueous-phase, saturation (S_{aq}) are related to each other in an air–water system. The NAPL–water capillary pressure (P_{ow}) and the aqueous-phase saturation are related to each other in an air–NAPL–water system. The air–NAPL capillary pressure (P_{ao}) and total-liquid saturation (S_l) are related to each other in an air–NAPL–water system. The total-liquid saturation is the sum of the NAPL and aqueous-phase saturations.

There have been several proposed models to describe mathematically the relations between capillary pressures and fluid saturations. The two most common models for describing S–P relations are those by Brooks and Corey, and van Genuchten. The models were initially proposed for describing S–P relations in air–water systems. Since then they have been adapted to describe S–P relations in air–NAPL–water systems. The Brooks–Corey model for an air–water system is:

$$\bar{S}_{aq} = 1 \quad \text{for } P_{aw} \le P_d^{aw}$$
$$\bar{S}_{aq} = \left(\frac{P_d^{aw}}{P_{aw}}\right)^{\lambda} \quad \text{for } P_{aw} > P_d^{aw} \qquad [7]$$

where

$$\bar{S}_{aq} = \frac{S_{aq} - S_r}{1 - S_r}$$

in which $\bar{S}_{aq}$ is the effective aqueous-phase saturation, P_d^{aw} is the displacement pressure in an air–water

system, λ is a parameter related to the pore-size distribution, and S_r is the residual aqueous-phase saturation.

The van Genuchten model for an air–water system is:

$$\bar{S}_{aq} = [1 + (\alpha_{aw} P_{aw})^n]^{-m} \quad \text{for } P_{aw} \geq 0 \qquad [8]$$

where

$$m = \frac{n-1}{n}$$

and α_{aw} and n are curve-fitting parameters.

The extension of the above equations to describe S–P relations in air–NAPL–water systems was proposed by Parker *et al*. To minimize the number of parameters needed to describe three-fluid S–P relations, a scaling parameter was used. A scaled S–P function was defined that would apply to air–water and air–NAPL–water systems. An implicit assumption in the approach is that the effects of clay–fluid interactions on S–P relations are similar in air–NAPL–water systems to those in air–water systems. The scaling format used by Parker *et al*., which has similarities to scaling methods proposed in the petroleum industry, is:

$$\begin{aligned}
\bar{S}_{aq}^{aw}(\beta_{aw} P_{aw}) &= S^*(P^*) \\
\bar{S}_{aq}^{ow}(\beta_{ow} P_{ow}) &= S^*(P^*) \\
\bar{S}_{o}^{ao}(\beta_{ao} P_{ao}) &= S^*(P^*) \\
\bar{S}_{aq}^{aow}(\beta_{ow} P_{ow}) &= S^*(P^*) \\
\bar{S}_{l}^{aow}(\beta_{ao} P_{ao}) &= S^*(P^*)
\end{aligned} \qquad [9]$$

in which $\bar{S}_{aq}^{aw}$ and $\bar{S}_{aq}^{aow}$ are the effective aqueous-phase saturations in the air–water and in the air–NAPL–water fluid systems, respectively; $\bar{S}_{l}^{aow}$ is the effective total-liquid saturation in the air–NAPL–water system; $\bar{S}_{aq}^{ow}$ is the effective aqueous-phase saturation in a two-phase NAPL–water system; $\bar{S}_{o}^{ao}$ is the effective NAPL saturation in a two-phase air–NAPL system; β_{aw}, β_{ow}, and β_{ao} are the air–water, NAPL–water, and air–NAPL scaling factors, respectively; P_{aw}, P_{ow}, and P_{ao} are the air–water, NAPL–water, and air–NAPL capillary pressures, respectively; and $S^*(P^*)$ is the scaled S–P relation. The scaling format given in eqn [9] has been tested for nonhysteretic S–P data (i.e., main drainage branch S–P relations) and the results (**Figure 1**) validated the basic concepts of the scaling format.

Describing NAPL movement in soils requires that air–water and air–NAPL–water systems be considered simultaneously in any model. As NAPL moves into a soil or as it drains, volatilizes, or dissolves, there must be no discontinuities in fluid contents as a function of capillary pressure. To prevent any such discontinuities, the scaling factors must satisfy:

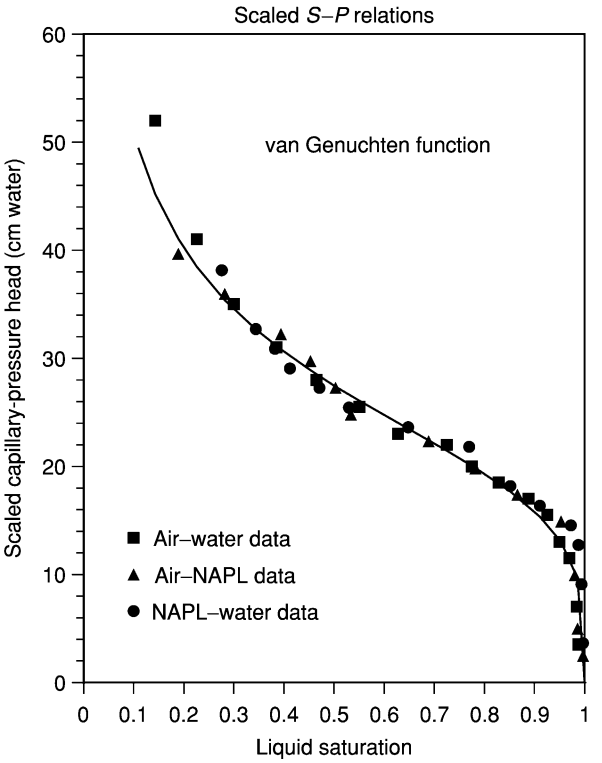


Figure 1 Scaling of capillary pressure–saturation measurements of different fluid systems using Lenhard's scaling format. (Adapted from Lenhard RJ (1994) Scaling fluid content–pressure relations of different fluid systems in porous media. In: Morel-Seytoux HJ (ed.) *Proceedings of 14th Hydrology Days*, pp. 223–235. Atherton, CA: Hydrology Days Publications, with permission.)

$$\frac{1}{\beta_{aw}} = \frac{1}{\beta_{ow}} + \frac{1}{\beta_{ao}} \qquad [10]$$

The scaling factors can be determined from S–P data in two-fluid or three-fluid systems or predicted from ratios of interfacial tensions. The β_{ow} and β_{ao} scaling factors are determined when the $\bar{S}_{aq}^{aow}(\beta_{ow} P_{ow})$ and $\bar{S}_{l}^{aow}(\beta_{ao} P_{ao})$ or $\bar{S}_{aq}^{ow}(\beta_{ow} P_{ow})$ and $\bar{S}_{aq}^{ao}(\beta_{ao} P_{ao})$ relations, respectively, are scaled to the $\bar{S}_{aq}^{aw}(\beta_{aw} P_{aw})$ relations in which β_{aw} is defined as 1. To ensure that eqn [10] is followed when β_{ow} and β_{ao} are determined from regression analyses of S–P data, β_{ao} should be set to:

$$\beta_{ao} = \frac{\beta_{ow}}{\beta_{ow} - 1} \qquad [11]$$

By following this procedure, then only one scaling factor (β_{ow}) needs to be determined in the regression analyses. If $\bar{S}_{aq}^{aow}(\beta_{ow} P_{ow})$ and $\bar{S}_{l}^{aow}(\beta_{ao} P_{ao})$ relations are unavailable, then the β_{ao} and β_{ow} scaling factors can be estimated from ratios of air–NAPL and NAPL–water interfacial tensions. To ensure that no discontinuities in fluid contents exist as a function of capillary pressure, the β_{ao} and β_{ow} scaling factors can be determined from:

$$\beta_{ao} = \frac{\sigma_{ao} + \sigma_{ow}}{\sigma_{ao}}$$
$$\beta_{ow} = \frac{\sigma_{ao} + \sigma_{ow}}{\sigma_{ow}} \quad [12]$$

Figure 2 shows the distribution of air, NAPL, and water above the water-saturated region in a soil when the scaling factors conform to eqn [10] and when they do not. The discontinuity in water content at the transition between the air–water and the air–NAPL–water regions can be seen in Figure 2b when eqn [10] is not followed.

The relationship between fluid content and capillary pressure is very complex and not unique; it depends on the history of fluid movements. This phenomenon is called hysteresis. In air–water systems, there are only two possible combinations of fluid movement. Either water is imbibing and air is draining, or water is draining and air is imbibing. In air–NAPL–water systems, there are six possible combinations of fluid movement, which makes predicting, or modeling, three-fluid *S–P* behavior more difficult. Because of the difficulty of describing hysteresis, many modelers neglect it.

Major reasons for hysteresis in *S–P* relations are fluid entrapment, pore-geometry effects, and contact angle changes with direction of fluid movement (draining versus imbibing). As nonwetting fluid is displaced by more wetting fluid, the nonwetting fluid can become trapped and isolated in the soil pores by various mechanisms. The net result causes the wetting fluid to occupy larger pores on imbibition saturation paths than it would on drainage saturation paths for the same wetting-fluid saturation. When the fluid–fluid interfaces are in larger pores, the radii of curvature of the interfaces will be larger (see eqn [3]), which means the capillary pressure will be smaller (see eqn [1]). Pore-geometry effects include the phenomenon that some pores initially filled with fluid may not refill at the same capillary pressure at which they drained because the direction of fluid movement is different. Sometimes this phenomenon is called the ink-bottle effect. Contact angle effects result because the angle a fluid makes with the soil particle surfaces may be different depending if the fluid is advancing or retreating over the soil or fluid surfaces. A different contact angle for the same fluid content will yield a different capillary pressure. However, the change in contact angles in soils with direction of fluid movement is very complex, because the soil particle surfaces are very irregular at the microscopic scale. The magnitude of hysteresis can vary significantly in soils.

Parker and Lenhard have proposed a method to model hysteretic air–NAPL–water *S–P* relations in which water is the wetting fluid. An apparent

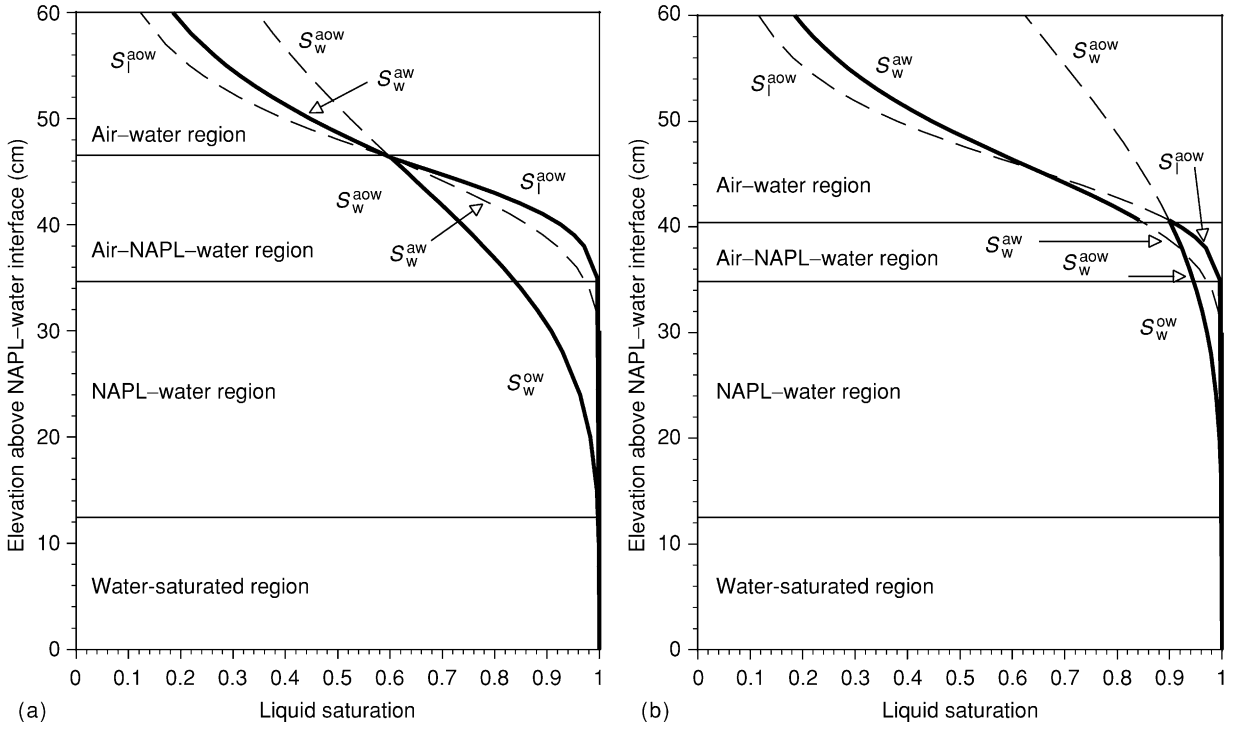


Figure 2 Distributions of air nonaqueous-phase liquids (NAPL), and water above the water-saturated region in which (a) the predictions were obtained using eqns [10] and [12] and (b) the predictions were obtained by only best-fitting the pertinent *S–P* data. Reproduced with permission from Lenhard RJ, Oostrom M, and Dane JH (2002) Prediction of capillary pressure-relative permeability relations. In: Dane JH and Topp GC (eds) Methods of Soils Analysis, Part 4 – Physical Methods. Soil Science Society of America Book Series 5.

saturation was defined to account for fluid entrapment. For an air–water system, the entrapped air in water can be determined from the apparent aqueous-phase saturation and the aqueous-phase saturation-path history. For an air–NAPL–water system, the entrapped NAPL in water can be determined from the apparent aqueous-phase saturation and the aqueous-phase saturation-path history. Also, for an air–NAPL–water system, the entrapped air in NAPL and water can be determined from the apparent total-liquid and aqueous-phase saturations and the saturation-path histories.

Parker and Lenhard have defined and parameterized two main *S–P* paths that describe relations between apparent saturations and capillary pressures. One path is for main drainage *S–P* relations, and the other path is for main imbibition relations. To describe *S–P* relations between the NAPL–water capillary pressure and the apparent aqueous-phase saturation, and relations between the air–NAPL capillary pressure and the apparent total-liquid saturation, the two main *S–P* paths of an air–water system and the scaling format given in eqn [9] are used. To describe any drainage *S–P* path, the main drainage *S–P* relations are scaled to pass through the pertinent saturation-path reversal points (imbibition to drainage and drainage to imbibition). To describe any imbibition *S–P* path, the main imbibition *S–P* relations are scaled to pass through the pertinent saturation-path reversal points. This approach has been modified to allow concurrent entrapment of NAPL and air by water. The hysteretic *S–P* model has also been tested against experimental data. In Figure 3, the air–water and air–NAPL–water *S–P* model is tested against the water saturation-path experimental data. In Figure 4, the *S–P* model is tested against the total-liquid saturation-path experimental data. In both cases, the model is able to describe the experimental data very well.

Recent modifications to the hysteretic *S–P* model involve accounting for residual NAPL in an air–NAPL–water system. Experimental investigations have shown that some NAPL will not drain from the soil pores in the vadose zone, even after long periods of time. The NAPL that does not drain from the soil pores is assumed to be immobile and has been called residual NAPL. Modifications to the hysteretic air–NAPL–water *S–P* model have been made to account for residual NAPL. The total NAPL saturation was defined to consist of mobile (free), immobile (residual), and water-trapped components. The amount of each component is calculated from the capillary pressures and the saturation-path histories.

Movement of Immiscible Fluids in Soils

The dimensions of the soil pores containing fluids are a very important factor for determining how readily those fluids move throughout the pore spaces in a

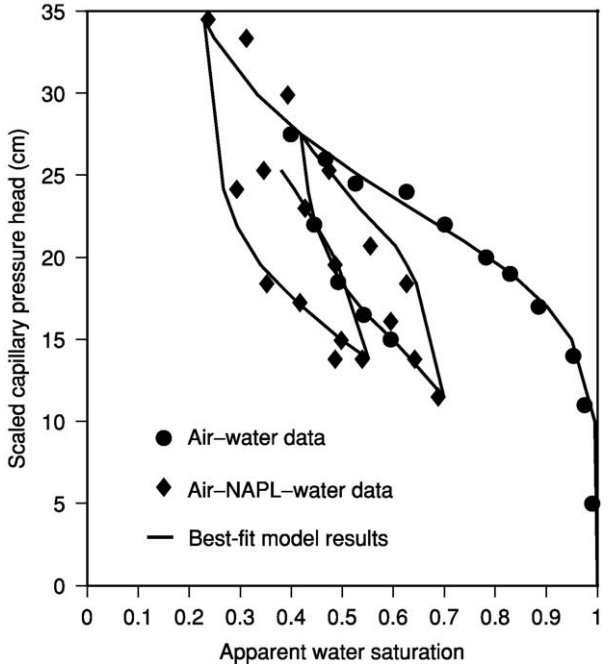


Figure 3 Scaling hysteretic water *S–P* relations of different fluid systems. (Adapted from Lenhard RJ (1992) Measurement and modeling of three-phase saturation–pressure hysteresis. *Journal of Contaminant Hydrology* 9: 243–269.)

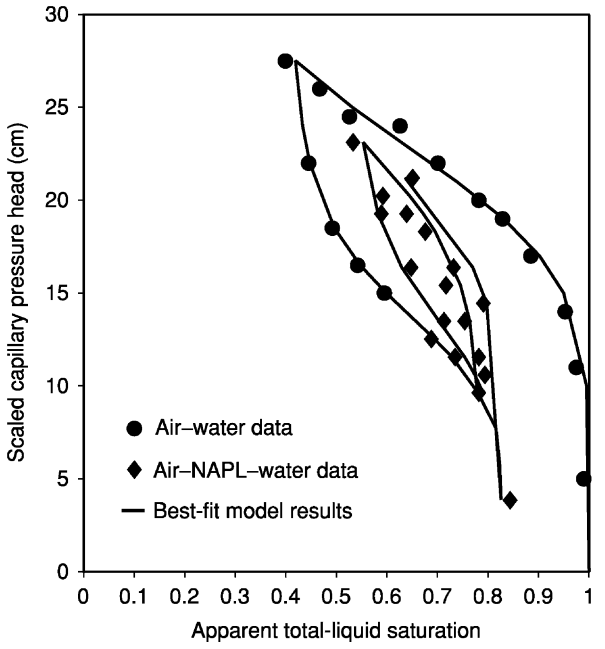


Figure 4 Scaling hysteretic total-liquid *S–P* relations of different fluid systems. (Adapted from Lenhard RJ (1992) Measurement and modeling of three-phase saturation–pressure hysteresis. *Journal of Contaminant Hydrology* 9: 243–269.)

soil. The Poiseuille equation shows that the velocity of fluid flowing through a circular tube varies with the square of the radius of the tube. Therefore, the fluid velocity in a tube having twice the radius of another tube is 4 times as great for the same gradient in piezometric pressure. The latter is an important factor governing how readily fluids move throughout the pore spaces in a soil. Because the geometry of soil pores is far more complicated than circular tubes, calculating the fluid velocity is more complex than the Poiseuille equation. However, the same principles apply in that larger soil pores are able to conduct fluid with less resistance than smaller pores.

The description of multifluid flow in soils is based on a modification of the Darcy law for saturated water flow, which was developed in 1856. In 1907, Buckingham modified the Darcy law to describe unsaturated water flow by replacing the saturated hydraulic conductivity with an unsaturated hydraulic conductivity that is a function of water content or capillary pressure. Over the years, the modification made by Buckingham and the recognition that the hydraulic conductivity consists of fluid-dependent and -independent elements has yielded what is commonly referred to as a generalized form of the Darcy law:

$$q_i = -\frac{k_{ri}k}{\mu_i}\left(\frac{\partial P_i}{\partial x_j} + \rho_i \mathbf{g}\right) \qquad [13]$$

in which q_i (meters per second) is the volumetric flux density of fluid i per cross-sectional area of a porous medium, k_{ri} is the relative permeability of fluid i, k (square meters) is the intrinsic permeability, μ_i (newton-seconds per square meter) is the viscosity of fluid i, P_i is the pressure of fluid i, ρ_i (kilograms per cubic meter) is the mass density of fluid i, $\mathbf{g}$ (meters per second squared) is the gravitational vector, and x_j (meters) is a symbol representing the Cartesian coordinates x, y, and z. The term in brackets in eqn [13] is the gradient of piezometric pressure. The relative permeability given in eqn [13] is a ratio of the permeability (square meters) at the relevant fluid saturation (i.e., effective permeability) divided by the intrinsic permeability.

A key parameter in eqn [13] that is very difficult and time-consuming to measure is the relative permeability, k_r. Almost all relative permeability measurements in air–NAPL–water systems have been conducted when the water saturation is at its residual value and for specific saturation paths. In the petroleum industry, it is common to measure the NAPL (oil) and water relative permeabilities for saturation paths similar to those that will be used during oil recovery. The measurements are made on samples that are carefully extracted from the petroleum reservoir and stored so as not to change the wettability of the samples. Measurements are generally conducted at temperatures and pressures of the petroleum reservoir. In determining cleanup strategies for NAPL-contaminated soils, measurements of relative permeability are typically not conducted, because they are expensive and no standard measurement techniques have been developed for soils. Whereas the water content in petroleum reservoirs is commonly at the residual water saturation, the water content in soils is usually much higher. Therefore, air–NAPL–water relative permeability measurement techniques used in the petroleum industry are not applicable to soils. Additionally, the cost for the elaborate relative permeability measurements in the petroleum industry cannot be justified for conducting analyses of NAPL-contaminated soils. Relative permeabilities that are needed to predict fluid flow of air–NAPL–water systems in soils are normally predicted.

The stationary fluid method has been applied to determine NAPL effective permeability values in NAPL–water systems. The method allows for direct measurements during steady-state flow conditions (constant capillary pressure and fluid saturations) throughout the sample. After each steady-state flow condition, the water saturation can be either increased or decreased.

Prediction of relative permeabilities is based on theory developed by petroleum engineers and soil scientists. In 1949, Purcell formulated a method to predict relative permeability using the porosity and S–P relations. Quickly thereafter, several investigators in the early 1950s developed models for predicting relative permeability using concepts proposed by Kozeny-Carman and Purcell. These early efforts focused on predicting relative permeabilities in two-fluid systems. It was not until the mid-1950s that investigators began to develop models for air–NAPL–water systems.

Investigations in the petroleum industry have shown that the water relative permeability in an air–NAPL–water system depends only on the water saturation. Similar investigations have also shown that the air relative permeability depends only on the air saturation. In an air–NAPL–water system, the air saturation is equal to $1 - S_l$; therefore, the air relative permeability is also a function of the total-liquid saturation. The NAPL relative permeability has been found to be a complex function of the water and NAPL (or total-liquid) saturations. To demonstrate how the NAPL relative permeability varies with water and NAPL saturations, predictions of the NAPL relative permeability using equations developed by Lenhard and Parker are shown in **Figure 5** for water saturations of 0.2 and 0.5. Hysteresis exists in k–S relations because of

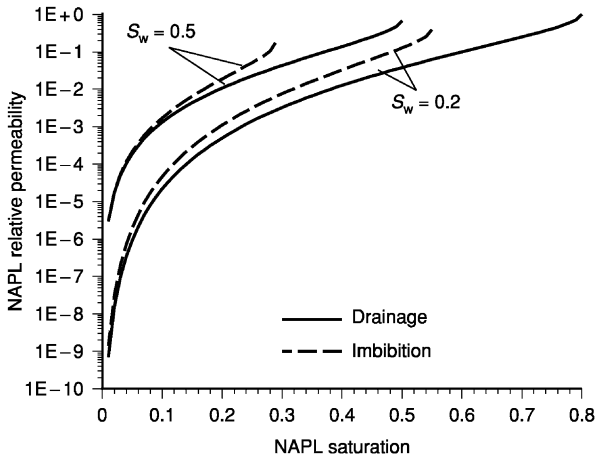


Figure 5 Predicted hysteretic NAPL relative permeability at two different water saturations. (Adapted from Lenhard RJ and Parker JC (1987) A model for hysteretic constitutive relations governing multiphase flow. 2. Permeability–saturation relations. *Water Resources Research* 23: 2197–2206.)

entrapment of fluids with lesser wettability by fluids with greater wettability. For example, air and NAPL can be entrapped and occluded by water, but air can only be entrapped and occluded by NAPL in strongly water-wet porous media. Water theoretically cannot be entrapped by NAPL or air in strongly water-wet porous media, because it is the wetting fluid (i.e., the fluid with the greatest wettability). It can be seen that significantly different NAPL relative permeability relations occur depending on the water and NAPL saturations. Furthermore, there can be large differences in the predicted NAPL relative permeability because of hysteresis. The difference is greater for larger total-liquid saturations.

Partitioning of NAPL Components Among the Fluid and Solid Phases

NAPL components in soils are present in various forms. They can exist in the aqueous, vapor, and solid phases as well as in a distinct liquid phase (i.e., NAPL). To model the subsurface behavior of NAPL components, the partitioning of the components among the liquid, gas, and solid phases needs to be considered. Mass transfer of NAPL components in strongly water-wet porous media involves three main processes: volatilization into the gas phase from the aqueous phase and NAPL; dissolution into the aqueous phase from the gas phase and NAPL; and sorption from the aqueous phase to the solid phase. Because water is always the wetting fluid in strongly water-wet porous media, it coats the soil grains, and gas-to-solid and NAPL-to-solid mass transfer are not possible. NAPL components in the gas phase and the NAPL must first partition to the aqueous phase before sorbing on solid surfaces.

Researchers have applied several different models to describe mass transfer of compounds between aqueous, gas, and solid phases. Because many mass-transfer processes are assumed to occur much faster than the convective flow of the fluids, mass transfer is commonly modeled as being instantaneous and local equilibrium is assumed. The simplest equilibrium model involves a sorption isotherm in which the concentration of a compound in the aqueous phase is linearly related to the concentration adsorbed to the solid surfaces. The so-called linear equilibrium isotherm is described by the equation:

$$C_s^i = K_d^i C_{aq}^i \qquad [14]$$

in which C_s^i is the mass of the sorbed NAPL component i per unit dry mass of solids (kilograms per kilogram), K_d^i is the partitioning coefficient (cubic meters per kilogram) of NAPL component i, and C_{aq}^i is the mass of dissolved NAPL component i per unit volume of the aqueous phase (kilograms per cubic meter).

If the organic carbon fraction of soils is greater than 0.01 (kilograms per kilogram), then sorption of dissolved NAPL components is dominated by the organic fraction. In this case, the partitioning coefficient, K_d^i, can be estimated by:

$$K_d^i = f_{oc} K_{oc}^i \qquad [15]$$

in which f_{oc} is the organic carbon fraction of the solid phase (kilograms per kilogram) and K_{oc}^i is the organic carbon partitioning coefficient (meters cubed per kilogram) of NAPL component i, which can be estimated from solubility data and molecular structure.

The partitioning of NAPL components between the gas phase and the NAPL, and between the gas phase and the aqueous phase, are also commonly modeled using instantaneous linear partitioning. The Raoult law is utilized to describe the partitioning of each NAPL component between the gas phase and the NAPL when the NAPL consists of multiple components. A different partitioning coefficient is needed for each NAPL component:

$$C_g^i = K_{go}^i C_o^i \qquad [16]$$

in which C_g^i is the concentration of NAPL component i in the gas phase (kilograms per cubic meter), K_{go}^i is the linear gas–NAPL partitioning coefficient (cubic meters per cubic meter) for NAPL component i, and C_o^i is the concentration of NAPL component i in the NAPL (kilograms per cubic meter). Similarly, the Henry law is used to describe the partitioning of NAPL mass between the gas and aqueous phases:

$$C_g^i = K_{gaq}^i C_{aq}^i \qquad [17]$$

in which K^{i}_{gaq} is the linear gas–aqueous phase partitioning coefficient (cubic meter per cubic meter) for NAPL component *i*. The concentrations of the NAPL components in the gas phase can be converted to partial pressures, and vice versa, using either the ideal or some nonideal gas law. The Henry law and Raoult law appear to be the same. Both describe linear relationships for idealized gas–liquid partitioning; however, there is a difference. The proportionality constant in the Raoult law is the partial pressure of a solvent component in its pure form, while the proportionality constant in the Henry law depends on the standard state of a solute in a mixture, generally at infinite dilution. Consequently, ideal Henry law behavior is observed for low component concentrations, while ideal Raoult law applies to a component that is nearly pure (high concentration). The two are thermodynamically consistent; however, significant deviations are frequently observed at intermediate concentration (mole fraction) ranges.

Linear partitioning coefficients are also used to describe the partitioning of NAPL mass between the NAPL and the aqueous phase:

$$C^{i}_{aq} = K^{i}_{aqo} C^{i}_{o} \qquad [18]$$

in which K^{i}_{aqo} is the linear NAPL–aqueous phase partitioning coefficient (cubic meters per cubic meter) for NAPL component *i*. Because it is common to assume that C^{i}_{aq} will be at its equilibrium solubility concentration, eqn [18] can be rewritten as:

$$C^{i}_{aq} = K^{i}_{aqo} \chi^{i}_{o} \rho_{o} \qquad [19]$$

in which χ^{i}_{o} is the mole fraction of NAPL component *i* in the NAPL and ρ_o is the mass density of the NAPL (kilograms per cubic meter). As NAPL components partition to fluid and solid phases at different rates, ρ_o changes with time.

If mass transfer cannot be assumed to be instantaneous, rate-limited mass transfer needs to be used to describe changes in NAPL component partitioning. Rate-limited interphase mass transfer is often approximated by first-order kinetic relationships representing the combined effect of all physical nonequilibrium processes affecting the migration between phases. The driving force is the concentration gradient across the interfaces separating the phases.

In nonequilibrium first-order mass-transfer relationships across gas–water and water–solid interfaces, the interphase exchange term, I_{aqo} (kilograms per cubic meter per second) for first-order dissolution of NAPL into the aqueous phase is usually given by:

$$I^{i}_{aqo} = k^{i}_{aqo} a_{aqo} \left(\chi^{i}_{o} \rho_{o} - C^{i}_{aq} \right) = k^{i*}_{aqo} \left(\chi^{i}_{o} \rho_{o} - C^{i}_{aq} \right) \qquad [20]$$

where k_{aqo} (meters per second) and k^{*}_{aqo} (per second) are mass-transfer coefficients between the aqueous phase and NAPL, and a_{aqo} (per meter) is the specific interfacial area separating the phases. As shown in eqn [20], NAPL dissolution rates are dependent on interfacial contact area. Due to difficulties in obtaining interfacial areas experimentally, this parameter is usually not explicitly obtained, but implicitly contained in k^{*}_{aqo}. The relation between rate-limited NAPL dissolution and interfacial area has been the focus of several research projects.

Summary

The behavior of immiscible fluids (NAPLs) in soils is very complex, and there are some important physical and chemical processes that govern the fate of NAPLs and their components in soils. To predict the fate of NAPL components in soils, the convective, dispersive, and diffusive movement of the NAPL components, as well as their chemical and biological transformations and interphase mass transfer, must be described. The resulting mathematical equations are nonlinear partial-differential equations that cannot be solved analytically for general conditions. Therefore, numerical techniques are used to estimate the solutions. Besides difficulties in describing all of the relevant processes that affect the fate of NAPL in soils, there are difficulties in solving the complex system of mathematical equations for predicting NAPL fate in soils. Additional research, both experimental and theoretical, will be required to achieve a more complete understanding of these processes.

Acknowledgment

The above work has been authored, in part, by a contractor of the US government under DOE contract DE-AC07-99ID13727.

See also: **Capillarity**; **Hydrocarbons**; **Hydrodynamics in Soils**; **Hysteresis**

Further Reading

Adamson AW (1990) *Physical Chemistry of Surfaces*. New York: John Wiley.

Anderson WG (1987) Wettability literature survey. 4. Effects of wettability on capillary pressure. *Journal of Petroleum Technology* 39: 1283–1300.

Armstrong JE, Frind EO, and McClellan RD (1994) Nonequilibrium mass transfer between the vapor, aqueous, and solid phases in unsaturated soils during vapor extraction. *Water Resources Research* 30: 355–368.

Brooks RH and Corey AT (1966) Properties of porous media affecting fluid flow. *Journal of Irrigation and Drainage Division Proceedings of the ASCE* 92: 61–88.

Busby RD, Lenhard RJ, and Rolston DE (1995) An investigation of saturation-capillary pressure relations in two- and three-fluid systems for several NAPLs in different porous media. *Ground Water* 33: 570–578.

Corey AT (1994) *Mechanics of Immiscible Fluids in Porous Media*. Highlands Ranch, CO: Water Resources Publications.

Dane JH and Topp GC (2002) *Methods of Soil Analysis,* part 4, Physical Methods. Madison, WI: Soil Science Society of America.

Dane JH, Hofstee C, and Corey AT (1998) Simultaneous measurement of capillary pressure, saturation, and effective permeability of immiscible liquids in porous media. *Water Resources Research* 34: 3687–3692.

Dullien FAL (1992) *Porous Media: Fluid Transport and Pore Structure*. New York: Academic Press.

Fetter CW (1993) *Contaminant Hydrogeology*. New York: Macmillan.

Hillel D (1980) *Fundamentals of Soil Physics*. New York: Academic Press.

Hofstee C, Oostrom M, Dane JH, and Walker RW (1998) Infiltration and redistribution of perchloroethylene in partially saturated, stratified porous media. *Journal of Contaminant Hydrology* 34: 293–313.

Honarpour M, Koederitz L, and Harvey AH (1986) *Relative Permeability & Petroleum Reservoirs*. Boca Raton, FL: CRC Press.

Imhoff PT, Jaffe PR, and Pinder GF (1994) An experimental study of complete dissolution of a nonaqueous phase liquid in saturated porous media. *Water Resources Research* 30: 307–320.

Jury WA, Gardner WR, and Gardner WH (1991) *Soil Physics*. New York: John Wiley.

Lenhard RJ (1992) Measurement and modeling of three-phase saturation-pressure hysteresis. *Journal of Contaminant Hydrology* 9: 243–269.

Lenhard RJ (1994) Scaling fluid content–pressure relations of different fluid systems in porous media. In: Morel-Seytoux HJ (ed.) *Proceedings of 14th Annual Hydrology Days*, pp. 223–235. Atherton, CA: Hydrology Days Publications.

Lenhard RJ and Parker JC (1987) Measurement and prediction of saturation–pressure relationships in three-phase porous media systems. *Journal of Contaminant Hydrology* 1: 407–424.

Lenhard RJ and Parker JC (1987) A model for hysteretic constitutive relations governing multiphase flow. 2. Permeability–saturation relations. *Water Resources Research* 23: 2197–2206.

Lenhard RJ and Parker JC (1988) Experimental validation of the theory of extending two-phase saturation-pressure relations to three-fluid phase systems for monotonic drainage paths. *Water Resources Research* 24: 373–380.

Lenhard RJ, Parker JC, and Kaluarachchi JJ (1989) A model for hysteretic constitutive relations governing multiphase flow. 3. Refinements and numerical simulations. *Water Resources Research* 25: 1727–1736.

Mayer AS and Miller CT (1996) The influence of mass transfer characteristics and porous media heterogeneity on nonaqueous phase dissolution. *Water Resources Research* 32: 1551–1568.

Oostrom M and Lenhard RJ (1998) Comparison of relative permeability-saturation-pressure parametric models for infiltration and redistribution of a light nonaqueous-phase liquid in sandy porous media. *Advances in Water Resources* 21: 145–157.

Oostrom M, Hofstee C, Lenhard RJ, and Wietsma TW (2003) Flow behavior and residual saturation formation of liquid carbon tetrachloride in unsaturated heterogeneous porous media. *Journal of Contaminant Hydrology* 64: 93–112.

Parker JC and Lenhard RJ (1987) A model for hysteretic constitutive relations governing multiphase flow. 1. Saturation–pressure relations. *Water Resources Research* 23: 2187–2196.

Parker JC, Lenhard RJ, and Kuppusamy T (1987) A parametric model for constitutive properties governing multiphase flow in porous media. *Water Resources Research* 23: 618–624.

Powers SE, Abriola LM, and Weber WJ Jr (1994) An experimental investigation of nonaqueous phase liquid dissolution in saturated subsurface systems: transient mass transfer rates. *Water Resources Research* 30: 321–332.

Sposito G (1989) *The Chemistry of Soils*. New York: Oxford University Press.

Tiab D and Donaldson EC (1996) *Petrophysics*. Gulf Publishing.

van Genuchten MT (1980) A closed-form equation for predicting the hydraulic conductivity of unsaturated soils. *Soil Science Society of America Journal* 44: 892–898.

Willhite GP (1986) *Waterflooding*. Richardson, TX: Society of Petroleum Engineers.

INCEPTISOLS

A Palmer, Massey University, Palmerston North, New Zealand

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Inceptisols are found in a wide range of environments and include a variety of soils. They occur in cool to very warm and humid to subhumid climates, from the arctic to the tropics. In temperate areas with ustic or udic soil-moisture regimes, Inceptisols tend to be characterized by weak soil weathering, just sufficient for mild soil development. This is often expressed as an alteration of material from its original condition to the extent that silicate clays are formed, oxides are liberated, or soil structure has formed. Inceptisols often occur on more resistant quartz-rich deposits where the rate and degree of weathering are somewhat retarded. Elsewhere Inceptisols are merely soils that do not meet the criteria to be members of any of the other soil orders.

Inceptisols have also undergone major changes in classification since *Soil Taxonomy* was published. Soils formed in volcanic ash (formerly called Andepts) have now been grouped in the order Andisols, and less importance has been placed on soils with low organic matter content (formerly Ochrepts) compared with those with higher organic matter content in their topsoil (formerly Umbrepts). Now, more than ever, Inceptisols can be regarded as the 'sweepings' of *Soil Taxonomy*.

Occurrence

The main factors that determine the occurrence of Inceptisols are soil temperature, landscape position, and soil-moisture regime:

Soil Temperature Regime: Cryepts;
Landscape Position: Aquepts;
Soil-Moisture Regime: Xerepts (dry) – Ustepts – Udepts (moist).

Inceptisols are classified in the normal *Soil Taxonomy* manner of a cascade, where the aquic soil-moisture regime (poorly drained) takes precedence, followed by human disturbance, cold soil temperatures, and finally other soil-moisture regimes:

1. Permanent or intermittent high water table: Aquepts;
2. Profiles relating to human disturbance: Anthrepts;
3. Cold but not permanently frozen soil temperatures: Cryepts;
4. Dry season with soil-moisture deficit: Ustepts;
5. Pronounced soil-moisture deficit in summer: Xerepts;
6. Other Inceptisols that remain moist for most of the year: Udepts.

It is therefore convenient to discuss the Inceptisols under these headings.

Inceptisols with Permanent or Intermittent High Water Tables

Aquepts are the Inceptisols with a permanently or intermittently high water table. The water table may rise through the profile or be perched above an impermeable layer or pan when rainfall exceeds evapotranspiration. Inceptisols are common on alluvial plains, above normal flood level, where time has been adequate for formation of a cambic horizon. However, the Aquepts often occupy former back-swamp positions, or the toe positions of alluvial fans where water tables are high.

In temperate areas affected by land destabilization during cool climates of the late Pleistocene, loess is a common soil parent material on river terraces, fans, marine terraces, and gently sloping downlands. Particularly in areas with a ustic soil-moisture regime, where the loess is subject to annual wetting and drying cycles, fragipans form in the subsoil. Many of these loessial soils with fragipans develop perched water tables when rainfall is high and evapotranspiration rate is low (usually winter). The Inceptisols that develop are Fragiaquepts, but in these conditions soils also commonly exhibit clay translocation and formation of an argillic horizon to the extent that they fall into the Fragiaqualf category. On loess-covered marine terraces in the Manawatu region of New Zealand (Figure 1), the Tokomaru series may classify as either a Fragiaqualf or a Fragiaquept, because clay translocation in the Bgt horizon is borderline between the two orders.

However, these are not the only Aquepts in the Manawatu landscape. During the last glacial there was wholesale destabilization of the landscape. In particular the adjacent Mesozoic greywacke-argillite ranges (up to 1600 m) lost their forest cover; scree slopes fed huge quantities of detritus into river headwaters. The Tertiary and Pleistocene mudstone-sandstone hill country also became a grass-shrubland rather than forest, also resulting in wholesale erosion.

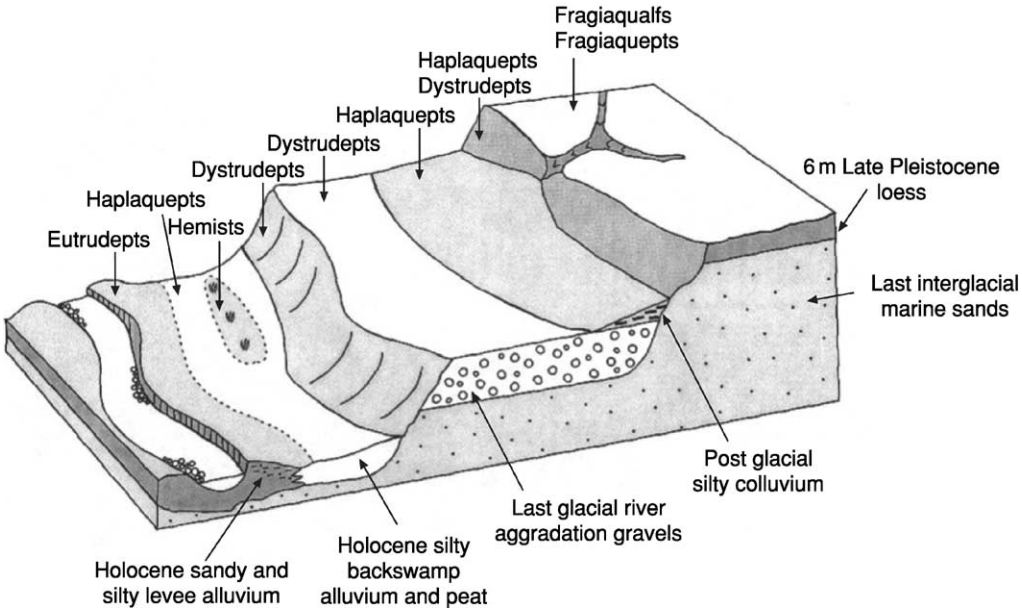


Figure 1 Soil-landscape diagram for part of the Manawatu region, southwest North Island, New Zealand. The rainfall is 1000 mm per annum with an ustic soil moisture regime. The upper terrace is comprised of Last Interglacial marine sands, subsequently covered with Last Glacial quartzo-feldspathic loess. In the prevailing climate Fragiaqualfs and Fragiaquepts form. The middle terrace is a river aggradation terrace from the last cooling of the Last Glacial. The inner side has been covered by postglacial colluvium in which Haplaquepts have formed, while Dystrudepts form on the stony gravel surface. On the lower (Holocene) alluvial terrace, Eutrudepts have formed on well-drained levees and Hemists and Haplaquepts on back-swamp positions.

As a result the local rivers became wide, braided beds and aggraded, forming large, low-angle fans that grade back high into the ranges. The main soils formed on the gravels and well-drained loamy alluvial cover are Dystrudepts.

However, as the climate warmed after 15 000 years BP forests recloaked the hill and mountain land, and the supply of gravels slowed. This led to the rivers downcutting, forming a series of degradation terraces. Preservation of the last glacial terrace was further enhanced by regional uplift. Fine-grained colluvium from adjacent terraces or hill country was washed on to part of the gravel terrace. Today, this colluvium is poorly drained, with a perched water table, and Haplaquepts have formed. On the Holocene degradation terraces, there is a repeated sequence of Eutrudepts on sandy textured levees, levee slopes, crevasse splays and point bars, and Haplaquepts and Hemists on back-swamp positions. On flood plains are Psamments, Orthents, and Aquepts.

All the Aquepts in this region pose problems for land use. They must be artificially drained but, even so, are difficult soils for horticulture and cropping. Loss of soil structure and formation of plough pans are the main problem. In pastoral farming, animal hooves and other traffic cause severe pugging (disturbance of the topsoil layer) in winter. Paradoxically, the finer-textured Aquepts are also drought prone in summer, because these soils have a pore-size distribution unsuited to storing adequate plant-available water.

Aquepts occupy similar back-swamp positions on many of the worlds major rivers, particularly in lower reaches; for example, Aquepts of the Indo-Ganges Plain and Aquepts of the lower Mississippi Valley.

Anthrepts

Anthrepts are the Inceptisols that are found in areas of prolonged cropping, manuring, and effluent disposal associated with human activities. They have either anthropic epipedons (Haplanthrepts) or plaggen epipedons (Plagganthrepts).

Anthropic epipedons result from prolonged irrigation and fertilizer application to produce crops, and from disposal of bones, shells, and wastes as a result of human occupation. The anthropic epipedon has the same organic matter content, thickness, structure and colour as mollic epipedons, but has higher phosphate content. Anthropic epipedons on irrigated land usually occur in arid areas, where application of water and fertilizer has produced an artificial epipedon that would not be there were it not for the water. Examples of these soils can be found in dry, irrigated areas of the USA (e.g., Idaho, California, Texas), Mexico, the Middle East, and Asia. Anthropic

epipedons resulting from disposal of bones, shells, and wastes (middens) are found in many parts of the world but are very localized, e.g., North America from former Indian occupation, the Pacific Islands and New Zealand from Polynesian or Melanesian occupation and Europe.

Plaggen epipedons are produced when the manure and bedding materials (sawdust, straw, pine needles) from housed or penned animals is applied to local soils, resulting in overthickened topsoils on otherwise less-fertile soils.

Cryepts

Cryepts occur on the margins of tundra at high latitudes, and at high altitude on major mountain ranges and plateaux. Soils that have permafrost or gelic materials within the top 1 m are now called Gelisols. Therefore Cryepts are common in coastal areas near the Alaska Peninsula, in northern Saskatchewan, and Western Siberia, bordering Gelisols of the Arctic regions. Cryepts are also common on high mountain ranges such as the Rocky Mountains in west-central USA, the Urals, Himalayas, and other ranges and plateaux of Central Asia and the Andes.

Cryepts have formed mostly under coniferous or mixed coniferous hardwood forest. Parent materials are till, solifluction deposits, loess, colluvium, or bedrock regolith of late Pleistocene or Holocene age. It is most common for ochric epipedons to overlie cambic horizons.

Ustepts

Ustepts are the Inceptisols that receive a seasonal rainfall, are not poorly drained, and most commonly have ochric epipedons and cambic horizons. The discriminating criteria at great group level are the presence of duripans or petrocalcic horizons, and whether the parent material is leached or contains free calcium carbonate:

With a duripan: Durustepts;
With a calcic or petrocalcic horizon: Calciustepts;
Leached or noncalcareous parent materials: Dystrustepts;
Other Ustepts: Haplustepts.

Soils with ustic soil-moisture regimes have commonly formed on early Holocene or Pleistocene parent materials, under native grasslands, or open forests near the grassland–forest transition. These areas can have had a complex history of forestation and deforestation during climate cycles of the Pleistocene and early Holocene, often assisted by fire during droughty periods. In the western USA particularly on the Great Plains, Ustepts commonly grade laterally into vast areas of Mollisols, where the history of grassland vegetation has been more uniform, and where stable forms of organic matter have been able to accumulate to greater depth. These Ustepts usually have umbric epipedons and might have truncated topsoils, not meeting the depth criteria for a mollic epipedon. In western Texas and New Mexico, there is a transition from Argids and Calcids in arid areas to Calciustepts and Ustolls as rainfall increases on the grassland Praries toward the east. On river terraces dating back to the early Pleistocene and Tertiary on the Great Plains in northern Texas and the Oklahoma panhandle, soils are formed in fine eolian and fluvial sediments. The dominant soils are Paleustolls and Argiustolls, with Psamments on the river flood plains. However, some younger terraces, and eroded soils or soils developed in reworked materials on the higher terraces are often Calciustepts.

In Hawkes Bay, on the east coast of the North Island of New Zealand, Inceptisols form on the low river terraces (Figure 2). The Twyford series (Haplustept) forms on levees and crevasse splays of current and abandoned river channels. The Karamu (Haplustepts) and Hastings series (Humaquepts) occupy backswamp positions, and Kaiapo series (Humaquepts) are found in the low-lying basins between levee systems. The Hastings Series is most extensive and has been deeply drained so that it is a most productive soil for pip- and stonefruit, vegetables, and other field crops. The Twyford and Karamu soils can be equally productive but require irrigation to ensure plant survival. Haplustepts of the Ngatarawa series are also found on older gravelly alluvium that dates back to the early Holocene and late Pleistocene. The most extensive land use on these soils is for vineyards.

Xerepts

These soils occur in areas with a 'Mediterranean climate,' typified by moist and cool winters, and warm dry summers. The winter precipitation is effective, because it coincides with minimum evapotranspiration rates. They are usually well drained and occur in frigid to thermic soil temperature regimes. Most have ochric or umbric epipedons and cambic horizons, but duripans, fragipans, and calcic horizons occur in some great groups:

Xerepts with duripans: Durixerepts;
Other Xerepts with calcic or petrocalcic horizons: Calcixerepts;
Xerepts with fragipans: Fragixerepts;
Leached Xerepts with moderate or low base saturation: Dystroxerepts;
Other Xerepts: Haploxerepts.

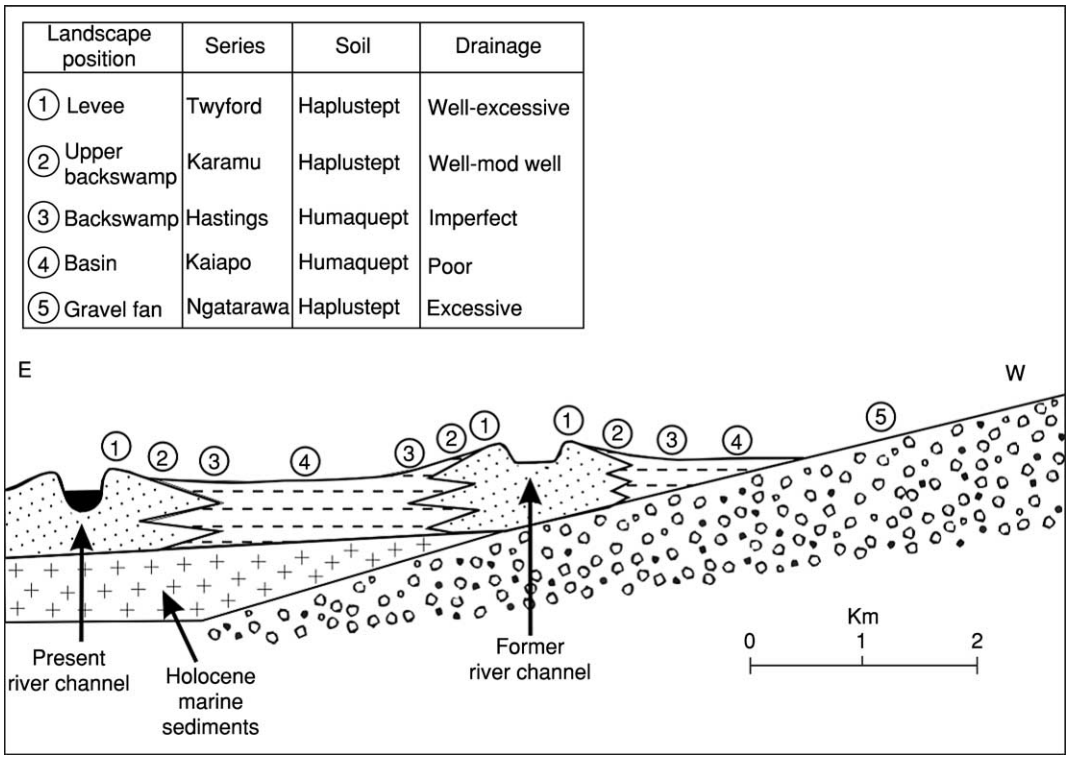


Figure 2 A cross-section of part of the Heretaunga plains on the east coast of the North Island of New Zealand. The rainfall is 780 mm with an ustic soil moisture regime. The plains were an embayment of the sea until being infilled by alluvium in the last 2000 years. Late Last Glacial to early Holocene gravels are overlain by transgressive marine sediments and then by the river alluvium. A complicated pattern of levees, backswamps, and interchannel basins has formed. The soils formed relate closely to geomorphology.

Xerepts are common, as the climate connotation suggests, in Mediterranean countries, particularly Spain, Italy, Turkey, and northern Tunisia, Algeria, and Morocco. In the USA Xerepts are most common in the western states of California, Oregon, Washington, Idaho, and Utah. Open mixed coniferous forest and grassland dominates in cooler climate, whereas mixed oak, shrub, and grassland is more common in thermic temperature regimes.

Udepts

Udepts are the Inceptisols that remain moist for most of the year. Dry periods are not long enough in most years to have soil-moisture deficits for as long as 3 months, i.e., they occur in a udic soil-moisture regime. Most have developed under a forest or forest–shrubland grassland vegetation. They are usually well or moderately well drained, with ochric or umbric epipedons over cambic horizons. The various great groups are named for the presence of other horizons:

Presence of a near-surface sulfuric horizon: Sulfudepts;
Presence of a duripan: Durudepts;
Presence of a fragipan: Fragiudepts;
Free carbonates or high base status: Eutrudepts;
Remaining Udepts: Dystrudepts.

Dystrudepts and Fragiudepts are common in the northeastern USA, on non-calcareous glacial till in Pensylvania and New York, and on acid shales and sandstones in the Appalachian Mountains and Allegheney Plateau. In forested areas, on acid or coarse-textured parent materials, Dystrudepts commonly form a mosaic with Spodosols. On the west coast of the South Island of New Zealand, in the Smoky Mountains, at the southern end of the Appalachian mountains, there is a similar mosaic of Spodosols and Drystrudepts. Both soils are deeply weathered, as indicated by destruction of feldspars and formation of gibbsite, kandite, and vermiculite-chlorite in the soil B horizon. The only observable difference is that Spodosols are more frequent on coarse-textured but less well drained sites, with slightly different flora.

The soils that form often relate to variations in parent material. In the soils formed on Wisconsin till in New Jersey, Dystrudepts have been developed on till derived from gneissic rocks, whereas adjacent soils formed in quartzose till and calcareous till are Spodosols and Alfisols, respectively. There has been

considerable translocation of materials in the Spodosols and Alfisols; whereas, in the iron-rich gneissic material, iron release has been gradual and the residual iron has coated local resistant particles.

Inceptisols may also occur in landscapes dominated by other orders representing strongly weathered soils, by occurring on less-stable positions in the landscape. In the Carolinas, Dystrudepts occur on steep slopes in the granite-gneiss terrain above the fall line. On more stable parts of the granite-gneiss terrain are Udalfs and Udults, while on coastal plain sediments below the fall line are Psamments, Spodosols, Udults, and Aquults. At the southern end of the Appalachians, in Tennessee, Dystrudepts occur on upland positions on shale, where mass wasting, overland flow, and soil creep inhibit soil formation. On adjacent, stable shoulder slopes, Hapludults occur, reflecting time enough to form a strongly leached Bt rather than a Bw horizon. On toeslope and floodplain bottom lands are Endoaquepts, while Hapludalfs occur on alluvial terraces.

In southeastern Brazil on steep, pre-Paleozoic gneiss and granite, deep Hapludox is found on stable summits, Dystrudepts on shoulder slopes, shallow Hapludox on backslopes, Kandiudults on footslopes, and Eutrudepts on toeslopes. The shoulder slope and backslope positions in this landscape are considered to be unstable retreating slopes, enabling soil rejuvenation. Similarly, a little farther south, Inceptisols are found on a retreating escarpment of Cretaceous subarkosic sandstone, whereas Oxisols occur on old fluvial terraces and pediments below the escarpment, and on the stable summit above the escarpment. Shoulder slopes and footslopes are occupied by Alfisols and Ultisols.

In the southern South Island of New Zealand, a climosequence of Inceptisols occurs in late Pleistocene loess deposits (Figure 3). As rainfall increases from 600 to 1200 mm, the soils change from Lamellic Haplustepts (600 mm) to Aeric Fragiaquepts and Aquic Haplustepts (700 mm); Aeric Fragiaquepts and Aquic Fragiudepts (850 mm); Aeric Fragiaqupts (950 mm); Aquic Dystrudepts (1100 mm); and Typic Dystrudepts (1200 mm). A feature of this sequence is first the deterioration then improvement of drainage as rainfall increases. The reason lies in the development of a gammate fragipan, which causes winter moisture to perch in the subsoil above, and then its disappearance at higher rainfall. Profiles above 1000 mm do not dry out substantially in most summers, whereas profiles in the range 700–1000 mm experience strong fluctuations of soil moisture, with winter surpluses and summer deficits.

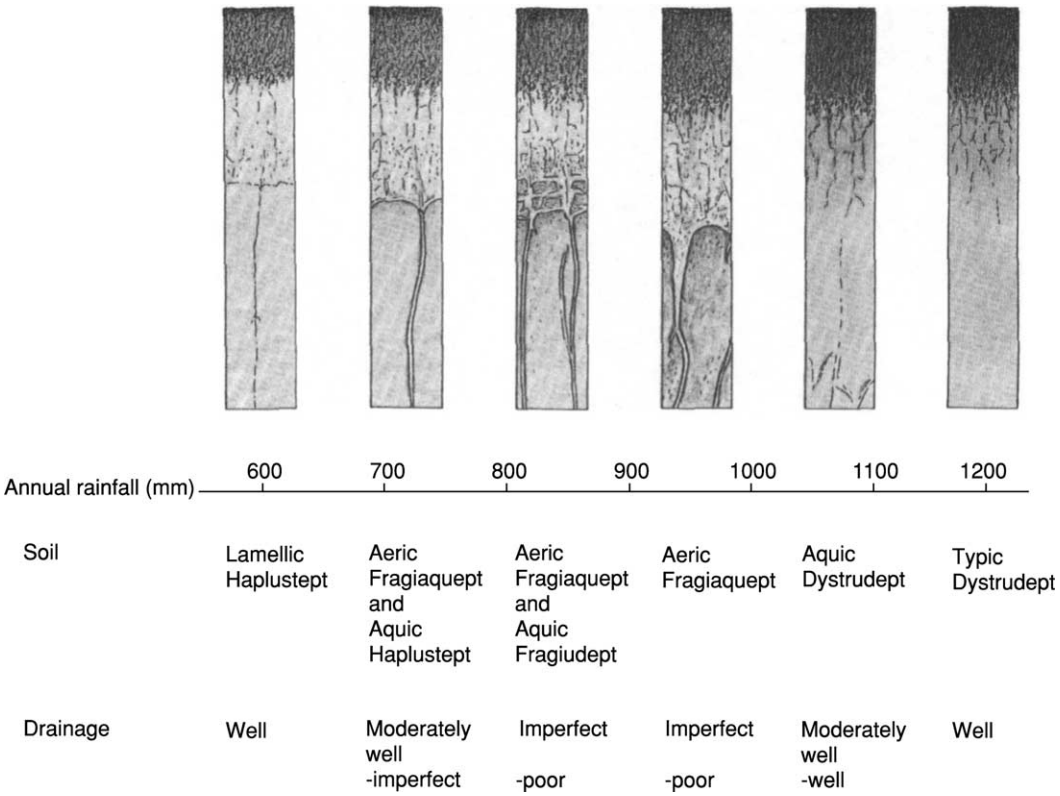


Figure 3 A climosequence of soils developed in Last Glacial loess in Southland, New Zealand. As rainfall increases, the drainage first deteriorates and then improves, influenced by the formation of an impermeable gammate fragipan at intermediate rainfalls (adapted with permission of the NZSSS).

At Otaki in the southwestern North Island of New Zealand, at a rainfall of approximately 1050 mm, a complex pattern of poorly to well-drained soils occurs on late last glacial and early Holocene loamy alluvium. The soils occur as a complex on gently undulating, last glacial river aggradation terraces. Soil series are defined according to their drainage status. The soils are Endoaquepts (poorly drained); Eutrudepts (imperfectly drained); and moderately well and well-drained Dystrudepts. Various reasons can be proposed for this soil pattern, but none have been proven to date. The soil pattern has a grain or orientation that suggests an influence of fluvial deposition, but no relationship has been established with either texture, depth to underlying gravels, or landscape position.

The Dystrudepts often have a trace of allophane, and the suggestion has been made that these soils contain more volcanic ash than adjacent soils, because allophane is a recognized weathering product of volcanic glass. Soils rich in allophane have good structure and are generally well drained. The drainage status is reflected in the chemical and physical properties of the soils, with well-drained soils having lower bulk density, higher macroporosity, and lower pH and base saturation than poorly drained soils. The well-drained soil also has less mica clay, more chlorite, hydroxy-interlayered vermiculite instead of vermiculite, and less kandite than the poorly drained soil.

Another similar soil complex of poorly drained to well-drained soils occurs in the same district on flat to gently undulating loess-covered terraces. In this complex, the soils are: poorly drained Typic Haplaquepts; imperfectly drained Typic Haplaquepts; moderately well drained Typic Dystrudepts, and well drained Umbric Dystrudepts. The latter soil contains appreciable (up to 10%) allophane in its upper horizons, but not enough to be an Andisol. Available mineralogy suggests that the loess was initially uniform, dominated by quartzo-feldspathic components, but with a small tephric component. Early in the weathering process, a critical threshold was reached, where in some locations, the volcanic glass formed allophane, while in other places formed halloysite. Where allophane formed, leaching of silica continued, more allophane formed, the allophane promoted stable, subangular blocky, granule and crumb structure, moderate macroporosity, and well-drained profiles. Where halloysite formed, leaching of silica was inhibited, blocky structure formed in a less-porous, denser soil. The soil distribution and soil properties suggest that with time conditions have more and more favored weathering to halloysite clay, and formation of Haplaquepts rather than Dystrudepts.

The described soil complexes occur where the rainfall is between 1050 and 1150 mm. At lower rainfall, the same alluvium and loess weathers to Fragiaqualfs, Fragiudepts, and Haplaquepts, while, at higher rainfall, Dystrudepts and Hapludands dominate, i.e., soils receiving higher rainfall have better drainage than those with lower rainfall.

See also: **Classification Systems:** USA; **Pedology:** Basic Principles

Further Reading

Almond PC and Tonkin PJ (1999) Pedogenesis by upbuilding in an extreme leaching and weathering environment, and slow loess accretion, south Westland, New Zealand. *Geoderma* 92: 1–36.

Anjos LH, Fernandes MR, Peteira MG, and Franzmeier DP (1998) Landscape and pedogenesis of an oxisol-inceptisol-ultisol sequence in southeastern Brazil. *Soil Science Society of America Journal* 62: 1651–1658.

Aslan A and Autin WJ (1998) Holocene flood-plain soil formation in the southern lower Mississippi Valley, implications for interpreting alluvial paleosols. *Geological Society of America Bulletin* 110: 433–449.

Blume HP and Kalk E (1986) Bronzezeitlicher Auftragsboden bei Rantum auf Sylt. *Zeitschrift fur Pflanzenernahrung und Bodenkunde* 149: 608–613.

Bruce JG (1983) Effect of climate on morphological features of soils from loess in the southern part of South Island, New Zealand. *Australian Journal of Soil Research* 21: 359–371.

Buol SW, Hole FD, and McCracken RJ (1980) *Soil Genesis and Classification*. Ames: Iowa State University Press.

Carter BJ, Ward PA, and Shannon JT (1990) Soil and geomorphic evolution within the Rolling Red Plains using Pleistocene volcanic ash deposits. *Geomorphology* 3: 471–488.

Dipak Sarkar Sahoo AK (2000) Aquepts of the Indo-Gangetic Plain of Bihar and their suitability for some major crops. *Journal of the Indian Society of Soil Science* 48: 561–566.

Duncan MM and Franzmeier DP (1999) Role of free silicon, aluminum, and iron in fragipan formation. *Soil Science Society of America Journal* 63: 923–929.

Engel RJ, Witty JE, and Eswaran H (1997) The classification, distribution, and extent of soils with a xeric moisture regime in the United States. *Catena* 28: 203–209.

Foss JE, Moormann FR, and Rieger S (1983) Inceptisols. In: Wilding LP, Smeck NE, and Hall GF (eds) *Pedogenesis and Soil Taxonomy. II. The Soil Orders*. Amsterdam: Elsevier.

Holliday VT (1990) Soils and landscape evolution of eolian plains: the southern High Plains of Texas and New Mexico. *Geomorphology* 3: 489–515.

Karlstrom ET (2000) Use of soils to identify glacial deposits of various ages east of Glacier National Park, Montana, USA. *Arctic Antarctic and Alpine Research* 32: 179–188.

Krebs RD and Tedrow JFC (1957) Genesis of three soils derived from Wisconsin till in New Jersey. *Soil Science* 83: 207–218.

Lepsch IF, Buol SW, and Daniels RB (1977) Soil-landscape relationships in Occidental Plateau of Sao Paulo State, Brazil: I. Geomorphic surfaces and soil mapping units. *Soil Science Society of America Journal* 41: 104–109.

Lepsch IF, Buol SW, and Daniels RB (1977) Soil-landscape relationships in Occidental Plateau of Sao Paulo State, Brazil: II. Soil morphology, genesis, and classification. *Soil Science Society of America Journal* 41: 109–115.

McCracken RJ, Shanks RE, and Clebsch EEB (1962) Soil morphology and genesis at higher elevations of the Great Smokey Mountains. *Soil Society of America Proceedings* 26: 384–388.

McFadgen BG (1980) Maori plaggen soils of New Zealand, their origin and properties. *Journal of the Royal Society of New Zealand* 10: 3–19.

Molloy L (1998) *Soils in the New Zealand Landscape: The Living Mantle*. Lincoln: New Zealand Society of Soil Science.

Pape JC (1970) Plaggen soils in the Netherlands. *Geoderma* 4: 229–255.

Phillips DH, Foss JE, Stiles CA, Trettin CC, and Luxmoore RJ (2001) Soil–landscape relationships at the lower reaches of a watershed at Bear Creek: near Oak Ridge Tennessee. *Catena* 44: 205–222.

Pillans B, McGlone M, Palmer AS, Mildenhall D, Alloway BV, and Berger G (1993) The last Glacial maximum in central and southern North Island, New Zealand: a paleoenvironmental reconstruction using the Kawakawa Tephra Formation as a chronostratigraphic marker. *Palaeogeography, Palaeoclimatology, Palaeoecology* 101: 283–304.

Ruhe RV and Olson CG (1980) Clay-mineralogy indicators of glacial and non-glacial sources of Wisconsinan loesses in southern Indiana. *Geoderma* 24: 283–297.

Soil Survey Staff (1975) *Soil Taxonomy: A Basic System of Soil Classification for Making and Interpreting Soil Surveys*. Agriculture Handbook. United States Department of Agriculture. Washington: US Government Printing Office.

Soil Survey Staff (1999) *Soil Taxonomy: A Basic System of Soil Classification for Making and Interpreting Soil Surveys*, 2nd ed. Agriculture Handbook 436. United States Department of Agriculture. Washington: US Government Printing Office.

Ugolini FC and Schlichte AK (1973) The effect of Holocene environmental changes on selected western Washington soils. *Soil Science* 116: 218–227.

Viereck LA, Dyrness CT, Cleve K, and van Foote MJ (1983) Vegetation, soils, and forest productivity in selected forest types in interior Alaska. *Canadian Journal of Forest Research* 13: 703–720.

Yaalon DH (1997) Soils in the Mediterranean region: what makes them different? *Catena* 28: 157–169.

Industrial Pollution *See* **Pollution:** Industrial

INFILTRATION

T P A Ferré and A W Warrick, University of Arizona, Tucson, AZ, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Infiltration is defined as the flow of water from aboveground into the subsurface. The topic of infiltration has received a great deal of attention because of its importance to topics as widely ranging as irrigation, contaminant transport, groundwater recharge, and ecosystem viability. More generally, a quantitative understanding of this process is vital to our ability to relate surface and subsurface processes in describing the hydrologic cycle.

While the definition of infiltration is simple, it can involve all aspects of flow through a variably saturated porous medium, ranging in complexity from steady-state, saturated flow in a homogeneous, isotropic medium to transient, unstable, unsaturated flow through an anisotropic, heterogeneous medium. The rate and pattern of infiltration vary with the distribution and rate with which water is supplied at the ground surface, the depth of the water table, the hydraulic properties of the subsurface materials, and the antecedent moisture content distribution. The following discussion focuses on infiltration into a homogeneous, isotropic medium as a basis for understanding the infiltration process.

Plug Flow

Constant Infiltration at the Ground Surface

In its simplest form, infiltration can be visualized as a one-dimensional, downward displacement process. Water application at the ground surface can be considered in two general ways. The simplest ground-surface boundary condition to envision, physically, is the application of water at a constant rate, such as infiltration during a constant-intensity rainfall. This is known as a type II, Neumann boundary condition. A type I, Dirichlet boundary condition, describes the application of water at a constant pressure head (or water content).

For simplicity, we begin by considering water supplied at a constant rate at the ground surface (type II ground-surface boundary condition). This rate can be defined as a flux density, with units of length per time, and is equivalent to the flow of water (volume per time) over an area. If the rate of supply is greater than the saturated hydraulic conductivity of the medium, water will eventually pond at the ground surface and flow overland. If the rate of water delivery is lower than the saturated hydraulic conductivity, all of the water will infiltrate. If the spatial area over which water is supplied is relatively large, infiltration can be approximated as one-dimensional downward. The water content of the medium will increase until the hydraulic conductivity of the medium at that water content is equal to the rate of water supply and infiltration will proceed under a unit gradient condition (i.e., the magnitude of the hydraulic head gradient is 1). As water is continually supplied, this wetted region will advance downward. Assuming that the water velocity in front of the advancing water is nearly zero, which is reasonable unless the antecedent water content is quite high, the added water will advance as a sharp wetting front. The depth to the wetting front, z_f, after a time t, can be determined from the antecedent volumetric water content, θ_i, the water content behind the wetting front, θ_{wet}, and the flux density, q, by a simple mass balance calculation:

$$z_f = \frac{qt}{\theta_{wet} - \theta_i} \quad [1]$$

The rate of advance of the wetting front, v_f, can be determined as the derivative of z_f with time:

$$v_f = \frac{q}{\theta_{wet} - \theta_i} \quad [2]$$

Notice that if the initial water content and the water content behind the wetting front are constant with depth and, given that the applied flux is constant in time, the velocity of the wetting front is constant with time. The rate of advance of the wetting front increases with an increase in the flux or the antecedent water content and with a decrease in the water content behind the wetting front.

As a first approximation, infiltrating water is generally assumed to displace the antecedent resident pore water completely. For example, consider that the water that is present in the medium before infiltration begins is dyed blue and that the infiltrating water is dyed red. There is a transition from a higher water content to a lower water content at a depth of z_f; but the water on either side of the wetting front is blue. At the same time, t, there is a transition from red water to blue water at a depth, z'_f, that is equal to:

$$z'_f = \frac{qt}{\theta_{wet}} \quad [3]$$

The rate of advance of the applied water front, v'_f, is:

$$v'_f = \frac{q}{\theta_{wet}} \quad [4]$$

Notice that if the water content behind the wetting front is constant with depth and the applied flux is constant in time, the velocity of the wetting front is also constant with time. The rate of advance of the applied water is independent of the antecedent water content. This simple representation of infiltration can be extended to consider flow through a layered subsurface in which each layer can have a different value for θ_i and θ_{wet}.

Constant Pressure Head at the Ground Surface

Now we consider the advance of a wetting front in response to a type I ground-surface boundary condition. As shown above, the velocity of the wetting front can be expressed as:

$$\frac{dz_f}{dt} = v_f = \frac{q}{\theta_{wet} - \theta_i} \quad [5]$$

To consider the case of the advance of the wetting front in response to a constant pressure head applied at the ground surface, ψ_0, the pressure head at the wetting front is defined as ψ_f. Then the flux across the ground surface is:

$$q = -K_{wet}\frac{\psi_f - \psi_0 - z_f}{z_f} \quad [6]$$

where K_{wet} is the hydraulic conductivity of the medium behind the wetting front. Substituting for the flux gives:

$$v_f = \frac{dz_f}{dt} = -K_{wet}\frac{\psi_f - \psi_0 - z_f}{z_f(\theta_{wet} - \theta_i)} \quad [7]$$

Integrating this expression gives an expression for the time, t, for the wetting front to advance to a depth z_f as:

$$t = \frac{(\theta_{wet} - \theta_i)}{K_{wet}}\left(z_f + (\psi_f - \psi_0)\ln\left(1 + \frac{z_f}{\psi_0 - \psi_f}\right)\right) \quad [8]$$

Green and Ampt derived a similar expression. For short times, gravity can be ignored and the results simplify to:

$$z_f = \left[\frac{2(\psi_0 - \psi_f)K_{wet}}{\theta_{wet} - \theta_i}\right]^{0.5} t^{0.5} \quad [9]$$

The rate of advance of the wetting front at time t is described as the derivative of z_f with t:

$$v_f = 0.5\left[\frac{2(\psi_0 - \psi_f)K_{wet}}{\theta_{wet} - \theta_i}\right]^{0.5} t^{-0.5} \quad [10]$$

Notice that, unlike the advance of a wetting front in response to a constant applied flux, the rate of advance of the wetting front in response to a type I ground-surface boundary condition changes with time. The rate of advance of the wetting front is highest initially when the distance from the surface to the wetting front is smallest, leading to the highest pressure-head gradient. This indicates that the rate of infiltration at the ground surface is highest at early times. At long times, gravity dominates infiltration and the infiltration rate at the ground surface decreases to a constant rate, q. As a result, at later times, the rate of advance of the wetting front decreases to $q/(\theta_{wet} - \theta_i)$. That is, for larger times the wetting front in response to a type I ground-surface boundary condition advances at the same rate as a wetting front in response to a type II boundary condition with an equivalent applied flux.

Richards Equation – The Shape of the Wetting Front

The preceding treatments are sufficient to provide simplified descriptions of the one-dimensional advance of a wetting front into a homogeneous, drained soil under time-invariant boundary conditions. However, both of these analyses are based on the assumption of a step change in water content at the wetting front. A more complete, detailed description of the infiltration process requires solution of the flow equation, including constitutive relationships describing the water content–pressure head relationship for the soils into which water is infiltrating.

In general, the Richards equation is a nonlinear, parabolic partial differential equation that describes water movement through a variably saturated medium under the assumption of zero air resistance. The equation and the accompanying constitutive relationships can be written in many forms. For example, the mixed form describing one-dimensional downward water flow based on the water content, θ, and pressure head, ψ, is:

$$\frac{\partial\theta}{\partial t} = \frac{\partial}{\partial z}\left[K(\theta)\left[\frac{\partial\psi}{\partial z} + 1\right]\right] \quad [11]$$

In the preceding examples, the change in the volumetric water content was abrupt, taking one of two values, θ_{wet} or θ_i. In reality, the water content can assume any value between the residual water content and the porosity; the water content is a function of the capillary pressure and the antecedent conditions (hysteresis). If the water content varies smoothly across the wetting front, then the hydraulic conductivity will also take a continuous range of values. The nature of the dependence of the hydraulic conductivity on the water content controls the deviation of the shape of the wetting front from the idealized sharp boundary assumed above.

The following sections show examples of the infiltration process calculated using HYDRUS-2D, one of several available numerical models that solve the Richards equation for a range of boundary conditions and soil properties. The hydraulic properties for silt and sand were taken from the soil catalogue available through ROSETTA, which is included in HYDRUS-2D. These cases have been chosen to illustrate the influences of soil properties and water application rates on infiltration.

The Advance of a Wetting Front

For one-dimensional infiltration into a soil profile overlying a deep water table in response to a constant applied flux at the ground surface, the water content and the water pressure at the ground surface increase from their initial values to constant, higher values. These values are equal to the water content and pressure head at which the hydraulic conductivity of the soil equals the infiltration rate. The wetting front remains sharp because the hydraulic conductivity of the relatively dry medium below the wetting front is much lower than the hydraulic conductivity above the wetting front (solid lines in Figure 1a). As a result, a large gradient is needed across the wetting front to drive water into the soil. The pressure heads correspond, having the general shape of the solid lines in Figure 1b. The exact corresponding values of pressure head and water contents will depend on the soil-water characteristic curve, i.e., the relationship of θ to ψ.

For one-dimensional infiltration into a soil profile overlying a deep water table in response to a constant,

increased pressure head at the ground surface, the water content and the water pressure at the ground-surface increase from their initial values to constant, higher values. For comparison, consider a ground surface boundary condition that corresponds with the constant infiltration rate applied in creating Figure 1. As for a type II ground-surface boundary condition, the wetting front remains sharp because the hydraulic conductivity of the relatively dry medium below the wetting front is much lower than the hydraulic conductivity above the wetting front (dashed lines in Figure 1a). As a result, a large gradient is needed across the wetting front to drive water into the soil. The pressure heads correspond, having the general shape of the dashed lines in Figure 1b. The exact corresponding values of pressure head and water contents depend on the soil-water characteristic curve, i.e., the relationship of θ to ψ. Notice that the wetting front has advanced to a greater depth at any given time for the constant pressure head ground-surface boundary condition than for the corresponding constant flux boundary condition. This is due to an early high pressure head gradient, leading to high initial fluxes under the constant-head boundary condition. At late times, the wetting front advances at the same rate for both boundary conditions, leading to a constant distance between the solid and dashed lines on Figure 1.

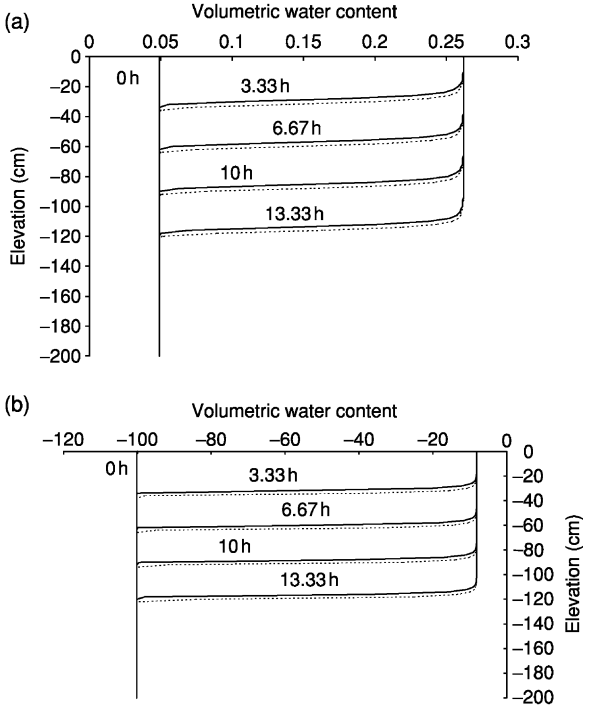


Figure 1 (a) Water-content and (b) pressure-head profiles after 3.3, 6.7, 10, 13.3, and 16.7 h of infiltration into a drained sand (−100-cm pressure head throughout). Solid lines show the response to a type II ground-surface boundary condition with a constant applied flux of 0.0005 cm s^{-1}. Dashed lines show the response to a type I ground-surface boundary condition with a constant applied pressure head of −8.08 cm.

The Effects of Infiltration Rate

For simplicity, the remaining examples will only consider infiltration in response to a constant applied flux. For these conditions, for a given soil, the water content behind the wetting front increases with increases in the rate of infiltration until full saturation is reached. The wetting front velocity has been shown (above) to advance at a rate that is proportional to the applied flux and inversely proportional to the water content behind the wetting front. The rate of advance of the wetting front increases with higher rates of water application. However, note that the difference in water pressure at the surface (Figure 2b) is small compared with the difference in water contents (Figure 2a) for the range of water application rates chosen.

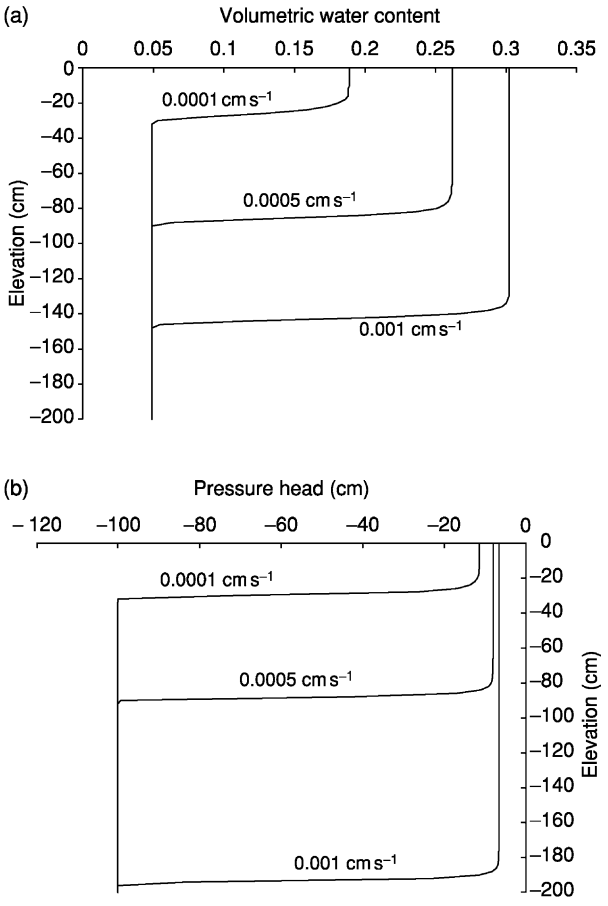


Figure 2 (a) Water-content and (b) pressure-head profiles after 10 h of infiltration into a drained sand at rates of 0.001, 0.0005, or 0.0001 cm s^{-1}.

The Effects of the Antecedent Water Content

As infiltration progresses, the downward-moving water continually displaces preexisting pore water and fills drained pores. For a sufficiently long profile that has no flow across the bottom boundary, the total volume of water added is equal to the total change in water stored in the profile. If the antecedent water content is high, the change in storage per unit length is small (Figure 3). As a result, the wetting front will progress deeper for a given total volume of added water.

The Effects of the Saturated Hydraulic Conductivity

If one medium has a higher saturated hydraulic conductivity than another and all other hydraulic properties are the same, then the first soil is able to transmit a given flux at a lower water content than the second soil (Figure 4). As a result, for a given rate of water application, the water content behind the wetting front is lower for the first, more conductive soil. To accommodate the same total change in water storage throughout the profile in response to a given volume of applied water, the wetting front moves deeper in the more conductive soil.

The Effects of Soil Texture

The texture of a soil strongly influences its hydraulic properties. For example, at low pressure heads, a fine-textured material tends to have a higher water content than a more coarse material. As a result, the antecedent water content may be higher in a silt or clay than in a sand. This can lead to a fine-textured soil having a higher hydraulic conductivity and a lower storage capacity than a coarse-textured soil for the same initial pressure head (Figure 5). At higher water contents, a coarse soil has a higher hydraulic conductivity than a fine soil. Therefore, at a given rate of water application, the water content behind the wetting front is lower for a sand than for a silt or clay. The balance of these effects for a given rate of water application and for given initial conditions determines the relative rate of advance of a wetting front through any two soils.

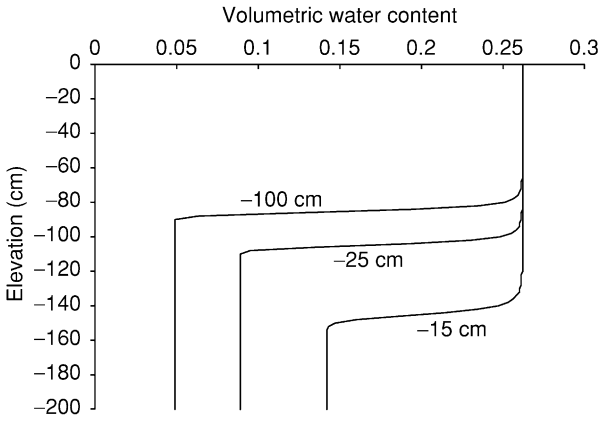


Figure 3 Water-content profiles after 10 h of infiltration at a rate of 0.0005 cm s^{-1} into a sand with an initial pressure head of −15, −25, or −100 cm.

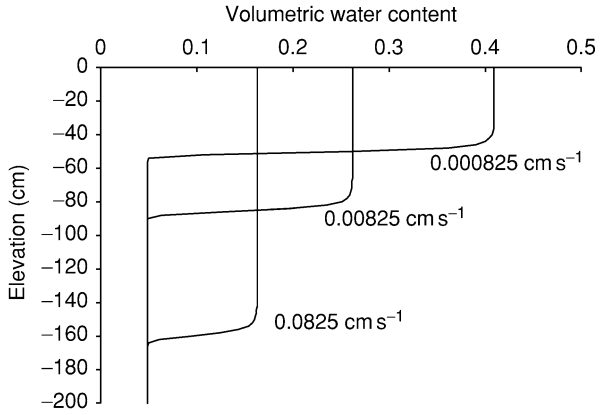


Figure 4 Water-content profiles after 10 h of infiltration at a rate of 0.0005 cm s^{-1} into a sand with a saturated hydraulic conductivity of 0.00086, 0.0086, or 0.086 cm s^{-1} and an initial pressure head of −100 cm.

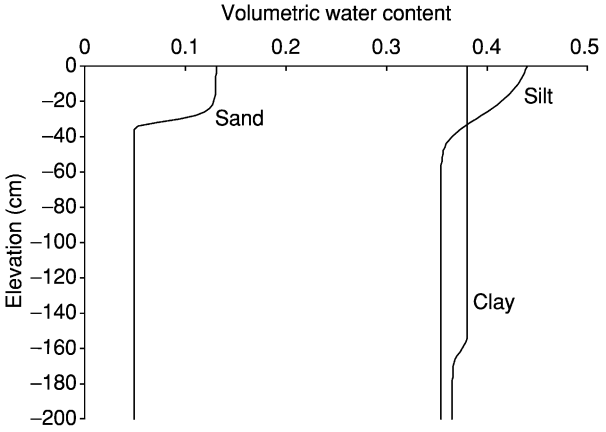


Figure 5 Water-content profiles after 44.4 h of infiltration at a rate of 0.000015 cm s^{-1} into a sand, silt, or clay at an initial pressure head of −100 cm.

The Effects of Gravity

The preceding discussions have considered only vertical infiltration. However, water can also move horizontally across vertical soil surfaces (e.g., the walls of a water-filled surface depression). All of the influences described above for vertical flow have

the same effect on horizontal advance in response to water applications controlled by the rate of application at a soil surface. Definite differences arise when water is applied at the soil surface at a constant pressure head (type I boundary condition). In that case, horizontal flow responds only to the pressure gradients caused by the surface-water pressure head, while vertical flow responds to this gradient and to gravity. After a short time, when the pressure-head gradient is high across the surface, the rates of horizontal and vertical flow are nearly equal. However, later the pressure-head gradient approaches zero while the gravitational gradient is always 1. As a result, horizontal flow decreases toward zero while vertical flow reaches a steady nonzero rate. This difference in the horizontal and vertical flow in response to the same applied pressure-head is critical to an understanding of multidimensional infiltration.

Two-Dimensional Infiltration

The preceding discussions have focused on one-dimensional, vertical infiltration. While this simplified representation is applicable to many common conditions (e.g., rainfall or sprinkler irrigation over a flat field), other conditions require a consideration of multidimensional flow. Multidimensional flow can be understood as a balance between horizontal and vertical flow. For instance, a wetting front will proceed farther downward if flow is one-dimensional vertical, than if flow is multidimensional, because in the latter case the amount of water moving vertically downward will be less as a result of water 'lost' to horizontal flow. Figure 6 shows the water-content profiles beneath a point of constant infiltration for both one-dimensional (solid lines) and two-dimensional (dashed lines) flow regimes. Regardless of the soil texture, the wetting front moves deeper under one-dimensional flow conditions. This difference is more pronounced for the finer-textured soils, because their relatively low hydraulic conductivities preclude rapid, downward flow. As a result, the effects of gravity are minimized, and horizontal and vertical flow diverge less slowly. A measure of the effects of horizontal flow on the downward advance of the wetting front can be determined from the total change in water storage directly beneath the point of water application for one- and two-dimensional simulations. For the examples shown in Figure 6, the change in storage directly beneath the point of water application for the sand undergoing two-dimensional flow is 74% of that for the one-dimensional case; the silt and clay show 51% and 32%, respectively. This emphasizes the relatively large impact of horizontal flow on finer-grained, less-permeable media.

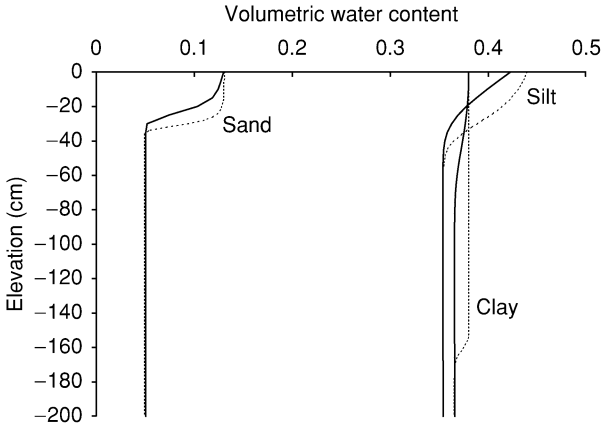


Figure 6 Water-content profiles after 44.4 h of infiltration at a rate of 0.000 015 cm s^{-1} into a sand, silt, or clay at an initial pressure head of −100 cm. Dashed lines show the response calculated for one-dimensional flow. Solid lines show the response directly beneath the point of water application when flow is two-dimensional.

There is a further difference between one-dimensional and multidimensional flow. If flow is one-dimensional, the wetting front continues to advance downward as long as water is applied at the ground surface. Specifically, because the water content behind the wetting front is nearly constant, the flux across the wetting front is equal to the flux across the ground surface. The wetted region can be seen to advance due to this flux across the wetting front, which results in water filling pores that were previously filled with air. In contrast, under multidimensional flow, the wetted region continues to expand in multiple directions. As a result, the surface area of the wetted region continues to increase. The area of the surface bounding the wetted region increases in direct proportion to the flux across the boundary of the wetted region. As a result, at long times, the area of the boundary of the wetted region becomes so large that the flux across the boundary approaches zero at every location and the size of the wetted region becomes essentially constant in time. This gives the unusual result that there can be a constant water content distribution with both depth and time at long times during multidimensional infiltration.

Further Considerations

The preceding discussions have been limited to infiltration into a homogeneous, isotropic porous medium overlying a deep water table. Infiltration in the presence of a shallow water table obeys the same general principles of Darcy's law and the Richards equation. When a shallow water table is present, the moisture profiles (and pressure profiles)

must merge vertically with the saturated profile and capillary fringe. More complex aspects of infiltration arise due to the effects of horizontal and subhorizontal layering, capillary barriers, preferential flow, unstable finger-flow, anisotropy, dual porosity, changes in hydraulic properties due to clogging with time, and the effects of air resistance during infiltration.

See also: **Compaction**; **Crusts:** Structural; **Hydrodynamics in Soils**; **Water Cycle**

Further Reading

Green WH and Ampt GA (1911) Studies on soil physics. I. Flow of air and water through soils. *Journal of Agricultural Science* 4: 1–24.

Hillel D (1998) *Environmental Soil Physics*. New York: Academic Press.

Jury WA, Gardner WR, and Gardner WH (1991) *Soil Physics*, 5th edn. New York: John Wiley.

Philip JR (1969) Theory of infiltration. *Advances in Hydroscience* 5: 215–305.

Richards LA (1931) Capillary conduction of liquid through porous media. *Physics* 1: 318–333.

Schaap MG, Leij FJ, and van Genuchten MT (2001) ROSETTA: a computer program for estimating soil hydraulic parameters with hierarchical pedotransfer functions. *Journal of Hydrology* 251: 163–176.

Simunek J, Šejna M, and van Genuchten MT (1999) The HYDRUS-2D software package for simulating two-dimensional movement of water, heat, and multiple solutes in variably saturated media. Version 2.0, IGWMC-TPS-53. International Ground Water Modeling Center. Golden, CO: Colorado School of Mines.

Smith RW (2002) *Infiltration Theory for Hydrologic Applications*. Water Resources Monograph 15. Washington, DC: American Geophysical Union.

Warrick AW (2003) *Soil Water Dynamics*. New York: Oxford University Press.

Infrared Spectroscopy *See* **Fourier Transform Infrared Spectroscopy**

IRON NUTRITION

K Mengel and H Kosegarten, Justus Liebig University, Giessen, Germany

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Iron (Fe) is an indispensable element for all living organisms, in which it is present in many enzymes. Some of the most important enzymic reactions depend on changes between the divalent and trivalent Fe forms to enable the transfer of electrons as occurs in the respiratory and photosynthetic electron (e^-) transport chains. Fe taken up from the soil by plants is for many animal species the exclusive nutritional source of Fe, and in vegetables and fruits it is also of high importance in the human diet. Apart from organic soils containing virtually no inorganic soil matter, Fe is present in soils in abundant quantities. Its availability to plants, however, depends much on ecologic conditions. In aerobic soils its concentration in the soil solution is extremely low and decreases with an increase in soil pH. Of a still higher importance for Fe availability to plant roots is the soil microbial activity by which siderophores are produced which dissolve mineral Fe. Under anaerobic soil conditions, mineral Fe(III) is reduced and the Fe^{2+} produced is soluble and may attain toxic concentrations under soil flooding. Fe nutrition of crops may be insufficient in calcareous and alkaline soils caused by an inhibition of Fe(III) reduction in the root apoplast. Iron deficiency affects crop growth and quality.

Iron in the Soil and its Acquisition by Plants

Of all the chemical elements composing the Earth's crust, Fe is present fourth-highest by weight, with mean concentration of approximately 50 mg kg^{-1}. The greatest part of soil Fe occurs in the crystal lattices of numerous primary soil minerals. In 2:1 phyllosilicates such as mica, smectite, and illite, it is located in the center of the octahedra in divalent and trivalent form, depending on the redox conditions of the soil. Weathering of these primary Fe-containing minerals

results in the production of various Fe oxides such as goethite, hematite, and ferrihydrate. These oxides are the most important direct Fe sources for plants and soil microorganisms and their solubility decreases in the following sequence:

$$\begin{aligned}&Fe(OH)_3 \text{ amorphous} > Fe(OH)_3\\ &\quad\text{in soils(semiamorphous)} > \gamma\text{-}Fe_2O_3 \text{ maghemite}\\ &\quad> \gamma\text{-FeOOH lepidocrocite} > \alpha\text{-}Fe_2O_3 \text{ hematite}\\ &\quad> \alpha\text{-FeOOH goethite}\end{aligned}$$

Goethite is the most abundant Fe mineral in soils. Another Fe oxide of interest is ferrihydrite ($Fe_2O_3 \cdot 9H_2O$). Heterotrophic soil microorganisms are able to oxidize the organic component of Fe chelates, during which process the Fe^{3+} released is hydrolyzed and precipitated:

$$Fe^{3+} + 3H_2O \leftrightarrow Fe^{III}(OH)_3 + 3H^+$$

The resulting ferrihydroxide then gives rise to other Fe oxides, particularly ferrihydrite. Under aerobic soil conditions, the solublity of all Fe(III) oxides in water is extremely low and increases as soil pH decreases. Accordingly a pH as low as 3 is required for inorganic Fe solubility sufficient to cover plant demand.

Since most higher plants that grow on aerobic soils do so over a pH range well above 3 and are usually adequately supplied with iron, it can be concluded that organic, water-soluble Fe compounds (siderophores) must contribute significantly to the Fe supply of plants from soil. Total soluble Fe concentration in soils should be at least 1 μmol l^{-1}. As shown below in more detail, numerous microbial species and also some higher plant species produce organic molecules (siderophores) that chelate the Fe of insoluble inorganic Fe compounds, thus forming soluble Fe compounds (Fe siderophores), which are indispensable for the Fe supply of plants in aerobic soils.

The situation is quite different under anaerobic soil conditions, where soil microorganisms use the Fe(III) of the various Fe(III) oxides as a terminal e^- acceptor in the respiratory chain as a substitute for O_2:

$$Fe^{III}(OH)_3 + e^- \rightarrow Fe^{2+} + 3OH^-$$

The Fe^{2+} is water-soluble and may accumulate to toxic concentrations, especially in flooded soils. Plant toxicity begins at redox potentials of less than 0.5 V.

In aerobic soils, Fe availability to plants depends crucially on the formation of Fe-chelating organic molecules, so-called siderophores. These are produced by soil bacteria, soil fungi, and some plant species. More than 100 different siderophores are known that are produced by soil fungi and soil bacteria. Numerous siderophores are derived from hydroxamate, which binds Fe(III) with a coordinative and an ionic bond. These siderophores are water-soluble and stable over a wide pH range and thus also stable in the alkaline pH range characteristic of calcareous and alkaline soils, where the solubility of Fe oxides is extremely low. Grasses are able to synthesize siderophores in their roots (phytosiderophores). Phytosiderophores such as mugineic acid and its derivatives avenic acid, 3-hydroxymugineic acid, and distichonic acid are nonprotein amino acids that are capable of binding ferric Fe. As shown in **Figure 1**, they have six functional groups by which they bind Fe(III) in a hexadentate form, with three carboxyl groups, two N atoms, and one OH. Phytosiderophores are secreted by root tips a few millimeters behind the apex. Secretion of phytosiderophores depends on the Fe nutritional status of plants and is particularly high if plants are suffering from an insufficient Fe supply.

Phytosiderophores thus released by roots react with the Fe(III) present in soil minerals to form an Fe(III) siderophore complex, which is soluble in the soil solution and can be translocated by mass flow or diffusion toward the plant roots. There is no firm evidence yet that the Fe(III) siderophore complex is translocated across the plasmalemma of the root cortical cells into

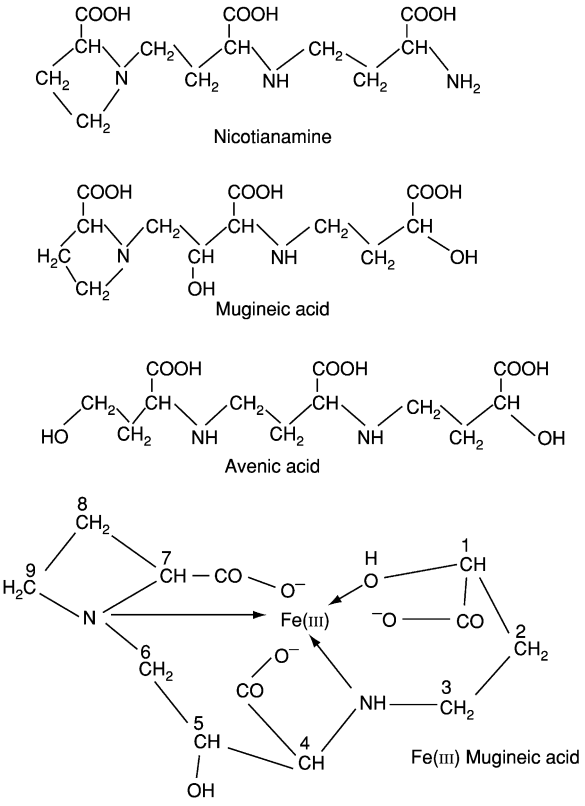


Figure 1 Important siderophores and the hexadentate Fe(III) mugineic acid complex.

the cytosol, because until now neither the transporter nor the genes coding for such a system have been found. Due to their capability for producing siderophores, grasses are better equipped for mining soils for Fe than are dicotyledenous species. Nicotianamine, shown in Figure 1, chelates Fe^{2+}. This complex is of no importance for the mobility of Fe in soils but is essential for the transport of Fe in the cytosol.

One may distinguish between two strategies by which plant species are able to cope with an insufficient Fe supply. In strategy-I plants, the dicotyledons respond by increased H^+ secretion from the roots, thus mobilizing soil Fe and depressing apoplastic pH, promoting Fe uptake. The grasses (strategy-II species) respond by enhanced release of phytosiderophores from the roots. Monocotyledonous species differ in their potential to release phytosiderophores and similarly there may be differences between cultivars and genotypes in their abilities to synthesize and release siderophores. Siderophores including phytosiderophores are decomposed by soil microorganisms. There is direct evidence that microbial activity in soils is indispensable for the acquisition of Fe by dicotyledonous species and is also of some importance for grasses.

Both dicotyledons and grasses form rhizodermal transfer cells under the conditions of insufficient Fe supply. These cells are characterized by invaginations and thus by an enlargement of the plasma membrane, and therefore by an increase in Fe-uptake systems, including H^+ pumps. The apoplastic pH is of crucial importance for Fe uptake. The increased H^+ secretion by roots under conditions of insufficient Fe promotes Fe uptake as a consequence of the decrease in the apoplastic pH. Proton secretion by roots may also promote the solubilization of inorganic soil Fe. In most cases, however, H^+ ions diffusing from the roots into the rhizosphere are quickly neutralized. This is particularly true for calcareous and alkaline soils high in pH.

There exist various methods for the determination of Fe bioavailability. One of the most used techniques is that of soil extraction by diethylenetriaminepentaacetic acid (DTPA). This extractant reacts with soluble and exchangeable Fe, as well as with inorganic insoluble Fe such as Fe oxides. The problem with all Fe extraction tests is that the bioavailability of soil Fe depends more on the presence and activity of siderophores than on the physicochemical solubility and extractability of soil Fe.

Iron Uptake, Long-Distance Transport in the Plant, and Storage in the Cell

Iron has a unique uptake system, as the transport across the plasma membrane is associated with the reduction of Fe(III). Uptake takes place in the roots and in the mesophyll. It is supposed that the principal uptake mechanism of both is the same; the roots, however, are supplied by a wide range of different siderophores with chelated Fe(III). In the leaves it is mainly Fe(III) dicitrate which is translocated via the xylem, through the larger and minor vein system, and the free space of cell walls (apoplast) into the leaf tissues, where the Fe is taken up into the cells. The term 'uptake' means primarily the transport of Fe across the plasma membrane. Fe(III) reductase and the Fe^{2+} channel are essential components of the uptake system, as shown in Figure 2. Fe(III) reductase is a transmembrane enzyme, contacting with one part the cytosol and with the other the apoplast. The cytosol supplies the reductase with reducing power in the form of reduced nicotinamide adenine dinucleotide phosphate (NAD(P)H). The reducing enzyme activity is highly sensitive to the apoplastic pH and decreases as the pH increases, as shown by direct apoplastic pH measurements in sunflower leaves. Generally the optimum apoplastic pH range is provided by the activity of the plasmalemma proton pump (Figure 2). Processes that change the apoplastic pH, such as the H^+–nitrate cotransport across the plasmalemma, raise the apoplastic pH and thus depress Fe(III) reduction and Fe^{2+} uptake. Highest reduction rates were found at pH <5.44. Basically the same relationship between pH in the apoplast and the Fe(III) reduction is supposed to exist in roots. Here it is particularly HCO_3^- that increases the apoplastic pH and thus restricts or even blocks Fe(III) reduction. Monohydrogen carbonate (HCO_3^-) is the anion species dominant in soils high in pH, such as calcareous and alkaline soils, and these are the soils that are prone to Fe deficiency. As shown in Figure 2, the apoplastic pH is dependent on the H^+ pump activity. Accordingly cultivars resistant to Fe chlorosis release more H^+ from the roots than do the susceptible ones.

Iron uptake of roots is restricted to the root tips, and the Fe uptake systems are located in the plasma membranes of the epidermal cell layer. From this

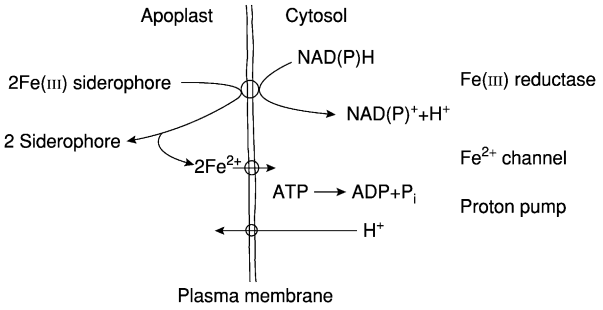


Figure 2 Scheme of Fe^{2+} uptake across the plasmalemma.

follows that the surface area from which Fe is taken up is small in comparison with the total surface area exposed to the soil by young roots from which most other plant nutrients are absorbed. There are indications that in leaves the Fe uptake systems are also exclusively located in the meristematic tissues. The Fe^{2+} produced in the apoplast passes through a specific channel of the plasma membrane into the cytosol. This takes place by diffusion and thus depends on the electrochemical gradient for Fe^{2+} across the plasmalemma. In flooded soils, where a surplus of Fe^{2+} may be produced by microbial respiratory activity, relatively high rates of Fe^{2+} may diffuse into the cytosol and thereby induce Fe toxicity. Toxicity may arise, because the excess Fe^{2+} easily reacts with O_2, forming the highly toxic superoxide radical O_2^-:

$$Fe^{2+} + O_2 \rightarrow O_2^- + Fe^{3+}$$

Under normal conditions the cytosolic Fe^{2+} concentration is maintained at a low level by the translocation of Fe^{2+} into the plastids, where it is oxidized by O_2, as shown in Figure 3. The resulting superoxide radical is detoxified by superoxide dismutase. The Fe^{3+} produced in the oxidation process reacts with water and thus forms the Fe(III) hydroxide:

$$Fe^{3+} + 3H_2O \rightarrow Fe^{III}(OH)_3 + 3H^+$$

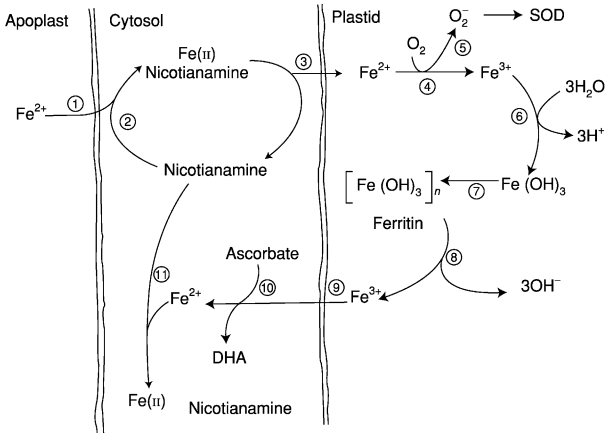


Figure 3 Putative scheme of Fe uptake into the cytosol and into the plastid, Fe incorporation into ferritin, and Fe release from ferritin into the cytosol:

1. Fe uptake across the plasmalemma;
2. Fe binding by nicotianamine;
3. Fe^{2+} transport across the plastid membrane;
4. Fe^{2+} oxidation to Fe^{3+};
5. Formation of the superoxide radical;
6. Formation of $Fe(OH)_3$;
7. Incorporation of $Fe(OH)_3$ into ferritin;
8. Release of Fe^{3+} by ferritin;
9. Transport of Fe^{3+} across the plastid membrane;
10. Reduction of Fe^{3+} by ascorbate to dehydroxyascorbate (DHA);
11. Binding of Fe^{2+} by nicotianamine.

The Fe(III) hydroxide produced is the basic unit for the formation of ferritin, which is an important Fe storage form in plants, animals, and microorganisms. The core of ferritin consists mainly of Fe(III) hydroxide and is surrounded by a protein layer, apoferritin. The core comprises several thousand Fe(III) atoms. This Fe store is characterized by the storage and the release of Fe according to physiological demand and thus represents an Fe buffer. In higher plants, chloroplasts may store remarkably high quantities of Fe so that their Fe concentrations may be 5–10 times higher than those in the corresponding leaf tissue.

Under conditions of physiological Fe demand, Fe^{3+} diffusing out of the plastid is reduced by ascorbate, and the resulting Fe^{2+} forms an Fe(II) chelate with nicotianamine (Figure 3).

Nicotianamine is a nonprotein amino acid which is structurally related to siderophores (Figure 1) and forms stable complexes with Fe^{2+}. Fe–nicotianamine is not involved in membrane transport and therefore cannot be excreted into the xylem. Fe^{2+} nicotianamine is essential for the symplastic transport from cell to cell. Fe–nicotianamine is phloem-mobile. Until now, however, it has not been known whether symplastic phloem-loading of Fe plays a major role, since older leaves hardly provide sufficient Fe to younger leaves. Indeed it is for this reason that Fe-deficiency symptoms first appear in younger leaves.

Biochemical Functions of Fe in Plants

In most of the Fe-containing enzymes, the functional mechanism is based on the valency change of Fe. The enzyme with the highest negative standard redox potential is ferredoxin ($E'_o = -0.43$ V). Ferredoxins have a low molecular weight of approximately 6–12 kDa. They are stable 2Fe–2S proteins, with a prosthetic group that contains two Fe and two inorganic S atoms, which bridge the two Fe atoms. Fe is further bound by two bonds to cysteine residues. A soluble form of ferredoxin exists as well as a form integrated in the thyalkoid membrane, which is also known as the 2Fe–2S complex. The redox reaction is a one-e^- step process, which means that in the reaction only one and not two e^- are released or taken up during electron transport. In the oxidized form, both Fe constituents are present as Fe(III), and in the reduced form only one Fe is present as Fe(II).

During photosynthesis the soluble ferredoxin present in the chloroplast receives the e^- directly from the membrane-bound ferredoxin of photosystem I and transfers it to other important enzymes related to nutrient assimilation, such as $NADP^+$ (CO_2 assimilation), nitrite reductase (N assimilation), sulfate reductase (S assimilation), heme proteins (cyclic

photophosphorylation), as well as to O_2. Reduction of O_2 by ferredoxin takes place under high light intensity associated with a surplus of photosynthetic e^-. In this case the superoxide radical is formed, which, if not detoxified completely by superoxide dismutase, may lead to the damage of photosynthetic pigments, known as photoxidation. Ferredoxin is also essential for the N_2 assimilation of free-living and symbiotically living N_2-fixing bacteria. The ferredoxin provides the e^- for the N_2 reduction (nitrogenase). Besides the binuclear Fe–S complexes, tetranuclear Fe–S complexes (4Fe–4S) also exist that catalyze redox reactions in an analogous way. The core of nitrogenase consists of an Fe4–S4 cluster and an Fe3–Mo–S4 cluster. In the latter, one Fe is substituted by one Mo.

The enzymes discussed above belong to the nonheme Fe group. The group of heme Fe enzymes is characterized by heme as the prosthetic group with Fe in the center of the tetrapyrrole ring, as occurs in the cytochromes which mediate e^- transport in the photosynthetic and respiratory e^- transport chains. Other heme enzymes include catalase and peroxidase.

Some heme enzymes do not react by a valency change but function as O_2 transporters such as myoglobin and hemoglobin in animal organisms and leghemoglobin in the symbiotic nitrogen fixation of N_2-fixing bacteria and their host plants. The leghemoglobin brings about a controlled delivery of O_2 to the N_2-fixing bacteroid. This is essential because the nitrogenase is very sensitive to O_2.

The pathway for the synthesis of chlorophyll requires Fe, therefore insufficient Fe supply frequently results in chlorosis, which means a lack of chlorophyll. Ribonucleotide reductase is another Fe-containing enzyme of universal importance, because it catalyzes the reduction of ribonucleotide to deoxyribonucleotide. These are the building blocks for DNA. Insufficient DNA affects cell division and thus growth.

Lime-Induced Chlorosis

Absolute Fe deficiency in plants grown in the field is very rare. However, it may occur on organic soils with a very low proportion of mineral soil matter. Lime-induced chlorosis occurs almost exclusively on calcareous and alkaline soils. It has been recognized since the nineteenth century and may cause considerable loss in crop production. Deficiency symptoms are initially the yellowing of intercostal areas in the youngest leaves, which proceeds until the total leaf is almost golden-yellow except for small green stripes along the ribs. In extreme cases the whole foliage of a plant may become yellow (chlorotic).

Calcareous and alkaline soils have a high potential to produce HCO_3^-, which may diffuse into the root apoplast and neutralize protons released by the plasmalemma H^+ pump, thus increasing the apoplastic pH, a process associated with a decrease in Fe(III) reductase activity and Fe^{2+} uptake. This interpretation is in agreement with earlier findings in which 24 soils were investigated differing in carbonate concentration in relation to chlorosis of sorghum. Of the various parameters tested, the intensity of visual chlorosis in sorghum leaves was most closely correlated ($P<0.001$) with soil pH in the KCl extract and with the alkalinity of soils measured as the concentration of HCO_3^-, which was in the range of 1.0–4.5 mol l^{-1} in the soil extract. Iron uptake of plants was negatively correlated ($P < 0.001$) with the pH in the KCl extract and with the HCO_3^- concentration in soils. The latter also depends on the water saturation of soils, high saturation increasing the partial pressure of CO_2 in soils and thus the dissolution of CO_2 and the production of HCO_3^-. Under such highly saturating conditions, chlorosis may occur and is known as 'bad-weather chlorosis.'

Frequently, chlorotic plants have high Fe concentrations in their roots. It has been suggested that this high Fe concentration is due to adhering soil particles on the root surface. However, plants grown in solution culture where no adhering soil particles are present may also have Fe concentrations in the roots ranging from 300 to 600 mg kg^{-1} dry weight, which is several times higher than the Fe concentration in leaves. This observation shows that plants are capable of storing high amounts of Fe in their roots. Such high Fe concentrations in roots occur when the Fe translocation from roots to shoots is restricted. By means of EDX analysis, it has been shown that the Fe concentration in the radial cell walls of the first cell layer of the root cortex parenchyma of maize grown in a calcareous soil (Mollisol) is more than 10 times higher than that of plants grown in an acid organic soil (Histosol). In the radial cell wall of the third cell layer, Fe concentrations are lower and there are no major differences between the two soils. Interestingly, the DTPA-extractable Fe concentration in the Histosol amounted to 499 and in the Mollisol to 4.2 mg Fe kg^{-1} soil. Nevertheless, Fe accumulates in the root apoplast of the plants grown on the calcareous soil. Since the Fe is located in the radial cell wall, this Fe accumulation cannot be caused by soil contamination. This phenomenon of Fe accumulation is of particular interest, because selective Fe^{2+} channels are exclusively located in the plasmalemma of the root epidermis. If the root apoplastic pH is high because of abundant HCO_3^-, $Fe^{(III)}$ reduction and Fe uptake are restricted. Finally Fe siderophores are

decomposed by soil microorganisms and the resulting Fe^{3+} is bound to cell walls. Toxic Al concentrations may also cause extremely high Fe concentrations in roots, whereas the translocation of Fe into the shoot is restricted. Since toxic cationic Al species inhibit the plasmalemma H^+ pump, the resulting high apoplastic pH presumably depresses Fe(III) reduction and hence Fe^{2+} uptake.

A further characteristic of Fe chlorosis is that the intensity of the chlorosis, i.e., the chlorophyll concentration, is not correlated with the Fe concentration in leaves. Leaves with the same Fe concentration may show chlorosis whereas others of the same species may not. Based on pH measurements and Fe(III) reduction in leaves of sunflowers, a pH >5.5 in the leaf apoplast depresses Fe(III) reduction and thus Fe uptake into the cytosol. Hence in chlorotic leaves with a relatively high Fe concentration a substantial part of this Fe may be located in the leaf apoplast and therefore is of no physiological use. Nitrate nutrition promoting leaf chlorosis increases the apoplastic pH in leaves because of H^+–nitrate cotransport. Accordingly, an acid foliar spray has led to a regreening of chlorotic leaves without increasing the leaf Fe concentration. In parallel investigations it has been shown that citric acid application decreases the apoplastic pH by approximately 0.5 pH units. This finding is in accord with earlier investigations in which a different *in situ* pattern was found for Fe between chlorotic and green leaves by Mössbauer spectrometry. In chlorotic leaves obtained with nitrate nutrition, Fe was present exclusively as Fe(III), whereas, in green leaves obtained with ammonium nutrition, more than 20% of the Fe occurred as Fe(II) and the rest was Fe(III).

In solution culture experiments with sunflower, Fe chlorosis is found when plants are exclusively supplied with nitrate. In the presence of ammonium in the nutrient solution, however, no chlorosis appears. Since ammonium is taken up preferentially relative to nitrate, nitrate may be translocated to the leaves only at minute rates, so that a major pH increase in the leaf apoplast due to H^+–nitrate cotransport will not appear. The phenomenon of leaf chlorosis being associated with high Fe concentration in leaves is only found on soils with a high pH. On such soils ammonium does not accumulate in the soil, because the equilibrium $NH_3 + H^+ \leftrightarrow NH_4^+$ is shifted to the ammonia side of the equation and ammonia is released from the soil by evaporation. Hence, under such conditions, plants may be exclusively supplied with nitrate, which may lead to a high apoplastic leaf pH, restricting Fe(III) reduction and leading to an accumulation of Fe in the leaf apoplast. Such conditions may especially appear under dry weather, which favors the volatilization of NH_3 and CO_2 from the soil. The CO_2 release leads to a decrease in the soil HCO_3^- concentration and hence benefits the Fe^{2+} uptake by roots; the lack of NH_4^+ promotes nitrate uptake of the roots and thus the nitrate transport to the leaves, with its adverse effect on Fe uptake from the leaf apoplast into the cytosol. Under such conditions Fe may accumulate in the leaf apoplast the longer such conditions prevail. Therefore high leaf Fe concentrations in chlorotic leaves appear mainly in older leaves and later in the season. Iron accumulation is enforced by a depression in leaf growth because of a lack in Fe in the cytosol. Iron deficiency on calcareous soils not only is evident in leaf chlorosis, but may also result in a serious growth reduction without visible symptoms of chlorosis.

In natural vegetation there is hardly any evidence of plant species with lime-induced chlorosis, since genotypes have been developed during evolution which have adapted to the environmental conditions, as shown for *Lathyrus* species. *Lathyrus montanus* is susceptible to Fe chlorosis, whereas *L. pratense*, evolved on calcareous soils, is not. This is also true for *Vitis* species, of which rootstocks are used in wine production. Rootstocks of *Vitis labrusca* and *V. riparia*, both evolved in the north and northeast of the USA, are susceptible to chlorosis; but *V. berlandieri*, which evolved on calcareous soils in Texas and Mexico, is resistant. The physiological processes relating to the chlorosis resistance are not yet completely understood. The rootstock of the resistant vine cultivar Faber releases more than twice as much H^+ by its roots than the chlorosis-susceptible cultivar Huxel. This H^+ release reflects the apoplastic pH in roots and therefore also the conditions for Fe(III) reduction as considered above. Even wheat cultivars, which supposedly acquire Fe mainly by phytosiderophores, differ considerably in their resistance to chlorosis. When sunflowers and *Vicia faba* were grown under the same conditions in solution-culture with nitrate $+HCO_3^-$ the sunflowers became chlorotic, but *V. faba* did not. This behavior is related to the effect of nitrate $+HCO_3^-$ on the pH of the xylem sap, which was raised in sunflowers but not in *Vicia*. Supposedly *Vicia* must pump relatively high rates of H^+ into the xylem sap so that the pH remains low even in the presence of nitrate $+HCO_3^-$ supply.

Iron Deficiency and Toxicity

Growing dicotyledon species in a sterile soil results in growth-retarded plants with low Fe concentrations in the leaves. With sunflowers, typical leaf chlorosis appears. Rape plants do not develop chlorosis, but

growth is depressed and the leaves remain very small. The addition of an Fe chelate to the soil of the sterile grown plants produces vigorous growth, proving that the growth restriction is attributable to Fe deficiency. Also, vines may suffer from an inadequate supply of Fe without showing chlorotic symptoms. Chlorosis is a very spectacular symptom of Fe deficiency, but with only symptoms of reduction in growth, in plants of normal habit and green leaves, it is hardly possible to recognize a lack of Fe as a cause of deficiency. Supposedly the poor growth is caused by an insufficent synthesis of deoxyribonucleotide and thus restricts deoxyribonucleic acid (DNA) synthesis.

Plants deficient in Fe have reduced activities of catalase, superoxide dismutase, and ascorbic acid oxidase. Iron deficiency causes a disturbance of thyalkoid formation in chloroplasts and thus affects photosynthetic activity. Iron toxicity in plants may occur on flooded soils because of a high Fe^{2+} concentration so that Fe^{2+} diffuses uncontrolled into the cytosol. Symptoms of Fe toxicity in rice are known as 'bronzing,' which in the beginning is characterized by tiny brown spots on the leaves which later develop into a uniform brown color. The Fe concentration in such leaves is high and in the range of 300–1000 mg $Fe\,kg^{-1}$ dry matter. Iron toxicity in rice is known in various rice-growing areas and occurs especially on heavy soils.

Iron Application

Curing Fe chlorosis of plants grown on calcareous or alkaline soils may present difficulties, since chlorosis in most cases is a physiological deficiency and roots or even leaves may be rich in Fe but still suffering from insufficient Fe in the cytoplasm. Generally Fe chelates are supplied as a soil fertilizer or as a foliar spray. The most important Fe chelate in this respect is EDDHA (ethylenediaminedi(*o*-hydroxyphenylacetic)acid), which is stable over a pH range of 4–10. For tree crops a foliar spray may cure the deficiency.

Soil applications of mixtures of Fe salts with organic matter such as manure, compost, peat, and sewage sludge may help to cure Fe chlorosis of crops. From the salts, vivianite ($Fe_3^{II}(PO_4)_2$) is of particular interest, as its bivalent Fe can be directly taken up by roots. The problem, however, is that in aerobic soils Fe^{2+} is quickly oxidized to Fe^{3+} and as such reacts immediately with water to $Fe(III)(OH)_3$, which precipitates.

List of Technical Nomenclature

E'_o	Standard redox potential (volts)

Further Reading

Abadia J (ed.) (1995) *Iron Nutrition in Soils and Plants*. Dordrecht, the Netherlands: Kluwer Academic.

Becker R, Grün M, and Scholz G (1992) Nicotianamine and the distribution of iron in the apoplasm and symplasm of tomato (*Lycopersicon esculentum* Mill.). *Planta* 187: 48–52.

Chen Y and Barak P (1982) Iron nutrition of plants in calcareous soils. *Advances in Agronomy* 25: 217–240.

Chen Y and Hadar Y (eds) (1991) *Iron Nutrition and Interactions in Plants*. Dordrecht, the Netherlands: Kluwer Academic.

Crowley DE, Wang YC, Reid CPP, and Szaniszlo PJ (1991) Mechanism of iron acqusition from siderophores by microorganisms and plants. In: Chen Y and Hadar Y (eds) *Iron Nutrition and Interactions in Plants*, pp. 213–232. Dordrecht, the Netherlands: Kluwer Academic.

Kosegarten H and Koyro HW (2001) Apoplastic accumulation of iron in the epidermis of maize (*Zea mays*) roots grown in calcareous soil. *Physiologia Plantarum* 113: 515–522.

Kosegarten H, Hoffmann B, and Mengel K (1999) Apoplastic pH and Fe^{3+} reduction in intact sunflower leaves. *Plant Physiology* 121: 1069–1079.

Lindsay W (1991) Iron oxide solubilization by organic matter and its effect on iron availability. In: Chen Y and Hadar Y (eds) *Iron Nutrition and Interactions in Plants*, pp. 29–36. Dordrecht, the Netherlands: Kluwer Academic.

Loeppert RH and Hallmark CT (1985) Indigenous soil properties influencing the availability of iron in calcareous soils. *Soil Science Society of America Journal* 597–603.

Loeppert RH and Inskeep WP (1996) Iron. In: Bigham JM (ed.) *Methods of Soil Analysis, part 3, Chemical Methods*, pp. 639–664. Madison, WI: Soil Science Society of America.

Machold O, Meisel W, and Schnorr H (1968) Bestimmung der Bindungsform des Eisens in Blättern durch Mössbauer-Spektrometrie. *Naturwissenschaften* 10: 499–500.

Marschner H, Treeby M, and Römheld V (1989) Role of root induced changes in the rhizosphere, for iron acquisition in higher plants. *Zeitschrift für Pflanzenernährung und Bodenkunde* 152: 197–204.

Mengel K (1994) Iron availability in plant tissues – iron chlorosis on calcareous soils. *Plant and Soil* 165: 275–283.

Neilands JB and Leong SA (1986) Siderophores in relation to plant growth and disease. *Annual Review of Plant Physiology* 37: 187–208.

Römheld V and Marschner H (1986) Mobilization of iron in the rhizosphere of different plant species. *Advances in Plant Nutrition* 2: 155–204.

Rroco E, Kosegarten H, Harzay F, Imani J, and Mengel K (2003) The importance of soil microbial activity for the supply of iron to millet and rape. *European Journal of Agronomy* 19: 487–493.

Schwertmann U and Taylor RM (1989) Iron oxides. In: Dixon JB and Weed SB (eds) *Minerals in Soil Environments*, pp. 379–438. Madison, WI: Soil Science Society of America.

Tagliavini M and Rombola AD (2001) Iron deficiency and chlorosis in orchard and vineyard ecosystems. *European Journal of Agronomy* 15: 71–92.

IRRIGATION

Contents
Environmental Effects
Methods

Environmental Effects

S Topcu and C Kirda, University of Cukurova, Adana, Turkey

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

The delta plains of the Euphrates and the Tigris rivers in Mesopotamia, and the Nile in Egypt have witnessed numerous civilizations. With the early practice of irrigated agriculture, these soils have offered a prosperous life to the people of the region. Archeological remains of irrigation structures such as barrages, channels, and deep wells date back to as early as 4000–6000 BC. The real advancement of irrigated agriculture, however, started in the twentieth century with the development of large irrigation projects in Egypt, Pakistan, India, and the USA. Irrigation contributed significantly to the high crop yields achieved following the Green Revolution in the 1960s, which is commonly attributed to breeding of high-yielding crop cultivars and use of fertilizers and pesticides.

The demand for fresh water increased fourfold from 1950 to 1960 with the implementation of new irrigation projects and the increase in municipal and industrial water demands. Irrigated agriculture, complemented by fertilizer and pesticide use, has increased crop yields. In some instances, however, this has been at the risk of soil degradation and water pollution and loss.

Irrigation development projects facilitate improvement of infrastructure such as roads, telecommunications, and energy networks; improvement of public services in health care and education; promotion of investments in agribusiness and related areas; and creation of new employment opportunities. While the country-wide prices of agricultural produce tend to decrease under irrigated agriculture, the living standards of the whole population tend to be improved. Although only approximately 20% of the world's cultivable lands are under irrigation, they account for 40% of global agricultural production. It is projected that the population will reach 8 billion by the year 2025, necessitating a 40% increase in food production. Therefore, it is likely that irrigated agriculture will continue to expand.

The expansion of irrigation projects, however, without consideration of drainage, coupled with insufficient funds for proper implementation of technical and socioeconomic aspects, brings about the question of sustainability. Lack of funds for extension service, lack of training and experience, poor choice of irrigation methods, and excessive use of fertilizers and pesticides all contribute to the unfavorable impacts that irrigation can have not only on soil and water resources, but on the whole environment.

Soon after the transition to irrigated agriculture, the initially achieved increase in crop yields may start to decline, plant diseases that are initially nonexistent may become widespread, and the water table may begin to rise, leading to waterlogging and salinity at a later stage. History has witnessed the disappearance of numerous civilizations not only because of wars and long spells of drought or floods, but also as a result of the unintended consequences of irrigated agriculture. Because of the increased soil erosion from irrigated lands, sedimentation causes filling of irrigation channels and dams. Excess use of water triggers drainage and salinity problems, which may cause irreversible loss of soil fertility. Erosion and sedimentation, as well as salination, have gone hand-in-hand with the deterioration of farmlands and the eventual collapse of some social structures. The Babylonian and Sumerian empires in Mesopotamia are examples

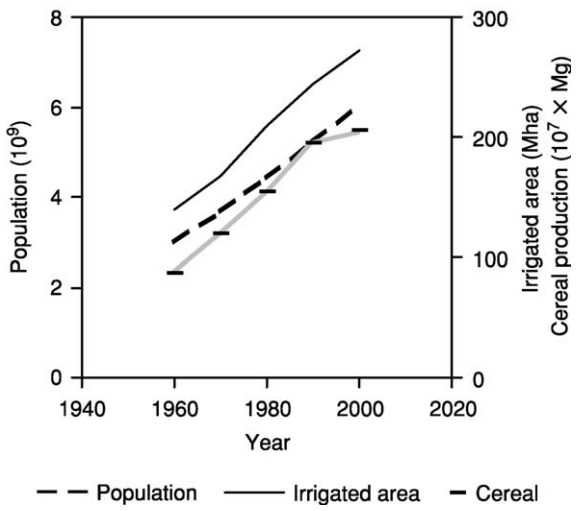


Figure 1 Population, irrigated areas, and annual cereal production (1960–2000). (Adapted from FAOSTAT–Agricultural Data, FAO on-line database.)

of past civilizations whose collapse, in part, was caused by poorly practiced irrigation.

The replenishment of water in the plant root zone during periods when natural rainfall is not sufficient for good plant growth is the classic definition of irrigation. Irrigation is essential to ensure high crop yields in arid and semiarid regions. Supplementary irrigation in humid and subhumid regions not only guarantees high crop yields, but also improves the quality of crops. It is well documented that crop yields under irrigation increase from 100 to 500% compared with yields obtained under rain-fed conditions in arid regions. Starting with the 1950s, irrigated areas increased from a mere 94 Mha to 271.7 Mha in the year 2000. Figure 1 shows the expansion of irrigated areas in the last 50 years and compares it with the global increases in population and cereal production. Because governments had committed to investing in large irrigation projects, the rate of expansion of irrigated areas was 2% annually until the late 1970s. Starting from the early 1980s, it leveled off to around 1% per year in the 1990s. Although irrigated agriculture is practiced in nearly all continents, 70% of the globally irrigated areas exist in Pakistan, China, India, and other developing countries in arid and semiarid regions.

The expansion of irrigated areas has brought about a significant increase in the production of food crops, particularly cereals. Cereal production shows a strong association with the increasing acreage of irrigated areas (Figure 1). Although 31% of the world's cereal-planted area is under irrigation, 42% of global cereal production originates from these areas. Irrigated areas in South and Southeast Asia, and in Near-East Asian countries produce, respectively, 50%, 65%, and 75% of the total cereal production in the respective regions. Irrigated agriculture and its continuing expansion are essential in meeting the needs of increasing populations in the developing counties.

In the future, the largest expansion of irrigated areas will take place in East and Near-East Asian countries and in North Africa where farmlands are limited and irrigation is the only means of increasing agricultural production. However, it is expected that the increase in irrigated areas will be only marginal in Sub-Saharan Africa and Latin America. It has been estimated that three-quarters of the farmlands in East and Near-East Asian countries and North Africa will have irrigation projects implemented in the near future. Similarly, irrigated agriculture will extend to more than 90% of the cultivable lands in South Asia. Thus, it is expected that irrigated areas will reach 330 Mha by the year 2025. The major constraint to the expansion of irrigated areas is the increased demand for municipal and industrial water allocations in competition with agriculture.

Although the benefits of irrigated agriculture are clear, there may be major environmental drawbacks if proper irrigation management is not implemented. Mismanagement of water can cause degradation of soils and pollution of water resources.

Environmental Impact of Irrigated Agriculture

The natural hydrology of watersheds is disturbed during storage and conveyance of water allocated for irrigation, and during discharge of drainage and returning waters from irrigated areas. Overexploitation of groundwater, changing flow regimes of rivers, and raising of water tables (which may trigger salt accumulation) are some of the environmental issues to be addressed. Use of high amounts of pesticides and fertilizers may be a source of pollution of water resources. Furthermore, large bodies of water stored in dams and diversion channels can cause significant changes in local climate.

Water Shortage

Irrigated agriculture uses the highest portion of the world's water resources. The agricultural sector accounted for 66% of the total water withdrawn from both surface water and groundwater resources in 1995 (Figure 2). However, water withdrawal and consumption show great variability in different continents and countries. For example, total water withdrawal in Asia is 57%, of which 80% is used in

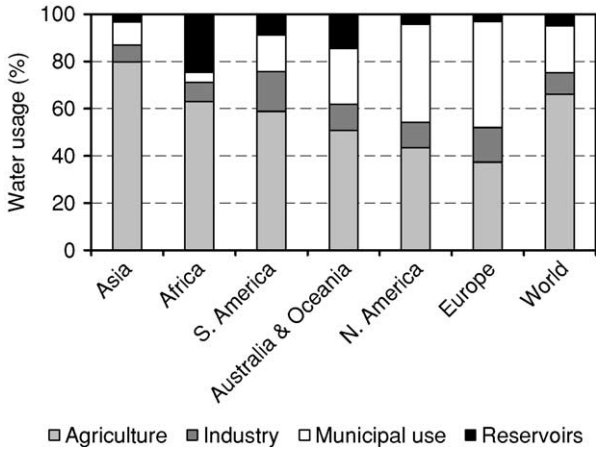


Figure 2 Percentage water consumption of annually withdrawn water resources by sector and continent (1995). (Adapted from Shiklomanov LA (1999) *World Water Resources at the Beginning of the 21st Century*. IHP Report. Paris, France: UNESCO.)

irrigated agriculture. While only 1% of renewable water resources is used in Latin America, the consumption exceeds the renewable limit in Libya and Saudi Arabia.

Actual plant water consumption is only a small portion of the total water diverted to irrigated agriculture because of losses in conveyence networks, use of irrigation methods of low water-application efficiency, excess water application due to unrealistically low water-pricing, and the like. A survey of irrigation networks carried out in 90 different developing countries in 1998 showed that water-use efficiency was as low as 38%. Water allocation for agriculture decreased from 78% to 66% during the last 50 years, in spite of the increasing demand of the agricultural sector. Because of overexploitation of groundwater resources for irrigation in many countries, water table depths have increased. For example, the water table depth in Northern China increased from 8 m to 50 m from 1967 to 2000. Similarly, it was reported that water levels of 10 aquifers in Mexico decreased annually 1.8–3.3 m. More than one million deep wells are opened annually in the Indian peninsula and Pakistan. The total water withdrawal from these wells exceeds the annual recharge rate. Overexploitation of deep wells close to coastal areas triggers seawater intrusion and thereby causes degradation of water quality, which in turn endangers the sustainability of irrigated agriculture.

In the coming decades, three-quarters of the world's population will face water scarcity. Therefore, innovations are needed to sustain high crop yields with lesser amounts of water allocated for irrigated agriculture. There are so-called deficit irrigation practices to sustain high crop yields when irrigation water supplies are scarce and limited. Research for developing new innovative techniques to decrease irrigation water requirements is still continuing. To this effect, a technique called the partial root drying practice, where only half of the plant root zone is wetted by applying water to one side of the rows while keeping the other half dry, is an example of newly emerging innovative irrigation practices to save water. The wetted half of the roots is alternately changed in the subsequent irrigations. The technique is easily practiced in furrow-irrigated crops. The new technique is valued highly because it increases crop water-use efficiency and, in accord with increasing public pressure, decreases the allocations of water resources to irrigated agriculture.

Waterlogging and Salinity

Although irrigated areas are extended 5 Mha annually, the net increase is only limited to 2 Mha. This is because of the annually abandoned areas of 2–3 Mha due to salinity and drainage problems. Therefore, there has been only marginal increase in overall irrigated areas in recent years. It is saddening to note that huge irrigation investments made from limited national resources are wasted because of salinity and water-logging. Among the causes are excess water applications, seepage from conveyance networks, lack of administrative organizations, and limited budget allocations for timely maintenance and management programs both at farm and system levels of irrigation schemes. In some cases, soil fertility has decreased to levels that are even lower than those before irrigation was introduced.

Excess use of irrigation water can cause a rise in the water table, and the resulting waterlogging problem can become a major constraint to high crop yields. Soil salinity and waterlogging are interrelated and both occur due to lack of drainage and excess irrigation-water application. Salinity problems associated with irrigated agriculture have been recognized for a long time. These problems, however, are often ignored in the planning stage of irrigation projects. This is not because of the lack of technical understanding, but mainly because of populist political considerations and commitments to extend irrigated acreage to larger areas without drainage (which may hardly provide any political gain). The drainage systems are largely installed underground, and therefore they have the least political appeal. The usefulness of the drainage systems is hard to appreciate in the short term, since they function as preventive measures for soil salination that may occur in the future. Of the 11 Mha of irrigated land in 12 different irrigation projects in India, 2 Mha of good farmlands are waterlogged. One of the major problems confronted by

farmers of the Indus Plain in Pakistan is the rise of the water table and the resulting increase in soil salinity. Similarly, after completion of the High Aswan Dam in Egypt, 60% of irrigated farmlands began to have salinity and waterlogging problems.

While water is evaporated from the soil surface and is transpired through plants, the salts contained in irrigation water accumulate in the soil profile. In arid and semiarid regions where annual precipitation is not enough to leach accumulated salts to deeper zones of the soil profile, the soil surface becomes covered with a white salt crust. In addition to the actual crop water requirement, a certain fraction of excess water, called the leaching fraction, must be included in the total applied irrigation water. The excess applied water leaches the salts left in the soil profile and thereby prevents the accumulation of salts. However, if the area lacks properly designed drainage systems, the water table soon rises to the soil surface, continuing evaporation takes place, and the salt accumulation speeds up. The lack of drainage in irrigated areas may lead to soil salination and eventually to the loss of soil fertility, if timely preventive measures are not taken.

Soil salinity that develops because of mismanaged farm irrigation systems is called secondary salinity. Of the world's 1500 Mha of total cultivable land, salt-affected soils make up 340 Mha (23%) and sodium-affected alkaline soils make up 560 Mha (37%). The loss of soil fertility due to soil salinity is a problem that is widespread in nearly all continents. Of irrigated areas, 50% in Uzbekistan, 30% in Brazil and Mexico, 27% in Egypt, 26% in Pakistan, and 13% in China are salt-affected soils with continuing declines of soil fertility.

In many instances, the natural drainage rate may be sufficient to prevent salt accumulation. In the absence of natural drainage, manmade drainage must be installed. In either case, it is essential to have a good drainage outlet to remove saline seeps from agricultural areas. In most natural systems, drainage from upland agricultural areas finds its way into rivers and then to the seas. There may be, however, closed basins without natural outlets. The Dead Sea in the Jordan Valley, the Great Salt Lake in the USA, and the Aral Lake in Central Asia are examples of closed basins storing saline seeps from surrounding irrigated farmlands.

Pollutants Attributable to Irrigation

Sediment load carried in returning waters from large-scale irrigation projects and erosion occurring in watersheds may be a source of major pollutants and result in serious environmental consequences. There may be two different sources of soil erosion: (1) erosion occurring during floods that silts water reservoirs, (2) erosion in poorly planned irrigation project areas where land capability classifications are ignored and surface irrigation methods are practiced with no land grading. Sediments originating from the erosion of fertile surface soil in agricultural areas are carried to downstream areas in returning irrigation waters.

Water reservoirs, manmade small lakes, and large dams may be filled with silt load before they complete their economic life, if forestation is not planned and implemented. Silt and sediment found in water reservoirs may carry pesticide residues (with serious health-hazard effects) and plant nutrients (that may cause algae blooms). Heavy metals are also carried into water reservoirs by way of soil erosion.

Plant nutrients Use of fertilizers is one means of achieving high crop yields under irrigation. However, excess or improper use of fertilizers may pollute air, water, and soil resources. A high percentage of applied fertilizer remains in the soil after the harvest of crops, although this varies depending on the fertilizer practice that is followed, crop species, rainfall, and irrigation management.

With the introduction of new crop cultivars, starting in the 1960s, that showed good response to fertilizer applications, the use of chemical fertilizers showed a significant increase (Figure 3), particularly in Europe and North America. The share of fertilizer consumption in the developing world increased from 20% in the 1970s to 43% in the 1990s. Starting in the early 1980s, fertilizer use no longer showed an increasing trend in America and Europe, with growing public awareness about pollution and degradation of water resources.

Over 60% of applied nitrogen fertilizer may remain in the soil after the harvest of crops. The trend of increasing nitrate and phosphorus concentrations in the watersheds of rivers in South and

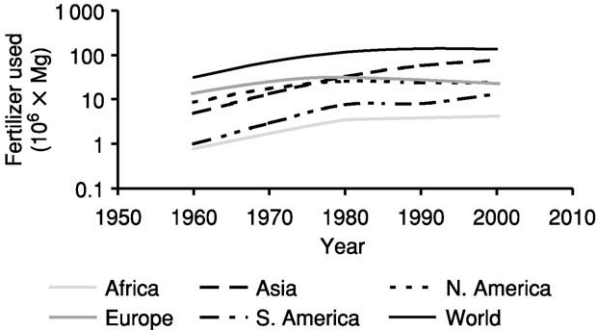


Figure 3 Annual fertilizer consumption by continent (1960–2000). (Adapted from FAOSTAT–Agricultural Data, FAO on-line data base.)

Central Asia since the mid-1970s is attributed to increasing fertilizer usage. Although excess use of potassium may not pose a pollution threat, nitrogen fertilizers can cause pollution in groundwater and surface water resources if care is not taken in their use. Nitrate denitrification occurring in waterlogged areas produces nitrous oxides and other forms of volatilization products that contribute to the atmospheric greenhouse gases responsible for global warming.

Nitrate fertilizers have high solubility, and the end-product nitrate ion is easily leached due to its high mobility with excess irrigation-water application. Hence, nitrate fertilizers mix with groundwater sources. Returning waters from irrigated areas with high sediment loads may also contain high nitrate concentrations. Preventive measures for soil erosion in agricultural areas help to minimize nitrate and phosphorus concentrations in rivers and lakes. The increase in nitrate and phosphorus in lakes is the main cause of eutrophication, algal blooms, and depletion of oxygen that in turn lead to mass fish die-offs. High nitrate concentration in drinking water is also a public health threat. While a high nitrate concentration in drinking water prevents fattening of cattle, it may cause a deadly health disorder called methemoglobinemia in humans (particularly infants). The maximum allowed nitrate concentration in drinking water has been set at $100\,\mathrm{mg\,l^{-1}}$ by the World Health Organization. The European Union restricted it at an even lower limit of $50\,\mathrm{mg\,l^{-1}}$. Agricultural activities are listed among the major causes of river pollution in the USA.

Pesticides Evaporation from large bodies of water stored in dams, networks of channels, and large irrigated fields contributes to increasing humidity, which in turn creates a favorable environment for the spread of plant diseases and pests. Growers, understandably, prefer crops of high cash value under irrigated agriculture. Therefore, it is not uncommon that crop rotation is completely ignored, with the consequences that continuous monoculture of high-value cash crops has become a widespread practice under irrigation.

Pesticides are often heavily applied against weeds and insects under irrigated agriculture. The use of pesticides, in addition to fertilizers, as major inputs of irrigated agriculture dates back to as early as the fifteenth century when arsenic salts were used as herbicides and insecticides. Later, in the sixteenth century, organic pesticides containing nicotine were used as pesticides.

The environmental effects of pesticides were first brought to light in regard to dichlorodiphenyltrichloroethane (DDT), which was first introduced in 1939. The widespread use of DDT continued until the 1970s, when it was banned because of its well-known persistent environmental effects. Despite long years of testing of newly developed pesticides by industry and governmental regulation of their use, pesticides have caused problems such as the development of resistance in target organisms, accumulation in food chains, nonselective killing of all insects, and persistence in the environment.

World pesticide usage exceeded 2.5 million tonnes in 1999. Herbicides accounted for the larger proportion of total usage of 925 000 t, followed by other pesticide usage such as fumigants, nematicides, rodenticides of 756 000 t. Insecticide and fungicide usages were 643 000 t and 252 000 t, respectively. While the share of pesticide use in the developing countries was only 20% in the 1980s, there is an increasing trend in Asia and Latin America, which started in the 1990s. With increased public awareness of health hazards of pesticide residues, they are now more sparingly used in the developed countries and therefore their global use is no longer increasing.

The contribution of pesticide use to overall crop production increase has been estimated to be about 20%. It is therefore unrealistic to assume that their use will cease or be significantly reduced in the future. There should be, however, a rational basis for determining at what point the cost of pesticides becomes prohibitive in economic and environmental terms. Only a small fraction of applied pesticides affects the target organisms and the rest often remains in the soil, contaminates surface water and groundwater resources, or mixes with the air.

Irrigation practices play a major role in the accumulation of pesticide residues encountered in water resources. Residues remaining in the soil are carried into water resources through the leaching of the soil profile with irrigation water. Vaporized pesticide residues mixed in the atmosphere may be carried long distances. The pesticide residues found in arctic mammals were identical to some pesticides used in the tropics and in subtropical areas. Adverse environmental effects of pesticides are further amplified when there is lack of proper training of agricultural labor in the handling and use of pesticides. Death tolls resulting from acute pesticide poisoning are at alarming levels among agricultural labor. It is a widespread problem in Southeast Asia, including Indonesia, Malaysia, and Thailand, where large amounts of pesticides are used in paddy fields. In 1986 in Sri Lanka, 66% of the reported poisoning deaths resulted from the wrong use or mishandling of pesticides.

Heavy metals Fertilizers and pesticides that are used in irrigated agriculture contain a wide array of heavy

metals, including arsenic, boron, mercury, copper, cadmium, lead, nickel, and zinc. While boron is often added to the soil with irrigation water, arsenic is included in many insecticides and herbicides. Similarly, many fungicides contain mercury, copper, and cadmium. Phosphorus fertilizers may contain large amounts of heavy metals; they are the major source of cadmium pollution in soils. Heavy metals tend to be adsorbed by clay minerals in the soil profile and therefore are consequently less likely to be leached and mixed with groundwater. However, runoff and tail waters carrying sediments from irrigated areas often contain heavy metals that are discharged into drainage channels and then mixed with other surface water resources. For example, selenium pollution in the San Francisco Delta area shows how sediments carried away from agricultural areas may cause serious environmental damage.

The use of municipal and industrial wastewater in irrigation is another source of heavy-metal pollution in agricultural areas. The discharge of 1 m^3 of untreated municipal or industrial wastewater can contaminate 8–10 m^3 of the clean water source. Hence, the potential for the pollution of water resources is large if one considers that the annual wastewater discharges in Asia, North America, Europe, and Africa are 590, 431, 326, and 55 km^3, respectively.

Public health concerns Socioeconomic benefits and improvement of agricultural infrastructure are major gains that may result from the implementation of large-scale irrigation projects. However, immediately following irrigation development projects, there may be major demographic changes as a high number of people from rural areas move to cities or agrobusiness centers to take advantage of emerging employment opportunities. There may also be undesirable changes such as an increase in pests and parasites carrying diseases, related to the large bodies of water stored in dams and water-conveying open-channel systems and related climate changes in the project area.

Among the diseases spread by parasites and insects feeding in water sources are malaria, schistosomiasis (bilharzias), yellow fever, brain fever (Japanese encephalitis), and river blindness. Brain fever is largely found in the rice fields of South, Southeast, and West Asia; lymphatic filariasis, in all continents; river blindness, largely in Africa and Central America; and malaria and schistosomiasis, in all continents (but largely in Africa). At the beginning of the twenty-first century, two million people die each year because of malaria. Following the construction of the Diama dam in Senegal, there was a sudden burst of schistosomiasis incidence. Similarly, the number of schistosomiasis patients increased threefold after the completion of the Aswan dam. Malaria incidence has increased because of seepage and waterlogging in the irrigated plains of Punjab and Haryana, in India. Similarly, in Latin America, as much as a threefold increase in malaria incidence was noted during a short span of time following implementation of open-channel irrigation (Figure 4).

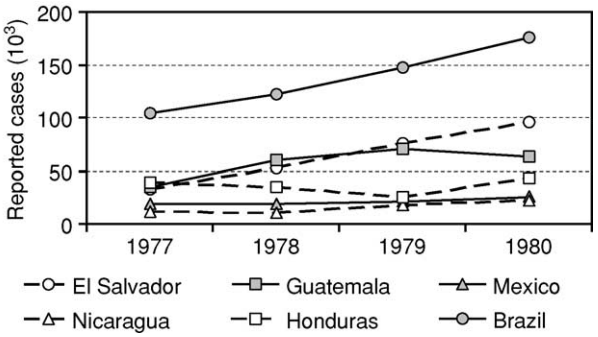


Figure 4 Incidence of malaria due to introduction of canal irrigation in some Latin American countries. (Adapted from WHO (1983) *Report of the Experts on Environmental Management for Vector Control*. VBC/83.4. Geneva, Switzerland: WHO.)

Municipal and industrial wastewaters are often recycled and rerouted into irrigation schemes in water-scarce regions. Although treatment of such water sources is common practice in the developed world, examples of water treatment in developing countries are rare. Wastewaters from poultry farms, cattle ranges, and household sewage may pose a serious health hazard if used for the irrigation of freshly consumed vegetables. Among the diseases spread because of the use of untreated wastewater for irrigation are cholera, typhoid, ascariasis, amebiasis, giardiasis, and enteroinvasive *Escherichia coli*. Five million people die annually from diseases related to the practice of irrigation with untreated wastewater.

Sustainability

The history of agriculture has shown that sustainable irrigated agriculture requires adequate salt balance and drainage. The technology to accomplish this exists, but its use is limited by numerous factors. These factors include the lack of economic incentives for growers to improve their existing irrigation structures; need for investment in pressured irrigation systems and drainage works; lack of training on and experience with effective management practices; public opposition to the possible effects of irrigation return flows on the environment; and administrative constraints such as low budget allocations for timely maintenance and management of irrigation schemes.

Recent developments in irrigation technology are helping to improve irrigation efficiency and to reduce the drainage water volume that requires disposal. For example, sprinkler and drip irrigation systems facilitate the fine-tuned control of fertilizer applications in order to minimize fertilizer residue remaining in the soil after the harvest of crops. The new irrigation technologies that facilitate the application of fertilizers along with irrigation water should be adopted and used widely. Irrigation water requirements are least, for example, with drip systems that supply water and fertilizers directly into the root zone where they are needed. The crop utilizes nearly all of the fertilizer applied, leaving the least residue in the soil after harvest.

Over-use of pesticides is now being replaced with integrated pest management practices that take advantage of the biological control of insects and pests in addition to the use of chemical sprays. However, efforts are still needed to create public awareness of the importance of conservation of soil and water for meeting the needs of present and future generations.

Training farm advisers and extension workers on effective management practices of land preparation and soil tillage, and use of pesticides and fertilizers will help to minimize pollution attributable to agricultural activities. Adopting advanced irrigation technologies supported by governmental subsidies or low-interest loans (if needed), and drafting and enforcing legal codes for preventing pollution of water and soil resources will contribute to the environment-friendly and sustainable practice of irrigated agriculture. Presently available technical knowledge on irrigation is adequate for developing sustainable agricultural practices to keep pace with the food and fiber demands of the increasing world population. To this effect, despite the drawbacks, irrigated agriculture has an important role to play in world food production.

See also: **Civilization, Role of Soils**; **Drainage, Surface and Subsurface**; **Salination Processes**; **Salinity:** Management; **Salt-Affected Soils, Reclamation**

Further Reading

Alexandratos N (ed.) (1995) *World Agriculture: Towards 2010. An FAO Study*. Chichester, UK: FAO/John Wiley.

Barrios A (2000) *Agriculture and Water Quality*. CAE Working Paper Series, WP00-2. DeKalb, IL: American Farmland Trust's Center for Agriculture in the Environment.

Daugherty TC and Hall AW (1995) *Environmental Impact Assessment of Irrigation and Drainage Projects*. FAO Irrigation and Drainage Paper 53. Rome, Italy: FAO.

Donaldson D, Kiely Timothy, and Grube A (2002) *Pesticides Industry Sales and Usage. 1998 and 1999 Market Estimates*. Biological and Economic Analysis Division Office of Pesticide Programs, Office of Preventation, Pesticides, and Toxic Substances, US Environmental Protection Agency. Washington, DC: US Government Printing Office.

FAO (2002) *Deficit Irrigation Practices*. Water Reports 22. Rome, Italy: FAO.

FAO (2002) *World Agriculture Towards 2015/2030. Summary Report*. Rome, Italy: FAO.

Hornsby AG (1990) Pollution and public health problems related to irrigation. In: Stewart BA and Nielsen DR (eds) *Irrigation of Agricultural Crops*, pp. 1173–1188. Agronomy Monograph No. 30. Madison, WI: American Society of Agronomy.

Kirda C, Moutonnet P, Hera C, and Nielsen DR (eds) (1999) *Crop Yield Response to Deficit Irrigation*. Dordrecht, The Netherlands: Kluwer Academic Publishers.

Marr PD (1969) The social context of irrigation. In: Hagan RM, Haise HR, and Edminster TW (eds) *Irrigation of Agricultural Lands*, pp. 12–22. Agronomy Monograph No. 11. Madison, WI: American Society of Agronomy.

Shah T, Molden D, Sakthivadivel R, and Seckler D (2000) *The Global Groundwater Situation: Overview of Opportunities and Challenges*. Colombo, Sri Lanka: International Water Management Institute.

Shiklomanov LA (1999) *World Water Resources at the Beginning of the 21st Century*. IHP Report. Paris, France: UNESCO.

Tanji KK (ed.) (1990) *Agricultural Salinity Assessment and Management*. Manuals and Reports on Engineering Practice No. 71. New York: American Society of Civil Engineers.

Umali DL (1993) *Irrigation Induced Salinity: A Growing Problem for Development and the Environment*. World Bank Technical Paper 215. Washington, DC: The World Bank.

Wagner RH (1971) *Environment and Man*. New York: Norton.

Methods

D L Bjorneberg and R E Sojka, USDA Agricultural Research Service, Kimberly, ID, USA

Published by Elsevier Ltd.

Introduction

Irrigation is the process of applying water to soil, primarily to meet the water needs of growing plants. Water from rivers, reservoirs, lakes, or aquifers is pumped or flows by gravity through pipes, canals, ditches, or even natural streams. Applying water to

fields enhances the magnitude, quality, and reliability of crop production – approximately 30% of the world's food is grown on irrigated land, which accounts for only about 15% of the world's land used for crop production.

Various irrigation methods have been developed over time to meet the irrigation needs of certain crops in specific areas. The three main methods of irrigation are surface, sprinkler, and drip or microirrigation. In surface irrigation, water flows over the soil by gravity. Sprinkler irrigation applies water to soil by sprinkling or spraying water droplets from fixed or moving systems. Microirrigation applies frequent, small applications by dripping, bubbling, or spraying. A fourth, and minor, irrigation method is subirrigation, where the water table is raised to or held near the plant root zone using ditches or subsurface drains to supply the water.

More than 90% of the approximately 270 Mha of irrigated crop land in the world is surface-irrigated; less than 1% is irrigated by microirrigation. In the USA, sprinkler and microirrigation are used on a greater percentage of irrigated cropland, primarily to reduce labor and improve control of water application. Approximately 45% of the 20 Mha of cropland irrigated in the USA is sprinkler-irrigated, and microirrigation is used on approximately 4% of the irrigated cropland.

Surface Irrigation

Surface irrigation entails water flowing by gravity over soil. Water is usually supplied by gravity through canals, pipes, or ditches from the water source to the field. In some locations, however, water may need to be pumped from the source to a field at a higher elevation. Types of surface irrigation systems include furrow, basin, and border irrigation. Surface irrigation systems are typically used for field crops, pastures, and orchards. Efficiency of surface irrigation systems varies tremendously because of variations in soil type, field uniformity, and management. Surface irrigation is often considered less efficient than sprinkler irrigation or microirrigation, because soil, not a pipe, is the water-conveyance system for surface irrigation (Table 1). However, a well-managed surface irrigation system on a uniform soil with a runoff-reuse system can approach 90% application efficiency.

Table 1 Application efficiencies for irrigation systems

System type	*Application efficiency (%)*
Surface irrigation[a]	
Furrow	50–70
Level basin	60–80
Border	60–75
Sprinkler irrigation	
Solid set	60–85
Set move	60–75
Moving[b]	75–95
Traveling gun	55–65
Microirrigation[c]	80–95
Subirrigation	50–80

[a]Surface irrigation efficiencies can be greater if runoff is reused.
[b]Includes center-pivot, linear move, and low-energy precision application (LEPA) systems.
[c]Efficiency can decrease to 50% with poor management.

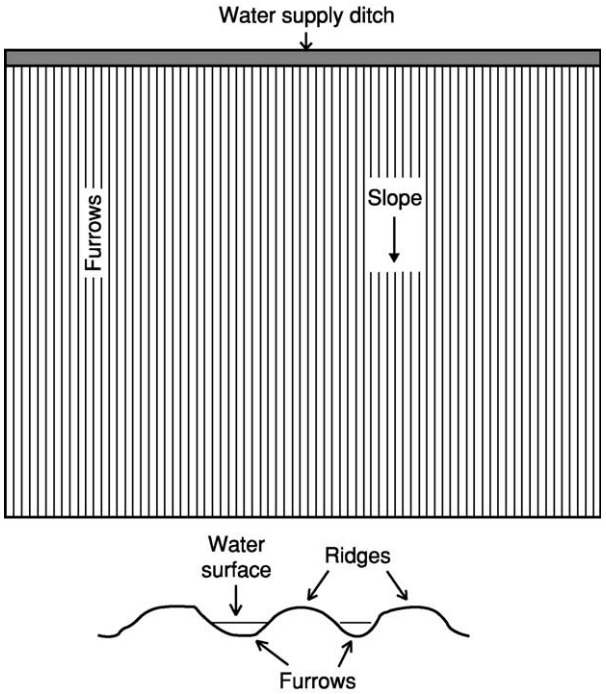


Figure 1 Schematic of a furrow-irrigated field, where water flows in evenly spaced furrows or corrugates.

Furrow Irrigation

In furrow irrigation, water flows in evenly spaced furrows or corrugates that are typically 0.1–0.3 m wide (Figure 1) on fields with slopes of 0.1–3%. Water commonly flows in furrows for 12–24 h during an irrigation, but shorter or longer durations may be used depending on furrow length and soil, water, and management considerations. Inflow rates for individual furrows vary from approximately 10 to 100 l min^{-1}, again depending on soil, slope, field length, and management. Ideally, water should advance across the field in approximately 25% of the total irrigation time to irrigate the field uniformly. Since soil erosion increases as field slope and inflow rate increase, flow rate must be carefully managed on fields with steeper slopes (greater than 1%). Low inflow rates and long irrigation durations may be needed to apply the desired amount of water during

Figure 2 Furrow-irrigated sugar beet field, with water supplied by siphon tubes from a concrete ditch.

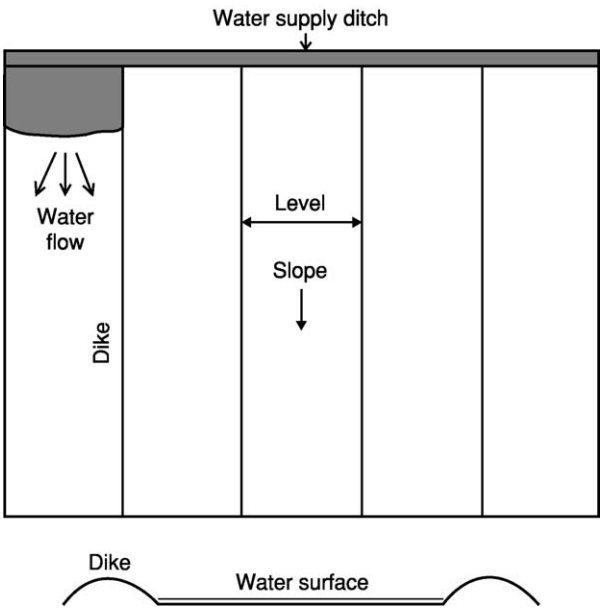


Figure 3 Schematic of a border-irrigated field, where a uniform sheet of water flows between dikes.

an irrigation on soils with low infiltration rate. Conversely, higher inflow rates are often needed on fields with low slopes and/or high infiltration rates in order for the water to flow across the field and uniformly irrigate the upper and lower ends of the field.

Inflow to irrigation furrows may be supplied from gated pipes or ditches (earthen or concrete). Siphon tubes are frequently used to convey and regulate water flow from ditches to individual furrows (Figure 2). By creating a siphon, water flows through the tube, over the ditch bank, and into the furrow, as long as the tube outlet is lower than the water elevation in the ditch. Furrow inflow rate is controlled by tube diameter and the elevation difference between the ditch-water level and tube outlet. Gated pipes distribute water to furrows through evenly spaced outlets on the pipes. With earthen ditches, water can flow through a breach or other opening in the ditch bank to individual furrows or a smaller feed ditch that distributes water to several furrows. It is much more difficult to regulate flow through a breach in an earthen ditch than through siphon tubes or pipe gates.

Furrow irrigation requires lower capital investment, less technical knowledge, and greater labor than most other irrigation systems. Fields can be irrigated without leveling or grading, because water flows in furrows. Furrow irrigation is not well suited to automation, because water flow rate must be adjusted for each furrow for each irrigation.

Basin and Border Irrigation

Basin and border irrigation systems are similar in that both involve a uniform sheet of water flowing over the soil. The general difference is that basin irrigation involves applying water to a nearly level field and may include ponding for extended time periods. With border irrigation, water flows between dikes that divide a sloping field into rectangular strips with free drainage at the end (Figure 3). The purpose of the dikes is to contain water as it flows across the field, unlike basin irrigation where the dikes pond the water.

Basins can be as small as a few square meters for a single tree or as large as several hectares with greater than $100\,l\,s^{-1}$ inflow rates. Basin size is a balance between soil infiltration rate, slope, and water supply. Water depth in basins varies from approximately 5 to 20 cm, with typical depths of 10–15 cm. Efficient basin irrigation requires a level soil surface with uniform soil texture and adequate water supply so the basin is quickly and uniformly covered with water. If the basin is not level, the higher elevation areas will receive less water than the low areas. If the basin inflow rate is inadequate, water will slowly advance, causing large differences in infiltration opportunity time within the basin.

A special type of basin irrigation is a drain-back level basin. Drain-back level basins have a series of parallel basins that receive inflow from a shallow, 5–10-m-wide ditch. After the first basin is filled, a gate opens to start filling the adjacent basin, which is at a lower elevation. Water near the inflow end of the first basin drains back to the inflow ditch and flows to the next basin. This procedure is repeated until every basin has been irrigated. The drain-back phase improves uniformity by reducing the amount of water that infiltrates near the inflow end and initially increases the inflow rate to the next basin, which increases the advance rate.

Border irrigation systems are better suited for sloping fields than basin systems, because water flows between dikes rather than ponds within basins. The irrigated areas between dikes can be 3–30 m wide and up to 400 m long. The field slope between dikes (perpendicular to water flow direction) should be nearly level so water flows uniformly down the field. The slope along the dikes can be similar to furrow irrigation, but border systems often have slopes of less than 0.5%.

Water can be supplied to borders and basins from open ditches with gates, breaches, or siphon tubes or from above- or belowground pipes. Typical inflow rates vary from 10 to 100 l s^{-1}, but vary widely depending on the size of basin or border, soil texture, and slope. Border and basin irrigation require less labor than furrow irrigation because water is supplied to a larger area with a single outlet.

Sprinkler Irrigation

Sprinkler irrigation applies water to soil by spraying or sprinkling water through the air on to the soil surface. Water is pressurized and delivered to the irrigation system by a mainline pipe, which is often buried so it does not interfere with farming operations. Three main categories of sprinkler irrigation systems are solid-set, set-move, and moving. Sprinkler irrigation is used for a wide variety of plants, including field crops, orchard trees, turf, and pasture. Sprinkler systems are also installed for applying wastewater and protecting plants from frost.

Solid-set systems may be installed for a single season for certain field crops or permanently for turf, orchards, or permanent crops. Set-move systems are manually or mechanically moved to another part of the field after the irrigation set is complete in the present location. Moving systems such as center pivots or traveling guns apply water as the system slowly travels through the field.

Sprinkler irrigation is often more efficient than surface irrigation, because water application is more controlled (Table 1). In hot and/or windy areas, however, sprinkler irrigation can have significant losses to evaporation and wind drift. Maintenance is also important for efficient sprinkler irrigation. Worn nozzles and leaking pipe connections reduce application uniformity and system efficiency.

Solid-Set Sprinkler Systems

Most solid-set sprinkler irrigation systems are designed to apply frequent, small amounts of water to meet plant water needs (e.g., daily). Water application rates can vary from approximately 4–6 mm h^{-1} for field crops to 5–30 mm h^{-1} for turf applications. Overhead costs are greater for solid-set systems than other sprinkler systems, because the entire irrigated area must be equipped with sprinklers and pipe. However, permanently installed systems can be automated to reduce labor and allow irrigation at any hour of the day, which reduces the opportunity for plants to be stressed. When properly designed, solid-set systems have high application uniformity. While solid-set systems are most commonly used with turf, and landscape and permanent crops, these systems are also used for some high-value annual crops with low tolerance for water stress.

Solid-set system designs are as varied as the applications: small sprinklers may irrigate 20 m^2, or large, gun-type sprinklers may be spaced 50 m apart. Plastic pipe is frequently used for buried applications, but it is also used in some aboveground applications. Aluminum pipe (50–100 mm in diameter) is often used for field crops when the system is installed after planting and removed before harvest. Most systems are divided into zones so a portion of the area is irrigated at one time. Solid-set systems used for frost control, however, must be designed to water the entire area simultaneously.

Set-Move Sprinkler Systems

Set-move sprinkler irrigation systems are designed to apply water slowly during the irrigation set (e.g., 4–6 mm h^{-1}), which often lasts 8–24 h. After completing the irrigation set, the sprinkler system is moved to an adjacent area for the next set (Figure 4). Adequate water has to be applied during an irrigation set to meet the crop's water needs until the system is moved back to the area, often in 7–10 days.

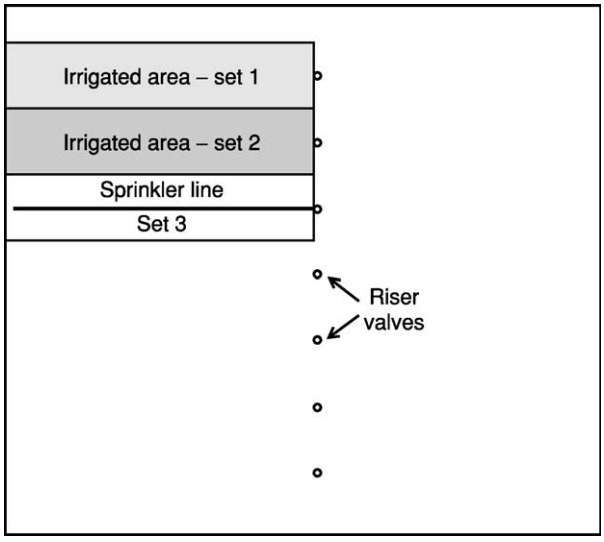


Figure 4 Schematic of a set-move-irrigated field, where the irrigation system is moved to the adjacent area after completing an irrigation set.

The common types of set-move irrigation systems are hand-move and side-roll systems. Hand-move systems can be a single sprinkler or a line of sprinklers. A line of hand-move sprinklers, sometimes called 'handlines,' is typically composed of 9- or 12-m-long pieces of 75- or 100-mm-diameter aluminum pipe with a sprinkler mounted on one end or in the center. Individual pipes are connected to form an irrigation line, usually not more than 400 m long. After an irrigation set is completed, the line is disconnected, and each piece is moved by hand 10–20 m to the next set. A slight variation to the handline is the dragline or end-pull system. These systems, which are less common, have special connections between sprinkler pipes that allow the irrigation line to be pulled by a tractor to the next set.

Side-roll systems, also called wheel lines, are similar in principle to handlines except a large-diameter wheel (1.5–3 m in diameter) is mounted in the center or on the end of each piece of aluminum pipe (100–125 mm in diameter) to elevate the sprinkler. The sprinkler pipe is the axle for the side-roll. When an irrigation set is completed and the pipe has drained, the wheel line, powered by an engine, is rolled to the next position. Self-leveling sprinklers are used so the side-roll does not have to be exactly positioned for the sprinklers to operate correctly.

Moving Sprinkler Systems

Moving irrigation systems include center-pivot, linear-move, and traveling gun systems. Center-pivot and linear-move systems are similar in design and appearance. These systems consist of one or more spans of sprinkler pipe elevated by 'A-frame' towers. Span length varies from 30 to 65 m. Towers, powered by hydraulic or electric motors, elevate the sprinkler pipe 2–4 m above the ground.

The center pivot has a stationary pivot point so the towers move in a circle (Figure 5). Water and power are supplied to the system through the pivot point. A typical center pivot in the USA has eight spans, a total length of approximately 400 m, and irrigates 50–60 ha. Center pivots are extremely popular, because water is uniformly applied to a large area with little labor. Furthermore, once a circular field has been irrigated, the center pivot is in position to start the next irrigation. System cost per irrigated area is reduced by increasing the total length of a center pivot, because irrigated area per unit length increases with distance from the pivot point. Consequently, water application rate increases with distance from the pivot point, because each span must irrigate a larger area per revolution (a 50-m span at the pivot point irrigates 0.8 ha, while a 50-m span that is 350 m from the pivot point irrigates 12 ha). Application rates often exceed the infiltration rate of the soil under the outer spans of center pivots. Thus, the opportunity for runoff increases as center-pivot length increases. Center pivots can also irrigate fields with rolling terrain that are difficult or impossible to irrigate by surface-irrigation methods; however, these conditions can create additional management challenges.

Linear-move systems have a control unit on one end, or in the center on longer systems, that moves the towers in a straight line to irrigate rectangular-shaped fields. Power is supplied by an electrical drag cord or by an engine-powered generator mounted on the control unit. Water is typically supplied to the drive unit by a drag hose connected to a buried or aboveground pipe. Drive units can be equipped with a pump so water can be supplied from an open ditch flowing parallel to the travel direction, but this is not commonly used, because it requires a nearly level ditch. System cost per irrigated area is reduced by increasing the distance the system travels. Since hose length is limited to approximately 150 m, the system can move 300 m before the hose is connected to the next riser. Similar to set-move systems, adequate water must be applied to meet crop needs until the linear-move can irrigate the area again.

Early center pivots had impact sprinklers mounted on top of the irrigation pipe that required 500–600 kPa. Most new systems have low-pressure nozzles (70–200 kPa) mounted on tubes that extend below the

Figure 5 Center-pivot irrigation system.

irrigation pipe, so nozzle height varies from 1 to 3 m above the soil. Manufacturers make numerous types of low-pressure nozzles with fixed or spinning spray plates that provide a wide range of application rates and water-droplet sizes to meet field conditions and operator preferences. A common feature to all nozzle types is a pressure regulator, which maintains constant nozzle pressure as the system travels across a field with varied elevation.

A traveling gun has a large-capacity sprinkler on a cart that is pulled across the field by a cable or by the water-supply hose. These systems irrigate an area 50–100 m wide and up to 400 m long. A traveling gun can be considered a moving, set-move system because water is applied as the cart moves across the field and then the system is moved to another area in the field for the next irrigation set. For cable-tow systems, a winch on the cart winds the cable, pulling the cart and a soft hose across the field. A hose-reel system pulls the cart as a hard plastic (polyethylene) hose is wound around a reel on a trailer anchored at the end of the run. The reel or winch is powered by an engine or a water turbine. Smaller versions of traveling guns are available for irrigating athletics fields, small pastures, or arenas.

Microirrigation

Microirrigation applies water at low rates and pressures to discrete areas so irrigation water reaches the root zone with minimal losses. Water drips from emitters in plastic pipe or tape, or bubbles or sprays from small emitters that only wet a portion of the soil surface. Microirrigation systems are popular for permanently installed systems that irrigate trees, vineyards, orchards, and shrubs. These systems are typically automated so that water is applied frequently (e.g., daily) to maintain optimum soil-water content near the plants. Filtration is important for microirrigation, because sediment and algae can plug the small openings on drip emitters, bubblers, and microsprays. Chemical treatment may also be necessary to reduce salt or mineral deposits that can plug emitters.

Drip irrigation emitters are preinstalled within polyethylene pipe at regular intervals or emitters are attached to the outside of the pipe at desired locations. Emitter flow rates typically vary from 2 to 7.5 l h^{-1}. Pressure-compensating emitters maintain a constant flow rate as pressure varies from approximately 70 to 200 kPa (Figure 6). This type of system is common in vineyards.

Drip tape is thin-walled (0.1–0.375 mm) plastic tubing (10–20 mm in diameter), with outlets at 100–600-mm intervals (Figure 7). Flow rates can vary from 100 to 400 l h^{-1} per 100 m of length.

Figure 6 Pressure-compensating drip emitter watering ornamental plants.

Figure 7 Drip tape installed on the soil surface to irrigate edible beans.

Typical operating pressure for drip tape is 35–100 kPa. Drip tape is commonly installed below the soil surface, where there is less opportunity for damage. Buried drip tape can be used for several seasons or retrieved after a single season, depending on crop types and farming practices.

Bubblers and microsprays are often used for irrigating trees, shrubs, or ornamental plants. Bubblers discharge water with low energy to flood a small area. Flow rates up to 100 l h^{-1} can apply water to 4-m-diameter areas, depending on nozzle size, type, and pressure. Microsprays apply a fine spray or mist with similar flow rates and wetted areas as bubblers.

Subirrigation

Subirrigation applies water below the soil surface to raise the water table into or near the plant root

Table 2 Typical advantages and disadvantages of irrigation systems

System type	*Advantages*	*Disadvantages*
Surface irrigation		
Furrow	Low capital and maintenance costs; water flows in small channels	High labor; less water control; soil erosion; possible runoff and percolation losses
Level basin	Efficient with good design; less labor than furrow	Ponded water; sloping fields must be leveled
Border	Less labor and less runoff than furrow; easier to manage infiltration depth	Water flows over entire soil surface
Sprinkler irrigation		
Solid set	Good water control; possible to automate and frequently irrigate; fits odd-shaped fields	High capital cost; system may interfere with field operations
Set-move	Lower capital cost than other sprinkler systems	More labor than other sprinkler systems; poor uniformity in windy conditions; greater application depth
Moving[a]	High uniformity; low labor	High capital and maintenance costs; not suitable for odd-shaped fields; potential wind and evaporation losses
Traveling gun	Lower capital cost than other sprinkler systems	Higher operating cost; wind and evaporation losses
Microirrigation	Excellent water control; frequent applications possible	Higher capital cost; requires clean water or treatment and filtration

[a]Includes center-pivot, linear-move, and low-energy precision application systems.

zone. Subirrigation is not often used in arid or semiarid irrigated areas; it is typically used in conjunction with subsurface drainage. Subsurface drainage lowers the water table and removes excess water through open ditches or perforated pipe. Water-table depth can be controlled by installing a weir on the drainage system. During wet periods, the water table is lowered so the root zone remains unsaturated. During dry periods, water is pumped into the drainage system to raise the water table and provide additional water for plant growth. In some situations, drained water can be stored for use when irrigating.

Salinity Hazards

Salinity problems are more closely associated with soil and water chemistry than the type of irrigation system; however, the irrigation system can accentuate salinity problems and affect how salinity is managed. Basin and border irrigation can uniformly leach salts from the soil when infiltration is uniform. Sprinkler irrigation can also uniformly leach salts, but sprinkling can injure sensitive plants when highly saline water is applied to foliage. Furrow irrigation leaches salt from soil below the furrow, while increasing salt concentrations in the bed. Tilling before planting crops tends to minimize salinity problems by mixing salts in the soil. Salt concentration tends to increase radially from a drip emitter or laterally from a line source such as drip tape. Well-managed drip irrigation can minimize salt-induced stress by maintaining a high soil-water content with frequent irrigations.

Choosing an Irrigation System

Choosing an irrigation system is a difficult task. Irrigation systems are as varied as the people who use them. The right selection for a user depends on soil, water, and climatic conditions, as well as crop types, user knowledge and preference, capital and operating costs, and infrastructure availability. No system is best for all situations. Some typical advantages and disadvantages of irrigation systems are shown in Table 2. Sprinkler or microirrigation are often better choices than surface irrigation on sandy soil, where excessive percolation is a problem. Surface irrigation may be better in arid, windy areas, where wind and evaporation losses can be significant. Surface irrigation offers less control of application depth, so small, frequent irrigations are not practical for water-sensitive crops, which are better suited to microirrigation, solid-set, or center-pivot systems.

List of Technical Nomenclature

ha	hectares
kPa	kilopascals
l h^{-1}	liters per hour
l min^{-1}	liters per minute
l s^{-1}	liters per second
Mha	million hectares
m	meters
m^2	square meters
mm	millimeters

See also: **Crop Water Requirements**; **Evapotranspiration**; **Fertigation**; **Infiltration**; **Irrigation:** Environmental Effects; **Salinity:** Management; **Water Harvesting**; **Water-Use Efficiency**

Further Reading

Brouwer C, Prins K, Kay M, and Heibloem M (1988) *Irrigation Water Management: Irrigation Methods*. Training manual No. 5. Rome, Italy: Food and Agriculture Organization of the UN.

Cuenca RH (1989) *Irrigation System Design: An Engineering Approach*. Englewood Cliffs, NJ: Prentice-Hall.

FAO *Aquastat – Information System on Water and Agriculture*. Rome, Italy: Food and Agriculture Organization of the UN.

Hoffman GJ, Howell TA, and Solomon KH (1990) *Management of Farm Irrigation Systems*. St. Joseph, MI: American Society of Agricultural Engineers.

Jensen ME (1983) *Design and Operation of Farm Irrigation Systems*. ASAE Monograph No. 3. St. Joseph, MI: American Society of Agricultural Engineers.

USDA (1997) *1997 Census of Agriculture*, vol. 1. *National, State, and County Tables*. National Agricultural Statistics Service. Washington, DC: US Government Printing Office.

ISOTOPES IN SOIL AND PLANT INVESTIGATIONS

K Reichardt and O O S Bacchi, Center for Nuclear Energy in Agriculture, Piracicaba, Brazil

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Isotopes, both radioactive and stable, have been used successfully as tracers in environmental studies involving the soil–plant–atmosphere system. Each chemical element of the periodic table has isotopes, which are atoms of the same species, behaving chemically and biologically in the same manner, differing only in some of their physical properties. They have in their nuclei the same number of protons, which defines the element, but different number of neutrons, which makes them different in terms of mass. Most of them are natural, i.e., present in nature since the formation of the Earth, and some are artificial, laboratory-made through nuclear reactions. When stable, they differ only by their atomic weight, and, when radioactive (unstable), they emit radiation. The important thing is that they differ in a measurable way and can be detected. Stable isotopes are detected by mass spectrometry, with instruments capable of distinguishing atomic weights, and radioisotopes, by radiation detection, made by a wide range of detectors, depending on the type of the radiation. One important feature of these methodologies is that they are able to measure extremely low amounts of isotopes.

Radioisotopes as Tracers

Radioisotopes are unstable isotopes that approach stability by emitting radiation. This radiation can easily be detected even in extremely low amounts. Radioisotopes can be traced at concentrations as low as 10^{-11} (1/100 billion). Each radioisotope has its own characteristics and emits one or more radiation types (mainly α (alpha), β^- (beta negative), β^+ (beta positive), γ (gamma), and n (neutron)), with one or more energies, and at a given rate that depends on their half-lives. Since this rate decreases exponentially with time, the half-life T is defined as the time in which any emission rate decreases by half. The type and energy of radiation affect detection of radioisotopes and a short half-life can limit their use in long-term experiments.

Radioisotopes can be found naturally, for example, ^{3}H, ^{14}C, ^{40}K, ^{226}Ra, and ^{235}U, some of them being continuously produced by natural processes occurring in the upper atmosphere which are induced by solar or cosmic radiation. In this case the production rate of the radioisotope is in equilibrium with its decay rate. Some radioisotopes have half-lives longer than the age of our planet, therefore they are present naturally in the environment. The great majority of radioisotopes used as tracers are, however, artificially produced in nuclear reactors. Table 1 presents the principal tracer isotopes used in environmental studies, mainly in plant biology and soil studies.

To use isotopes as tracers, they are added to the respective stable isotope in amounts that allow their detection to the end of the experiment. Whenever possible they should be added in the same chemical form as the stable compound, e.g., if a phosphorus study is performed using superphosphate, the radioactive ^{32}P has to be present also in the form of superphosphate. Sometimes this is costly, as is the case in herbicide studies in which the radioactive ^{14}C has to be incorporated into the same molecular species.

Table 1 Some tracer radioisotopes for soil, plant, and environmental studies

Most abundant isotope	Tracer radioisotope	Radiation energy (MeV)		Half-life
		Beta	Gamma	
Calcium-40	^{45}Ca	0.254		153 days
Carbon-12	^{14}C	0.156		5720 years
Cesium-133	^{137}Cs	0.52; 1.18	0.662	30 years
Cobalt-59	^{60}Co	0.31	1.17; 1.33	5.27 years
Hydrogen-1	^{3}H	0.0181		12.26 years
Iodine-127	^{131}I	0.61; 0.25; 0.85	0.36; 0.08; 0.72	8.05 years
Magnesium-24	^{28}Mg	0.45	0.032; 1.35; 0.95	21.3 h
Manganese-55	^{52}Mn	0.6	0.94; 1.46	5.7 days
Phosphorus-31	^{32}P	1.71		14.3 days
Potassium-39	^{40}K	1.32	1.46	1.3×10^{9} years
Rubidium-85	^{86}Rb	0.7; 1.77	1.08	18.7 days
Strontium-88	^{89}Sr	1.47		50.4 days
Sulphur-32	^{35}S	0.168		86.7 days
Zinc-64	^{65}Zn	0.33	1.11	245 days

The strength of radioactive materials is measured in terms of activity ($1\,\text{Beq} = 1$ disintegration s^{-1}; $1\,\text{Ci} = 3.7 \times 10^{10}$ disintegrations s^{-1}). Each detected radiation corresponds to one count, and therefore activities are also expressed as counts per second (cps) or counts per minute (cpm). Since the measured activity depends on the size of the sample, the specific activity (S) is defined as the activity per unit of mass, typically expressed as cps per milligram or cpm per microgram.

From the above it can be seen that radioisotopes can be used as tracers in a wide range of experiments. Sometimes one looks only for the presence of the isotope in a given part of the system, and in other situations one needs to measure amounts absorbed or transferred, or rates of specific processes. The basis of most applications lies in the principle of isotope dilution, which states that, for a given constant amount of radioactivity, the specific activity is inversely proportional to the total amount of test element or substance present in the system. The technique has several advantages, being especially useful in obtaining analytical data when chemical separations are difficult, time-consuming, or impossible. It allows the estimation of uptake of a nutrient by a plant from fertilizer, or the recognition and quantitative determination of intermediates in metabolism studies. In a fertilizer study, if a plant receives ^{32}P-labeled superphosphate, a distinction can be made between the soil P and the fertilizer P absorbed by the plant. The percentage phosphorus derived from the fertilizer (%Pdff) can be calculated as follows:

$$\%\ \text{Pdff} = \frac{S \text{ of } {}^{32}\text{P in plant material}}{S \text{ of } {}^{32}\text{P in fertilizer}} 100 \qquad [1]$$

For example, if the plant material of a crop grown on a soil fertilized with a ^{32}P-labeled fertilizer presents a specific activity S of 35 758 cpm μmol^{-1} P, and the applied fertilizer has a specific activity of 102 441 cpm μmol^{-1} P, one can say that approximately 35% of the phosphorus present in the plant comes from the fertilizer. Such experiments are used to study fertilizer-use efficiency, optimization of fertilizer placement and rate, evaluation of varieties, and to answer several other important questions raised during soil–plant management.

Radioactive tracers have been extensively used in studies concerning environmental pollution, dynamics of nutrients in the soil–plant–atmosphere system, the fate of herbicides and other agrochemicals, etc. One important application is the use of fallout ^{137}Cs as a label for soil erosion. The nuclear tests made in the middle of the twentieth century have spread over the whole surface of the Earth a fairly constant amount of ^{137}Cs, an element that is strongly bound to the soil, remaining in the surface layer. When soil is washed downslope, the cesium is carried on soil particles, and in this way it is possible to identify erosion areas in which the ^{137}Cs content is lower than expected, and deposition areas with higher ^{137}Cs content.

Stable Isotopes as Tracers

Radioactive isotopes of suitable half-life do not exist for some elements, in which case stable isotopes may be used. The quality of stable isotopes is defined in terms of their relative abundance. For example, the ^{15}N abundance of a sample containing nitrogen is the ratio of ^{15}N atoms to the total of N atoms in the sample, which in the case of nitrogen, which has only two isotopes, is equal to (^{15}N + ^{14}N), i.e.:

$$\%\ {}^{15}\text{N abundance} = \frac{\text{atoms } {}^{15}\text{N}}{\text{atoms } {}^{15}\text{N} + \text{atoms } {}^{14}\text{N}} 100 \qquad [2]$$

The natural environment, due to the intensive nitrogen-cycling in biological materials over billions of years, has a fairly constant ^{15}N abundance of 0.365%. Therefore, amounts of ^{15}N in enriched samples can conveniently be expressed as percentage of ^{15}N atoms in excess of the natural abundance of 0.365.

Dynamic processes in which the flow of compounds is affected by atomic mass affect the relative abundance of isotopes. This fact is used in the laboratory to isotopically enrich compounds. Using batteries of long resin-exchange columns, ammonium sulfate can be enriched with ^{15}N. In this way, this compound, which is a nitrogen fertilizer, can be enriched with ^{15}N to an abundance of, e.g., 10.365%, having therefore an atom excess of 10.000%. This fertilizer can be used as a nitrogen tracer in agronomic studies. If a soil is fertilized with this compound, a plant growing on this soil will have two sources from where to draw nitrogen, the native soil N, with an abundance of 0.365%, and the fertilizer N, with an abundance of 10.365%. The tracer makes it possible, for instance, to find out the proportion by which each source is used by the plant, and so find out the efficiency of the fertilizer use, the strength of the soil as a N source, and several other agronomic features of the system. An important concept is the amount of nitrogen in the plant derived from the fertilizer, termed Ndff, calculated according to:

$$\%\ \text{Ndff} = \frac{\text{atom}\ \%\ ^{15}\text{N excess in plant material}}{\text{atom}\ \%\ ^{15}\text{N excess in fertilizer}} 100 \quad [3]$$

If in the above example the analysis of the plant results in an abundance of 7.855%, or in an atom excess of 7.490%, the Ndff will be 65%, meaning that, of all nitrogen in the plant, 65% comes from the fertilizer and 35% from the soil.

Another good example of the use of stable isotopes as tracers is the evaluation of atmospheric nitrogen fixation by legumes. Legume crops have the ability to fix atmospheric N_2, which has also the natural abundance of 0.365%, an agronomically very important fact because this means they have a lower demand for fertilizer N. When legumes are grown together with other nonfixing crops, under the same conditions and fertilized with ^{15}N-enriched material, the legume has three N sources: soil, fertilizer, and atmosphere, while the nonfixing crop has only the soil and the fertilizer. Therefore, at a given time, the enrichment of the legume is lower in comparison with the nonfixing plant. By difference it is possible to measure the amount of nitrogen derived from the atmosphere, Ndfa, absorbed by the legume:

$$\%\ \text{Ndfa} = \left[1 - \frac{\text{atom}\ \%\ ^{15}\text{N excess in legume}}{\text{atom}\ \%\ ^{15}\text{N excess in nonfixing crop}}\right] 100 \quad [4]$$

Another way of using stable isotopes as tracers is through natural variations in their abundance. As already mentioned, dynamic processes lead to isotopic discrimination, and, since these processes are active all over the environment, small but measurable variations in isotopic abundance occur in nature. Plants discriminate ^{15}N from ^{14}N, ^{13}C from ^{12}C. Processes occurring in the soil can also discriminate these isotopes and those of water, i.e., H and O. The enormous cycling of water due to the processes of evaporation, condensation, sublimation, exchange, and others induces a large range in natural water isotopic compositions. The water in an evaporation tank becomes increasingly enriched in the heavier hydrogen and oxygen isotopes, which are left behind at a slightly higher rate than the lighter isotopes.

Since the variations in abundance are rather small, a more sensitive measure is used, which is the isotopic ratio given in delta (δ) values. For the $^{13}C/^{12}C$ ratio, the delta value of a given sample is expressed by:

$$\delta\ ^{13}\text{C}\% = \left(\frac{^{13}\text{C}/^{12}\text{C in sample}}{^{13}\text{C}/^{12}\text{C in standard}} - 1\right) 1000 \quad [5]$$

Convenient and reproducible standards are internationally set and are available for calibration purposes.

The most commonly used isotopic ratios are $^{13}C/^{12}C$, $^{15}N/^{14}N$, $^{2}H/^{1}H$, and $^{18}O/^{16}O$ (Table 2). They are used in a great variety of environmental studies. Plants, for instance, discriminate isotopes in

Table 2 Principal stable tracer isotopes used for soil, plant, and environmental studies

Element	*Isotope*	*Abundance*
Carbon	^{12}C	99.985
	^{13}C	0.015
	^{14}C	Traces
Hydrogen	^{1}H	98.89
	^{2}H	1.11
	^{3}H	Traces
Nitrogen	^{14}N	99.635
	^{15}N	0.365
Oxygen	^{16}O	99.759
	^{17}O	0.037
	^{18}O	0.204
Sulfur	^{32}S	95.0
	^{33}S	0.76
	^{34}S	4.22
	^{36}S	0.014

the photosynthesis process. Thus C_3 (Calvin cycle) species have $\delta^{13}C\%$ values of -25 to -30, while C_4 (Hatch and Slack pathway) species, of -11 to -14. This makes it possible to study carbon cycling in different ecosystems, as is the case in comparisons of natural forests C_3 with introduced pastures C_4. Differences in the pasture intake by cattle can also be studied by analyzing the isotopic ratios in their feces.

Hydrologic studies use the $^{18}O/^{16}O$ ratio in many different ways. This ratio is different for sea water (taken as standard), fresh water (to the extend that different water bodies can present consistent differences), rainfall, water, soil water, etc.

Use of Radiation in Soil–Plant Investigations

Gamma Attenuation

Gamma rays are electromagnetic waves of the same nature as light, having much shorter wavelength, and consequently much higher frequency and energy. They are emitted by unstable radioactive nuclei and penetrate into matter, opening the possibility to explore the inside of materials. A gamma-ray beam is attenuated during the travel through matter, and this process is used to characterize the matter in question. Several isotopes are used as gamma sources, but the most frequently employed ones are ^{60}Co, ^{137}Cs, and ^{241}Am. ^{60}Co has two gamma rays of high energy: 1.17 MeV and 1.33 MeV; ^{137}Cs has one gamma of 661 KeV, and ^{241}Am also one gamma of low energy, 60 KeV. Beams of these gamma rays are used to study soil physical properties such as bulk density (d; kilograms per cubic meter) and water content (θ; cubic meters per cubic meter). The gamma attenuation follows the Beer law and, applied to a moist soil sample, is as follows (see Figure 1):

$$I = I_0 \exp[-x(\mu_s d + \mu_w \theta)] \qquad [6]$$

The incident I_0 and the emergent I beams can easily be measured with scintillation detectors coupled to electronic equipment that have an output in cps or cpm. The mass attenuation coefficients μ_s of the soil material and μ_w (square meters per kilogram) of water have to be known, but are also easily measured experimentally. The sample thickness x (meters) has also to be measured, and in this way d can be calculated when the soil water content θ is known, or θ is calculated when d is known. Using a double-energy gamma-ray beam, both d and θ can be measured simultaneously.

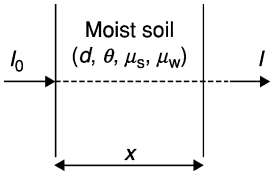


Figure 1

The great advantage of this technique is that it is nondestructive, the gamma-ray beam passes through the sample with no interference. Therefore, it is possible to make repeated measurements at the very same point, like following soil-moisture changes with time in a pot with a growing plant. Since the sample thickness x has to be known, the samples have to be regular-shaped or placed in containers. Although being developed for the laboratory, a few instruments use this technique in the field; they are gamma probes, for soil surface or greater depths, when introduced in the soil using aluminum access tubes.

Tomography

Computed tomography (CT) is a fairly new technique in soil science. It is basically the same as gamma- or X-ray attenuation, but the result is a picture of a cross-section of the sample showing a bidimensional distribution of d and/or θ. While the sample rotates along the plane of the desired cross-section, a large number of I_0 and I measurements are made without the need to know each value of x. A computer program processes the data in order to obtain the image. Only small, first-generation tomographs are used in soil research, but, in the near future, third-generation instruments will give instantaneous pictures of d and θ distributions in soil samples.

Neutron Moderation

This physical process has been used throughout the second half of the twentieth century to measure successfully soil water contents in the field. When a source of fast neutrons (high energy, 2–10 MeV) is placed into a soil, these neutrons become moderated or slowed down by collisions with the atoms of the surroundings, i.e., they lose energy down to the order of fractions of electronvolts. The most suitable neutron source used today is of ^{241}Am. The moderation process is proportional to the H content of the sample, and since H is part of soil water, the number of slow neutrons is proportional to soil-water content. A detector, sensitive only to slow neutrons, gives us their count rate in cps or cpm. An adequate calibration between the count rate and θ, for samples of known θ, makes it possible to estimate soil-water contents in other situations.

Neutron probes are commercially available instruments to measure soil-water content in the field. There are surface and depth neutron probes, and also some combined with gamma probes, to measure

simultaneously d and θ. They are widely used to measure soil-water availability to crops, irrigation-water needs, and water balances of crops and natural ecosystems.

Sterilization and Mutation Breeding

At very high intensities, gamma and other types of radiation have a sterilizing effect and can be used to sterilize materials. Soils and other laboratory materials are sterilized by irradiation. For this, mostly gamma but also neutron sources of high intensity are confined in special irradiation chambers. The dose to be applied depends on the distance between sample and source, strength, type, and energy of the radiation, and exposure time. In biological materials, radiation has also the property of inducing mutations. In many situations plant breeders induce mutations in plants, to increase the probability of obtaining beneficial genetic changes in agronomically important crops.

Other Nuclear Techniques

Activation Analysis

If there are some stable trace elements of interest in a sample that are difficult to measure or which cannot be easily extracted for detection, the activation analysis can be useful. It is a method of quantitative analysis in which the sample is exposed to a flux of bombarding particles, usually slow neutrons in a reactor or fast neutrons in a neutron generator, in order to provoke nuclear reactions with the elements of interest within the sample. The stable isotopes to be determined are thereby converted into radioactive isotopes that can be assayed by usual detection techniques. This technique has been used with success in a wide range of studies, for example in determining trace elements in soils, fertilizers, and plants, and determining the provenance of archeological specimens.

X-Ray Fluorescence

The principle of X-ray fluorescence spectrography is that characteristic secondary X-rays are produced when cathode rays of high energy interact with a target of a particular element. This means that every element emits its own secondary X-ray spectrum when bombarded by primary high-energy X-rays. In the case of complex samples such as soil, it is necessary to separate and identify each element of interest, recognizing their specific secondary X-rays. The sample of the material to be analyzed is irradiated with a beam of high-energy X-rays, and the fluorescent X-rays produced in the material are directed through a crystal, following the Bragg law, and counted by a specific counter. Spectral analysis is made automatically by computed multichannel systems. The most sensitive elements suitable for this analysis lie between ^{21}Sc and ^{56}Ba, being ideal for ^{30}Zn, ^{38}Sr, and ^{42}Mo.

Autoradiography

Ionizing particles emitted by radioisotopes have the same photochemical effect on photographic emulsions as do light rays. If the radioisotope of interest emits ionizing radiation of adequate energy, i.e., with a relatively short range to induce a clear image on the photographic emulsion, an autoradiograph can be made. The plant specimen containing the radioisotope, such as root, leaf, seedling, is closely applied to a photographic film and left for an appropriate time. An image is obtained after development of the film, which permits ready location and visualization of the distribution and concentration of the tracer.

C-14 dating The radioisotope ^{14}C has a half-life of 5720 years and, since it is continuously produced by sunlight-induced nuclear reactions that take place in the atmosphere, its concentration is fairly constant in atmospheric CO_2. Metabolic processes in living organisms also ensure constant C and ^{14}C concentrations in their bodies; however, after death, the carbon exchange with external sources stops, and the ^{14}C concentration decreases gradually according to the exponential radioactive disintegration law. Therefore, the ^{14}C concentration of paleobiological materials will be much lower and proportional to their age. This technique is called carbon dating, and has been used successfully in many situations, including determinations of organic carbon turnover rates in soil.

See also: **Nitrogen in Soils:** Cycle; Plant Uptake

Further Reading

Bacchi OOS, Reichardt K, and Calvache M (2002) *Neutron and Gamma Probes and Their Use in Agronomy*. IAEA Tecdoc. Vienna, Austria: International Atomic Energy Agency.

Gusmán MEJ (1989) *Nucleonica Basica*. Bogotá, Colombia: Instituto de Asuntos Nucleares.

Hardarson G (1990) *Use of Nuclear Techniques in Studies of Soil–Plant Relationships*. Training Course Series No. 2. Vienna, Austria: International Atomic Energy Agency.

L'Annunziata MF (1987) *Radionuclide Tracers*. San Diego, CA: Academic Press.

Slater RJ (1990) *Radioisotopes in Biology: A Practical Approach*. New York: Oxford University Press.

Vose PB (1980) *Introduction to Nuclear Techniques in Agronomy and Plant Biology*. Oxford, UK: Pergamon Press.

ISOTROPY AND ANISOTROPY

T-C J Yeh and P Wierenga, University of Arizona, Tucson, AZ, USA
R Khaleel, Fluor Federal Services, Richland, WA, USA
R J Glass, Sandia National Laboratory, Albuquerque, NM, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

The directional behavior (isotropy or anisotropy) of soil hydraulic conductivity is of primary importance in analyzing rates and directions of water flow and solute transport in the vadose zone. While the hydraulic conductivity of a soil may vary from one location to another (heterogeneity), it can also vary from one direction to another at the same location (anisotropy). The directional behavior in the hydraulic conductivity of a soil governs many hydrologic processes, such as infiltration, recharge, evaporation, and runoff, as well as solute migration. Field soils with a large hydraulic-conductivity anisotropy may result in extensive lateral spreading of fluid solutes, and contaminants in the vadose zone. Such spreading can potentially impact groundwater recharge, irrigation scheduling, and plant root development and microbial activity.

Darcy Law

The Darcy law is the fundamental principle underlying the study of water flow in porous media. It states that the specific discharge of water in a porous medium is linearly proportional to the hydraulic head gradient and is in the direction of decreasing hydraulic head. The constant of proportionality is the hydraulic conductivity. A general form of Darcy law for three-dimensional flow through unsaturated media can be written as:

$$q_x = -K(h)\frac{\partial \phi}{\partial x}, \quad q_y = -K(h)\frac{\partial \phi}{\partial y}, \quad q_z = -K(h)\frac{\partial \phi}{\partial z} \qquad [1]$$

In eqn [1], q is the specific discharge (LT^{-1}) and subscripts x, y, and z denote directions. The hydraulic head, $\phi = h + z$, has a unit of length (L). The pressure head is h (L), which is positive for saturated media and is negative for unsaturated media, and z is the elevation head (L), which is positive upward. The hydraulic conductivity (LT^{-1}) is denoted by $K(h)$, which is the saturated hydraulic conductivity when the medium is fully saturated and is the unsaturated hydraulic conductivity when the media is partially saturated. The unsaturated hydraulic conductivity is a function of the pressure head, decreasing from the saturated hydraulic conductivity with increasing negative pressure head, or as the medium becomes less saturated.

At the pore scale, water flow in porous media takes place through a complex network of interconnected pores or openings. To describe such an intricate network in any exact mathematical manner is practically impossible. Consequently, the Darcy law considers only the average flow behavior over a certain volume of porous media, which must be greater than several pores. The volume over which the flow is averaged is defined as a 'control volume' (CV). Using this CV approach, the Darcy law essentially bypasses both the microscopic level, at which what happens to each fluid particle is considered, and the pore-scale level, at which the flow pattern within each pore and between pores is considered. Analysis of flow in porous media, based on the Darcy law, thus moves to the macroscopic level, at which only average phenomena over the control volume are considered. The property defined at a point in the mathematical models therefore represents an average property over a CV and, thus, the property at every point in space varies smoothly such that the differential calculus applies. The medium and flow are subsequently being considered as the darcian continua. This continuum concept is parallel to the continuum hypothesis in fluid mechanics and other branches of sciences.

Control Volume, Representative Control Volume, Homogeneity, and Heterogeneity

Using the CV concept, porous media can be classified as homogeneous or heterogeneous in terms of hydraulic conductivity. If the hydraulic conductivity defined over a CV is constant throughout the entire porous medium in spite of the location of the CV, the medium is said to be homogeneous. Mathematically, homogeneity means that the hydraulic conductivity of the medium does not depend on the location. Under this condition, the CV is a representative elementary control volume (REV), implying that the hydraulic conductivity measured over the CV in any part of the medium is representative of the entire medium. This definition of REV mandates the size of REV to be much smaller than the entire medium such that the hydraulic conductivity defined over the REV is constant regardless of the location of the REV.

On the other hand, if the hydraulic conductivity defined over a CV varies from location to location, the medium is then characterized as being heterogeneous. Then, the REV does not exist for this medium. Mathematically, 'heterogeneity' implies that the hydraulic conductivity of a medium is a function of the location.

Isotropy and Anisotropy

The hydraulic conductivity over a CV at any point in a medium can be either isotropic or anisotropic. The hydraulic conductivity isotropy implies that hydraulic conductivity values of a medium are independent of the direction. Mathematically, an isotropic hydraulic conductivity is a scalar, a quantity with magnitude but no direction. Conversely, if the hydraulic conductivity over a CV differs for different directions, the hydraulic conductivity of the CV is said to be anisotropic. For example, the hydraulic conductivity of the CV measured in the horizontal direction may be greater or smaller than in the vertical. In general, a porous medium may be homogeneous and nevertheless anisotropic, or it may be heterogeneous and yet isotropic at each location. Under most field conditions, soils exhibit both heterogeneity and anisotropy.

To explain the causes of anisotropy in hydraulic conductivity, steady-state, saturated flow through a series of inclined layers of homogeneous porous media can be considered (Figure 1a). Each layer has a different hydraulic conductivity value, but the hydraulic conductivity of each layer is isotropic. If the upper and the lower boundary of the media are extended to infinity, the hydraulic head on the left-hand side of the medium is greater than the hydraulic head on the right-hand side, and they are maintained constant. Analysis of flow through each individual layer based on the Darcy law under the above specified conditions reveals that the specific discharge vector follows a zigzag path (solid vectors in Figure 1a) from the left to the right boundary.

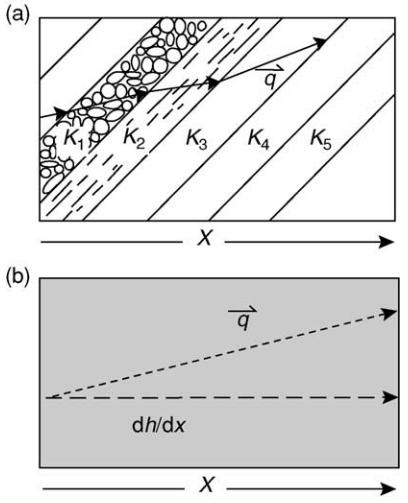


Figure 1 Schematics of flow through a layered soil (a), where K_1 is the hydraulic conductivity of each layer, $\vec{q}$ is the specific discharge vector within each layer, and an equivalent homogeneous and anisotropic soil (b) where dh/dx is the average hydraulic gradient, and $\vec{q}$ is the average specific discharge vector.

The series of layers of homogeneous porous media can be visualized as an equivalent homogeneous medium at a macroscopic level (Figure 1b) – homogenization – where the average (macroscopic) hydraulic gradient is uniform in the x direction (long dashed line). The average (macroscopic) specific discharge vector (short dashed line in Figure 1b) is no longer in the same direction as the average gradient. To produce such conditions, the hydraulic conductivity in the Darcy law for the three-dimensional homogeneous media thus becomes a hydraulic conductivity tensor with nine components. If a coordinate system is chosen in which the x-axis is perpendicular to the layers and the y- and z-axis are parallel to the layers, the number of components of the tensor can be reduced to three. The Darcy law in three dimensions for the equivalent homogeneous medium is then written as:

$$q_x = -K_{xx}(h)\frac{\partial \phi}{\partial x}, \quad q_y = -K_{yy}(h)\frac{\partial \phi}{\partial y}, \quad q_z = -K_{zz}(h)\frac{\partial \phi}{\partial z} \qquad [2]$$

In eqn [2], the hydraulic conductivity is the macroscopic hydraulic conductivity – no longer the same as the local-scale hydraulic conductivity of each layer. In addition, the macroscopic hydraulic conductivity values, $K_{xx}(h)$, $K_{yy}(h)$, and $K_{zz}(h)$, in x, y, and z directions, respectively, differ, depending on the direction. In general, the macroscopic hydraulic conductivity in the direction parallel to layers is greater than the hydraulic conductivity in the direction perpendicular to the layers. That is to say, the macroscopic hydraulic conductivity for the equivalent homogeneous medium for Figure 1a is not isotropic but anisotropic. Based on the preceding discussion, the macroscopic hydraulic conductivity anisotropy of the equivalent homogeneous medium is a result of several factors. They are namely variation in Ks of layers (local-scale heterogeneity), layers (structures), and most importantly, the homogenization (averaging) of the layered medium as an equivalent homogeneous medium.

Scales and Anisotropy

Anisotropy in the hydraulic conductivity varies with observation scale (i.e., the volume over which

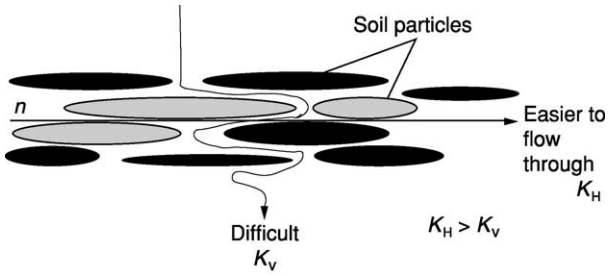


Figure 2 Pore-scale hydraulic-conductivity anisotropy, where K_v is the effective hydraulic conductivity in the vertical direction and K_H is the effective hydraulic conductivity in the horizontal direction.

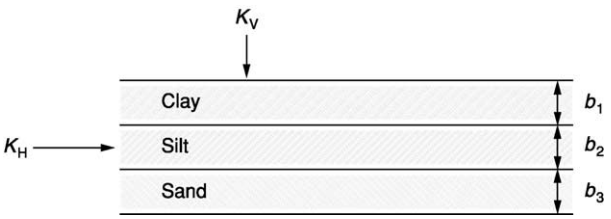


Figure 3 Field-scale hydraulic-conductivity anisotropy, caused by strata of different materials. K_v is the effective hydraulic conductivity in the vertical direction and K_H is the effective hydraulic conductivity in the horizontal direction. The thickness of each layer is denoted by b_1.

homogenization takes effect), as well as the scale of heterogeneity within the observation scale. Pore-scale hydraulic-conductivity anisotropy arises because the macroscopic hydraulic conductivity of a soil is determined over a certain volume of soil (e.g., a soil core). Within the soil volume, depositional processes cause flat particles (minerals) to orient themselves with the longest dimension parallel to the plane on which they settle. This produces flow channels parallel to the bedding plane, which allows fluid flow with little resistance. Fluid flow in the direction perpendicular to the flat surface of particles, however, must detour and take more tortuous and longer paths than for flow parallel to the bedding plane. Therefore, under the same hydraulic gradient, more flow can occur through the soil core if the gradient is parallel to the bedding plane than for perpendicular to the bedding. The hydraulic conductivity in the direction parallel to the bedding (K_H in Figure 2) is thus greater than in the direction perpendicular to the bedding (K_v in Figure 2). The soil core thus possesses a pore-scale anisotropy in hydraulic conductivity.

Field-scale hydraulic-conductivity anisotropy arises because when the hydraulic conductivity is determined in a field situation, the Darcy law assumes that there is homogeneity of the medium over a large CV. In essence, the soil is homogenized in a large CV that probably includes numerous large-scale structural heterogeneities such as stratification, cross-bedding, and clay lenses. For example, a stratified medium might be comprised of layers of clay, silt, and sand (Figure 3). Although within each layer the local-scale hydraulic conductivity may be considered homogeneous and isotropic, the macroscopic hydraulic conductivity for the equivalent homogeneous medium, an average over the conductivity values of the three layers, becomes anisotropic. In saturated media, the macroscopic hydraulic conductivity of the equivalent homogeneous medium in the direction parallel to bedding is given by a weighted arithmetic mean of hydraulic conductivity values of the layers:

$$K_H = \sum_{i=1}^{n} b_i K_i / \sum_{i=1}^{n} b_i \qquad [3]$$

In eqn [3], K_H is the macroscopic hydraulic conductivity in the direction parallel to bedding for the equivalent homogeneous medium; b_i and K_i are the thickness and the local-scale hydraulic conductivity of the ith layer, respectively; the number of layers is denoted by n. On the other hand, the macroscopic hydraulic conductivity in the direction perpendicular to bedding, K_v, for the equivalent homogeneous medium is given by a weighted harmonic mean of hydraulic conductivity values of the layers:

$$K_v = \sum_{i=1}^{n} b_i / \sum_{i=1}^{n} (b_i / K_i) \qquad [4]$$

For soil and geologic formations, where heterogeneity is not perfectly stratified, and the local-scale hydraulic conductivity exhibits a complex variation, stochastic methods have been used to derive the macroscopic hydraulic-conductivity anisotropy of the equivalent homogeneous medium. Based on stochastic analysis, the macroscopic hydraulic-conductivity anisotropy depends on several physical properties of the formations. These include the variance of local-scale hydraulic conductivity (the variation in hydraulic conductivity due to heterogeneity) and the correlation scale in different directions of the local-scale hydraulic conductivity (the average dimensions of the heterogeneity). The hydraulic-conductivity anisotropy of a fully saturated medium is therefore considered as an intrinsic property of the medium. The field-scale hydraulic-conductivity anisotropy is generally more significant than the pore-scale anisotropy for most geologic media.

Pressure Head- or Moisture-Dependent Anisotropy

For flow through unsaturated media, the unsaturated hydraulic-conductivity function can also be either

isotropic or anisotropic. In contrast to the isotropy and anisotropy in saturated hydraulic conductivity, the isotropy in unsaturated hydraulic conductivity means that the unsaturated hydraulic-conductivity function is the same in all directions. The unsaturated hydraulic conductivity is anisotropic otherwise.

Few studies in the past have investigated the unsaturated hydraulic-conductivity anisotropy, especially at the pore scale. The anisotropy in unsaturated hydraulic conductivity nevertheless has often been considered as being an intrinsic property of porous media, the same as the anisotropy in saturated hydraulic conductivity. Therefore, the unsaturated hydraulic-conductivity anisotropy has been treated by scaling the unsaturated hydraulic conductivity in different directions with the anisotropy of the saturated hydraulic-conductivity. The anisotropy of the unsaturated hydraulic conductivity thus remains constant over the full range of saturation or pressure head – a constant anisotropy.

Only recently, the anisotropy in the unsaturated hydraulic conductivity at field scales has been explored. A direct averaging approach has been used as a means to approximate the anisotropy. Following the approach for saturated flow, all local-scale unsaturated hydraulic-conductivity values at a given pressure head are arithmetically averaged. This average value is then used to represent the macroscopic unsaturated hydraulic conductivity for flow parallel to bedding at that pressure head:

$$K_{\mathrm{H}}(h) = \sum_{i=1}^{n} b_i K_i(h) / \sum_{i=1}^{n} b_i \qquad [5]$$

In eqn [5], $K_{\mathrm{H}}(h)$ is the macroscopic unsaturated hydraulic conductivity in the direction parallel to bedding at a given pressure head, h. The parameters b_i and K_i are the thickness and the local-scale hydraulic conductivity of the ith layer, respectively; the number of layers is denoted by n. On the other hand, the macroscopic unsaturated hydraulic conductivity in the direction perpendicular to bedding, K_{v}, for the equivalent homogeneous medium at a given pressure head is given by:

$$K_{\mathrm{v}}(h) = \sum_{i=1}^{n} b_i / \sum_{i=1}^{n} [b_i / K_i(h)] \qquad [6]$$

That is, all local-scale, unsaturated hydraulic conductivity values at a given pressure head are harmonically weighted. This average value is then used to represent the macroscopic unsaturated hydraulic conductivity for flow perpendicular to bedding at that pressure head. The approximations by eqns [5] and [6], however, follow the saturated flow case and do not consider the nonlinear pressure head distribution during flow in an unsaturated medium. Consequently, anisotropy based on eqns [5] and [6] can be unrealistically large.

The macroscopic anisotropy for unsaturated hydraulic conductivity at field scales has also been derived by using a stochastic approach, which considers the nonlinear pressure-head distribution in flow processes. The stochastic formulas relate the macroscopic unsaturated hydraulic conductivity to spatial statistics of parameters of the local-scale unsaturated hydraulic conductivity. For example, if the local-scale, unsaturated hydraulic conductivity follows an exponential function:

$$K(h) = K_{\mathrm{s}} \exp(\beta h) \qquad [7]$$

where K_{s} is the saturated hydraulic conductivity and β is the pore-size distribution factor (L^{-1}), the macroscopic unsaturated hydraulic conductivity for flow parallel to bedding, K_{H}, is given as:

$$K_{\mathrm{H}}(h) = \exp\left[F + \frac{\sigma_{\mathrm{f}}^2}{2(1+B\lambda_{\mathrm{z}})} - \left(B - \frac{(2\lambda_{\mathrm{z}} - |h|)}{2(1+B\lambda_{\mathrm{z}})}\sigma_\beta^2\right)|h|\right] \qquad [8]$$

The macroscopic unsaturated hydraulic conductivity for flow perpendicular to bedding, K_{v}, is then expressed as:

$$K_{\mathrm{v}}(h) = \exp\left[F - \frac{\sigma_{\mathrm{f}}^2}{2(1+B\lambda_{\mathrm{z}})} - \left(B - \frac{(2\lambda_{\mathrm{z}} + |h|)}{2(1+B\lambda_{\mathrm{z}})}\sigma_\beta^2\right)|h|\right] \qquad [9]$$

In eqns [8] and [9], F is the mean value of $\ln K_{\mathrm{s}}$ (the natural logarithm of K_{s}), B is the mean value of β, σ_{f}^2 is the variance of $\ln K_{\mathrm{s}}$, σ_β^2 is the variance of β, and h is the mean pressure head. The formulation in eqns [8] and [9] assumes that the flow domain is perfectly stratified (λ_{z} is the average vertical thickness of the stratification (L), or the vertical correlation scale) and $\ln K_{\mathrm{s}}$ and β are uncorrelated. A ratio of eqn [8] to eqn [9] gives the anisotropy of the macroscopic-level, unsaturated hydraulic conductivity:

$$\frac{K_{\mathrm{H}}(h)}{K_{\mathrm{v}}(h)} = \exp\left[\frac{\sigma_{\mathrm{f}}^2 + \sigma_\beta^2 h^2}{(1+B\lambda_{\mathrm{z}})}\right] \qquad [10]$$

Equation [10] suggests a pressure head- or moisture-dependent anisotropy for unsaturated media. That is, the anisotropy for unsaturated K may increase with decreasing moisture content or with increasing pressure head. It may also decrease from the ratio at saturation to a ratio of 1 (isotropy) and then increase at lower degrees of saturation if K_{s} and β are

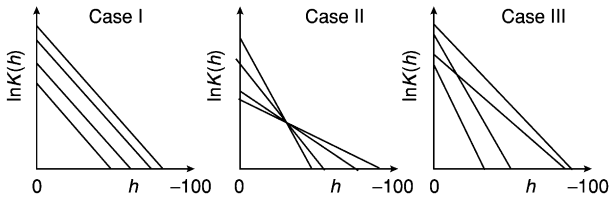


Figure 4 Local-scale, natural logarithm of unsaturated hydraulic-conductivity, ln$K(h)$, as a function of pressure head, h, in layered soils for three cases.

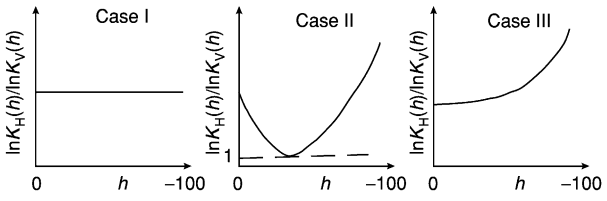


Figure 5 The corresponding anisotropy in the effective unsaturated hydraulic conductivity for the three cases shown in **Figure 4**. The natural logarithm of unsaturated hydraulic-conductivity in the vertical and horizontal direction is denoted by ln$K_v(h)$ and ln$K_H(h)$, respectively.

correlated. Also shown in eqn [10] is that several factors control the macroscopic unsaturated hydraulic-conductivity anisotropy. The factors are the mean, variance, and correlation scale of the saturated hydraulic conductivity and the pore-size distribution factor of the unsaturated hydraulic conductivity at local scales. More importantly, eqn [10] depicts the dependence of the anisotropy on the average pressure head or moisture content over the unsaturated media. Similar to hysteresis in moisture-retention curves, the moisture-dependent anisotropy may be subject to hysteresis effects.

While the direct averaging approach is strictly applicable to saturated media, it can be used to explain the cause of the moisture-dependent anisotropy. Consider the following three cases (**Figure 4**): For case I, a formation consists of many layers of porous media, in each of which the local unsaturated hydraulic conductivity has different K_s but the same β; in case II, each layer has a different K_s and β-value, but the values of K_s and β in each layer are positively correlated; case III considers each layer has a different K_s and β-value, but the K_s and β-values are independent. An arithmetic mean of all local unsaturated hydraulic conductivity values at a given pressure head yields the macroscopic unsaturated hydraulic conductivity for flow parallel to bedding at that given pressure head. In contrast, a harmonic mean is used to yield the macroscopic unsaturated hydraulic conductivity for flow perpendicular to bedding at the given pressure head.

Recognizing that a harmonic mean weighs heavily on the smallest $K(h)$, the anisotropy ratio can be plotted as a function of the pressure head for the three cases (**Figure 5**). Again, the anisotropy ratio is defined as the ratio of the macroscopic unsaturated hydraulic conductivity for flow parallel to bedding to the conductivity for flow perpendicular to bedding. Because the variation in $K(h)$ remains the same over the full range of the pressure head in case I, the anisotropy remains constant. In case II, the variation in $K(h)$ decreases toward a cross-over point and then increases with the pressure head. The corresponding anisotropy ratio first decreases, becomes unity (isotropic) at the crossover point, and then increases again. In case III, on the other hand, the variation in $K(h)$ always increases with the pressure and, in turn, the anisotropy grows with the pressure head.

While the moisture-dependent anisotropy has been derived for field-scale soils, similar moisture-dependent anisotropy at the pore scale has also been observed in the relative permeability of oil and gas measurements on sandstone cores. The sandstone, when dry or fully saturated, appeared to be homogeneous and isotropic. When the material was partially desaturated, however, thin and regular spaced strata became apparent. The air pemeability of the dry core was approximately twice as great parallel to the bedding planes as perpendicular to the bedding planes. Evidently, the material was quite uniform, but it was not isotropic. The effect of the anisotropy was to increase greatly the critical gas saturation and make the oil relative-permeability curve steeper when flow was across the bedding planes.

Influences of Anisotropy on Movement of Moisture Plumes in Soils

To illustrate impacts of the unsaturated hydraulic-conductivity anisotropy on water movement in unsaturated soils, the moisture content distributions after infiltration from a surface source in three different soils have been simulated. **Figure 6a** shows the simulated moisture-content distribution in a homogeneous soil with isotropic unsaturated hydraulic conductivity. In this homogeneous and isotropic soil, water moves predominantly in the direction of gravity. The moisture content distribution in **Figure 6b** corresponds to the simulated result for the same surface infiltration event in a homogeneous soil profile with a constant anisotropy in unsaturated hydraulic conductivity. **Figure 6c** shows the simulated moisture-content distribution in a homogeneous soil with a moisture-dependent anisotropy in unsaturated hydraulic conductivity. The moisture content distributions in **Figure 6b** and **c** show significant

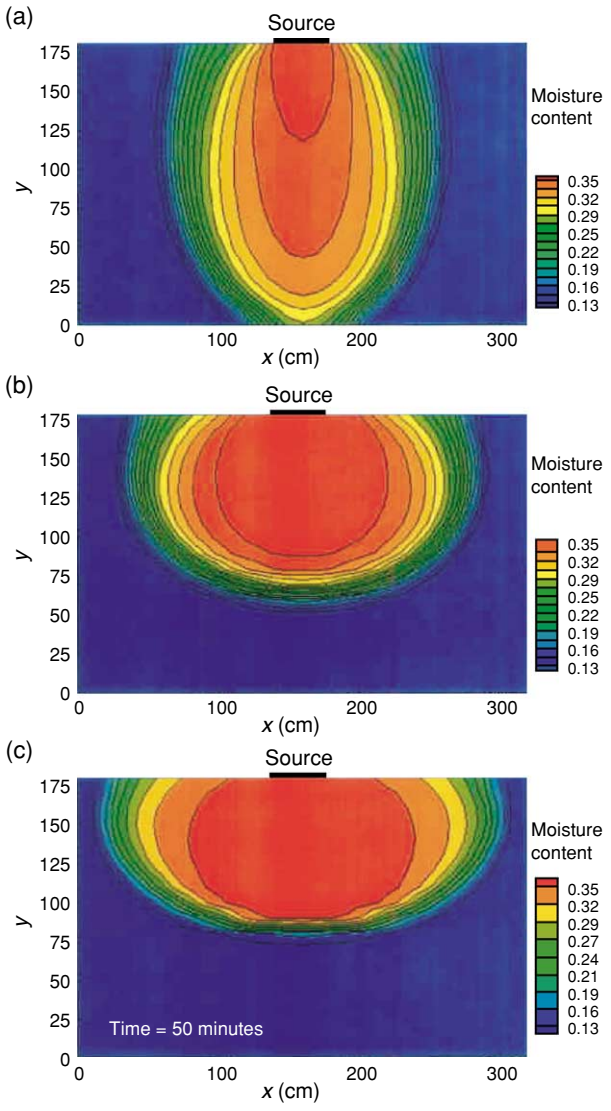


Figure 6 Simulated moisture-content contours in soil profiles after infiltration from a surface source: (a) a homogeneous and isotropic soil; (b) a homogeneous soil with a constant anisotropy; and (c) a homogeneous soil with a moisture-dependent anisotropy. The *x* and *y* denote the horizontal and vertical dimension of the soil profile, respectively.

lateral movement of water. The anisotropy in unsaturated hydraulic conductivity in the two anisotropic soils evidently inhibits vertical movement of water but enhances its lateral spreading. However, greater lateral spreading of water in the soil profile with the moisture-dependent anisotropy is evident than in the soil with the constant anisotropy.

Effects of the anisotropy in unsaturated hydraulic conductivity have also been observed in numerous field experiments. For example, Figure 7 is a snapshot of moisture and tracer plumes in an experiment conducted in Rio Bravo alluvial deposits in Albuquerque, New Mexico. The media, according to the photo, is

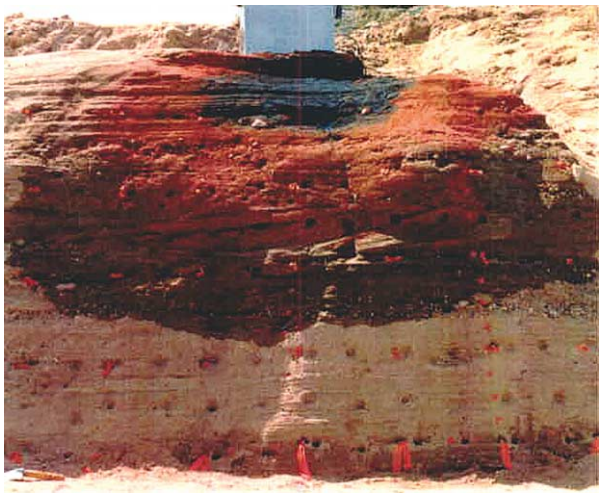


Figure 7 Moisture (dark area) and tracer (red and blue) plumes in the Rio Bravo deposits in Albuquerque, New Mexico, USA.

highly stratified, consisting of thin layers of materials of different textures. Moisture and tracer plumes spread great lateral distances and do not follow the prediction relying on a homogeneous isotropic concept – a manifestation of macroscopic anisotropy in unsaturated hydraulic conductivity. Similar phenomena have been reported in field observations of pollutant migration from a septic tank drain field in glacial outwash deposits in the Spokane Valley, Washington. At the Hanford site near southeastern Washington, extensive lateral water movement but limited vertical movement is also observed in high-level radioactive waste leakage in the unsaturated zone. Field observations in the loess-loam area north of Beer-Sheba, Israel show that rain infiltrates the soil to a limited depth with no net recharge of the groundwater – an effect of lateral flow in unsaturated soil.

More recently, detailed soil-water tracer experiments in the shallow soils of a sandy hillslope at an arid site near Socorro, New Mexico, indicate that significant horizontal downslope flow components develop despite the presence of a near-vertical downward hydraulic gradient. This suggests that, at *in situ* pressure heads, the dune sands behave as a highly anisotropic medium, although field and laboratory permeability analysis of the sands shows that the sands are nearly isotropic at complete saturation. The field and laboratory measurements have provided strong supporting evidence for the concept that, for a layered heterogeneous porous media, the macroscopic media anisotropy varies as the state (pressure, moisture, and saturation) of the media varies.

Finally, while the anisotropy is important to vadose zone processes, its effects on surface hydrology processes have also been reported. The anisotropy of

unsaturated soils and the slope of the land surface can produce strong lateral flow components, which, in turn, cause moisture accumulation in concave parts of the landscape, leading to the point of saturation. The concentration of water in concave areas explains some rainfall runoff and erosion phenomena that were previously unexplained by classic concepts of infiltration.

List of Technical Nomenclature

$\boldsymbol{\beta}$	pore-size distribution parameter (L^{-1})
$\boldsymbol{\lambda_z}$	vertical correlation scale, average vertical thickness of the stratification (L)
$\boldsymbol{\sigma^2_\beta}$	variance of β (L^{-2})
$\boldsymbol{\sigma^2_f}$	variance of $\ln K_s$
$\boldsymbol{\phi}$	hydraulic head (L)
$\boldsymbol{b_I}$	thickness of the ith layer (LT^{-1})
$\boldsymbol{B}$	mean value of β
$\boldsymbol{F}$	mean value of $\ln K_s$
$\boldsymbol{h}$	pressure head (L)
$\boldsymbol{K(h)}$	hydraulic conductivity (LT^{-1})
$\boldsymbol{K_H}$	hydraulic conductivity for flow parallel to bedding (LT^{-1})
$\boldsymbol{K_i}$	hydraulic conductivity of ith layer (LT^{-1})
$\boldsymbol{K_s}$	saturated hydraulic conductivity (LT^{-1})
$\boldsymbol{K_v}$	hydraulic conductivity for flow perpendicular to bedding (LT^{-1})
$\boldsymbol{K_{xx}}$	hydraulic-conductivity tensor component in the x direction (LT^{-1})
$\boldsymbol{K_{yy}}$	hydraulic-conductivity tensor component in the y direction (LT^{-1})
$\boldsymbol{K_{zz}}$	hydraulic-conductivity tensor component in the z direction (LT^{-1})
$\boldsymbol{\ln K_s}$	natural logarithm of K_s
$\boldsymbol{n}$	number of layers
$\boldsymbol{q_x}$	specific discharge in the x direction (LT^{-1})
$\boldsymbol{q_y}$	specific discharge in the y direction (LT^{-1})
$\boldsymbol{q_z}$	specific discharge in the z direction (LT^{-1})
$\boldsymbol{z}$	elevation head (L)

Further Reading

Bear J (1972) *Dynamics of Fluids in Porous Media*. New York: Elsevier.

Corey AT and Rathjens CH (1956) Effects of stratification on relative permeability. *Journal of Petroleum Technology* 69.

Crosby JW, Johnstone DL, Drake CH, and Fenston RL (1971) Migration of pollutants in a glacial outwash environment. 3. *Water Resources Research* 7: 713–720.

Gelhar LW (1993) *Stochastic Subsurface Hydrology*. Englewood Cliffs, NJ: Prentice Hall.

Mantoglou A and Gelhar LW (1987) Stochastic modeling of large-scale transient unsaturated flow. *Water Resources Research* 23: 37–46.

McCord JT, Stephens DB, and Wilson JL (1991) Hysteresis and state-dependent anisotropy in modelling unsaturated hillslope hydrologic processes. *Water Resources Research* 27: 1501–1518.

Mualem Y (1984) Anisotropy of unsaturated soils. *Soil Science Society of America Journal* 48: 505–509.

Palmquist WN Jr and Johnson AI (1962) Vadose flow in layered and nonlayered materials. *Geological Survey Professional Paper* 450: 142–143.

Polmann DJ, Mclaughlin D, Luis S, Gelhar LW, and Ababou R (1991) Stochastic modeling of large-scale flow in heterogeneous unsaturated soils. *Water Resources Research* 27: 1447–1458.

Stephens DB and Heermann S (1988) Dependence of anisotropy on saturation in a stratified sand. *Water Resources Research* 24: 770–778.

Ursino N, Gimmi T, and Flühler H (2001) Combined effects of heterogeneity, anisotropy, and saturation on steady flow and transport: a laboratory sand tank experiment. *Water Resources Research* 37: 201–208.

Wildenschild D and Jensen KH (1999) Laboratory investigations of effective flow behavior in unsaturated heterogeneous sands. *Water Resources Research* 35: 17–27.

Yeh T-CJ (1998) Scale issues of heterogeneity in vadose zone hydrology. In: Sposito G (ed.) *Scale Dependence and Scale Invariance*. New York: Cambridge University Press.

Yeh T-CJ, Gelhar LW, and Gutjahr AL (1985) Stochastic analysis of unsaturated flow in heterogeneous soils. 1. Statistically isotropic media. *Water Resources Research* 21: 447–456.

Yeh T-CJ, Gelhar LW, and Gutjahr AL (1985) Stochastic analysis of unsaturated flow in heterogeneous soils. 2. Statistically anisotropic media with variable α. *Water Resources Research* 21: 457–464.

Zaslavsky D and Sinai G (1981) Subsurface hydrology. IV. Flow in sloping layered soil. *Journal of the Hydraulics Division, Proceedings of the American Society of Civil Engineers* 107: 53–64.

J

JENNY, HANS

R Amundson, University of California–Berkeley, Berkeley, CA, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Hans Jenny, born 7 February 1899 in Basel, Switzerland, had a career that spanned nearly 70 years, ending only with his death on 9 January 1992. His youth was marked by a rigorous academic training, balanced by outdoor activities and experience on Swiss farms. This background led to his acceptance in the Swiss Federal Institute (ETH) in Zurich, where Jenny received a diploma in agriculture in 1922 and a DSc in colloid chemistry under Georg Wiegner in 1927 for work on ion exchange reactions. A Rockefeller Fellowship to work with Nobel Laureate Selman Waksman at Rutgers University in New Jersey, USA was followed by a position at the University of Missouri. In 1936, Jenny was appointed to a professorial position at Berkeley, which he held (as an emeritus after 1967) until his death. Jenny made significant contributions to the fields of colloid chemistry, ion exchange, and pedology. His most famous achievement is the book *Factors of Soil Formation* (1941), in which the diverse ideas of earlier scholars were combined with the abstract formalism of physical chemistry to produce a quantitative and revolutionary view of soils and ecosystems. He was the author of a biography of Eugene Hilgard, which renewed for many an insight into the significant contributions of Jenny's Berkeley predecessor. In later years, after 'retirement' in 1967, Jenny became noted for his mesmerizing lectures on soil and landscape painting and for his efforts, along with his wife Jean, on soil preservation. In 1982, at the age of 83, Jenny trekked Mount Kilimanjaro, collecting soil samples to test his hypothesis on the effect of latitude on soil formation, and subsequently performed the chemical analyses himself. Finally, he remained devoted to the science and preservation of his beloved Pygmy Forest, and the region of ancient, impoverished soils on the Ecological Staircase of Mendocino County, California. Today, Jenny's concepts on soil formation reverberate through the natural sciences. His work on soil organic matter established an observational and modeling framework that is the standard for modern soil carbon-cycling research, and his forward-looking ideas on soil esthetics and preservation are becoming topics of general research and application in the conservation arena.

Hans Jenny was born to Johann Jenny, a business accountant, who was himself the son of Johann, the owner of a chimneysweep business. Jenny's mother, Marie T. Althaus, was the daughter of a family of wood-carving factory owners in Meiringen. Jenny's first years in grammar school were in Basel. During his grammar-school period, Jenny's entire family moved to Meiringen, to live with Jenny's mother's family, because Jenny's father's business expertise was needed in the factory. Jenny remembered this period fondly. His grandfathers and uncles were artists, and they discussed and created modern art, an experience that early helped forge a lifelong interest in art and esthetics in the young Jenny. This period also introduced the young Jenny to farm life, and he spent much time on the farm adjacent to his grandparents' home (**Figure 1**).

Upon the family's return to Basel, his parents decided that he should attend a junior high school that focused on college preparation. As Jenny recollected:

> rigorous training began that transformed a näive adolescent into a sophisticated, somewhat smart-alecky young intellectual. . . . [but] throughout this period World War I raged at the Swiss frontiers.

Classes began early in the day, and homework kept the young students busy until after midnight. The demands on Jenny through the high-school period were traumatic. Great stress was placed on studies, particularly on rote memorization in modern languages, history, science, and mathematics. Jenny later recalled that he had a hard time memorizing words, and that because of broad-ranging extracurricular activities and tension in his parents' relationship he was just an "average student. . . and I know I had my share of feelings of frustration, and of being

Figure 1 Hans Jenny, early 1900s.

Figure 2 Georg Wiegner (left) and Hans Jenny (right), around 1925.

mediocre, which was considered worse than being a criminal." Jenny's 'mediocre' performance in school abruptly changed when he later entered the university.

In high school, Jenny joined the Boy Scouts, which proved to be an opportunity for him to excel in a different arena of activity and to develop leadership qualities. Jenny's particular troop emphasized an 'ascetic' approach to scouting: fasting, no display of merit badges, and no vices. Jenny was good with ropes, compasses, etc., skills which he attributed to his previous experiences on the farm. At the same time, the war raged around Switzerland, and Jenny would hear the thunder of cannons at night and see planes escaping into Basel pursued by antiaircraft fire. The wounded German and French soldiers in Basel, the families who had lost sons, and the Allied and German war propaganda left a strong impression on Jenny and his friends. Food shortages in Europe became prominent, and Jenny spent summers during his high-school years working on a farm in the Emmenthal Valley, eventually being placed in charge of horses, oxen, and plowing.

When his high-schooling ended in 1918, he was awarded a certificate to enter any Swiss university and chose to study agriculture. However, Jenny first spent a year working on farms in both German and French Switzerland. This period provided him with much practical experience about farming, and he entertained the notion of becoming a livestock manager following his studies at the university. The farming experience also had other effects on Jenny. The acquisition of physical skills, and the additional year of maturity, gave him a more serious perspective as a university student, and Jenny there focused his efforts on the educational opportunities that surrounded him. Many of the faculty at Zurich were internationally famous, and, as a student in agriculture, Jenny eventually took courses from Georg Wiegner in agricultural chemistry, an exposure that shifted his interests toward colloid chemistry and a postgraduate degree. Most notably, Jenny was attracted to Wiegner because he "expounded on principles, concepts, and ideas, rather than practical recipes."

In 1923, Jenny received his diploma in agriculture, and approached Wiegner about research in soil chemistry. Wiegner responded by insisting that Jenny first acquire a more solid foundation in chemistry. Wiegner was widely famous, and attracted students and postdoctoral scholars from across Europe as well as from overseas. Soon, Jenny became Wiegner's chief assistant in soil and colloid chemistry. In contrast to Wiegner's laboratory interests, Jenny began to spend weekends in the field, collecting and subsequently analyzing soil samples (Figure 2). This activity

brought him into contact with Josiah Braun-Blanquet, with whom he later collaborated on the relationship of soil formation and plant development.

As a result of Wiegner's preeminence in ion exchange, Jenny was encouraged to study the individual behavior of ions during exchange in relation to their degree of hydration. Jenny found that exchange behavior could not be explained by the activities of the ions, but instead must involve complex reactions between mineral surfaces and the hydrated forms of ions. Wiegner and his students moved forward on this problem from several fronts, including conducting the exchange reactions in alcohol, in which the ions were dehydrated. Jenny recalled that the laboratory environment in Wiegner's group was interactive and dynamic, and that "doing research was an exciting cross-fertilization experience." In fact, the field of colloid chemistry at that time held a central position in chemistry and physics, a situation that made research even more exciting for the young scientists in the laboratory. Among the great advancements in that time period was Einstein's Theory of Brownian Movement, which linked molecular theory with the behavior of colloidal particles, revealing a continuity of molecules and atoms and colloidal particles that provided an overall view of the corpuscular system of nature.

Yet, on a completely different scale, Jenny noted that he had developed a reputation, within a small circle, through his interest in linking laboratory and theoretical colloidal chemistry with soils in nature. Working with Dr. Braun-Blanquet, who was an innovator in the new field of plant sociology, they explored the Swiss National Park, linking soil properties with plant distribution. In this report, Jenny drew a speculative curve relating soil organic matter to climate, relying on some of Eugene Hilgard's analyses (whose position Jenny was later to fill at Berkeley). The book, entitled *Vegetation Development and Soil Formation*, was reportedly later called a classic by R. Tüxen in the 1950s, but Jenny jokingly called it a *Jugendsünde* (an illegitimate product of youth).

Near the end of Jenny's doctoral research, Dean Mann from Cornell University, representing the Rockefeller Foundation's International Education Board, visited Wiegner's laboratory and asked Jenny if he was interested in spending a year at an American university. This was something Jenny had already considered, and he asked if he could go to Berkeley to work with Dennis Hoagland in plant nutrition. Mann discouraged this idea, suggesting that Jenny select an Atlantic-coast institution. Shortly after this, Selman Waksman from Rutgers visited the Wiegner laboratory and suggested that Jenny apply to work in his laboratory in New Jersey. Jenny's application, with Waksman's support, was approved and in 1926 Jenny left for America. The voyage itself proved to be memorable for Jenny, for he was paid to travel first class (and was therefore obligated to purchase a tuxedo), but he spent much of the time being seasick rather than enjoying the amenities and improving his English.

The short year at Rutgers was stimulating for Jenny. Waksman proposed that Jenny conduct litter-decomposition experiments, examine the colloidal character of the humus produced, and characterize the microbes involved. In later years Jenny remarked that the questions and approaches were sound, but the dismal conditions of Waksman's basement laboratory (occupants were forced to wear overboots during rain storms) and the sloppy behavior of the students made Jenny decide to switch his research to the effect of ion type on plant transpiration. This work was conducted with Professor Shive. Jenny remarked that Waksman, who was probably disappointed that Jenny switched laboratories, remained a professional and personal friend until his death, and that, although Jenny never completed his research with Waksman, he was stimulated by the ideas and discussions that Waksman's group engaged in. Jenny ultimately recognized that laboratory conditions were not always indicative of success, for he remarked: "Years later I realized that laboratory life-styles are not crucial, for it was Waksman who got a Nobel prize, not Wiegner."

The American lifestyle encountered by Jenny proved to confirm some of his European biases and debunk others. Jenny was shocked by the gum-chewing of students, and their lack of interest in international affairs, philosophy, and art. He said: "What surprised me right away was that everybody put their feet on the table or another thing, so I drew that .... [entitled] A Landscape in the USA!" (**Figure 3**). On the other hand, Jenny was greatly impressed by the experimental ingenuity of American chemists, and by facets of American chemistry which Europeans had ignored or neglected.

While at Rutgers, Jenny and a Danish colleague, J.H. Blom, were asked to translate submissions to the upcoming First International Congress of Soil Science, organized by Dean Jacob Lippman at Rutgers, which was to be held in Washington, DC. Jenny later remarked that he and Blom were "quite malicious" and "had fun in the translations, giving a certain rhythm as that of a hexameter or a pentameter, making fun of pompous writing." As the Congress approached, Wiegner was able to have Jenny appointed as an official Swiss delegate, which reduced the financial burden and also enabled him to

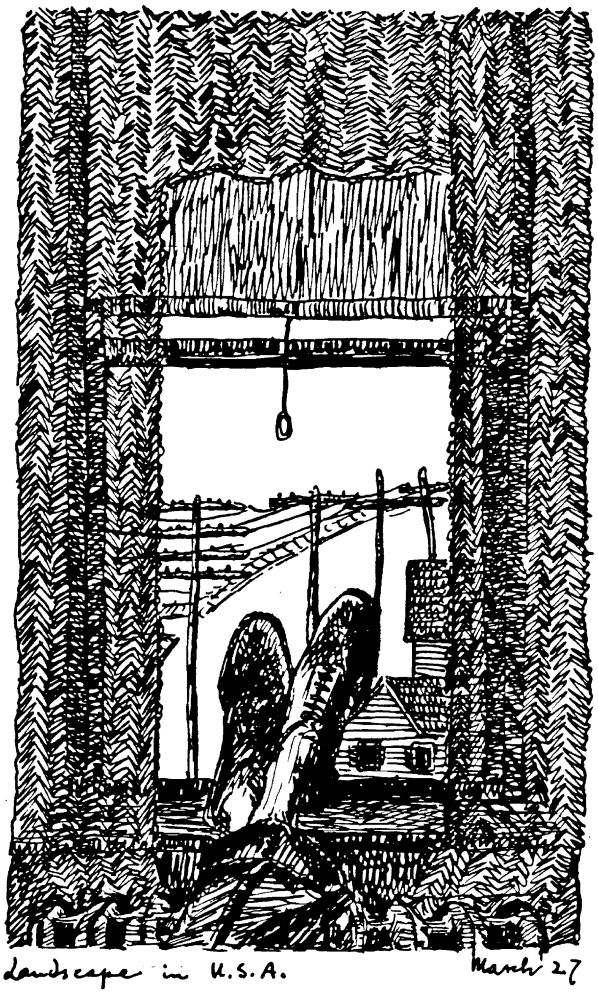


Figure 3 Hans Jenny's first sketch in the USA, 1927.

attend the important postcongress, transcontinental soils excursion.

The postcongress excursion appears to have been a remarkable trip, bringing forth both cultural and scientific integration. A select group of scientists of differing ages and backgrounds were confined to a private train, with Pullman cars, that traversed North America from the Atlantic to the Pacific. Jenny later remarked on the tensions and conflicts between the sometimes abrasive youth and the older, established leaders – as well as between scientists of differing backgrounds. Importantly, the trip, which passed through enormous ranges of climate, opened Jenny's mind further to the relationships between soil types: "The rolling plains, I fancied, must harbor the secret of mathematical soil functions. At times I could hardly sleep thinking about it." During the Congress, Robert Bradfield, from the University of Missouri, decided to go to Wiegner's laboratory for a sabbatical, and offered his laboratory and teaching position to Jenny for the year, which Jenny "accepted with great pleasure."

In the autumn of 1927, Jenny arrived in Columbia, Missouri. The physically isolated but vibrant academic setting enabled him to follow through on the ideas which were conceived on the excursion earlier that summer (**Figure 4**). Professor K.K. Krusekopf informed Jenny of the availability of soil C and N data from various states, and Jenny then began assembling data from across the Plains states, first arranging it by temperature (Canada to Louisiana). Jenny developed equations to describe the trends

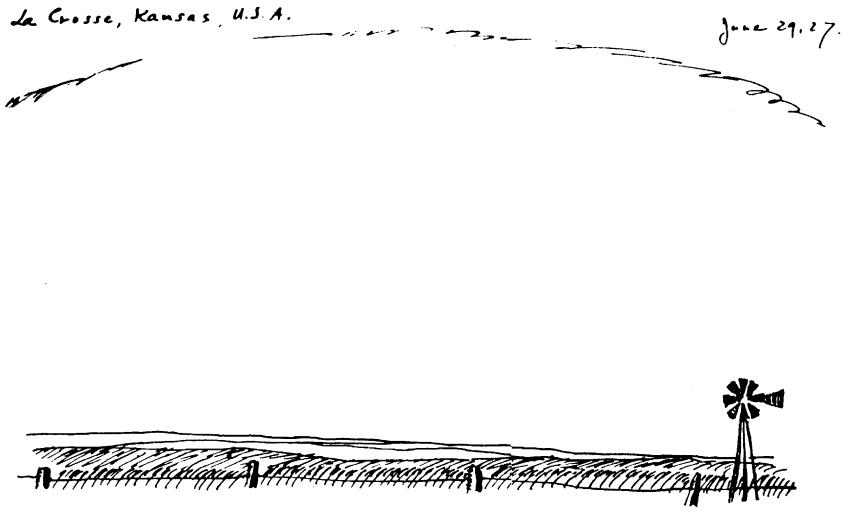


Figure 4 Jenny's sketch of Kansas made during the Transcontinental Excursion, 1927.

("primitive modeling" as he later described it) and showed it to his colleagues, who encouraged him and suggested he present the paper that November at the meetings of the American Soil Survey Association in Chicago. The paper, which was subsequently published, quantitatively explored the climatic effects on soil N, and this effort earned Jenny a research award from the American Society of Agronomy in 1931. In this and subsequent papers, data were aligned in graphs and described mathematically, and Jenny remarked:

> I enjoyed seeing field data aligned by equations and derived aesthetic pleasure from the shapes of the curves. Several pedologists, however, accused me of trying to be erudite.

In addition to his work on the climatic effects on soil N, Professor M.F. Miller suggested to Jenny that he examine the effects of cultivation on soil organic matter. A study between a native pasture soil and an adjacent cultivated field, along with data derived from the literature, produced a seminal paper and report on the loss of C and N during cultivation, and its rate. Jenny later remarked that this study added the factor of 'time' to his growing list of what would later become the famous factors of soil formation.

In addition to Jenny's pedological research at Missouri, he was greatly engaged in renewing and expanding his research on colloids and ion exchange. This interest was partially the result of large conceptual differences held by Robert Bradfield, Jenny's Missouri colleague, and Wiegner. One of the major experimental differences in Missouri was that Jenny worked with natural clays, in contrast to artificially made colloidal gels used in Zurich. A result of this work was a statistical model of ion exchange, one derived during late hours at night in his basement office during the hot and humid Missouri summer. About this period, and his parallel work in pedology and colloid chemistry, Jenny asked himself: "Am I doing the right thing... but both avenues proved interesting and beneficial to me, so why should I give up either one? I could have two sweethearts." An important, but not widely known additional area that Jenny also explored at this time was the differential behavior of K and Na during chemical weathering, work in a topic now widely recognized as geochemistry. This work, published in the Missouri Agricultural Experiment Station Bulletin was, as Jenny noted, not widely read. But Jenny sent a copy to V.M. Goldschmidt, who later cited the work in his discourses on geochemistry.

With the Great Depression still in full force, the University of Missouri was hit by financial difficulties. Young faculty were laid off, though Jenny (then an assistant professor) was spared. In 1934, faculty members were forced to take a year's leave of absence without pay, and Jenny's turn was set for 1934–35. With the assistance of his old Zurich friend Max Kleiber (then at the University of California–Davis), Jenny was able to obtain an appointment at the Citrus Research Station at Riverside, California. There, Jenny collaborated with W.P. Kelley, an arid-lands soil chemist and the codiscoverer of the crystalline nature of soil clay minerals. Jenny also, with E.R. Parker, developed a study to examine the role of tillage on water infiltration, and the time required for adverse tillage effects to disappear. While at Riverside, Kelley encouraged Jenny to visit Berkeley, where he delivered one or two lectures in the Department of Plant Nutrition. Soon, Jenny was offered the position of Associate Professor of Soil Chemistry and Morphology, which he accepted.

Upon his arrival at Berkeley, Jenny became a member of the Plant Nutrition Department, under the leadership of the famous plant nutritionist Dennis Hoagland. The Department contained numerous faculty, including Roy Overstreet. Overstreet, a soil physical chemist, was a colleague with whom Jenny was able critically to discuss his ideas and with whom he conducted joint research. Jenny repeatedly recounted the great scientific and personal pleasure that his discussions with Overstreet produced, and Jenny gives great credit to Overstreet for rigorously challenging his ideas and concepts. At the time of Jenny's arrival, the Department of Soil Technology, headed by Charles Shaw, contained people devoted to more pragmatic soil research agendas. Jenny notes that conflicts between the two groups existed, but that Jenny's interests in pedology led him to interact with Shaw, who told Jenny he (Jenny) "was going to be the new Dokuchaev." Shaw suddenly died, and Jenny and Overstreet, along with other soil researchers (soils people), then moved to a new Department of Soils.

The breadth and depth of Jenny's scientific accomplishments at Berkeley are truly remarkable. He later described his preceding years in Missouri as his '*Sturm und Drang*' (storm and stress) period because, while he was rapidly developing ideas (and publishing) in colloidal chemistry and pedology, he was haunted with insecurities ("Sometimes you have nightmares ... is what you are doing all wrong?"). Berkeley provided Jenny with colleagues and collaborators who challenged him and questioned him, but encouraged him to become more confident in his emerging research.

Upon his arrival, Jenny continued to pursue his twin 'sweethearts' of colloid chemistry and pedology.

It is instructive to remember that his hiring was based on his expertise in colloids, and he immediately began teaching a rigorous and widely attended course on the topic – one attended by students from across the campus. In turn, in preparation for his courses and research, Jenny himself also audited classes on optical mineralogy, geography, and mathematics. In terms of research in colloidal chemistry/ion exchange, some of Jenny's achievements are remarkable.

First, Jenny's ion exchange expertise was applied to a problem, brought to him by Hoagland, regarding the use of ammonia gas for fertilization. Jenny demonstrated that ammonia could be applied directly to soil and that, through ion exchange, the ammonia would be held on exchange complexes until is was nitrified or bioaccumulated. Jenny wrote a memorandum to Hoagland illustrating the manner in which the gas could be injected into the soil, and the adsorption mechanisms involved. Jenny shared the memorandum with a representative of Shell Oil, who himself then patented the idea shortly thereafter. Though Jenny recognized the missed financial opportunities and prestige for the university and himself, he was pleased to see that basic scientific research in ion exchange could yield such widespread and useful applications to agriculture.

Second, following his earlier work at Missouri and before, and stimulated by his new colleagues (particularly Hoagland), Jenny and Overstreet set out conceptually and experimentally to explore the concept of 'contact exchange,' or the direct ion exchange between roots and soil colloids, in ion uptake by plants. Their work involved refined concepts of the nature of soil water (or the differing character of soil water adjacent to colloids and roots) and clever experiments with roots and gels to illustrate exchange reactions.

Additionally, in relation to contact exchange, Jenny and Overstreet (with the help of other colleagues) quickly recognized the utility of radioisotopes, then being produced by Berkeley's new Radiation Laboratory under the direction of Ernest Lawrence, in their research. Jenny recalled the intellectual excitement and potential that the Radiation Lab caused on the campus at the time:

> I remember a lecture given by J. Robert Oppenheimer where he actually drank a glass of water with radioactive sodium chloride in it, and then showed two or three minutes later that his fingers were radioactive. I guess he wouldn't do that anymore...

Jenny and Overstreet received solutions of radioactive potassium (treated by Lawrence's cyclotron) on a preferred basis, partly because both knew and talked to Lawrence, but also because the graduate student (Niels Edlefsen) who helped Lawrence build his first cyclotron (which resulted in Lawrence's Nobel Prize) later became a soil physicist, and "soils [were] known at the time to be very progressive."

While Chair of both his Department at Berkeley, and the smaller one at Davis, Jenny traveled weekly to meet his Davis colleagues. There, he discussed his model of monovalent ion exchange with Professor Lannes Davis. Jenny encouraged Davis and Overstreet to investigate the problem further, and they began to develop a thermodynamic-based model of ion exchange. Jenny encouraged and supported this research, but did not participate, because he felt the thermodynamic approach did not reveal the colloid chemical phenomena he wished to illuminate, and Jenny's own skills in mathematics and statistical mechanics were insufficient for him effectively to push his own approach further.

Jenny and his colleagues, inspired by exchange and ion uptake, focused on the nature and meaning of pH measurements in colloidal suspensions and mineral surfaces. Also, to examine better the mechanisms by which roots might interact with soil particles, electron microscopy of the 'mucigel' surrounding the root, and its reactivity, was undertaken. These and other studies too numerous to mention in this brief summary continued until Jenny's retirement.

It is probably fair to conclude, at least from his reputation today, that Jenny's greatest scientific achievement is his classic book *Factors of Soil Formation. A System of Quantitative Pedology.* It is also probably his most misunderstood achievement, a fact reflected in that few people ever cite the full name of the publication (and the second sentence is an essential component of the book), and that many recent pedological texts continue to state incorrectly the premise of the book, and its famous equation:

$$s = \mathrm{f}(\mathrm{cl}, \mathrm{o}, \mathrm{r}, \mathrm{p}, \mathrm{t}, \ldots)$$

where $s =$ is soil properties and cl, o, r, p, t are regional climate, potential biota, topography, parent material, and time, respectively. As Jenny once remarked, "It looks easy, but it's not."

The origins of the book appear to have begun back in Jenny's Swiss education and research, but he notes that the key to rigorously expanding the ideas, and putting them in a book, were driven by his assignment to teach pedology at Berkeley. Unlike his colloid chemistry course, in which he was an expert and part of the scientific mainstream, he entered pedology as an outsider, with somewhat novel views. He strove to develop a logical and integrative means of explaining soil geographical distributions. As many (including Jenny) have noted, his factors included those

first discussed by Dokuchaev. But what most do not appreciate is the conceptual gulf between those earlier formulations and Jenny's – a fact that Jenny made clear in the very first sentence of his book: "As a science grows, its underlying concepts change, although the words remain the same." Jenny conceived of soils as a physical system and defined state factors as independent variables that define the properties of the system, bringing to pedology the formalisms of physical chemistry. The completed manuscript, submitted to McGraw-Hill, was at first rejected. But intervention by Robert Bradfield, Jenny's long-time associate from his Missouri days, succeeded in achieving its publication in 1941.

Today, *Factors* remains a vibrant and hotly debated work, because it is, from cover to cover, a presentation and illumination of a theory, in the same vein as other books of its type, such as Darwin's *Origin of Species*. As with Darwin, subsequent reaction was both negative and positive. Yet, as with *Origin of Species*, the ultimate benchmark for its success is gaged in its ability to explain and understand nature. The book *Factors* allows pedologists and other natural scientists to view the world as a natural outdoor experiment, providing them with the intellectual tools to unravel its history and its processes. Scores of papers and books have now been published based on the 'factorial' concepts, and this research, which now spans pedology, ecology, and geology, shows no sign of abatement. It is one of the crowning achievements of twentieth-century soil science.

Jenny retired in 1967 (**Figure 5**), though his creative activity might be viewed as having merely shifted, not abated. First, Jenny (with the help of his wife Jean) must be credited with the first rigorous analysis of the role of soils in landscape painting. He lectured widely on this for decades, and, for those fortunate enough to have attended these seminars, the emotional impact will never be forgotten. This lecture was recorded, in published form, as part of a Pontifical Academy of Sciences volume, conference proceedings which mainly focused on soil fertility, and ultimately its role in the sustainability of human populations. To summarize briefly Jenny's views on the esthetic value of soils, he wrote:

> Soil appeals to my senses. I like to dig in it and work it with my hands. I enjoy doing the soil-texture feel test with my fingers or kneading a clay soil, which is a short step from ceramics or sculpture. Soil has a pleasant smell. I like to sit on bare, sun-drenched ground and take in the fragrance of soil. As yet, neither touch nor smell sensations have been accorded aesthetic recognition, but colors delight painters, photographers, and writers, as well as you and me. In loess country, plowed fields on slopes show wide bands of attractive color gradations from dark browns to light yellows, caused by erosion of the surface soil. Warm brownish colors characterize fields and roofs in Cézanne's landscape paintings of southern France, and radiant red soils of the tropics dominate canvasses of Gauguin and Portinari. Soil profiles viewed in pits may reveal vivid color and structure patterns of layers or horizons. I have seen so many delicate shapes, forms, and colors in soil profiles that, to me, soils are beautiful.

For decades, Jenny had thought about and discussed additional factors of soil and ecosystem formation. Since his work in the tropics in the late 1940s, he had considered latitude to be a candidate as an additional soil-forming factor. To test the theory required the selection of sites (all with the same mean climates, parent materials, etc.), but at different latitudes. To accomplish this experimental design meant that sites at low latitudes must be found at high elevations in order to be comparable with higher-latitude sites. Mount Kilimanjaro, in Tanzania, offered an opportunity. So, in 1982, at the age of 83, Jenny and Dr. Jennifer Harden, of the USGS, traveled to Tanzania and hiked and sampled soils to an elevation of more than 13 000 f. Jenny measured the C and N content of the soils himself, upon their return. Though not published alone, the samples and data later formed a central part of an MS thesis by, fittingly,

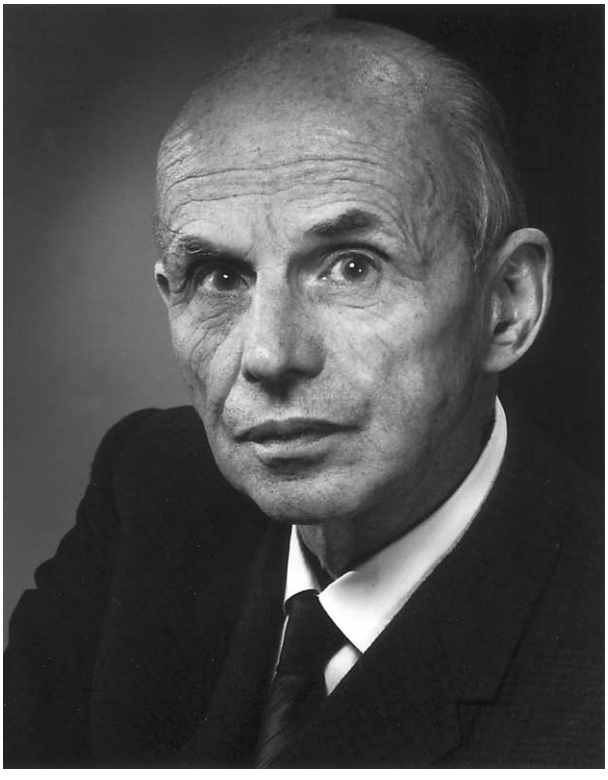


Figure 5 Hans Jenny, 1967, the year of his retirement.

a visiting student (Annelies Uebersax) from the ETH in Zurich.

For those who knew Jenny during the last decades of his life, it was apparent that his intellectual passion was fueled by a somewhat intertwined attachment to the nutrient-impoverished soils of the marine terraces of the Mendocino coast (the Ecological Staircase), and an effort to preserve them, and other rare soilscapes, for future generations. The Jenny family maintained a farm home, in nearby Comptche, which served as an overnight lodging and rustic 'think tank' for a generation of students and visitors of all types. The length of geologic time encompassed by the terraces and the revealing trends in soil chemistry enriched Jenny's views of the effect of time on soil formation, and the fate of soils on the Earth's surface. The rarity of these ecosystems, and the low regard in which they were held, led Jenny and his wife on a long journey into the workings of the politics of land preservation. Yet, despite the novel view of 'preserving soil,' they succeeded, and Jug Handle State Reserve, near Caspar, California, was created as a state-owned park devoted to leading the park visitor on a guided walk through geologic time. The visitor center prominently displays the trends of soils and ecology on the sequence of marine terraces. It is a living memorial to Jenny's insights and efforts on 'soil diversity,' many of which are being openly discussed and analyzed in the scientific literature only now. In an interview late in life, Jenny was asked:

> Does soil have a right to be protected for any reason other than that based on what is best for humans?

Jenny replied:

> Today, the idea of stewardship of land is pitted against the belief in soil exploitation for personal gain and that soil is merely an economic commodity in the marketplace ... I place soils and ecosystems, the nature museums, on a par with art museums ... colleges and temples ... Society grants human beings the right to exist, regardless of whether we are useful or not ... and the same privilege has been extended to a few endangered plant and animal species. I wish society would extend the same right to soil.

The remarkable life and genius of Hans Jenny have left an indelible stamp on soil science, forever changing the way in which soils are thought of and studied, and his ideas have diffused so widely that they are now part of the high-school curriculum. Jenny often remarked that he was honored and humbled to fill the chair of Eugene Hilgard, his predecessor at Berkeley. (*See* **Hilgard, Eugene Woldemar**). Those feelings are now appreciated by those who must follow Jenny.

See also: **Hilgard, Eugene Woldemar**

Further Reading

Amundson R and Jenny H (1997) On a state factor model of ecosystems. *Bioscience* 47: 536–543.

Amundson R, Harden J, and Singer MJ (1994) *Factors of Soil Formation. A Fiftieth Anniversary Retrospective*. SSSA Special Publication No. 33. Madison, WI: Soil Science Society of America.

Jenny H (1929) Relation of temperature to the amount of nitrogen in soils. *Soil Science* 27: 169–188.

Jenny H (1936) Simple kinetic theory of ionic exchange. I. Ions of equal valency. *Journal of Physical Chemistry* 40: 501–507.

Jenny H (1961) *E. W. Hilgard and the Birth of Modern Soil Science*. Pisa, Italy: Collana della Rivista "Agrochimia."

Jenny H (1968) The image of soil in landscape art, old and new. *Pontifical Academy of Sciences Scripta Varia* 32: 947–979.

Jenny H (1989) *Hans Jenny. Soil Scientist, Teacher, and Scholar.* Regional Oral History Office, The Bancroft Library, University of California–Berkeley, CA.

Jenny H (1994) *Factors of Soil Formation. A System of Quantitative Pedology.* New York: Dover Press. (Reprint, with Foreword by R. Amundson, of the 1941 McGraw-Hill publication.)

Jenny H and Stuart K (1984) My friend, the soil. *Journal of Soil and Water Conservation* 39: 158–161.

Jenny H, Overstreet R, and Ayers AD (1939) Contact depletion of bare roots as revealed by radioactive indicators. *Soil Science* 48: 9–24.

Jenny H, Nielsen TR, Coleman NT, and Williams DE (1950) Concerning the measurement of pH, ion activities, and membrane potentials in colloidal systems. *Science* 112: 164–167.

K

KELLOGG, CHARLES

J D Helms, USDA Natural Resources Conservation Service, Washington, DC, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Charles Edwin Kellogg led one of the most distinguished careers in the history of soil science, a career marked by a dedication to assist land users through his knowledge of soils. Kellogg attended college during the formative years of the young discipline termed soil science. As a student and young professor he made landmark contributions in interpretations of soil-survey information for practical uses (Figure 1). He made significant contributions in 'pedology,' but his advice on using soil surveys to assist people defined his career. To accomplish that objective he redirected and refined the US Department of Agriculture (USDA) soil survey program, which he supervised from 1934 to 1971. In his later career he advised, domestically and internationally, on soils and sustainable agricultural development. He became knowledgeable about tropical soils and used that knowledge to assist developing countries.

Figure 1 Charles Edwin Kellogg, courtesy of John Tandarich.

Education and Educator

Charles Edwin Kellogg (August 2, 1902–March 9, 1980), soil scientist, was born near Ionia, Ionia County, Michigan, the son of Herbert Francis Kellogg, a farmer, and Eunice Irene Stocken. Following graduation from Palo High School, Kellogg enrolled at Michigan State College where he pursued studies in science and mathematics, and eventually specialized in the emerging field of soil science. As a student, Kellogg participated in several pioneering events in the use of soil data for planning. During the summers between 1923 and 1926, he researched and mapped soils for the land economic survey, an interdisciplinary project of the Michigan Department of Conservation, which brought together information on soils, botany, and economics, among other subjects. One objective of the survey was to help solve the problem of inefficient use of land, which vexed failing farmers and taxed the local economy and the state. From this experience and others, Kellogg learned to view soil properties in the context of their meaning to and interaction with the local community. Kellogg recognized Lee Roy Schoenmann, director of the land economic survey, as "the person who first showed me the soil," in the dedication to his book *The Soils That Support Us*. Throughout his career, Kellogg insisted on studying soils as part of the broader natural and economic setting.

The next event stimulating Kellogg's interest in the relationship between soils and land uses came quickly. While on a Soils Fellowship at Michigan State College (April 1926 to April 1928) sponsored by the Michigan State Highway Department, Kellogg noted failures of new cement-concrete highways in the state and made a study of the relationship between soil characteristics and appropriate highway design. In 1929, the Michigan Academy of Science, Arts and Letters published Kellogg's article 'Soil type as a factor in highway construction in Michigan.' The

work led to definite procedures for highway construction in accordance with the soil conditions as assessed by detailed soil surveys. The Michigan field manual of highway construction incorporated the study's recommendations. The Highway Research Board of the Public Roads Administration also issued a bulletin *Engineering Use of Agricultural Soil Maps*.

From April 1928 until December 1929, Kellogg directed soil surveys and carried out research on soil genesis for the Wisconsin Geological and Natural History Survey. While taking additional graduate courses at the University of Wisconsin, Kellogg reworked the classification of Wisconsin soils and used the results as his dissertation. Michigan State College of Agriculture and Applied Science awarded the PhD degree in 1930 for his dissertation, 'Preliminary study of the profiles of the principal soil types of Wisconsin.' The study was in the field of 'pedology,' a newly applied disciplinary term referring to descriptions and studies of soil development.

Upon completing his doctoral work, Kellogg continued to innovate in the area of interpreting soils information for utilitarian purposes. Kellogg took an assistant professorship at the North Dakota Agricultural College in January 1930, with responsibilities for the soil survey program and land classification. While at North Dakota, Kellogg advised on the design of the soil survey of McKenzie County, North Dakota, so that it could be used in land classification for rural tax assessment. From his North Dakota experience he published articles that described methods for rural land classification (*Journal of Land and Public Utility Economics*), proposed magnesium as a possible key to phosphorus problems in semiarid soils (*Journal of the American Society of Agronomy*), and described the morphology and genesis of the Solonetz soils of western North Dakota (*Journal of Soil Science*). In 1934, 'The place of soil in the biological complex' appeared in the *Scientific Monthly*. The article belonged in a genre that was to mark much of Kellogg's writing. It explained a scientific subject to the nontechnical, but educated, audience interested in public policy issues. Although Kellogg would soon depart the university, he remained an educator. During his tenure at North Dakota he inspired a number of students to make soil science a career, including Roy W. Simonson, Marlin G. Cline, A. Clifford Orvedal, and Andrew 'Andy' Aandahl.

Soil Survey Division, US Department of Agriculture

In late 1933, A. G. McCall, Chief of Soil Investigations in the Bureau of Chemistry and Soils, corresponded with Kellogg about a survey of the lands above Norris Dam in the Tennessee Valley Authority (TVA), an arrangement that did not come to fruition. Subsequently, Kellogg accepted a position in land classification in the Bureau, despite North Dakota state's bid to retain him. Roy Simonson, a Kellogg student, recalled that Henry G. Knight, Chief of the Bureau of Chemistry and Soils, visited the university to hear one of Kellogg's seminars shortly before his hiring in February 1934. It is likely that the Bureau's leadership had already identified Kellogg to succeed Curtis Marbut.

On July 1, 1934, Kellogg became Acting Chief, Division of Soil Survey, and then Chief of the Division on July 1, 1935. Kellogg manifested his strong commitment to public service through the soil survey. He reoriented the soil survey to focus on interpretations of soils data for practical uses. When the survey began under Milton Whitney's leadership, it offered the promise of helping farmers in efficient land use, perhaps an ambitious objective given the state of the science at the time. Subsequent soil-survey operations under Curtis F. Marbut highlighted viewing soil as natural bodies worthy of scientific study. Marbut took pride in establishing the soil survey as a scientific document, and consciously eschewed agricultural interpretations, leaving that task to other agricultural specialists. Kellogg believed that the soil scientists themselves should work with other technical specialties to develop the interpretations of the soil-survey.

To reach the goal of creating a more useful soil survey document required greater attention to the survey itself. The early years of Kellogg's leadership focused on tools to improve and advance the soil survey. Kellogg believed the survey needed more precision in soil definition – a belief that led to the first edition of the *Soil Survey Manual*. The manual appeared in 1937, about 3 years after Kellogg assumed leadership, and was the first major revision of field instructions in 23 years. A little earlier in 1936, he had issued a USDA publication *Development and Significance of the Great Soil Groups of the United States*, which is significant for its discussion, in qualitative terms, of the five factors of soil formation. Kellogg also commenced work on the soil classification system. The Russian concepts of soil formation, which Marbut introduced to the American audience, informed the soil classification system that first appeared in the 1938 Yearbook of Agriculture, *Soils and Men*.

Kellogg as Scientist-Administrator

Kellogg fitted the model of the scientist-administrator in the federal employ. Directly and indirectly, government-sponsored scientific investigations advance

many facets of American life, through agricultural improvement, natural resources exploration and exploitation, military capabilities, and health and medicine. The government-sponsored soil survey program had far-ranging impacts on the interrelated fields of soil survey, soil classification, and soil science. Many of the methods and terminology developed in the soil survey effort formed the lexicon of soil survey and classification. The soil bodies as defined for the soil survey – the soil types and soil series – became the reference points in discussions of soils in the USA. Partly as a matter of convenience, the soil classification system developed by the soil survey, in cooperation with university-based soil scientists, became the national system. As a scientist-administrator, Kellogg was tireless in his efforts to shape the production aspects of the county soil surveys and the scientific and technical research needed to produce those surveys.

Soil Taxonomy

After World War II, Kellogg directed that a new soil classification system be developed. The staff, under Guy D. Smith's leadership, cooperated with soil scientists internationally in developing the new classification system, *Soil Taxonomy*. Smith had a keen appreciation for quantification and the scientific method and had studied with Hans Jenny, who earlier had tried to quantify the factors of soil formation. Kellogg had once remarked that there was no science without measurement. Following upon the theme of more precision in definition, division staff made greater use of laboratory analysis in definitions. Quantifiable boundaries and parameters defined the various categories in the system. *Soil Taxonomy* also sought to correct a problem that had plagued previous soil classifications: the lowest units, the soil type, did not fall logically and consistently into the upper categories. To remedy this, soil type and soil series, the lowest categories, were classified based on their measurable properties.

Kellogg hoped that the new classification system would aid transfer of empirical experience and research results internationally. If lessons in successful soil-management techniques on a particular soil could be transferred to similar soils elsewhere, it could be extremely beneficial to developing countries. For quite some time soil scientists had been keying interpretations and management recommendations to soil type. Bringing greater quantification to the designation of soil types and soil series enhanced the potential for interpretations. Despite his aspirations for *Soil Taxonomy*, Kellogg recognized that any attempt to organize knowledge about natural systems needed flexibility to adapt in the face of new discoveries. Thus he warned against letting a system become so classical that change and revision became impossible. *Soil Taxonomy* became the *lingua franca* of scientific discussion of soils in the USA and in other countries.

Soil-Geomorphology Studies

Kellogg launched a group of soil-geomorphology studies in various climatic and physiographic regions to expedite and improve soil surveys. The studies demonstrated the intricate processes of landscape and soil formation. Soil scientists could extrapolate some of the approaches and information about processes of landscape formation to other locations. Kellogg also believed that the studies would help soil scientists predict more accurately the long-range effects of soil- and water-management practices on soil stability. The studies conducted under the leadership of Robert V. Ruhe rank as seminal literature in the field of geomorphology. In addition to innovative research design and the ability to interpret the results, geomorphologic studies such as these required funding and institutional support beyond the resources of most individual earth scientists. Through Kellogg's commitment in concentrating on these projects, he and his staff profoundly influenced a field of earth science. While the studies influenced the operations of the soil survey at the time, some believe that their benefits for continuing management of the soil survey have not yet been fully exploited.

Training Soil Scientists

The last two decades of Kellogg's soil survey career marked a tremendous acceleration in the production of soil surveys. The Department of Agriculture merged the Soil Survey Division into the Soil Conservation Service (SCS) on November 19, 1952. Kellogg acquired more staff and resources to produce county soil surveys. That task presented a challenge. Kellogg's soil survey division had constituted a small corps of highly trained individuals who had contributed to the literature of soil science and established themselves in the discipline. The task before him was to create a much larger corps of field soil scientists to make soil surveys.

Kellogg prided himself on gaining governmental endorsement of the term 'soil scientist' as the official job classification, but struggled with how to ensure that this cadre of specialists had access to the most current knowledge. With the collaboration of former students and colleagues at universities, Kellogg created soil science institutes, or intensive short courses, at selected universities to give soil scientists additional training in the latest developments in the discipline. A former student, Marlin Cline, directed the first soil

science institute at Cornell University. Subsequently, faculty at Iowa State University and Oregon State University hosted the soil science institute. In more recent times North Carolina State University, University of Florida, Texas A & M University, Alabama A & M University, and the University of California-Davis also hosted the institute. Kellogg also instructed the SCS soil scientists through internal memoranda and reports, which demonstrated the same care and precise, lucid writing style that characterized his articles for scholarly journals.

International Influence

After World War II, Kellogg advised farmers, scientists, and administrators internationally, tailoring his advice to fit the local educational, economic, and industrial conditions. He was nondoctrinaire in his advice. While he recognized developing countries would want to increase food production with fertilizers, he did not reflexively advocate western technology. Shifting cultivation and other indigenous methods might still be needed until fertilizer plants and agricultural infrastructure were in place. He emphasized that social and institutional barriers, in addition to scientific and technical problems, must be addressed in the developing world. His preference was not for any particular, idealized, 'fixed' pattern of land use, but for the use of science. He believed that technology, science, national needs, and markets would bring change and the soil scientists should be ready to provide the needed information for efficiency and conservation. In the post World War II period, he saw the need to increase food production, but in the context of improving other facets of the human condition worldwide. He believed that scientists should work for a balance between resource use and population. Kellogg never left the land-users out of the equation in his discussions. As western governments tried to bring agricultural and scientific advancement to the developing world, he urged that those governments also work for the freedom and security of the individual cultivator.

In time, foreign-aid specialists would recognize the faults in the post World War II technical advice provided to developing countries. Very early Kellogg saw some of the difficulties confronting the effort and the mistakes being made. He wrote that agricultural improvement was an essential component of general economic progress, not a separate activity. Kellogg noted the faulty strategy of trying to improve availability of manufactured consumer goods without also supporting farmers and agriculture. He attributed some of this to the fact that most of the planners in the developing countries hailed from the urban centers. Each country or region would need to develop its own specialists and scholars. Imperialism based on scientific expertise would be as objectionable as other types of imperialism.

Tropical Soils

In the post World War II international demand for western agricultural specialists, scientists once again had access to the tropics. Kellogg traveled, studied, wrote, and spoke on world soil resources and made tropical soils a specialty (Figure 2). On his many field

Figure 2 Charles E. Kellogg interviews farmers at Oyeko, Gold Coast (now Ghana) about cocoa production. Left of Kellogg (center) is the 'Chief Farmer' of the Gold Coast, Kwane Poku; at extreme right is Saakodee who interpreted for Kellogg. Reproduced from *African Journal K-2225* August 9, 1954. Lincoln, NE: Soil Survey Laboratory, Natural Resources Conservation Service.

trips Kellogg took copious notes, which his wife Lucille later typed into journals illustrated with properly captioned photographs. The journals remained unpublished, since Kellogg recognized that more research and analysis was needed to verify his observations, although he did incorporate some observations into his writings. In some cases, he had an opportunity for extended visits, including time for research and analysis that did lead to publication. His visit to the Belgian Congo is examined in *An Exploratory Study of Soil Groups in the Belgian Congo* (1949).

Although Kellogg might discourse on complex scientific questions, he practically always turned the discussion to questions of immediate concern to the local land-users in the developing countries. His articles emphasized that much remained to be learned about tropical soils and their management, an aspect he felt was masked by a veil of misunderstanding about tropical soils. To make the point, he employed a metaphor from art. Michelangelo supposedly said of a block of pure white marble, "A beautiful figure lies within, I have only to find it." Likewise, the agricultural scientists had to discover the combination to productivity in tropical soils. Many tropical soils had low nutrient levels, but also had great potential. As he did with domestic audiences, Kellogg emphasized that agricultural productivity was the result of interactions. Inherent soil fertility was only one element. Kellogg often cited Florida as an example of potential productivity of inherently infertile tropical soils. But he also observed that putting all of the elements in place in developing countries would take time. In the interim, western scientists should not condemn existing systems outright, but work within them to raise productivity.

Kellogg became a leading spokesperson on global soil resources. He believed that a general world soils map could have great benefit, especially in the developing world. Such a map had the potential to save time, money, and effort by making it possible to relate experience and research results from one part of the world to another. He believed there were sufficient soils worldwide that could maintain food production, if used wisely. At the same time, he was in the vanguard of the prime farmland protection movement, and he noted that development often took some of the best farmland. Use of soil information, he advised, could help guide expansion on to soils not so well suited for agriculture. He also spoke out on soil conservation. Conservation was not merely saving and maintaining soil in place or preventing degradation, but rather was the sustained and efficient use of soils that in turn contributed to improved living standards. Conservation needed to be efficient and sustaining within the larger economy.

Interactions in Agricultural Productivity

In the postwar period, the USA continued its acceleration of agricultural productivity. Kellogg was an active speaker and writer, who wove science and agriculture into public-policy recommendations. He reiterated several themes. He emphasized that the increases in agricultural productivity resulted from the interaction of individual elements of productivity – machinery, hybrid varieties, fertilizer, and water control. The individual farmer could not create this environment, favorable to productivity, alone. Many pillars supported productivity: education, extension, science and technology, industry, and markets. Interactions with particular soil types also played a role. Kellogg told soil scientists at the Seventh International Congress of Soil Science that he believed the most important principle good farmers had learned was that of the interaction among soil characteristics and among practices. This concept had particular pertinence to the developing world, where there was a great temptation to promote one item as a panacea to boost agricultural production.

Science and Policy

Two somewhat countervailing emphases are seen in Kellogg's ideas about science and public policy. He liked rigorous science and specialization, but recognized the importance of bringing in knowledge and opinions outside his discipline. So, while he saw the need for studying small soil bodies in detail, he also emphasized that interpretation of soils data be made in an interdisciplinary fashion. For example, he wrote articles about the relationship of soil use to economics and about how economics influenced potential uses. He found the distinction often made between basic and applied research as specious. He thought the USA should spend more time on basic research directed toward an objective. Unimaginative basic research was as objectionable as unimaginative applied research. While he emphasized interdisciplinary work, he warned against relying on generalists, untrained in the sciences. Rather the nation needed scientists and specialists who stood ready to work with other disciplines in an interdisciplinary fashion. His advocacy of objective-driven research influenced many, but by the end of his career, Kellogg nevertheless detected a tendency among some in soil science to study soil without any seeming advance in utilitarian knowledge.

Kellogg's published record reveals him as free from dogma, and nonideological in his advice. One is tempted to see this as a manifestation of his deep understanding of the complex interactions of soils, society, and government. According to his son,

Robert Kellogg, he felt passionate about the public service role of the federal employee, but he saw that role as limited in land-use decisions. The soil survey gained new authorities in the 1960s to provide information to local governments and other entities for community planning and resource development. In the midst of this triumph he added the cautionary note. He supported the activist role of the New Deal and its capacity for giving advice and assistance in planning, providing that the decisions be made at the state and local level. The soil survey did not make recommendations. Decisions about alternative uses of tracts of land would properly be made by individuals within the regulations and restraints established by local governments under state authorities.

Kellogg educated rural America and the agricultural establishment, not just on soils but also on broad rural and agricultural policy issues. He advised planners to take account of the changing nature of rural society. As agricultural productivity grew in the USA, there was an accompanying decrease in numbers of farmers and shifts of agricultural land use. Kellogg felt strongly that people needed to distinguish between farming and agriculture. While jobs in farming had declined, other jobs in agriculture had grown. Rural land was only partly cropland or agricultural land. Rural communities had opportunities to use their lands for recreation, forestry, housing, and other activities that provided jobs. He recommended that the land-grant university system could be of assistance by establishing 'natural resources institutes,' so long as they involved local people. While Kellogg favored an active government in research and education, he often returned to the theme that local people had the responsibility to make their own land-use decisions.

See also: **Calcium and Magnesium in Soils**; **Classification Systems:** USA; **Jenny, Hans**; **Marbut, Curtis Fletcher**; **Nutrient Management**; **Shifting Cultivation**; **Tropical Soils:** Humid Tropical; **World Soil Map**

Further Reading

Effland ABW and Effland WR (1992) Soil-geomorphology studies in the US Soil Survey. *Agricultural History* 66: 189–212.

Gardner DR (1998) *The National Cooperative Soil Survey of the United States*. Historical notes no. 7. Washington, DC: Natural Resources Conservation Service, US Department of Agriculture.

Helms D, Effland ABW, and Durana P (eds) (2002) *Profiles in the History of the US Soil Survey.* Ames: Iowa State Press.

Kellogg CE (1930) *Preliminary Study of the Profiles of the Principal Soil Types of Wisconsin*. PhD dissertation. East Lansing, Michigan: Michigan State College of Agriculture and Applied Science, Soils Department.

Kellogg CE (1933) A method for the classification of rural lands for assessment in Western North Dakota. *Journal of Land and Public Utility Economics* 9: 10–15.

Kellogg CE (1934) Morphology and genesis of the Solonetz soils of Western North Dakota. *Soil Science* 38(6): 483–501.

Kellogg CE (1934) The place of the soil in the biological complex. *Scientific Monthly* 34: 46–51.

Kellogg CE (1937) *Soil Survey Manual*. Washington, DC: US Department of Agriculture Miscellaneous Publication Number 274.

Kellogg CE (1943) *The Soils That Support Us: An Introduction to the Study of Soils and Their Use by Men*. New York: Macmillan.

Kellogg CE (1948) Conflicting doctrines about soils. *Scientific Monthly* 66: 475–487.

Kellogg CE (1948) Modern soil science. *American Scientist* 36: 517–536.

Kellogg CE (1957) We seek; we learn. In: *Soils: The Yearbook of Agriculture 1957*, pp. 1–11. Washington, DC: Government Printing Office.

Kellogg CE (1963) Why a new system of soil classification? *Soil Science* 96(July): 1–5.

Kellogg CE (1968) Fit suburbia to its soils. In: *Soil, Water, and Suburbia*. Washington, DC: US Department of Agriculture, US Department of Housing and Urban Development.

Kellogg CE (1971) *Reading for Soil Scientists, Together With a Library, Revised*. Washington, DC: Soil Conservation Service, US Department of Agriculture.

Kellogg CE and Davol FD (1949) *An Exploratory Study of Soil Groups in the Belgian Congo*. Serie Scientifique No. 46. Publications de l'Institut National pour l'Etude Agronomique du Congo Belge.

Kellogg CE and Knapp DC (1966) *The College of Agriculture: Science in the Public Service*. New York: McGraw-Hill.

Kellogg CE and Orvedal AC (1969) Potentially arable soils of the world and critical measures for their use. *Advances in Agronomy* 21: 109–170.

Kellogg RL (2003) Remembering Charles Kellogg. *Soil Survey Horizons* 44: 86–92.

Simonson RW (1987) *Historical Aspects of Soil Survey and Soil Classification*. Madison, Wisconsin: Soil Science Society of America.

Simonson RW (1989) *Historical Highlights of Soil Survey and Soil Classification with Emphasis on the United States, 1899–1970*. Technical paper no. 18. Wageningen, The Netherlands: International Soil Reference and Information Centre.

Soils and Men: Yearbook of Agriculture 1938 (1938) Washington, DC: Government Printing Office.

KINETIC MODELS

P M Jardine, Oak Ridge National Laboratory, Oak Ridge, TN, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

The vast complexity of subsurface chemical, physical, and microbial processes in soils, lends itself nicely to a system that is, more often than not, in a state of nonequilibrium. The mere physical structure of most soils, with its vast array of pore regimes, coupled with the transient flux of water and solutes during storm infiltration, results in a system that is often limited by physical and hydraulic nonequilibrium. Likewise, the complex mineralogical makeup of soils, coupled with water flow and media structure, often creates a system that is kinetically limited with regard to interfacial geochemical and microbiological reactions. Experimental recognition of the importance and frequency of nonequilibrium reactions and processes in soils has encouraged the reformulation of tradition equilibrium-based fate and transport models to incorporate kinetically limited chemical, physical, and microbial processes. Such mathematical models vary in complexity; encompassing single or multiple coupled processes that are useful for predicting either microscopic- or macroscopic-scale solute fate and transport behavior. The text will focus on time-dependent interactions of solutes at the solid–liquid interface in open, flowing systems. Kinetic models for soil chemical phenomena in closed systems have been well documented elsewhere.

Single-Component, Multiprocess Models

Physical Nonequilibrium

Two-region approach The occurrence of physical nonequilibrium during solute transport in soils often depends on the degree of interaction between macroscopic transport properties (i.e., water flux and preferential flow, hydrodynamic dispersion) and microscopic physical traits (i.e., diffusional mass transfer, aggregate size, length scale between pore classes), with the extent of interaction controlled by the physical structure of the media and the interconnectedness of soil water. A common mathematical representation of the time-dependent interaction of water and solutes between various pore classes within soils involves partitioning soil water into mobile macro- and mesopores (flowing region) and immobile micropores (stagnant region). The two domains are linked by an interaction term that considers the diffusional mass exchange of water and solutes from one pore class to another. Both macroscopic and microscopic descriptions of intrapore solute transfer have been formulated. The latter description uses a pore-scale, mechanistic approach for describing rate-limited water and solute distributions, whereas the macroscopic description of solute mass transfer use a lumped–parameter, empirical approach.

The macroscopic approach mathematically describes the interaction of solutes between mobile and immobile pore domains using an empirical first-order expression, where knowledge of the system geometry or pore structure is not necessary. Because of this, the physical significance of the rate coefficient is lost; however, the time-dependency of the solute mass-transfer process is known. The governing transport equations that describe this scenario are expressed as:

$$\theta_m R_m \frac{\partial C_m}{\partial t} + \theta_{im} R_{im} \frac{\partial C_{im}}{\partial t} = \theta_m D_m \frac{\partial^2 C_m}{\partial x^2} - \theta_m v_m \frac{\partial C_m}{\partial x} \quad [1]$$

$$\theta_{im} R_{im} \frac{\partial C_{im}}{\partial t} = \alpha (C_m - C_{im}) \quad [2]$$

where subscripts 'm' and 'im' denote mobile and immobile domains, respectively; C denotes the solution concentration of the solute; D is the hydrodynamic dispersion coefficient; v is the average pore-water velocity; $R = R_m + R_{im}$ is the retardation factor; α is the first-order mass-transfer coefficient; $\theta = \theta_m + \theta_{im}$ is the volumetric water content; x is soil distance; and t is time. As written, eqn [1] assumes a linear isotherm of the form $s = k_d C$, where s is the adsorbed solute concentration and k_d is the equilibrium distribution coefficient. Therefore, $R = 1 + (\rho k/\theta)$ where ρ is the porous media bulk density. The partitioning of solutes into mobile and immobile domains is frequently distinguished with a parameter denoted as $F = s_m/s$ that represents the mass fraction of adsorbed solutes that are in direct contact with the mobile liquid phase. The transport model described by eqns [1] and [2] is commonly referred to as the two-region or mobile–immobile model.

The microscopic approach mathematically describes the interaction of solutes between mobile and immobile pore domains using a more rigorous, mechanistic approach that considers the pore geometry and implicit structure of the soil media. Using this approach, the interchange of solute between mobile

and immobile zones is often specified by Fick's second law of diffusion, where mass exchange into spherical aggregates and plate-like matrix blocks has been investigated. For uniformly sized spherical aggregates of radius a, eqn [1] remains valid for mobile-phase transport, while eqn [2] is replaced by the spherical diffusion equation:

$$R_{im}\frac{\partial C_a}{\partial t} = \frac{D_a}{r^2}\frac{\partial}{\partial r}\left(r^2\frac{\partial C_a}{\partial r}\right) \quad (0 < r < a) \qquad [3]$$

where C_a is the solute concentration in the aggregate; r is the radial coordinate; and D_a is the effective matrix diffusion coefficient. The mean immobile solute concentration C_{im} in eqns [1] and [2] is now the mean solute concentration of the intraaggregate liquid phase:

$$C_{im}(x,t) = \frac{3}{a^2}\int_0^a r^2 C_a(x,r,t)\mathrm{d}r \qquad [4]$$

Another class of two-region flow and transport models also exist that assume both pore regimes are mobile (i.e. no immobile zones) with one pore-class dominated by advective–dispersive flow and the other pore-class serving primarily as a source/sink for slowly migrating solutes. Interaction between the pore regimes is described by an empirical first-order mass transfer coefficient.

Multiregion approach Since soils are typically a vast continuum of pore regimes of varying size and shape, the two-region approach may oversimplify fate and transport processes in highly structured soils known to possess multiregion flow regimes. Experimental confirmation of multiregion flow and transport in soils has resulted in the development of numerically advanced multiporosity, multipermeability, flow and transport mathematical models. A three-region visual conceptualization of this approach is provided in **Figure 1**, where a representative elementary volume (REV) at any point in the soil consists of three regions (e.g., macro-, meso-, micropores), each with its own flow and transport parameters. Intraregion mass transfer is described by flow from a physical point to a neighboring one, and interregion mass transfer between the various pore regimes is controlled by both advective and diffusive processes. Advective mass transfer is driven by hydraulic gradients that are caused by differences in fluid velocity in different-sized pores, and diffusive processes are caused by concentration gradients between pore classes. Another process inherent in multiregion flow and transport in structured systems is the time-dependent nature of both the advective and diffusive mass transfer rates between the various pore domains.

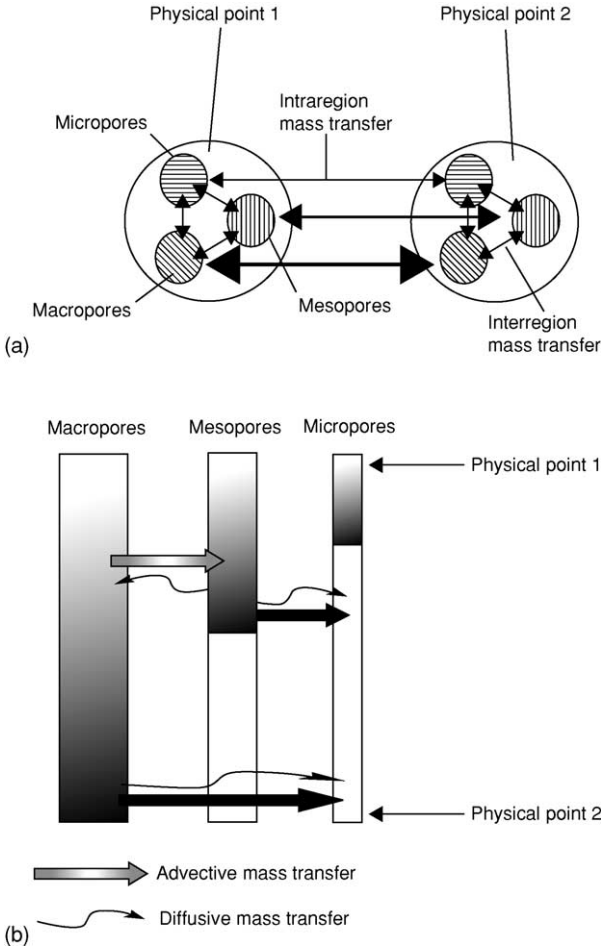


Figure 1 Triple-porosity, triple-permeability example of the multiregion flow and transport concept: (a) the representative elementary volume (REV) (large circle) at two physical points consists of three pore regions. Intraregion flow and transport is indicated by lines between large circles (REVs) and interregion transfer is depicted by lines within each large circle (REV); (b) both advective and diffusive mass transfer may occur in parallel or counter to each other. (Adapted from Gwo JP, Jardine PM, Wilson GV, and Yeh GT (1996) *Water Resources Research* 32: 561–570.)

Time-dependent mass transfer rates take into account changes in concentration gradients as solute mass is transferred between pore regimes. Time-dependent rate coefficients also indirectly account for variabilities in the size of soil aggregates and matrix blocks.

Chemical and Microbiological Nonequilibrium

The occurrence of chemical and microbiological nonequilibrium during solute transport in soils often depends on the mineral composition and physical makeup of the soil as well as the chemistry and interconnectedness of the pore water. Rate-limited geochemical and microbiological reactions such as adsorption, transformation, precipitation–dissolution,

and oxidation–reduction are commonplace in soil systems where transient storm infiltration events inevitably disrupt the equilibrium state of most any interfacial process.

Adsorption Nonequilibrium conditions during solute transport may be caused by kinetically limited adsorption reactions. Frequently, biphasic ion adsorption on soils is encountered where rapid reactions on readily accessible sites occur simultaneously with slow reactions on less-accessible sites. The energetics of adsorption may differ on these two types of sites depending on their location within the structure of the soil mineralogy. Certain locations exhibit structural constraints and increased charge densities (e.g., interlayer spaces) that influence the rate of chemical reaction during solute transport. Often, the more restrictive interlayer spaces of clay minerals exhibit larger solute adsorption energetics relative to easily accessible external mineral surfaces.

Disequilibrium due to chemical kinetic limitations on heterogeneous soil surfaces have been modeled using a two-site nonequilibrium approach, which assume solute adsorption on the two types of sites occur at different rates. Generally, empirical first-order and second-order expressions are utilized to describe the nonequilibrium adsorption process. The governing transport equation that describes solute adsorption on two different types of sites may be expressed as:

$$\frac{\partial C}{\partial t} + \frac{\rho}{\theta}\left(\frac{\partial s_1}{\partial t} + \frac{\partial s_2}{\partial t}\right) = D\frac{\partial^2 C}{\partial x^2} - v\frac{\partial C}{\partial x} \qquad [5]$$

where parameter definitions are analogous to those in eqn [1] and s_1 and s_2 are the adsorbed solute concentrations for type-1 and type-2 sites, respectively. The rate of solute interaction with the two different sites can be described by a variety of kinetic expressions listed below:

$$\text{Lumped first-order} \quad \frac{\partial s_i}{\partial t} = \alpha_i\left(s_i^* - s_i\right) \qquad [6]$$

$$\text{Reversible first-order} \quad \frac{\partial s_i}{\partial t} = \alpha_{bi}\, s_i^* - \alpha_{fi}\, s_i \qquad [7]$$

$$\text{Second-order} \quad \frac{\partial s_i}{\partial t} = \alpha_{fi}\psi_i\, C - \alpha_{bi}\, s_i \qquad [8]$$

$$\text{General-order} \quad \frac{\partial s_i}{\partial t} = \alpha_{fi}\, C^n - \alpha_{bi}\, s_i \qquad [9]$$

where s_i^* and s_i are the concentrations of adsorbed solute at equilibrium and any time t for type i sites, respectively; α_f and α_b are forward (adsorption) and backward (desorption) first-order rate coefficients for type-i sites, with $i = 1$ and 2 for type-1 and type-2 sites, respectively; ψ_i is the amount of vacant sites of type i; and n is the moles of reactant solute in solution, the 'reaction order.' Thus, both adsorption sites can be kinetically controlled or type-1 sites can be at equilibrium ($\alpha_1 \to \infty$) while type-2 sites remain kinetically controlled.

Transformation Nonequilibrium behavior during solute transport in soil may also result from time-dependent chemical and biological transformation reactions. Consideration of chemical fixation, dissolution, hydrolysis, and polymerization reactions in mathematical models are often necessary to correctly describe the transport of certain solutes in soil. Time-dependent biological reactions, which transform solutes into a variety of chemical species, may also need consideration during solute transport in some soils. Nitrification and denitrification processes that result from the application of N fertilizer sources to soils, are examples of complex subsurface processes that are influenced by time-dependent chemical and microbial reactions. Nitrogen fertilizers such as urea are rapidly hydrolyzed by the enzyme urease to form NH_4. The ammonium ion is a reactive cation in soils, which can become fixed within interlamellar spaces of certain 2:1 phyllosilicates, and is subject to microbial oxidation, resulting in the formation of NO_3. This oxidation reaction is known as nitrification and is typically governed by either zero-order or first-order kinetics. The microbial-mediated reduction of NO_3 (denitrification) also occurs in soils, resulting in time-dependent formation of N_2 and N_2O gases. Numerous transport models have been adapted to consider nonequilibrium nitrogen transformation reactions in soil during one- and three-dimensional flux of soil water.

Soil reactions involving potassium fixation and release can also be considered a time-dependent transformation process, since soil K^+ can be partitioned between solution, exchangeable, nonexchangeable (fixed), and primary mineral phases (Figure 2). Transformation of K from one phase to another is induced by transient soil conditions that promote K mineral weathering, leaching losses of soil solution K, and plant uptake of soil solution and exchangeable K. Exchangeable K is largely bound on readily accessible soil sites which may or may not be in local equilibrium with the solution phase. Many studies have shown that the adsorption and release of exchangeable K may be kinetically limited in the presence of 2:1 nonexpanding clay minerals. Nonexchangeable and mineral K retention and release reactions in soil are typically very time-dependent

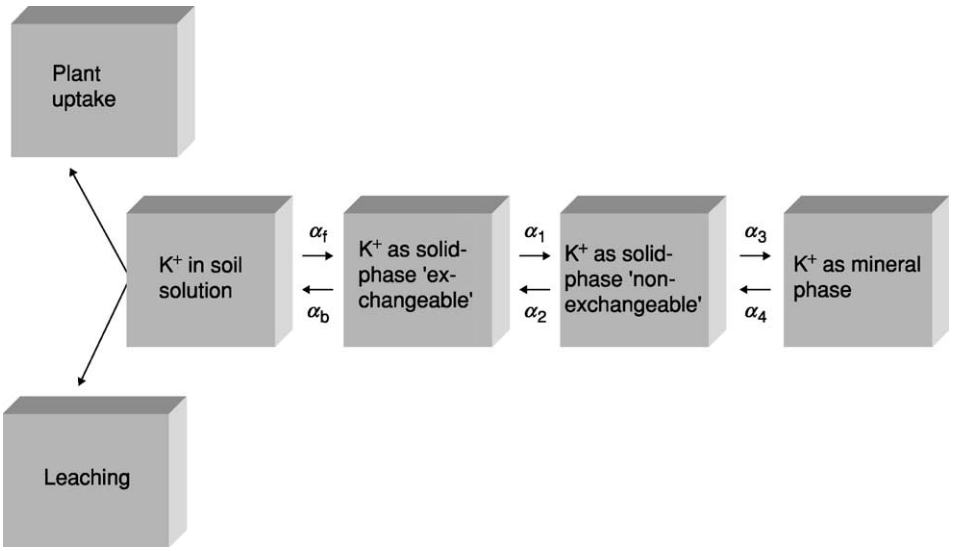


Figure 2 Schematic representation of the time-dependent reactions of potassium in soil. (Adapted from Selim HM, Mansell RS, and Zelazny LW (1976) Modeling reactions and transport of potassium in soils. *Soil Science* 122: 77–84.)

relative to transformations between solution and exchangeable K phases. Modeling K transport in soils, therefore, requires consideration of the nonequilibrium transformation reactions which occur between the various K phases.

A mathematical formulation of the conceptual model described in Figure 2 regarding the kinetics of K transformation during transport is described below. Transformations in the exchangeable phase (denoted by subscript 'ex') are represented by:

$$\frac{\partial s_{ex}}{\partial t} = \left(\frac{\theta}{\rho}\right)\alpha_f C^n + \alpha_2 s_{nex} - (\alpha_1 + \alpha_b)s_{ex} \qquad [10]$$

where parameter definitions are synonymous with eqns [1] and [5], α_f and α_b are the forward (adsorption) and backward (desorption) rate coefficients for reactions between the solution and exchangeable phases, α_1 and α_2 are the nonexchangeable fixation and release rate coefficients, respectively, and n is the order of the reaction for adsorption of K from solution to the exchangeable phase. Transformations in the nonexchangeable phase (denoted by subscript 'nex') is defined by:

$$\frac{\partial s_{nex}}{\partial t} = \alpha_1 s_{ex} + \alpha_4 s_{min} - (\alpha_2 + \alpha_3)s_{nex} \qquad [11]$$

where α_3 and α_4 are the immobilization and the mineralization rate coefficients describing time-dependent K reactions between nonexchangeable and mineral phases, and s_{min} is the amount of K in the mineral phase. Finally, transformations in the mineral phase are represented by:

$$\frac{\partial s_{min}}{\partial t} = \alpha_3 s_{nex} - \alpha_4 s_{min} \qquad [12]$$

Based on eqns [10–12], it can be assumed that the reactions between exchangeable and nonexchangeable as well as those between nonexchangeable and primary minerals are first-order kinetic reactions. Coupling eqns [10–12] with the one-dimensional convective–dispersive equation (one-site version of eqn [5]), one can describe the transformation kinetics of K during transport in soil.

The polymerization of solutes in soil, such as Al, may also be considered a time-dependent transformation process. The significance of Al transport through subsurface soil environments is of great concern to agriculturalist and environmentalist. Soil-solution Al is known to adversely effect terrestrial and aquatic ecosystems by severely limiting plant and fish production. It has also been linked to several human health disorders such as Alzheimer's disease and senile dementia. Therefore, the biogeochemical and hydrologic processes controlling the transport and cycling of Al are of considerable economic and scientific interest. The polymerization of Al tends to alleviate the adverse toxicity effects associated with Al^{3+}. Soil surfaces have been shown to serve as template for enhanced Al polymerization and these reactions have been modeled as multisite, first-order limited adsorption processes.

Solute transport in soils may be complicated by precipitation reactions which act as kinetic sinks for chemicals. Because of its agronomic and environmental significance, phosphorus (P) precipitation reactions during transport in soil have been most frequently

studied. Often P is a severely limited nutrient in terrestrial and aquatic ecosystems which adversely affects plant and fish production. Because of the significance of P precipitation reactions in soil, the biogeochemical and hydrologic processes controlling P transport must be clearly understood. Many of the principles associated with P reactions may in some instances be applicable to other types of precipitation reactions (i.e., Al, Fe, etc). Soil reactions involving P are generally modeled as first-order irreversible precipitation reactions described by:

$$Q = \alpha_c(\theta C) \qquad [13]$$

and first-order irreversible chemical immobilization reactions described by:

$$Q = \alpha_s(\rho s) \qquad [14]$$

where α_c and α_s are the rate coefficients for precipitation and chemical immobilization, respectively, and Q is the rate of solute consumption (sink). These kinetic sinks for P are often coupled to convective–dispersive transport with consideration of time-dependent, nonlinear adsorption to the soil surface. Certain modeling endeavors also consider the rate of P nucleation and crystal growth based on solubility product theory.

Biotransformation reactions in soils are often kinetically limited due to time-dependent microbial growth rates. Biologically mediated detoxification, mineralization, and electron-transfer reactions involving organic or inorganic solutes are often time-dependent processes that are mathematically described by either empirical zero- and first-order reactions, or Monod-type kinetics that account for microbial growth rates. Simple Monod kinetics consider the loss of substrate concentration, s (e.g., organic or inorganic pollutants) as a function of microbial growth rate, μ, where:

$$\mu = \frac{\mu_{max}[s]}{k_s + [s]} \qquad [15]$$

with μ_{max} is maximum specific growth rate, and k_s is substrate 0.5 saturation constant. When multiple substrates are available as an energy source for microbial growth, double Monod kinetics have been utilized, where a second substrate, the electron acceptor O_2, is used as here:

$$\mu = \left[\frac{\mu_{max}[s]}{k_s + [s]}\right]\left[\frac{\mu_{max,O_2}[O_2]}{k_{O_2} + [O_2]}\right] \qquad [16]$$

Monod kinetics have also been incorporated into rate expressions for substrate loss:

$$\frac{-d[s]}{dt} = \mu\frac{X}{Y} = \mu_{max}\frac{[s]}{k_s + [s]}\frac{X}{Y} = k_b\frac{[s]X}{k_s + [s]} \qquad [17]$$

where Y is yield (mass/mass), X is biomass per volume, and k_b is the biodegradation rate constant (t^{-1}). Any one of these expressions have been coupled with flow and transport equations in an effort to model time-dependent biotransformation processes in flowing, dynamic systems indicative of soil profiles.

Oxidation–Reduction Redox reactions at the solid–solution interface typically alter the geochemical behavior of many environmentally significant metals and organics. Soil solid-phase constituents often accelerate redox related charge transfer reactions and thus catalyze the oxidation or reduction of numerous redox-sensitive solutes. For example, common soil constituents such as Mn(III/IV) oxides and Fe(III) oxides are powerful oxidants that catalyze the oxidation of metals such as Co(II), Cr(III), and As(III). Other soil constituents such as organic matter and Fe(II)-bearing minerals, are known to be strong reductants with regard to metals such as Co(III) and Cr(VI). Models have been developed that couple transport and oxidation–reduction, and these models have been tested on systems involving redox-sensitive contaminants such as Cr and Co in the presence of mineral oxidants and reductants. As an example, the complex oxidation reactions of Co(II)-ethylenediaminetetraacetic acid (EDTA) by Mn oxides during transport have been modeled as shown in Figures 3 and 4. In this study, macroscopic transport experiments, coupled with the real-time spectroscopic interfacial interrogation methods, reveal that Mn(IV) oxides catalyze the time-dependent oxidation of Co(II)EDTA^{2-} to Co(III)EDTA^{-} (Figures 3 and 4). The redox reaction produces an intermediate Mn(III)-oxide phase, α-Mn_2O_3, on the surface of the Mn(IV) mineral that significantly passivated the redox reaction. Time-dependent reoxidation of Mn(III) to Mn(IV) is controlled by dissolved oxygen (DO). Successful modeling of these reactions involves coupling the advective–dispersive transport of Co(II)EDTA^{2-}, Co(III)EDTA^{-}, and DO with multiple, nonlinear kinetic expressions that describe: (1) oxidation of Co(II)EDTA to Co(III)EDTA by β-MnO_2, (2) inhibition of Co(II)EDTA oxidation due to precipitation of Mn_2O_3, an insoluble reaction product that occludes the β-MnO_2 surface, and (3) regeneration of the redox-reactive β-MnO_2 surface through oxidation of the Mn_2O_3 precipitate by dissolved oxygen (Figure 4). Comparison of experimental and calculated results demonstrates that the model describes the response of the coupled hydrologic and geochemical process to change in

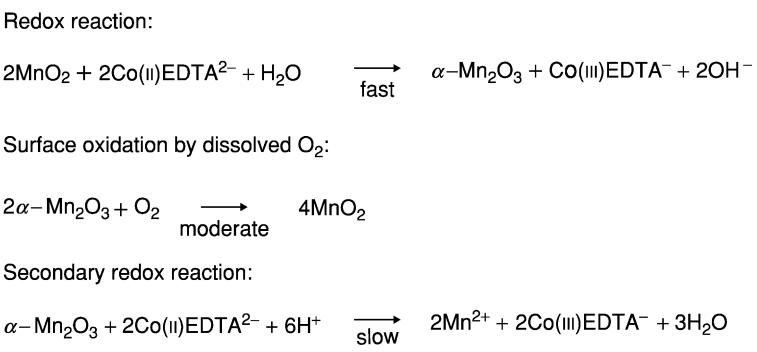


Figure 3 Spectroscopically confirmed reaction sequence describing the time-dependent, multireaction oxidation of Co(II)EDTA^{2-} to Co(III)EDTA^{-} in the presence of Mn oxides. (Adapted from Fendorf SE, Jardine RR, Patterson RR, Taylor DL, and Brooks SC (1999) Pyrolusite surface transformations measured in real-time during the reactive transport of Co(II)EDTA. *Geochimica et Cosmochimica Acta* 63: 3049–3057.)

flow rate, influent concentrations of Co(II)EDTA, and β-MnO_2 surface coverages.

Multicomponent, Multiprocess Models

In a system of multiple components, the identity of a component becomes chemically distinct from that of a species. Components are a set of linearly independent 'basic' chemical entities such that every species can be uniquely represented as a combination of those components. A species is then defined as the product of a chemical reaction, involving the components as reactants. Thus, numerical models that consider multiple components and multiple processes are computationally intensive, especially when aqueous- and solid-phase reaction kinetics are considered. The complexity of such models is often driven by the complexities observed in nature where enhanced interactions between experimentalist, mathematicians, modelers, and decision makers ultimately dictate the necessary processes that are needed to represent the desired system.

Nonequilibrium Processes Involving Dissolved Aqueous Phase Constituents

In recent years, a limited number of multicomponent reactive transport models have become available that couple hydrodynamic transport with multiprocess, time-dependent geochemical and microbial reactions. The equilibrium-based, multicomponent model HYDROGEOCHEM has been modified (HYDROGEOCHEM 2.0) to simulate mixed chemical kinetic and equilibrium reactions in an effort to deal with multispecies fate and transport that is controlled by time-dependent or thermodynamic-based reactions. Further modifications have incorporated time-dependent microbial reactions (HYDROBIOGEOCHEM), designed to solve a system of equations describing hydrologic transport and biogeochemical reactions in a reactive, multicomponent unsaturated or saturated system. The major transport processes are advection, dispersion–diffusion, and source–sinks. The major chemical processes are aqueous complexation, adsorption, ion-exchange, precipitation–dissolution, redox, and acid–base reactions. The major microbiological processes are biodegradation and microbial respiration. Rate expressions for kinetically controlled geochemical reactions are based on collision theory and can be represented generally by a stoichiometric equation of the following form:

$$\sum_{j\in M_x} \nu_j c_j \leftrightarrow \sum_{j\in M_x} \mu_j s_j \qquad [18]$$

where c_j and s_j are the reactants and products, respectively; ν_j and μ_j are the reactant and product stoichiometric coefficients, respectively; and M_x is the set of chemical species of interest. A generic rate equation corresponding to eqn [18] may then be written as:

$$\Omega = \alpha_f \prod_{j\in M_x} \left(\gamma_j c_j\right)^{\nu j} - \alpha_b \prod_{j\in M_x} \left(\gamma_j s_j\right)^{\mu j} \qquad [19]$$

where Ω, α_f, and α_b are the reaction rate, forward rate constant, and backward rate constant, respectively, and γ_j is the activity coefficient of species j according to the Davies formula. Microbial growth rates are represented by a modified Monod kinetic equation as follows:

$$\Omega = \Gamma I_g \left[\frac{S}{(k_s I_s + S)}\right]\left[\frac{A}{k_a + A}\right]\left[\frac{N}{k_n + N}\right] L \qquad [20]$$

where Γ is the microbial growth rate constant; k_s, k_a, and k_n are the half-saturation constants of substrate, electron acceptor, and nutrient, respectively; I_g and I_s are the inhibition factors for the growth rate and the half-saturation constant of the substrate, respectively; L is the metabolic lag coefficient; and S, A, and N are

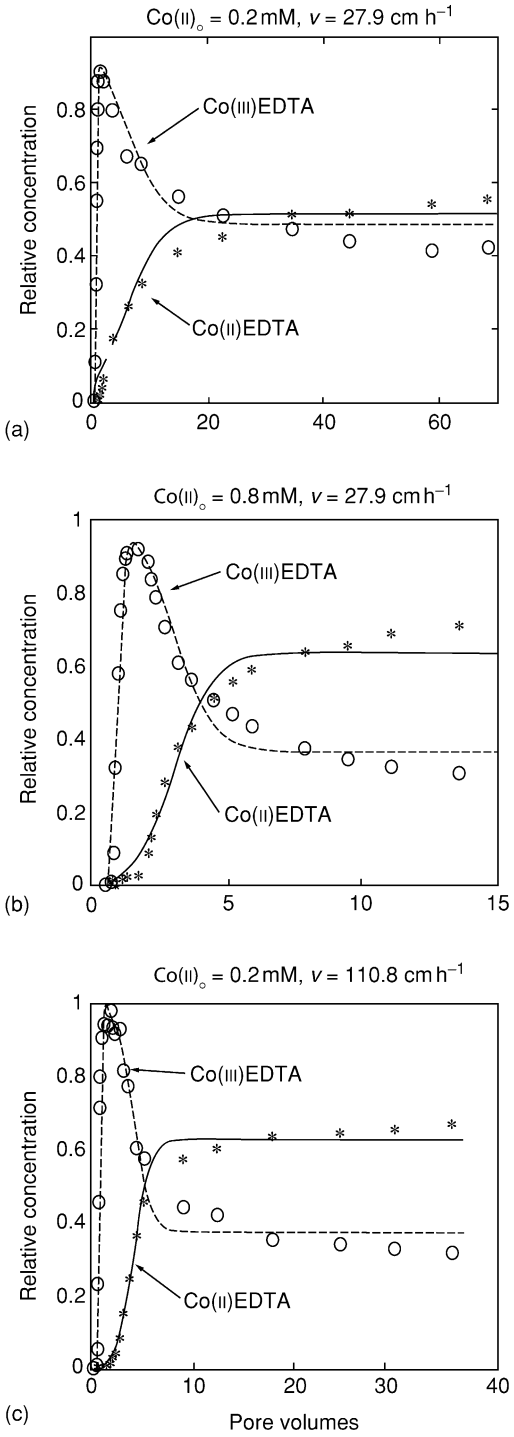


Figure 4 Observed and modeled fitted breakthrough curves for the displacement and subsequent oxidation of Co(II)EDTA^{2-} in columns of Mn(IV) oxide-coated sand for different influent concentrations and residence times. The model considered multiple nonlinear kinetic expressions to account for the complex series of oxidation and reduction reactions. (Reproduced from Saiers JE, Guha H, Jardine PM, and Brooks SC (2000) Development and evaluation of a mathematical model for the transport and oxidation–reduction of CoEDTA. *Water Resources Research* 36: 3151–3165.)

the concentrations of substrate, electron acceptor, and nutrient, respectively.

A high-performance computing version of HYDROBIOGEOCHEM has also been released (HBGC123D) which is a three-dimensional, shared-memory, parallel computer model that couples hydrogeologic and biogeochemical processes. The parallelized code offers the advantage of simulating large-scale, more realistic problems with improved performance and speed (Figure 5). Other multicomponent reactive transport models such as CRUNCH and FERACT are also available which couple multiple sets of kinetically controlled geochemical and biological processes with advective–dispersive flow. Still other kinetic-based flow models are available that target specific problems related to time-dependent, microbially driven redox reactions and complex, time-dependent metalorgano geochemical reactions in heterogeneous subsurface environments.

Nonequilibrium Processes Involving Dissolved Nonaqueous-Phase Constituents

Numerous multicomponent reactive transport models involving nonaqueous-phase constituents have also been developed that couple hydrodynamic transport with multiprocess, time-dependent geochemical and microbial reactions. The multiphase flow and multicomponent reactive transport simulator PARSim has been linked to a mixed chemical kinetic and equilibrium model (KEMOD) to allow simulation of multiple flowing phases with a full complement of reactions. The model (RPARSim/KEMOD) allows for a more general, nonequilibrium phase-transfer for KEMOD-style reactions where the reactants and products are in different phases. The model has been parallelized in order to enhance computational efficiency and the need to simulate larger-scale, more-realistic environmental problems. Other kinetic-based models designed to deal with subsurface dense, nonaqueous-phase liquid (DNAPL) issues, tend to emphasize substrate-limited biodegradation. The US Environmental Protection Agency (EPA) code BIOPLUME III is a two-dimensional contaminant transport model that couples DNAPL advective–dispersive transport with sorption, first-order decay, and biodegradation through instantaneous, zero-order, first-order, or Monod kinetics. The model is based on the US Geological Survey (USGS) BIOMOC code, where the hydrocarbon source and each active electron acceptor (O_2, NO_3^-, Fe(III), SO_4^{2-}, and CH_4) are simulated as separate plumes.

Another code, 'MT3D99,' which is an enhancement of MODFLOW, also couples advective–dispersive transport in soil systems with nonequilibrium sorption, time-dependent, nonaqueous-phase liquid dissolution,

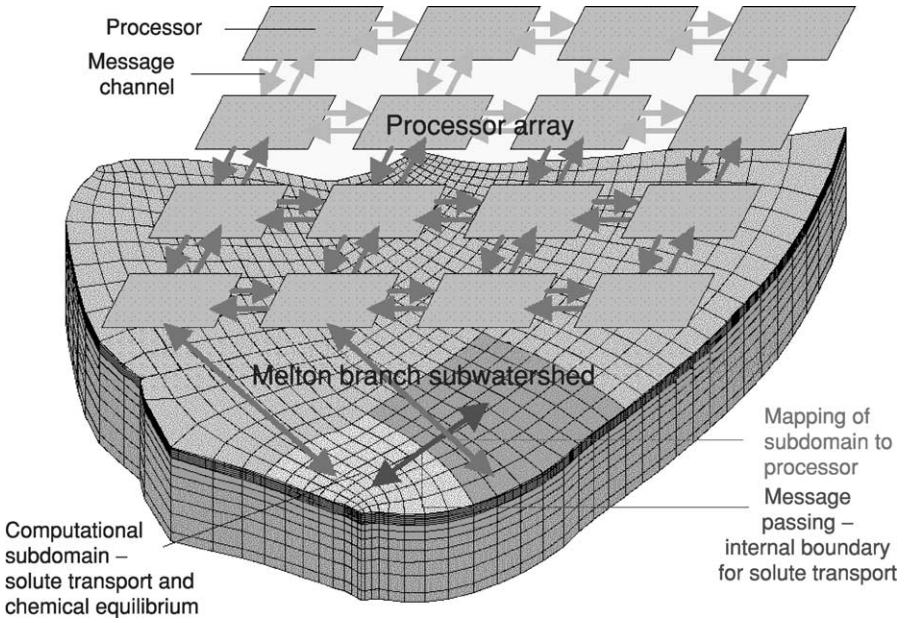


Figure 5 How solute fate and transport processes are modeled at the watershed scale using experimental subsurface transport facilities at the Oak Ridge National Laboratory. Computational subdomains are mapped to processors on a massively parallel computer. Since watershed-scale models are built with imperfect information of the subsurface and uncertainties associated with model parameters and processes, high-performance computing is necessary for quantifying parameter uncertainties when using mechanistically rigorous transport models at the field scale. Parallel-processing supercomputers provide the capability of solving large, realistic problems fast and economically. (Adapted from Gwo JP, Wilson GV, Jardine PM, and D'Azevedo EF (1999) Modeling subsurface contaminant reactions and transport at the watershed scale. Proceedings of the 1997 Chapman/Outreach Conference 'Assessment of Non-Point Source Pollution in the Vadose Zone.' Geophysical Monograph 108, American Geophysical Union, pp. 31–43.)

and rate-limited microbial processes. It considers Bioplume-type reactions, monad reactions, and daughter products, thus enabling the simulation of multispecies transport in a similar manner to the transport code RT3D. Kinetically limited hydrocarbon biodegradation using multiple electron acceptors and time-dependent transport of bacteria, electron acceptors, and hydrocarbons are explicitly considered. MT3D99 also maintains a dual-porosity option, where the soil media is divided into an advective-dominated mobile domain and a diffusion-dominated immobile domain. An empirical, first-order parameter accounts for mass transfer between the domains.

Although some multicomponent, multiporosity models do exist, typically they are limited by a small number of finite components. Full-blown, computationally rigorous, multicomponent, multiporosity, multipermeability models that consider coupled, time-dependent physical, chemical, and microbial reactions, have not yet been formulated or numerically implemented as of 2003. The utility of such models for simulating kinetically limited subsurface processes is conceptually appealing; however, parameter uncertainty resulting from limited experimental data at appropriate scales remains an obstacle for their practical use.

See also: **Cation Exchange**; **Chemical Equilibria**; **Surface Complexation Modeling**

Further Reading

Arbogast T, Bryant S, Dawson C *et al.* (1996) Computational methods for multiphase flow and transport problems arising in subsurface contaminant remediation. *Journal of Computational and Applied Mathematics* 74: 19–32.

Gwo JP, Jardine PM, Wilson GV, and Yeh GT (1995) A multiple-pore-region concept to modeling mass transfer in subsurface media. *Journal of Hydrology* 164: 217–237.

Gwo JP, D'Azevedo EF, Frenzel H *et al.* (2001) HBGC123D: A high performance computer model of coupled hydrological and biogeochemical processes. *Computers and Geosciences* 27: 1231–1242.

Hunter KS, Wang Y, and Van Cappellen P (1998) Kinetic modeling of microbially driven redox chemistry of subsurface environments: coupling transport, microbial metabolism and geochemistry. *Journal of Hydrology* 209: 53–80.

Hutson JL and Wagenet RJ (1995) A multiregion model describing water flow and solute transport in heterogeneous soils. *Soil Science Society of America Journal* 59: 743–751.

Jardine PM, Parker JC, and Zelazny LW (1985) Kinetics and mechanisms of aluminum adsorption on kaolinite using a two-site nonequilibrium transport model. *Soil Science Society of America Journal* 49: 867–873.

Saiers JE, Guha H, Jardine PM, and Brooks SC (2000) Development and evaluation of a mathematical model for the transport and oxidation-reduction of CoEDTA. *Water Resources Research* 36: 3151–3165.

Selim HM, Mansell RS, and Zelazny LW (1976) Modeling reactions and transport of potassium in soils. *Soil Science* 122: 77–84.

Steefel CI and Lichtner PC (1998) Multicomponent reactive transport in discrete fractures. I. Controls on reaction front geometry. *Journal of Hydrology* 209: 186–199.

Sudicky EA and Frind EO (1982) Contaminant transport in fractured porous media: analytical solutions for a system of parallel fractures. *Water Resources Research* 18: 1634–1642.

Szecsody JM, Zachara JM, Chilakapati A, Jardine PM, and Ferrency AS (1998) Importance of flow and particle-scale heterogeneity on Co(II/III)EDTA reactive transport. *Journal of Hydrology* 209: 112–136.

Tebes-Stevens C, Valocchi AJ, Van Briesen JM, and Rittmann BE (1998) Multicomponent transport with coupled geochemical and microbiological reactions: model description and example simulations. *Journal of Hydrology* 209: 8–26.

Wagenet RJ, Biggar JW, and Nielsen DR (1977) Tracing the transformations of urea fertilizer during leaching. *Soil Science Society of America Journal* 41: 896–902.

Yeh GT and Tripathi VS (1991) A model for simulating transport of reactive multispecies components – model development and demonstration. *Water Resources Research* 27: 3075–3094.

Yeh GT, Salvage KM, Gwo JP, Zachara JM, and Szecsody JE (1998) HYDROBIOGEOCHEM: a coupled model of HYDROlogic transport and mixed BIOGEOCHEMical kinetic/equilibrium reactions in saturated-unsaturated media. ORNL/TM-13668. Oak Ridge TN: Oak Ridge National Laboratory.

KIRKHAM, DON

D R Nielsen, University of California, Davis, CA, USA
R R van der Ploeg, University of Hannover, Germany

© 2005, Elsevier Ltd. All Rights Reserved.

Don Kirkham, American soil physicist and educator who made pioneering and fundamental contributions in solving soil water flow problems relevant to agriculture, was born in Provo, Utah, on February 11, 1908. He was the third child of six living children of Francis Washington and Alzina Robison Kirkham. His father, an educator from a Mormon family, had been an early missionary in New Zealand and wrote the first Maori grammar. The family moved to Salt Lake City in 1911, where the father served as president of a business college. In Salt Lake City, Don Kirkham attended Emerson School, and although absent a great deal of the time due to various children's diseases, he did well. His teachers recognized his ability and he skipped sixth grade. In the summers of 1919, 1920, 1921, 1922, and 1925, Kirkham worked at the family's West Jordan farm, incidentally acquiring experience to be useful in his later professional career in soil physics.

In December 1922, his family moved to Berkeley, California, for a stay of 1.5 years while his father worked for his PhD degree. Don attended Edison Junior High School (where he was elected vice-president of his class) and Berkeley High School. He graduated from East High School in Salt Lake City in June 1925, completing the scientific course. In the autumn of 1925, while attending the University of Utah, he developed Perthes disease (bone disease). His hip and left leg were kept in a cast from January 1926 to August 1926. He had to drop out of the University of Utah because it was not equipped for people on crutches. However, Kirkham had always had a deep interest in music. For Christmas when he was only 11 years old, he was given a clarinet and started music lessons with Lt Clarence J. Hawkins, just returned from World War I as director of the 145th Field Artillery band. Kirkham graduated from

Figure 1 Don Kirkham *c.* 1985.

the McCune School of Music and Art in Salt Lake City. When he graduated from the McCune School as teacher and soloist of clarinet, he played his graduation recital while in the cast. In the theory of music at this institution, he was high man in the final examination. In the summer of 1926, he joined the Musicians Union, the American Federation of Musicians, and earned considerable money playing and arranging band music and even composed a band number, the 'Ute Patrol.'

In 1927, he was called on a mission to Germany by Heber J. Grant, president of the Mormon Church. In Germany, he served as president of the Hamburg district, the largest district in point of membership in the Swiss-German mission. He had the reputation of knowing German grammar better than any others in the mission. At the end of his mission in 1930, his mother sent him the sum of $600, which had been left to her by an uncle. Kirkham used the money to travel extensively and frugally throughout Europe.

After his travels, he arrived in New York City in the autumn of 1930, where his family was then living, his father being the educational director of the National Child Welfare Association. He entered Columbia College that autumn. Three years later he graduated with honors in Physics. In the meantime he had won a Columbia College Scholarship and held scholarships in the Columbia Concert Band. In 1934, he received the Master's degree and in 1938 the PhD degree. While working on the doctor's degree, he was Assistant in the Ernest Kempton Adams (Physics) Precision Laboratory at Columbia University.

In September 1937, he arrived at Logan, Utah, where he had accepted a position to teach in the Physics and Mathematics Departments of the State Agricultural College. Three years later, as an Assistant Professor, he resigned to accept a position with the Navy Department, Bureau of Ordnance. The Navy was in need of physicists to protect ships against magnetic mines then employed by the Germans against British ships. He aided, first at Yorktown, Virginia, and then in New York City Harbor, in the development of a process for removing the permanent magnetism from ships called deperming, and in other techniques of degaussing, which was the general name employed for processes reducing ships' magnetism. He was in charge of the design, construction, and operation of degaussing facilities at New York Harbor. In January 1943 he was transferred to the Naval Ordnance Laboratory at Washington, DC, where he worked in torpedo research until September 1946.

From May 1946 to the middle of August of the same year, he was a civilian member of the scientific staff of Operations Crossroads, the Joint Army–Navy Task Force One. The mission of this Task Force was to test the atom bomb at Bikini Atoll in the Marshall Islands. Two bombs were exploded, one an air blast, the other an underwater blast. He witnessed both blasts from the USS Kenneth Whiting, a few miles outside the atoll. He went to Bikini by boat, being Senior Physicist of a group of Naval Ordnance Laboratory men, and was in the boarding party of scientists after the second and most powerful shot.

After his return to the USA from Bikini, he joined the staff at Iowa State College (now Iowa State University), one of the nation's leading agricultural schools, as Research Associate Professor in Soil Physics. His interest in soil physics had originated at Utah State Agricultural College, where he had been encouraged by Dr Willard Gardner, the eminent soil physicist there.

In 1938, Kirkham met and later married Mary Elizabeth ('Betty') Erwin Kirkham, who had taught English at Utah State. They had three children: Victoria, a professor in the Department of Romance Languages at the University of Pennsylvania; Mary Beth, a professor in the Department of Agronomy at Kansas State University; and Don Collier, a corporate lawyer and partner in the law firm of Cravath, Swaine & Moore in New York City.

Kirkham won many honors. He held the title of Charles F. Curtiss Professor of Agriculture at Iowa State University. He was the only professor to hold joint professorships in both the Department of Physics and the Department of Agronomy. He was a Fellow of the American Society of Agronomy, the American Physical Society, the Soil Science Society of America, and the American Geophysical Union. He was named Honorary Member of the Soil Science Society of America, the International Soil Science Society, the International Soil Tillage Research Organization, and the American Water Resources Association. He was a Fulbright-Hays Program grantee in the Netherlands, 1950–1951, and in Belgium, 1957–1958.

He received a Guggenheim Award to study in Belgium, 1957–1958. He was awarded an Honorary Doctor of Agriculture Degree from the Royal Agricultural University, Ghent, Belgium, in 1963 and an Honorary Doctor's Degree from Ohio State University in 1993. He was the First Inductee of the Drainage Hall of Fame at Ohio State University in 1979. He received the Horton Medal of the American Geophysical Union in 1995 and was a co-winner with Cornelius de Wit of the Netherlands of the 1983–1984 Wolf Foundation Prize for Agriculture. The citation read for 'innovative contributions to the quantitative understanding of soil, water and other environmental interactions influencing crop growth and yield.'

Rendering service to his profession, he traveled extensively and spent extended periods of time in the Netherlands, Belgium, Egypt, and Turkey. He served

on the editorial boards of both *Soil Science* and *Water Resources Research*. In addition to his exceptional abilities in research and teaching, Kirkham was a man of outstanding character and integrity. He was gentle, generous, and unpretentious, and recognized, without envy, skills and achievements of others. Don Kirkham died in Ames, Iowa, on March 7, 1998.

Kirkham's Scientific Contributions

Kirkham's early experiments at Utah State on land drainage pertained to land overlying an artesian aquifer under pressure. His results showed that spacing of horizontally laid drain tiles over such an area would not be economical. Vertically pumped drainage wells should be used. He made sand-tank models showing how the artesian drainage occurred and demonstrated that the seepage patterns depended on the boundary conditions rather than on whether the soil was a clay or a sand or a more permeable soil.

He next considered tile drainage of land over an impervious layer. He obtained equations for the streamlines, or water flow paths, and matched them against the sand-tank model flow paths obtained with dyes. A striking discovery of the theory was that by placing a drain line deeper and closer to an impervious layer, the flow would be less (rather than more) when compared with the flow from a less deep drain line in the soil. Details and amplified theory of this problem were published after World War II, in a paper which, although one in agriculture, received the 1948 mathematical award of the Iowa Section of the American Mathematical Society. The next year his paper on the seepage of water into drainage ditches was designated the best paper of the Iowa Academy of Science in 1949. In this paper he derived theoretically the amount of water that would flow into the drainage ditches. He gave the result in an explicit formula; other researchers in the past had failed to find an explicit formula.

In an early paper in 1945, Kirkham had attacked the problem of simultaneous upward and downward seepage flow into tile drains with excellent agreement of theoretical streamlines with experimental streamlines. But this paper, like others up to then, was for homogeneous soil. So Kirkham then analyzed the problem of nonhomogeneous soil and, with a resulting two-part paper, published in 1951, he received the American Geophysical Union's citation for the best paper in hydrology in 1951.

These earlier papers were for ponded water, that is, completely saturated soil, and thus represented an extreme condition. He next considered the more common condition of a curved water table resulting from successive recharges of water as from irrigation, or as from successive rains. By first neglecting the hydraulic friction in the soil under the water-table arch region and then later correcting for this neglect, he derived an explicit relation for the height of water table to be found between drain lines at certain depths and spacings. He compared his theoretical results with field data and found good agreement. The work was published in the *Transactions of the American Geophysical Union* in 1958. In a paper with a graduate student (the late Sadik Joksoz) as senior author, the solution to this drainage problem was put in a graphical form so that the various physical factors that entered into the problem could be considered in the graph and the solution obtained graphically for the depth and spacing of drains. This same student, in working for his PhD degree, considered stratified soil. Two other papers resulted, in which the ratio of hydraulic conductivities of a layered soil could be taken into account. This long, two-part paper was published in the Irrigation and Drainage Division of the American Society of Civil Engineers.

All the foregoing work was for a simplified steady-state condition of either ponded water or steady recharge to the drain tubes. Kirkham next analyzed the problem of a falling water table. He found equations that predicted where the water table arch would be after a certain time when the steady recharge was discontinued. He compared his theoretical results with actual field data that he had obtained with other scientists in the Netherlands many years before (in 1950–1951) and with other field data and published the work in 1964. There was excellent agreement.

In all the above work involving tile drains, it was assumed that the water entered the drain tube uniformly over its length. This would be true if there were an envelope of coarse material around the drain tube. Actually drain pipes were about 30–60 cm long and had about 1–6 mm spacing between adjacent drain pipes. (At least this had been the condition up until recent times when plastic tubes, perforated to let the water enter, became prevalent.) To see just how much the gaps between individual tile lengths or the perforations in tubes would control drainage flow if there were not a pervious envelope around them, Kirkham authored or co-authored several articles. A first one, published in 1951 in two parts, was co-authored by one of his joint soil physics–agricultural engineering PhD candidates (the late Glen O. Schwab), who conceived the idea of using plastic flexible tubing instead of clay pipes in soil. The two-part paper on the influence of perforations on drainage flow to drain tubes received a citation of merit from the American Society of Agricultural Engineers. Perforated flexible plastic tubing is now used almost

universally in drainage. Kirkham's theoretical work on the effect of perforations, cracks, and other openings in drain tubes continues to be of great importance.

In an artesian basin in Utah, it had been observed that tile drains would work in certain areas where it might be predicted that such drainage would be uneconomical. Kirkham reasoned that the drainage could be economical, provided there existed an upper layer of soil near the surface that was highly permeable compared with an existing stratum of soil overlying the artesian gravel. With a student and a colleague (Mohammed Najmaii and MD Dougal), he analyzed this problem in a paper published in the Irrigation and Drainage Division of the American Society of Civil Engineers in 1978. Just how much more permeable the upper soil layer must be, compared with that overlying the aquifer, was defined in the paper, which contains numerous graphs of a number of situations to be used in practical application.

All the above papers on water movement into horizontal or vertical drainage pipes depend on a knowledge of a soil coefficient called the hydraulic conductivity. This coefficient gives the ease with which the soil conducts water under an energy gradient. Early, Kirkham recognized that the use of soil cores would not give the hydraulic conductivity accurately because the samples were too small and because of their disturbed condition. He reasoned that, instead of measuring the hydraulic conductivity of soil in a core removed from the soil, measuring the conductivity of the soil around the location where the core was removed or its equivalent would be a better procedure. A practical way to proceed would be to auger a hole into the soil, while simultaneously driving a casing into the hole to seal off the walls of the hole against leakage, except for a cavity at its bottom. Flow would be allowed to come into the cavity, and, by rating this flow, the conductivity of the soil around the cavity could be determined. This method, called the piezometer method, is used for measuring the conductivity of field soil from layer to layer in the soil profile.

If a casing is not driven into a soil hole while it is being augered out, one obtains simply a cylindrical auger hole in the soil and it can be used to measure the hydraulic conductivity of the soil about the hole, provided certain mathematical formulas are found. Kirkham derived the formulas needed in papers published in 1948 and 1958. More detailed theory was later developed in 1971.

To solve a saturated soil-water movement problem, one wishes to know the potential energy in the soil and also the flow paths or streamlines of the water. Kirkham had determined the potential energy and the streamlines for a number of drainage problems.

Streamline functions and potential-energy functions, as Kirkham and his students had so far developed them, could not solve water movement problems where the flow domain was other than a relatively simple shape. Many seepage domains are, however, not of simple shape and to determine flow for them, a whole new mathematical method had to be developed. The needed new mathematical method, which he and his students developed, is called the modified Gram-Schmidt method. Kirkham's expressions are tabulated briefly in a paper published in the *Journal of Geophysical Research* in 1967 with W.L. Powers as senior author. The equations, which give the constants needed for this method of solving problems such as piezometer problems, are listed in more detail in the book *Advanced Soil Physics*, which Kirkham published with W.L. Powers in 1972. Numerous papers involving these functions have been published. Kirkham was an author or co-author in 25 such papers.

Agriculture is charged with polluting rivers and wells through runoff of dissolved chemical fertilizers, pesticides, or animal wastes. A number of pollution problems were solved and the results published by Kirkham and his students. A difficult one concerns the time required for a river that has been polluted to yield polluted seepage flow to wells near the river's side. Others concern the effect of an improper seal around a well and also the effect of a well casing and of a cover slab in stopping pollution. Complicated aquifer shapes and pollution patterns were considered in his publications.

All the foregoing problems have dealt with water-saturated soil. This is a realistic drainage situation because water cannot move from soil into a drainage facility such as a drain tube or a ditch unless the soil at an atmospheric-pressure outflow point is saturated. Nevertheless, water often moves in the soil in the unsaturated state. When Kirkham came to Iowa State University, his first paper was on horizontal unsaturated flow. In this paper, he showed that the wetting front moved by a square root of time law, but that, contrary to the statements of some other workers, this movement was not in accordance with constant-conductivity, heat-flow theory. This paper resulted in British scientists suggesting that a diffusivity or variable conductivity function needed to be used. The idea was followed through by other scientists, and numerous papers on the diffusion theory of capillary flow have subsequently been published.

In irrigated areas, salt moves to the soil surface by unsaturated flow and is removed by saturated leaching flow. In a laboratory experiment, a layer of salt was placed on a porous medium and leaching of the salt into simulated tile drains was performed. The concentration of the drainage water was analyzed as the salt was

gradually removed. A theoretical equation was derived for the salt concentration, published in 1972, and it agreed well with the experimental data.

Although Kirkham's papers were mainly on water movement in soil, he contributed to other aspects of soil physics. A paper on sediment transport by rivers and canals gave a theoretical equation that accorded with experimental data. He published papers on heat flow, on air and oxygen movement in soil, and on compaction. He and colleagues developed a neutron or subatomic particle soil probe for measuring soil moisture content. This probe, commonly called the 'neutron probe,' first described in 1952, is still widely used around the world to determine water content of soil. A portable rainfall-simulator infiltrometer was made and patented. Kirkham obtained other patents on soil physics equipment.

The wide application of Kirkham's work is evident from the numerous national and international journals in which his work is published, including the *Soil Science Society of America Journal*; *Water Resources Research*; *Soil Science*; *Transactions of the American Geophysical Union*; *Agronomy Journal*; *Journal of Geophysical Research*; *Transactions of the State Agricultural College, Ghent, Belgium*; *Journal of Irrigation and Drainage of the American Society of Civil Engineers*; *Agricultural Engineering*; *Physical Review*; *Journal of Applied Physics*; *American Math Monthly*; *Journal of Mathematics and Physics*; and many others.

Kirkham's research program attracted and inspired many students to do graduate work at Iowa State University. Under his supervision 72 students were guided towards advanced degrees in soil physics, agricultural engineering, and water resources. Through these students, his contributions to food production and the protection of land against deterioration are being applied in many countries around the world. Without the excellent and large input of capable graduate students into many papers in Kirkham's bibliography, the work could not have been done.

The strong relationships between Kirkham and his students and Kirkham's unique professional contributions are described in the following paragraph quoted from an article written by Dr Dale Swartzendruber, Professor of Agronomy at the University of Nebraska:

> Perusal of Dr Kirkham's research and publications reveals an interesting blend of diversity and specificity. Diversity has been provided by the considerable latitude he has allowed each graduate student in the choice of dissertation research problem. This resulted in problem selections that ranged broadly over the cardinal soil physics topics of soil water (saturated and unsaturated conditions), soil aeration, soil structure and strength, soil temperature and heat, plant–soil–water relationships, and water–solute interactions and processes as related to soil-water and groundwater pollution. Specificity in Dr Kirkham's research program has come from his own special genius – his unparalleled insight, capability, and skill in solving the mathematical boundary-value problems of soil-water movement, particularly as related to saturated flow and drainage. The numerous exact mathematical solutions he has provided establish Dr Kirkham as without peer in this important area of effort. Furthermore, even beyond the solutions as such, there is an additional benefit of great significance. By his direct, explicit, and unapologetic use of the mathematical–physical approach in research, he provided pioneering leadership that cleared the way for others to follow, not only in soil physics, but in related areas as well. For this he might well be called the father of mathematical soil physics.

Further Reading

Kirkham D (1940) Artificial drainage of land: stream-line experiments. The artesian basin. *Transactions of the American Geophysical Union* 21: 587–593.

Kirkham D (1958) Seepage of steady rainfall through soil into drains. *Transactions of the American Geophysical Union* 39: 892–908.

Kirkham D (1967) Explanation of paradoxes in Dupuit-Forchheimer seepage theory. *Water Resources Research* 3: 609–622.

Kirkham D and Powers WL (1972) *Advanced Soil Physics*. New York: Wiley-Interscience.

Kirkham D, van der Ploeg RR, and Horton R (1997) Potential theory for dual-depth subsurface drainage of ponded land. *Water Resources Research* 33: 1643–1654.

Swartzendruber D (1978) In recognition of Don Kirkham on his seventieth birthday. *Soil Science* 125(2): 65–67.

van der Ploeg RR, Horton R, and Kirkham D (1999) Steady flow to drains and wells. In: Skaggs RW and van Schilfgaarde J (eds) *Agricultural Drainage*. *Agronomy Monograph* 38, pp. 213–263. Madison, WI: ASA, CSSA.

L

Laminar and Turbulent Flow *See* **Hydrodynamics in Soils**

Landfills *See* **Waste Disposal on Land:** Municipal

LAND-USE CLASSIFICATION

J A LaGro Jr, University of Wisconsin-Madison, Madison, WI, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Land cover is the material cloaking the Earth's surface. Land cover includes water, soil and rock, vegetation, roads, buildings, and other constructed elements supporting human activities. Conversely, land use describes the type and intensity of human activities occurring in a given location. Land-use classification facilitates the monitoring and predicting of urban development patterns, management of water and other natural resources, and land-use planning and regulation.

Land and Society

Spatial variation in climate, physiography, and natural resources has influenced human settlement patterns throughout history. Civilizations have flourished in fertile valleys, along river and lake shores, in coastal areas, and near other highly productive ecosystems. Human population growth has precipitated extensive conversions of undeveloped areas to residential, commercial, and other 'built-up' uses. Factors contributing to population growth and dispersion include the expansion of transportation infrastructure, advances in telecommunication technologies, and a reduction in the differences between urban and rural lifestyles. These are among the many factors that reduce the influence of physiographic constraints on urban development patterns.

Landscapes are transformed when wetlands, prairies, forests, and other habitats are fragmented or replaced by buildings, roads, utility corridors, and other forms of development. These landscape changes impede the movement of wildlife and often occur too rapidly to permit the adaptation of native species through natural selection. Consequently, urban development poses a significant threat to species diversity and ecosystem resilience at both local and regional scales. Urban development may impair the functioning of biophysical systems and reduce these systems' capacity to neutralize pollutants, moderate microclimate, and provide a broad array of other important environmental 'services.' These land-use changes can fragment farms and impair the viability of agriculture on other nearby farms. Development associated with population growth and redistribution may also degrade cultural and historic resources, reduce local outdoor recreational opportunities, and destroy valued views and vistas. The environmental, aesthetic, and economic impacts of these land-use and land-cover changes depend upon the location, or spatial pattern, of development.

Spatial Information and Mapping

A geographic information system (GIS) consists of computer hardware and software, georeferenced

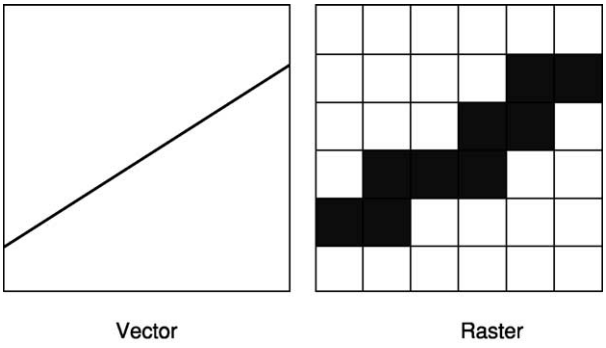


Figure 1 A linear map represented in vector and raster formats.

data on environmental or cultural attributes, and data about the data – or metadata. A GIS can facilitate the organization, analysis, and mapping of land-use and land-cover data. Digital spatial data may be in two forms: raster or vector (Figure 1). Raster data are grid-cell surfaces comprised of hundreds, thousands, or, with large data sets, millions of cells. Vector data, in contrast, consist of arcs, nodes, and polygons. Both types of spatial data are linked to tabular databases that store attribute information about the locations delineated by the grid cells or arcs, nodes, and polygons. The choice of GIS data models depends on the intended purpose of the analysis or map. Land-cover data in a raster format are widely used for environmental modeling and management. Land-use data in a vector format are useful for mapping irregularly shaped cultural features such as roads and land parcel boundaries.

Maps are efficient tools for conveying information about land and landscapes. Maps can show important spatial (or temporal) relationships among the physical, biological, and cultural attributes of a land parcel within its landscape context. Landscape attributes are portrayed with two common types of maps: (1) chloropleth maps; and (2) isopleth maps. A chloropleth map is a thematic map expressing attribute data in discrete categories. Land use, a cultural attribute, can be classified and mapped in a variety of general and more detailed categories. Because this attribute does not imply rank or quantity, land use is expressed on a nominal, or categorical, measurement scale.

An isopleth map, in contrast, displays numerical values of a single attribute. An isopleth map of elevation, for example, partitions into equal increments the entire range of elevations within the mapped area. An isopleth map of elevation might show several elevation classes, each of which represents the site's surface area where elevations fall within equal increments (e.g., 10 m). Each elevation class is simply an aggregated subset of elevations located within the mapped area.

Land-Use Classification

Land-use classification schemes typically address both land use and land cover. A major land-use classification system developed by the United States Geological Survey (USGS) has multiple levels of classification. The categories within these levels are arranged in a nested hierarchy. The most general or aggregated classification (level I) includes broad land-use categories, such as 'agriculture' or 'urban and built-up' land (Table 1). This level of classification is commonly used for regional and other large-scale applications. Within each level I class are a number of more detailed (level II) land-use and land-cover classes. For example, the 'urban and built-up' class includes 'residential,' 'commercial,' and 'industrial' subclasses. Within each of the level II classes, even more detailed classes (levels III and IV) can be defined and mapped. The classes within each level are mutually exclusive and exhaustive. That is, each location within the mapped area can be classified into one and only one class within each level. Together these four levels of classification comprise a hierarchical system for describing, monitoring, and predicting land-use and land-cover change. This standardized, multilevel classification system allows spatially explicit comparisons of land-use inventories conducted over time.

Classifying and mapping land cover is an integral step in understanding the Earth's biophysical systems. Data on the area and distribution of wildlife habitat, for example, are useful in managing and mitigating development impacts on protected and endangered species. Similarly, information on the type, area, and configuration of buildings, roads, and other impervious land-cover facilitates the modeling of storm-water runoff and watershed hydraulics and hydrology.

Different land-cover surfaces reflect sunlight in distinct wavelengths of the electromagnetic spectrum. These spectral 'signatures,' when recorded as remotely sensed images, permit the classification and mapping of individual land-cover types. Sources of data for the classification and mapping of land cover and land use include aerial photographs, satellite imagery, and maps of assorted social and cultural attributes (e.g., property boundaries, population density).

Land-Use Suitability

A location that is suitable for a particular urban land use is one that can accommodate the proposed use with the minimum amount of inputs or resources. This concept of land-use suitability is similar to the United States Natural Resources Conservation Service (NRCS) method for classifying the capability

Table 1 Example of a two-level land-use/land-cover classification system

Level I	*Level II*
1. Urban or built-up	11. Residential
	12. Commercial and services
	13. Industrial
	14. Transportation, communication, and utilities
	15. Industrial and commercial complexes
	16. Mixed urban and built-up land
2. Agriculture	21. Cropland and pasture
	22. Orchards, groves, vineyards, nurseries, and ornamental horticultural areas
	23. Confined feeding operations
3. Rangeland	31. Herbaceous rangeland
	32. Shrub and brush rangeland
	33. Mixed rangeland
4. Forest land	41. Deciduous forest land
	42. Evergreen forest land
	43. Mixed forest land
5. Water	51. Streams and canals
	52. Lakes
	53. Reservoirs
	54. Bays and estuaries
6. Wetlands	61. Forested wetlands
	62. Nonforested wetlands
7. Barren	71. Dry salt flats
	72. Beaches
	73. Sandy areas other than beaches
	74. Bare exposed rock
	75. Strip mines, quarries, and gravel pits

Source: After Anderson JR, Hardy EE, Roach JT, and Witmer E (1976) *A Land Use and Land Cover Classification System for Use with Remote Sensor Data*. Geological Survey Professional Paper 964. Washington, DC: US Government Printing Office.

of soils for agriculture. The NRCS ranks soils by the limitations they present to various types of agriculture. Prime soils, for example, require the fewest inputs for productive agriculture. Locations with these soils do not require irrigation or soil contouring or terracing. These areas also require comparatively smaller amounts of fertilizers and pesticides.

Land-use suitability analysis identifies and maps land areas with different levels of development suitability. The analyst must first identify the attributes that are most relevant to the proposed land uses. The suitability criteria for each land use must then be specified. These suitability criteria include the site conditions that are most desirable, or undesirable, for each land-use type. Suitability criteria for a given land use can vary substantially with differences in local and regional conditions.

The analysis of individual attribute maps is, in many cases, a process of partitioning the spatial distribution of attribute values. A proximity analysis of individual attribute layers may also identify locations within a defined distance of specified points (e.g., entrances), lines (e.g., roads, streams), or areas (e.g., ponds or lakes). A suitability analysis involves three discrete steps:

1. identify suitability criteria for each expected land use;
2. collect and map relevant site attribute data;
3. identify the site locations with attribute values that meet the land-use suitability criteria.

The analysis of an attribute layer may have several objectives, but the primary objective is to find areas that meet a given condition or set of conditions. These areas could be locations on the site where attribute values are:

1. greater than a specified minimum (e.g., elevation of at least 1 m (3 feet) above sea level);
2. less than a specified maximum (e.g., slope gradients less than 20%);
3. within a specified range (e.g., soil drainage is suitable for constructing building foundations and roads).

The analyst can map both favorable and unfavorable conditions for each proposed land use. This suitability map may be the product of either the union or intersection of two or more attribute maps. For example, the intersection of three attribute layers (slope, aspect, and soil drainage) could yield a map showing all locations in an area with slopes less than 20%, a southern aspect, and well-drained soils. This map would have two classes: suitable or unsuitable for the proposed land use.

Subsurface Conditions

Unstable soils can reduce a site's stability for construction of buildings, pedestrian and vehicle circulation systems, and other elements. Other important soil attributes that may be considered, depending on the site's intended uses, include:

1. fertility for plant growth;
2. permeability for drainage;
3. erosion potential;
4. depth to seasonally high water table;
5. depth to bedrock.

The composition and stability of the bedrock that lies beneath, and in some locations at, the Earth's surface are important attributes in land-use planning. The depth to bedrock, in particular, must be assessed if excavations are planned for building foundations or other structures.

When local groundwater is the source of a community's potable water, stringent efforts must be taken to prevent the contamination of this natural resource. Onsite wastewater-treatment systems may utilize soil as the treatment medium. Public health regulations may confine rural residential development to sites that are suitable for soil-based wastewater treatment. Conventional soil-absorption treatment systems function effectively on flat to gently sloping sites with relatively permeable soils containing a combination of sand, silt, and clay. Constraints that preclude the installation of these on-site systems include: high water table, shallow bedrock, steep slopes, flooding, slow substratum permeability, and slow soil permeability.

By increasing the percentage of impervious surfaces in a watershed, land development tends to increase total storm-water runoff volume, reduce groundwater recharge, and increase flooding potential. Increased runoff from land development can increase the risks of flooding, erosion, and habitat degradation. Without mitigation, new construction and other site-disturbing activities compact soils, add impervious land-cover surfaces, and reduce both storm-water infiltration and groundwater recharge. The configuration of impervious surfaces determines their effect on storm-water runoff. Disconnecting impervious surfaces, and separating them with pervious surfaces such as vegetated buffer strips and 'rain gardens,' can help facilitate storm-water infiltration on site. Rain gardens are small, bowl-shaped gardens, planted with perennial native plants and designed to absorb storm-water runoff from roofs, parking lots, and other impervious surfaces.

Poorly draining soils or a shallow water table are development constraints that must be identified early in the land-use planning process. These constraints can be overcome with sump pumps, foundation drains, and other site engineering techniques. But this approach increases development costs and, potentially, exacerbates development impacts on the environment. Sites with a history of prior agricultural use may have experienced significant losses of topsoil. Topsoil depletion increases the costs of re-establishing vegetative cover after construction.

Former commercial and industrial sites should be carefully investigated prior to land-use planning and site development. Gasoline stations, printing presses, dry cleaners, and other commercial sites may be contaminated by a variety of toxic substances. Heavy metals and other toxic substances pose substantial health risks, and if hazardous wastes are present on a site, the costs of remediation can be significant.

Surface Conditions

A topographic survey is essential for most land-use planning projects. Elevation changes create slopes that have both a gradient and an orientation, or aspect (the compass direction that the slope faces). A site's suitability for roads, walkways, buildings, and other structures is partially a function of the existing slope gradients on the site. Climate also affects the influence of slope on site suitability. The maximum slopes at which roads and walkways are safe for transportation are not the same in a climate with freezing temperatures and a climate where temperatures are always above freezing. Particularly in locations with seasonal freezing temperatures, pavement gradients must be relatively low (typically less than 12%) to prevent slipping on icy surfaces. Variations in slope gradient and aspect affect the levels of solar radiation striking the site's surface or objects on the site. This can influence the amount of energy expended in heating and cooling buildings on the site. Elevation changes also affect drainage patterns, which influence plant distributions.

Hurricanes, earthquakes, and other natural hazards pose significant risks to human health, safety, and welfare. Often the loss of life and property can be prevented, or at least mitigated, by reducing these risks. Yet vulnerability to natural hazards is not uniformly distributed across the landscape. Sites on the edges of water may be attractive properties for real-estate development. If the land is inherently unsuited for development, however, homes and other structures in these locations can be at serious risk. Rainfall associated with hurricanes and other long-lasting storms typically saturates soils, fills up surface water-storage capacity, and increases the risk of local flooding. The main threat to life and property is the storm surge (the water pushed toward the shore by the force of the

storm's winds); therefore, development in low-lying areas is especially vulnerable.

Preventing development of naturally hazardous sites, rather than attempting to control nature with technology, is both economically efficient and environmentally sound. Confining land development to sites without significant constraints reduces construction costs, protects inherent natural amenities, and allows the continuing function of critical environmental systems.

Land-Use Planning

Land-use changes are a direct result of the technologies, institutions, and values of human society. Critical environmental resources can be protected by directing growth and development to more suitable areas served by public infrastructure. At the community or municipal scale, land-use planning determines the locations of new roads, recreation areas, commercial, industrial, and residential developments, and other land uses. Municipal land-use planning is one way government protects public health, safety, and welfare. Land-use planning within the public sector is increasingly important in protecting environmental quality.

At the parcel or site scale, land-use planning determines the locations of individual buildings, pedestrian and vehicle circulation systems, and other built elements. Site-scale planning is initiated by individuals, corporations, and by agencies or departments of local, state, and federal governments. Land-use regulations, local real-estate markets, and cultural and physiographic attributes all influence, to varying degrees, a site's capacity to accommodate a particular land-use program. Zoning ordinances and other local codes may dictate permitted land uses, building heights and densities, and the location of buildings on the site. Federal and state regulations may also affect a myriad of land-use planning decisions.

Site Selection

Property value is affected by the size and shape of the parcel. Larger sites can accommodate more development and offer greater flexibility in the types and spatial arrangement of uses on the site. Property appraisers in the USA attempt to estimate value based on the property's 'highest and best use.' For residential and some commercial uses, scenic views to lakes, rivers, or other environmental features may contribute significantly to market demand and property value. Parks and other recreational facilities are amenities that also enhance property value. Long and linear properties have a much higher edge-to-interior ratio than properties that are more square in shape. Greater 'exposure' to the surrounding landscape may reduce the site's ability to buffer negative off-site impacts. However, if the adjacent areas are natural or cultural amenities, a parcel with a relatively high edge-to-interior ratio can be desirable, particularly if these amenities are in public ownership and are likely to remain so well into the near future.

Site-selection decisions involve consideration of economic, social, environmental, and even ethical, factors. Potential sites are evaluated and selected, or rejected, on the basis of each site's suitability for the intended land uses (i.e., program). Rugged topography, unstable soils, and shallow bedrock are just a few of the constraints that may deter development. Development on difficult sites, particularly if the site constraints are not avoided or the impacts are not mitigated, can have serious on-site and off-site consequences. These constraints also increase design and construction costs substantially.

Typically, one or two factors have a disproportionately large influence on the site-selection decision. These factors may be essential site attributes that, if absent, exclude a site from further consideration. For example, the primary site-selection criterion for a development of single-family homes in an non-sewered area may be the suitability of the soil for on-site wastewater-treatment systems. If the essential criteria are met, the site-evaluation process can proceed to consider other selection criteria. Severe constraints, however, would exclude the site from further consideration.

Programming

A land-development program is simply a description of the intended site uses. The programming phase often involves visiting the site, inspecting aerial photographs, performing market analyses or user-demand studies, and reviewing other available site and user information. The site's suitability for future development or restoration depends, however, on the land-use program under consideration. Site attributes relevant to a proposed nature conservation area, for example, would differ significantly from the site attributes relevant to a proposed housing development. A nature conservation program would place a high priority on a diversity of wildlife habitats, e.g., from dry upland south-facing slopes to lowland marshes and bogs. In contrast, the housing project might place a high priority on gently sloping areas without development constraints like shallow bedrock, shallow water table, or protected endangered species.

Site Inventory

The site inventory is an important step in understanding the physical, biological, and cultural linkages

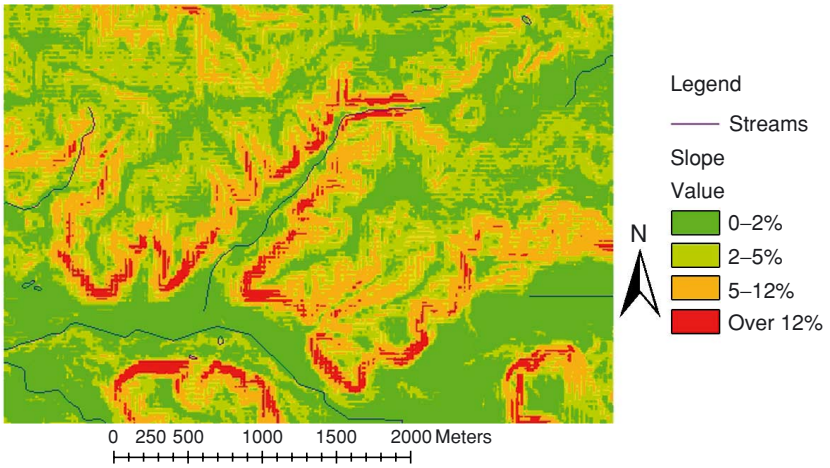


Figure 2 A topographic slope gradient map, showing four classes of slope gradients.

between the site and the surrounding landscape. The inventory is not, however, an open-ended investigation of biophysical or cultural phenomena. The project program, in conjunction with the site's physical, biological, and cultural features, determines which data should be collected in the site inventory. Consider, for example, a partially wooded and hilly site along a small, but navigable river. Several site attributes could be mapped, but all of these may not be relevant to any given program. A map showing the site's range of slope percentages (e.g., 0–2%, 5–12%) (Figure 2) could be potentially useful because some land uses are better suited for gentle slopes, and other uses are better suited for steeper slopes.

Sources of data for the site inventory include topographic surveys, aerial photographs, soils maps, and other maps created to support the land-use planning process. Ecologists, hydrologists, anthropologists, and other experts may participate in collecting, mapping, and analyzing the site-attribute data. The inventory documents the key biophysical and cultural attributes of the site and adjacent land. An inventory map of existing vegetation, for example, shows the existing site conditions for a single attribute – vegetation type. This map will remain valid whether the proposed uses of the site include intensive development or no development at all. Maps of site drainage patterns, steep slopes, or other site attributes are similarly unaffected by changes in the proposed land-use program.

Site Analysis

The site analysis is a program-driven assessment of a site's physical, biological, and cultural attributes. This analysis is tailored to each unique land-use program. Attribute maps in the site inventory provide the data needed for this program-specific analysis. The analysis of site attributes is made tractable by considering the assets and liabilities, or opportunities and constraints, that the site poses for the proposed land-use program. Information contained in physical, biological, and cultural attribute maps can be synthesized to create maps of the site's suitability for each expected land use (Figure 3). The liabilities (constraints) associated with a site and its context reduce the site's suitability for one or more of the proposed uses. Conversely, the assets (opportunities) associated with a site and its context increase the site's suitability for the proposed uses. Site amenities are features that have social, economic, ecological, or aesthetic value. Amenities warranting protection include features that enhance the visual quality of the site and the surrounding community. These include specimen trees, natural drainage swales and depressions, unique rock outcrops, scenic overlooks, and historic or culturally significant buildings and structures.

Plan Development and Implementation

Land-use suitability is a function of the site's soil conditions, slopes, microclimate, and other biophysical and cultural attributes. Land-use plans locate proposed activities, infrastructure, and other physical improvements on the site. Implementing a plan might include the construction of buildings, roads, walkways, or various other site structures. Site constraints increase the development's complexity and cost. By utilizing modern engineering technology, some land uses can be accommodated in locations that are inherently unsuited for those uses. For example, housing could be built in marshy areas, on highly erodible soils, and on a wide variety of other 'difficult' sites. Preparing these areas for development might require tree removal, extensive regrading and drainage, and modification of natural storm-water runoff and infiltration

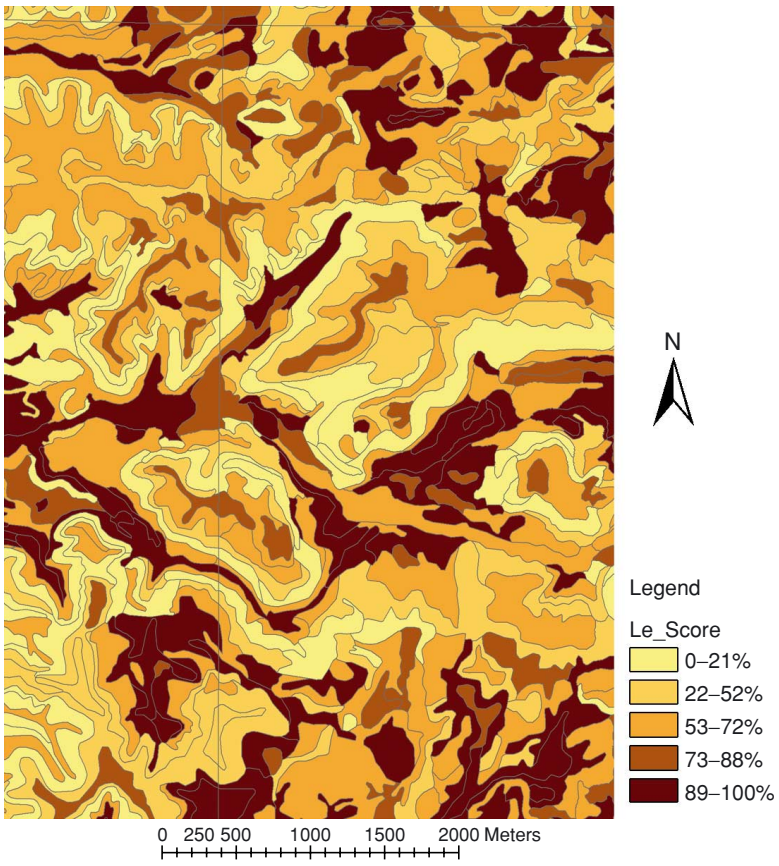


Figure 3 A land evaluation map showing a range of site suitability for residential development; higher scores indicate greater development suitability.

patterns. Yet development in unsuitable areas can destroy natural aesthetic amenities, degrade unique wildlife habitat, and alter long-standing hydrological regimes. This 'brute force' approach can also have significant negative environmental impacts.

Conclusion

Low-impact land-use planning protects the land's ecological 'infrastructure' from disturbances associated with development. This ecologically sensitive approach to planning can also involve environmental restoration. For example, restoring a riparian community along an eroded stream bank can stabilize soils, filter sediments and chemical pollutants from storm water, and create habitat for birds, insects, and other beneficial species. These land-planning decisions require accurate and spatially explicit information on the land's suitability for proposed uses. With accurate and comprehensive knowledge of the area's opportunities and constraints, land planners can mitigate, and often prevent, environmental impacts associated with population growth and redistribution.

See also: **Classification of Soils**; **Geographical Information Systems**

Further Reading

Anderson JR, Hardy EE, Roach JT, and Witmer RE (1976) *A Land Use and Land Cover Classification System for Use with Remote Sensor Data*. Geological Survey Professional Paper 964. Washington, DC: US Government Printing Office.

Beatley T and Manning K (1997) *The Ecology of Place: Planning for Environment, Economy, and Community*. Washington, DC: Island Press.

Forman RTT and Godron M (1986) *Landscape Ecology*. New York: John Wiley.

Goudie A (1994) *The Human Impact on the Natural Environment*, 4th edn. Cambridge, Massachusetts: The MIT Press.

LaGro JA Jr (2001) *Site Analysis: Linking Program and Concept in Land Planning and Design*. New York: John Wiley.

Lynch K and Hack G (1984) *Site Planning*. Cambridge, Massachusetts: The MIT Press.

McDonnell MJ and Pickett STA (eds) (1993) *Humans as Components of Ecosystems*. New York: Springer-Verlag.

McHarg IL (1969) *Design with Nature*. Published for the Natural History Press. Philadelphia: Falcon Press.
Mileti DS (1999) *Disasters by Design: A Reassessment of Natural Hazards in the United States*. Washington, DC: Joseph Henry Press.
Odum EP (1989) *Ecology and Our Endangered Life-Support Systems*. Sunderland, Massachusetts: Sinauer Associates.
Zonneveld IS and Forman RTT (eds) (1990) *Changing Landscapes: An Ecological Perspective*. New York: Springer-Verlag.

LAWES, JOHN BENNET AND GILBERT, JOSEPH HENRY

A E Johnston, Rothamsted Research, Harpenden, UK

© 2005, Elsevier Ltd. All Rights Reserved.

In agricultural science the names of Lawes and Gilbert are forever linked, and together they are regarded as the founding fathers of the scientific method in agricultural research. Gilbert took up his appointment with Lawes at Rothamsted on 1 June 1843, the date now regarded as the start of Rothamsted Experimental Station (now called Rothamsted Research). This was the beginning of one of the longest (57 years) and most productive scientific partnerships on record. Their skills and characters were totally complementary. Lawes, the landowner and practicing farmer, was versatile and farseeing, quick to appreciate the needs of the farmer and able to respond with sound practical and economic advice based on the results of the Rothamsted experiments. But the quality of these results owed much to Gilbert. Everything he undertook was performed with scrupulous accuracy and meticulous care to detail. Gilbert indeed gave scientific expression to the work of Rothamsted and established the experiments on farm animals and crops on a sound scientific basis. Together Lawes and Gilbert published some 150 scientific papers and 300 or so popular articles for farmers (Figure 1).

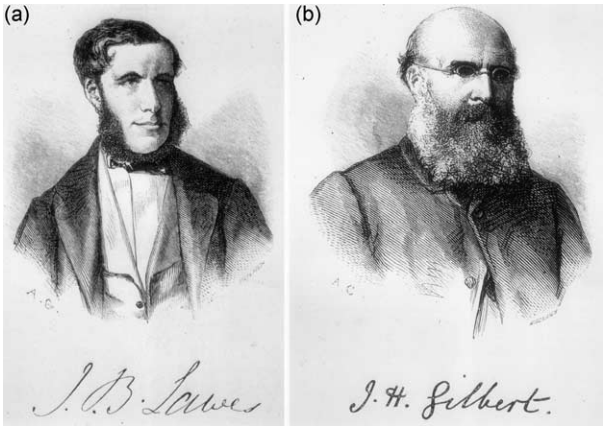


Figure 1 Lawes and Gilbert as young men: (a) J. B. Lawes; (b) J. H. Gilbert. Photographs probably taken in the early 1850s.

It is impossible to conceive that, in 1843, Lawes and Gilbert could have had the remotest idea that some of the field experiments they started on agricultural crops during the next 13 years would still be continued in the twenty-first century. Today these experiments are the oldest, continuous field experiments in the world and their importance is recognized worldwide. With modifications introduced only after very careful consideration, they still provide scientific data of immeasurable value not only to farming practice but also to the wider issues of the sustainability of farming systems and the interaction of agriculture and the environment. It is worth considering, therefore, the background to the start of these experiments and of their founding fathers.

John Bennet Lawes

Lawes (later Sir John Lawes) was born in 1814 at Rothamsted, a small agricultural estate near Harpenden, some 40 km north of London. When his father died in 1822, his widowed mother with her two children, John and Caroline, temporarily had to leave the manor house for financial reasons. During this period, John went first to Eton and then, as an undergraduate, to Oxford. When he was 20, Lawes' mother returned to live at the manor house. So, before taking his degree, Lawes decided to leave Oxford to manage the Rothamsted estate of approximately 100 ha on behalf of the family. He appears to have had few regrets at leaving university, because he found little to stimulate an inquisitive scientific mind in the formal teaching offered at that time.

While at Oxford, however, Lawes must have had contact with Professor C.G. Daubeny, FRS, whose experiments and writings were probably a major formative influence on Lawes. Daubeny simultaneously held the Chairs of Chemistry, Botany, and

Rural Economy and his research and writings from the 1820s onward ended a very barren period in the development of agricultural science in the UK. Two of Daubeny's interests are important to subsequent developments at Rothamsted. He was doing experiments in the Botanic Garden on the effect of a crop on the growth of a subsequent crop (allelopathy) and writing about the essential need for good experiments to help understand and improve agricultural practices.

Joseph Henry Gilbert

Born in 1817 and therefore very close in age to Lawes, Gilbert's background was very different to that of Lawes. As a boy, Gilbert (later Sir Henry Gilbert) had, according to his mother, only three interests: electricity, arithmetic, and chemistry. He studied chemistry at university, first at Glasgow then at University College London. After graduating he worked initially with Professor Anthony Todd Thompson, founder of the British Pharmacopoeia. Early in 1840 he went to work with Professor Justus von Liebig at Giessen, in Germany, and was awarded a doctorate within 12 months; apparently the criteria were different then from now! Returning to England, he was at University College London in 1841, before going to an industrial appointment in Manchester. Lawes may have met Gilbert when the latter was in London. Lawes had at least visited Thompson's laboratory and that of Thomas Graham, Professor of Chemistry and a world authority on phosphorus chemistry at that time. Later, in 1843, Graham, who was very friendly with Liebig, wrote to him asking him to receive Lawes at Giessen, but there is no record of a visit taking place.

The Agricultural Background in the Second Quarter of the Nineteenth Century

Lawes' and Gilbert's achievements need to be set against farming practice and the state of knowledge in the 1830s. Agronomic techniques had been changing only slowly during the previous centuries. Food production in the middle ages was based on the open field system. Although the land in each village belonged to a single landowner, the arable land was divided into two or three large fields and most villagers were allowed to cultivate small, scattered strips of land in each field. Each field was cropped in rotation, usually fallow followed by a cereal, or fallow followed first by a legume and then a cereal. Each villager aimed to produce at least enough grain to feed his family. In addition each villager's animals had access to common grazing land. Landlords started to create distinct fields, by enclosing, with hedges or fences, open grazing land in the fifteenth and sixteenth centuries. Later, in the latter half of the eighteenth and early nineteenth centuries, similar enclosure of the open arable fields occurred. Thus, landlords divided their land into discrete farms, each with part arable crops and part permanent grass. Although these enclosures allowed the opportunity for greater efficiency in the use of land, they created a large population who had no access to land to produce their own food.

This increased efficiency was aided by two inventions by Jethro Tull in the mid eighteenth century. The first was the seed drill and second, the horse-drawn hoe. This allowed seed to be placed at a uniform depth, improving germination, in regularly spaced rows, and the interrow space to be kept relatively free of weeds by mechanical hoeing. There then developed a four-course rotation of arable crops: roots, spring barley, a legume (either herbage, e.g., clover, or a grain legume, e.g., peas or beans) and winter wheat (the Norfolk four-course rotation). The introduction of the root crop (swedes, turnips, mangels) and the herbage legume allowed more animals to be kept over winter, with some return of plant nutrients via the farmyard manure (FYM) produced. However, supplies of FYM were limited and that which was available was usually applied to the root crop. Winter wheat, producing the staple food of the people, was grown only once in 4 years and must have obtained its nitrogen (N) from the residues of the leguminous crop that preceded it. This rotation became the backbone of arable cropping first in England and then in much of temperate Europe. Landlords insisted that their tenants followed it because it minimized nutrient losses from the farm and the build-up of pests and diseases, and it helped control weeds. There was one problem, however: the root crop was sown usually in early June and if it did not germinate and grow quickly the emerging seedlings could be killed by the turnip flea beetle. This decreased yields and the amount of food available for animals in the winter.

In the early years of the nineteenth century, there were still major gaps in knowledge of crop nutrition. Following the studies of de Saussure and others, it was generally accepted that plants acquired their mineral nutrients (phosphorus, P; potassium, K; calcium, Ca; magnesium, Mg; etc.) from the soil, but there was considerable uncertainty about the source of nitrogen (N). From analyses done in Alsace, France, it was known that FYM contained both minerals and nitrogen (N).

In Britain at about this time, N was becoming available as very impure ammonium sulfate, a by-product of the manufacture of town gas, and as sodium nitrate, imported from Chile. Potassium salts were available in limited supply, mainly from

Russia and North America. They were obtained by burning plant material (mainly trees) in large metal pots and extracting the soluble salts from the ash with water, hence the name 'potash,' the word frequently used for potassium in agriculture. The large natural deposits of K salts in Germany were not discovered until the 1860s.

In Britain in the early 1800s, it was observed that bone trimmings, from the manufacture of bone handles for steel cutlery, when applied to land speeded the germination of turnips and hence the yield. This led to a considerable trade in bones that was specific to the UK. During 1837–42 the estimated mean annual British domestic supply was 27 000 t, while imports averaged approximately 46 000 t. Justus von Liebig accused the British of sustaining these imports by robbing the battlefields of Europe and he maintained that the imports amounted to the manurial equivalent of an army of some 3.5 million men!

The Years at Rothamsted Immediately Before 1843

Returning to Rothamsted in 1834 to manage the estate, Lawes had a bedroom at the manor house converted to a laboratory and attempted to determine the 'active principles' of the many medicinal plants, including poppies, he started to grow on the farm. With only one untrained person to help, he was unsuccessful. However, a useful lesson was probably learnt – the need for skilled help and rigorous analytical methods. He also attempted to develop commercially a newly proposed method for the manufacture of calomel, mercurous chloride, then used as a purgative in medicine. Again he was without success, but he must have learnt some valuable lessons about 'trade' (running a commercial enterprise), not then normally a part of a Victorian country gentleman's education and training, but soon to be put to good use by him.

The Earliest Experiments

In 1837 and 1838, Lawes did some simple experiments using cabbages and turnips, both used for animal feed, and grown either in pots or on small plots on the farm fields. He compared a range of nitrogen-containing salts, each applied at the same rate of salt. He was clearly puzzled by the fact that the yields were not increased in proportion to the amount of N applied, but he did comment: "phosphate of ammonia to be one of the most powerful manures known." In a second article in the *Gardeners' Chronicle*, Lawes noted that applying ammonia to a soil deficient in P was useless and, where minerals were in adequate supply, the ammonia supplied by the atmosphere would be insufficient for the wants of a crop.

Early Observations on the Need for Phosphorus and the Production of Superphosphate

Once he was managing the farm, Lawes noted, as did other farmers in the locality, that crushed bones, seen to be so efficacious in certain parts of the country, did not increase the yield of his or his neighbors' turnips. Like others, however, Lawes found that if bones were treated with sulfuric acid to produce superphosphate of lime then, when this was applied to land, the yields of all the arable crops grown on the farm were greatly increased. Lawes did many experiments to determine the ratio of bones to acid and the acid strength to produce a product that was dry but hard so that it could be crushed to a fine powder and be spread by machine. In 1842, Lawes took out a patent for his process and besides bones he included mineral phosphates as a source of phosphorus. In 1843 he started production of superphosphate on a commercial scale at a factory at Deptford, London. This was the start of what was to become a successful fertilizer business. A much larger factory was opened at Barking Creek in 1852.

It was from the profits of his business enterprises (he also had a factory producing citric acid) that Lawes financed the agricultural experiments at Rothamsted. No doubt bearing in mind the pleas of Daubeny for well-conducted agricultural experiments, Lawes wisely decided that he could not be both a fledgling industrialist with a business in London and an aspiring agricultural researcher at Rothamsted 40 km away. In consequence Lawes appointed Gilbert to help set up and supervise the experiments on the nutrition of both animals and crops that he was keen to start.

The Start of the Experimental Program at Rothamsted

In many parts of the world today, Rothamsted is associated with those long-term, 'Classical' field experiments on arable crops and permanent grass started by Lawes and Gilbert between 1843 and 1856 that still continue (Table 1). Lawes, however,

Table 1 The principal field experiments started by Lawes and Gilbert at Rothamsted and the crops each investigated

Field	*Year experiment started*	*Crop*
Barnfield	1843	Turnips
Broadbalk	1843	Winter wheat
Geescroft	1847	Beans
Hoosfield	1849	Clover
Hoosfield	1852	Spring barley
Geescroft	1869	Oats
Agdell	1848	Crops in rotation
Park Grass	1856	Permanent pasture

was also interested in animal nutrition, and the pioneering experiments undertaken at Rothamsted laid the sure foundation for this branch of agricultural science. In fact, in the early years of the Rothamsted experiments, more effort was probably expended on these animal studies than on those with crops. Research on animals gradually diminished and ended in the 1890s, after Lawes and Gilbert had published some 50 scientific papers and articles on animal nutrition.

With regard to the field experiments on crops, Lawes had made some important decisions about their conduct before he appointed Gilbert. Probably the most important was to have separate experiments for each of the principal arable crops grown in rotation. This decision was probably a direct result of his contact with Daubeny whilst at Oxford and the results from the latter's studies of the effect of one crop on a subsequent crop. Over the period 1843–69, Lawes and Gilbert started an experiment on each of the arable crops grown most frequently in the Norfolk four-course rotation and, in 1856, an experiment on the nutrient requirements of permanent grassland (Table 1). Lawes and Gilbert quickly learnt that running a large-scale field experiment program not only required their skills in planning, designing, and supervising, but also, and equally important, the farm staff had to acquire the appropriate working skills for the field operations (Figure 2). The decision to grow each crop in monoculture illustrates Lawes' scientific approach, because the decision flew in the face of the accepted agricultural practice of crop rotation. Lawes justified the decision on the grounds that this not only would highlight the characteristic nutrient requirements of each crop, but also would emphasize differences between crops.

Another important decision taken at the start of the experiments identifies Lawes as a farmer. He decided to have large plots, believing that fellow farmers were more likely to accept the results if they could be seen to be consistent over a large area. Initially on Broadbalk there were 21 long, narrow plots (each about 322×7.5 m) on approximately 4.4 ha, the width being related to the 'lands' into which the field was divided for cultivation (Figure 3). These large plots were but one among the many legacies left us by Lawes, because it has been possible to subdivide the plots to test changes in treatment. Lawes must also have decided the treatments to be tested on the Barnfield experiment on root crops, because it was started very soon after Gilbert came to Rothamsted on 1 June 1843. Besides an unmanured control and FYM, the other treatments were probably based not only on his experience, but also on the views expressed by Liebig in 1840. In his now-famous report, commissioned by the British Association for the Advancement of Science, entitled *Organic Chemistry in its Application to Agriculture and Physiology*, which was presented in 1840 and also first published in that year, Liebig stressed the need to apply the mineral elements, i.e., plant nutrient elements other than nitrogen. The availability of simple, soluble inorganic salts of many of these mineral elements and of nitrogen made it possible for Lawes and Gilbert to test them singly and in various combinations, and compare the yields with those given by FYM as well as the unmanured control. The treatments initially tested on the winter wheat experiment started on Broadbalk in autumn 1843 (Figure 2) were very similar to those on Barnfield.

In the first year of the Barnfield experiment, only eight of 23 plots on Barnfield tested nitrogen and then only up to 15 kg N ha^{-1}. The effects of N were far from clear. The method of applying fertilizer and drilling the seed brought both into close proximity and germination was impaired, especially with ammonium salts. Where plants survived they were individually larger than with minerals only, but the increase in total yield was small because plant numbers were less. Similarly for the first crop of winter wheat sown in autumn 1843 on Broadbalk, nitrogen was tested on only five of the 21 plots. But in sharp contrast to the variable effects of N on Barnfield, winter wheat grain yields at harvest in 1844 were much increased by the modest amount (15 kg ha^{-1}) of nitrogen tested as an ammonium salt. Compared with the unmanured crop, minerals alone increased yield by 10%, minerals plus N by 35%. Lawes had not used cereals in his earlier experiments, because it was not normal practice to apply nutrients to wheat that followed a legume in the Norfolk four-course rotation. Whether or not this response to N was a surprise is unknown, but Lawes and Gilbert were quick to appreciate the significance of the result. For the next crop of winter wheat, drilled in autumn 1844, N was tested on 14 of the 21 plots and at four rates, 12, 24, 36, and 48 kg N ha^{-1}.

Lawes and Gilbert went on to prove that it was more important for crops, other than legumes, to have an adequate supply of readily available N than of minerals. Because soils at that time were usually deficient in P, yields with N, P, and K were always largest, but the effect of N was greater than that of P and K (Table 2). The experiments also showed that inorganic fertilizers applied in appropriate combinations and amounts could give the same yield as the large amount (35 t ha^{-1}) of FYM tested (Table 3). Lawes and Gilbert never said that fertilizers were better than FYM. What they foresaw was that fertilizers would have to be used to supplement the limited supplies of FYM if there was to be sufficient food

Quantities of Manure in lbs per Acre. 1st Season 1843-4.

Plot	Mineral Manures: Alkalies: Phosphate Potass	Phosphate Soda	Phosphate Magnesia	Silicate Potass	Superphosphate of Lime Manure	Organic Manures: Sulphate of Ammonia	Rape Cake	
Harpenden Butts	Mixture of equal parts of all the numbered Plots except Plot 2.							
0	Mixture of equal parts of all the numbered Plots except Plot 2.							
1					700		150	
2	14 Tons Farm-yard Manure							
3	Unmanured							
4	The ashes of 14 Tons of Dung burnt =				32 cwts per acre			
5					700			
6			420		350			
7		336			350			
8	364				350			
9					616	65		
10				224	560			
11					350		308	
12		168	210		350			
13	200		210		350			
14			210	275	350			
15	150		168	112	350			
16	75	65	84	112	350	65		
17	75	65	84	112	⁂ 350	65		⁂ Made with burnt bones.
18	75	65	84	112	350	65	154	
19		80	106	110	350	80		
20								

Figure 2 The manure sheet for the first crop of winter wheat sown autumn 1843 on Broadbalk and harvested in 1844.

for the rapidly expanding urban population. Their results also allowed them to say with some confidence that yields could be maintained by using fertilizers. Figure 4 shows that, over the subsequent years, yields of winter wheat on Broadbalk have not only been maintained but have increased where fertilizers are used. Similarly, yields on the other classical experiments have been increased as new, higher-yielding cultivars have been introduced. Clearly in these experiments the soil, climate, crop, and management have been in 'harmony'; excellent examples of the sustainability of crop production under these conditions.

In the second half of the nineteenth century, the demand for N fertilizers increased to such an extent that in 1898, Sir William Crookes, in his presidential address to the British Association, warned that the world's known reserves of combined N were being rapidly exhausted. Fortunately it was not long before scientists and chemical engineers discovered how to fix atmospheric nitrogen. This was done first by the Norwegians Birkeland and Eyde, who combined nitrogen and oxygen to produce nitrate in an energy-intensive process. Later the more effective process of combining nitrogen with hydrogen to form ammonium, the Haber–Bosch process, was developed.

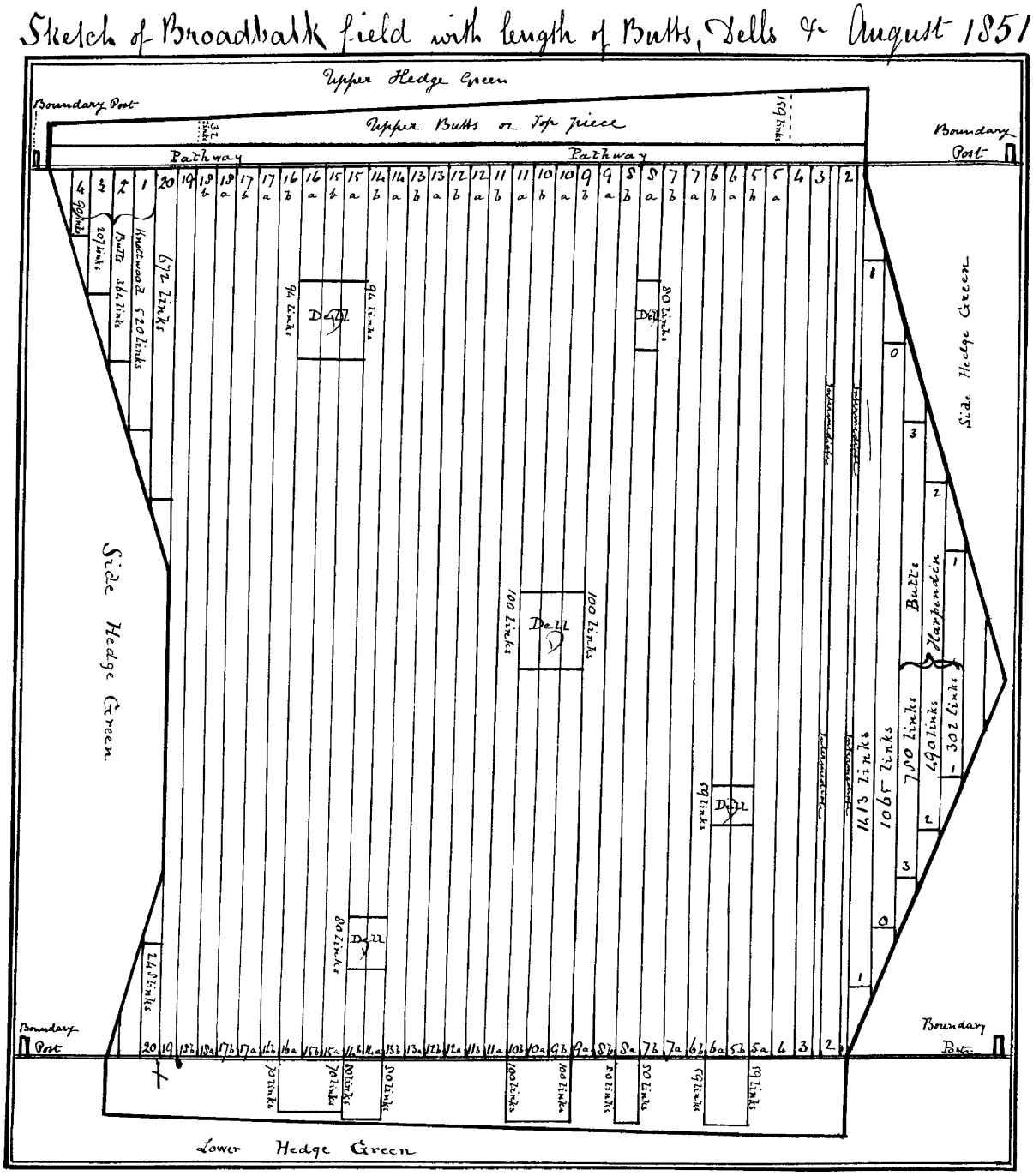


Figure 3 Division of the Broadbalk field at Rothamsted, 1851.

As they began to appreciate the questions to ask of their experiments, Lawes and Gilbert modified the treatments tested in the early years on Barnfield and especially on Broadbalk. The changes on the latter were designed to ascertain the direct and residual effect of fertilizers supplying ammonium and nitrate nitrogen, the major topic of debate with Justus von Liebig in Germany. They also decided that, although large plots were important, long narrow plots, as on Broadbalk and Barnfield, were not the best arrangement. When the experiment on spring barley was started on Hoosfield in 1852, the plots were squarer (42 × 17.5 m) than those on Broadbalk, and the treatments were arranged so that P and K were tested on strips running east–west, which were crossed at right angles with strips testing N. Outside of the main block of 16 plots (nil, P, K, PK × nil and three sources of N, ammonium and nitrate salts, and rape cake, an

Table 2 Yields (tonnes per hectare) of winter wheat, spring barley, mangels, and beans in the early years of the field experiments at Rothamsted

Experiment field name	Crop	Years	Yield ($t\,ha^{-1}$)			
			None	PK	N	NPK
Broadbalk	Winter wheat, grain	1852–61	1.12	1.29	1.63	2.52
Hoosfield	Spring barley, grain	1872–81	0.85	1.10	1.66	2.62
Barnfield	Mangels roots	1876–84	9.5	11.3	25.6	46.0
Geescroft	Field beans[a], grain	1847–62	1.11	1.50	NT	1.64

[a] *Vicia faba*.
NT, not tested.

Table 3 Yields (tonnes per hectare) of winter wheat and spring barley, grain at 85% dry matter, and roots of mangels and sugar beet at Rothamsted

Experiment field name	Crop	Period	Yield ($t\,ha^{-1}$)	
			FYM[a]	NPK fertilizers[b]
Broadbalk	Winter wheat	1852–56	2.41	2.52
		1902–11	2.62	2.76
		1955–64	2.97	2.85
		1970–75	5.80	5.47
Hoosfield	Spring barley	1856–61	2.85	2.91
		1902–11	2.96	2.52
		1952–61	3.51	2.50
		1964–67 (PA)[c]	4.60	3.36
		1964–67 (MB)[c]	5.00	5.00
Barnfield	Mangels[d]	1876–94	42.2	46.0
		1941–59	22.3	36.2
	Sugar beet	1946–59	15.6	20.1

[a] FYM 35 $t\,ha^{-1}$ containing on average 225 kg N.
[b] N fertilizer per hectare; barley, 48 kg; wheat, 144 kg; root crops, 96 kg.
[c] PA, Plumage Archer (cultivar) given 48 kg N ha^{-1}; MB, Maris Badger (cultivar) given 96 kg N ha^{-1}.
[d] *Beta vulgaris*.

organic N source), there were initially three plots: a second unmanured plot, one testing FYM, and the third testing the dried, powered soil ash added to the weighed amount of fertilizer for each plot so that there was a constant volume of material to be spread. A further improvement in experimental design was made in 1856 on Barnfield. Three of the original long, narrow plots were combined and assigned to test FYM, P, and K in various combinations, and these strip treatments were crossed at right angles by the same nitrogen treatments tested on Hoosfield. Thus only on Barnfield was fertilizer nitrogen applied with FYM, and this invariably gave the largest yields of mangels. The plot arrangement was not changed on Broadbalk, but there were various changes to the treatments before 1852, and from then little further change until recent times.

The Place of Long-Term Experiments in Agricultural Research and the Role of Lawes and Gilbert

Long-term or continuing experiments are the best practical way of assessing the sustainability of an agricultural system. Comparing different husbandry systems often allows the opportunity to explain failures if these occur and allows modifications to be introduced to maintain, and where appropriate, increase productivity. In this respect it is important to ask why Lawes and Gilbert continued with some of their experiments.

In part, the initial impetus to continue may have come from their various controversies with Liebig. Probably the most important argument was about the source of nitrogen for plants. In the first, 1840 edition, of his book, Liebig wrote: "Cultivated plants receive the same quantity of nitrogen from the atmosphere as trees, shrubs and other wild plants; but this is not sufficient for the purposes of agriculture." In 1843 Liebig changed the last phrase, and the sentence reads: "Cultivated plants receive the same quantity of nitrogen from the atmosphere as trees, shrubs and other wild plants; and this is quite sufficient for the purposes of agriculture." Results from the field experiments at Rothamsted supported the 1840 statement, not that of 1843. Yet for some 20 years Liebig stubbornly refused to retract his 1843 statement, and an acrimonious debate was continued in the pages of various journals.

Other arguments involved Liebig's frequently repeated statement that the cropping and manuring at Rothamsted would rapidly exhaust the soil of minerals and so diminish its productivity. This did not happen. Another source of disagreement was Liebig's assertion that it was only necessary to replace the quantity of minerals removed in the harvested crop. The Rothamsted experiments showed that this was not true; to get good yields, it was essential to apply more minerals than the crop contained. This was probably because at that time the soils contained little

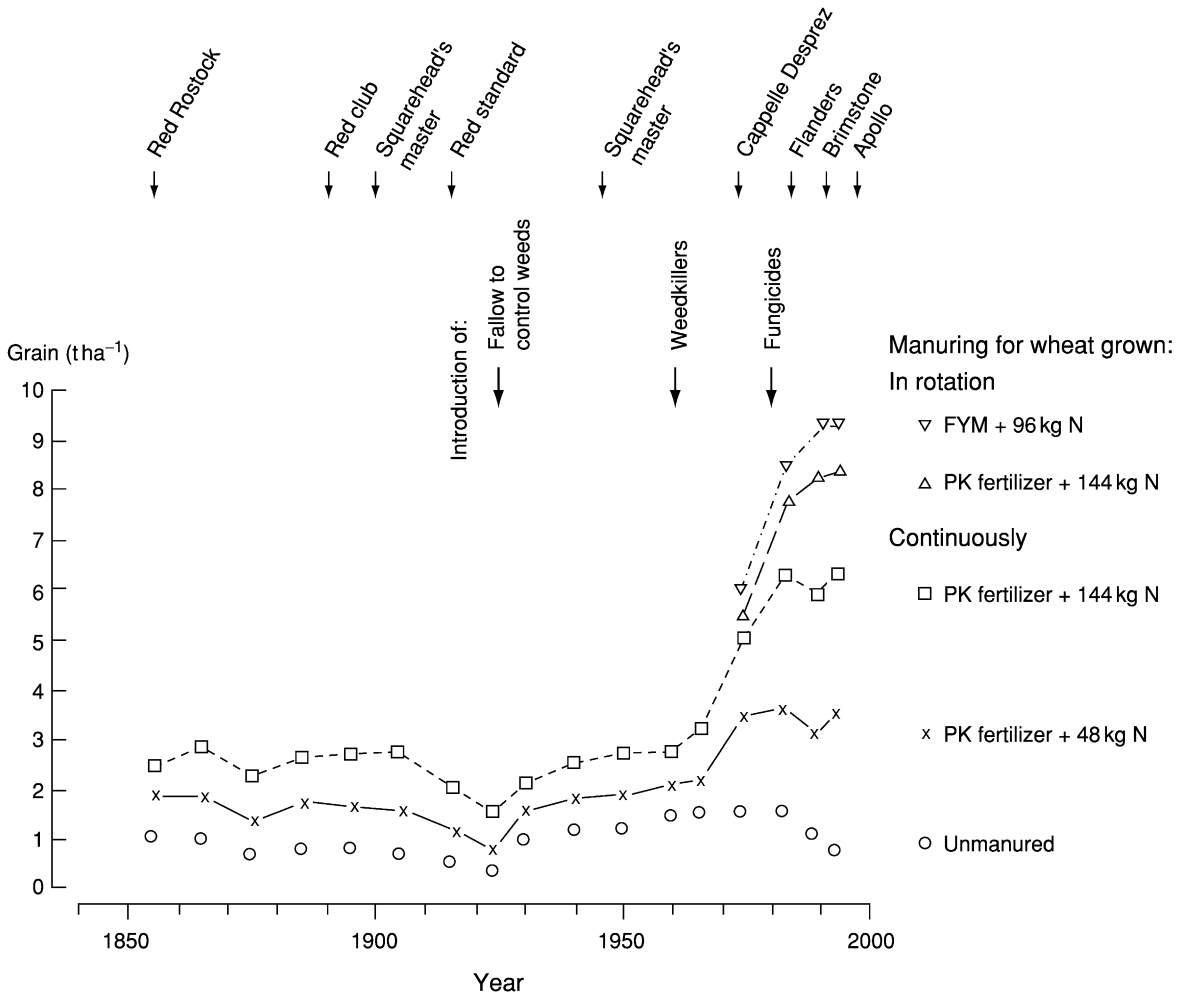


Figure 4 Yields of winter wheat cultivars (tonnes of grain per hectare) from the Broadbalk experiment, Rothamsted Research, Harpenden, UK.

plant-available P and K. Lawes and Gilbert may have needed to go on showing that they were not only right, but also right beyond all reasonable doubt. And so they kept the experiments going, lest Liebig claimed that the experiments were stopped because he was right and Lawes and Gilbert were wrong. Perhaps the benefits to agricultural science from the Rothamsted experiments owe something to Liebig! A lesson for today from these controversies may be that healthy competition, springing from a lack of consensus, may lead to greater and more rapid progress than complacency resulting from consensus!

Lawes was a great educationalist, using the results from the large-scale field experiments to offer sound advice to farmers. On 19 July 1855, in his response to the Earl of Chichester on the occasion of the presentation to him of the newly built Testimonial Laboratory, funded by public subscription, Lawes said "the object of these investigations (at Rothamsted) is not exactly to put money into my pocket but to give you the knowledge by which you may be able to put money into yours." Sir John Lawes did make a lot of money from his fertilizer business, as well as licencing others to make superphosphate using his process. He could readily have become a business tycoon, owning and living on a pleasant country estate, but his intense interest in agriculture decreed otherwise. As the field experiments continued, there was ever-greater opportunity both to 'fine tune' advice given previously and seek new interpretations of the accumulating data. In consequence, Lawes came out strongly on the importance of continuous investigations in agriculture and later, when summarizing the results of the first 50 years, Lawes and Gilbert said that the experiments had become more valuable with time.

Lawes maintained financially the Rothamsted experiments from his own resources and, using a third of the proceeds from the sale of the fertilizer business in 1872, he endowed trustees in 1889 to ensure that the experiments he had started with Gilbert might

continue. In 1935, money raised by public subscription allowed the trustees to purchase the estate when Lawes' descendants offered it for sale. The purchase ensured the security of the land with its world-renowned classical and other long-term field experiments. Today financial support comes from other sources, but agricultural and ecologic sciences, and the interface of both with the wider environment, reap great benefit from Lawes' generosity and Lawes and Gilbert's scientific objectivity and foresight. Rothamsted now has a well-documented database of results from these field experiments, extending over a period of more than 150 years, and a unique archive of crop and soil samples collected over the same period. The archives can be and have been mined for answers to questions Lawes and Gilbert could not have conceived. But besides the past there is the future: six of the experiments started between 1843 and 1856 still continue, all with carefully considered modifications. These experiments should have at least as much to offer in the future as they have yielded in the past.

Further Reading

Dyke GV (1980) *John Lawes of Rothamsted – Pioneer of Science, Farming and Industry.* Harpenden, UK: Hoos Press.

Dyke GV (1991) *John Bennet Lawes: The Record of His Genius*. Chichester, UK: John Wiley.

Johnston AE (1990) Liebig and the Rothamsted experiments. In: Judel GK and Winnewisser M (eds) *Symposium "150 Jahre Agrikulturchemie"*, pp. 37–64. Germany: Justus Liebig-Gesellschaft zu Giessen.

Johnston AE (1991) *Potential Changes in Soil Fertility From Arable Farming Including Organic Systems*, pp. 1–38. Proceedings No. 306. York, UK: International Fertiliser Society.

Johnston AE (1994) The Rothamsted Classical Experiments. In: Leigh RA and Johnston AE (eds) *Long-term Experiments in Agricultural and Ecological Sciences*, pp. 9–37. Wallingford, UK: CAB International.

Johnston AE (1997) The value of long-term experiments in agricultural, ecological and environmental research. *Advances in Agronomy* 59: 291–333.

Johnston AE and Powlson DS (1994) The setting up, conduct and applicability of long-term, continuing field experiments in agricultural research. In: Greenland DJ and Szabolcs I (eds) *Soil Resilience and Sustainable Land Use,* pp. 395–421. Wallingford, UK: CAB International.

Lawes JB (1842) Ammoniacal manure, p. 221. *The Gardeners Chronicle*. London, UK: Bradbury and Evans.

Lawes JB (1843) Ammonia, p. 692. *The Gardeners Chronicle*, London, UK: Bradbury and Evans.

Lawes JB (1882) The future of agricultural field experiments. *Agricultural Students Gazette, New Series* 1: 33–35.

Lawes JB and Gilbert JH (1895) The Rothamsted experiments over fifty years. *Transactions of the Highland and Agricultural Society of Scotland. Fifth Series* 7: 11–354.

Leigh RA and Johnston AE (eds) (1994) *Long-term Experiments in Agricultural and Ecological Sciences*. Wallingford, UK: CAB International.

Liebig J von (1840) *Organic Chemistry in its Application to Agriculture and Physiology.* London, UK: Taylor Walton.

Orwin CS (1938) *The Open Fields*. Oxford, UK: Clarendon Press.

Orwin CS (1949) *A History of English Farming*. London, UK: Thomas Nelson.

Pirie NW (1991) The role of Rothamsted in making nutrition a science. *The Biochemist* 13: 8–13.

Powlson DS and Johnston AE (1994) Long-term field experiments: their importance in understanding sustainable land use. In: Greenland DJ and Szabolcs I (eds) *Soil Resilience and Sustainable Land Use*, pp. 367–394. Wallingford, UK: CAB International.

Thompson FML (1968) The Second Agricultural Revolution 1815–1880. *Economic History Review,* 2nd Series. XXI: 62–77.

LEACHING PROCESSES

B E Clothier and S Green, HortResearch, Palmerston North, New Zealand

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Soil is the thin and biologically active skin of our "Green Planet." Soil, by virtue of its intrinsic fertility, plays a key role in sustaining our primary-production systems. To enhance the productivity of these agricultural systems, irrigation water, fertilizers, and pesticides are often applied to crops growing on soil. Also, soils are sometimes used for the disposal, either intentional or accidental, of certain waste products; and the soil surface is the depository of atmospheric compounds that might be either natural, such as salt, or compounds from human activities, such as sulfur. Since soil lies astride the main pathway along which rainwater leaches to groundwater or surface water bodies, the water quality of these receiving

water reserves is strongly influenced by land management, and the myriad of interlinked water-transport processes and chemical-exchange mechanisms that operate in the soil domain. Development of sustainable land-management strategies, and the design of environmental-protection protocols, demand that the leaching process be well understood.

Leaching encompasses the gamut of biophysical mechanisms and chemical exchange processes that can result in altered chemical concentrations in any pore water that is percolating through soil. There are two aspects to leaching: the quantity of the water that is moving through the soil and destined for subterranean or surface water bodies, and the chemical concentration, or quality, of that pore water. We use the term 'leaching' to encompass two meanings: the deep seepage of water, and the export of chemicals carried by that seepage.

There are controls on the quantity of drainage water quitting the base of the root zone (D, in millimeters). For risk assessment it is necessary to know both the concentration of a chemical in the leachate (C, in moles per cubic meter) and the total chemical load (per unit ground area) that is moving to the bodies that receive water (M, in moles per square meter; or, more commonly, kilograms per hectare). The total leaching of chemical, M (load), is the product of D (quantity) and C (quality). Thus, in the discussion of leaching processes, the transport of water and the chemical processes operating in soil are both considered.

Physically, soil is a porous medium comprising a complex arrangement of three phases: solid, liquid, and air. Much to the chagrin of scientists, convective and diffusive flow through soil does not behave in the uniform and isotropic way assumed by their simple theoretical descriptions. Rather, aggregates, cracks, and macropores serve to create an apparently chaotic flow regime that leads to rapid and far-reaching transport of chemicals. Modeling methods account for this preferential flow and are utilized for risk-assessment purposes.

Leachate Quantity

Globally, an average of some 720 mm of rainfall (RF, in millimeters) falls on soil each year. The many nooks and crannies in the soil provide it with a capillary attractiveness so that there is some tendency for the infiltrating rainwater to be absorbed by the soil, near the surface. The ever-present force of gravity tends to draw complementary water deeper into the profile beyond the grasp of the roots. On balance, approximately 410 mm of the global average rainfall is lost back as vapor to the atmosphere, either by direct free-water evaporation from the soil and directly off plant surfaces (E, in millimeters) or by transpiration through plants (T, in millimeters). Any transpired water must first have been captured from the soil by the plant's roots. The complement of evaporation, namely drainage, D, therefore averages approximately 310 mm per year, because, on an annual timescale, the change in soil-water storage is negligible. Drainage water acts as a vehicle, carrying with it passenger solutes that can leach the soil of chemicals that may compromise the quality of receiving waters. For productive systems, rainfall might be augmented by irrigation (I). Indeed, irrigation comprises approximately 80–85% of the world's consumption of fresh water.

Both for production purposes and environmental protection, it is critical to determine whether the partitioning of root-zone water and its dissolved solutes is either into the plants by uptake or despatch by drainage (D) toward groundwater. Roots are the arbiters of this partitioning, and so they determine water and solute fate (Figure 1). In simple terms, roots penetrate the soil to an effective depth, z_R (millimeters). From within this root zone they extract liquid water and some dissolved solutes. Thus rhizosphere processes critically affect both the quantity and quality of leachates leaving the root zone.

The water balance of the root zone, over a given time interval Δt, can be written to find the drainage at depth z_R as:

$$D = \mathrm{RF} + I - R - E - T - \Delta S \qquad [1]$$

where ΔS is the change, over Δt, of the storage of soil water in the root zone (millimeters). The variation in

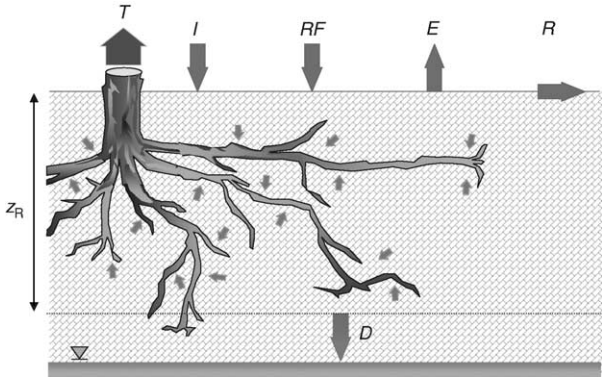


Figure 1 The water balance of the root zone. The effective depth of rooting is z_R. Inputs into the root zone include rainfall (RF), as well as irrigation (I) in some cases. Not all of the applied water might infiltrate, as there could be runoff (R). Evaporative losses are either by direct evaporation from the surface (E), or by transpiration from plants (T). Drainage of water from the base of the root zone toward groundwater, D, represents the quantity of leaching.

the storage of water can be found from the change in the soil-water content profile, $\theta(z)$ (cubic meters per cubic meter), over Δt, integrated from the surface down to the base of the root zone:

$$\Delta S = \int_0^{z_R} \Delta\theta(z)\mathrm{d}z \qquad [2]$$

Thus the quantity of leaching, D, is implicitly linked to the weather (RF, E, and T), as well as to the nature of the soil and the vegetation cover, plus the effective depth and distribution of plant roots.

At a mechanistic level, this interplay between soil-water diffusion and convection, and root-water extraction, can be described by the Darcy–Buckingham equation:

$$\frac{\partial\theta}{\partial t} = \frac{\mathrm{d}\theta}{\mathrm{d}h}\frac{\partial h}{\partial t} = \frac{\partial}{\partial z}\left(K(h)\frac{\partial h}{\partial z}\right) - \frac{\partial K(h)}{\partial z} - U(h(z),t) \qquad [3]$$

where h (in millimeters) is the soil-water pressure head, which is a function of soil's water content, and K (millimeters per second) is the pressure head-dependent, soil-water conductivity function. The first term on the right-hand side describes the diffusive nature of the capillary-induced movement of water through soil. The second term accounts for direct convection in response to gravity. The last term, U (millimeters per second), is the depth-dependent extraction of water by plant roots. Many forms have been proposed for U, and all rely on the theoretical insight provided by Gardner on the uptake of water by roots.

The Darcy–Buckingham equation is the theoretical scheme used to model and predict the partitioning of soil-water fate into either plant uptake (T), or drainage (D).

Leachate Quality

The chemical concentration of the soil's pore-water solution, C, depends on a complex of linked interactions between transient transport processes and chemical exchange mechanisms, as well as production, extraction, and degradation reactions. For any chemical species, the convection–dispersion equation describes the resident concentration at any time and location in the soil as:

$$\frac{\partial\theta C}{\partial t} + \frac{\partial\rho S}{\partial t} = \frac{\partial}{\partial z}\left(\theta D_s\frac{\partial C}{\partial z}\right) - \frac{\partial q_w C}{\partial z} + \sum\Gamma_i \qquad [4]$$

where S is the concentration of chemical that is retained by the soil matrix (moles per kilogram), ρ is the soil's bulk density (kilograms per cubic meter), D_s is the solute's diffusion–dispersion coefficient (millimeters squared per second), q_w is the water flux density (millimeters per second), and Γ_i is the source/sink term for the i reactions that can produce, extract, or degrade the chemical species concerned (moles per cubic meter per second). The structure of this formula (eqn [4]) highlights the inextricable link between water movement (q_w; eqn [3]), and chemical-transport processes.

Also, eqn [4] accounts for the link between the chemical concentration in the soil's solution (C) and the amount of chemical that is adsorbed to the surfaces of the soil's matrix (S). This relationship is called the adsorption isotherm. One commonly used form that describes this exchange process is the Langmuir isotherm:

$$S = S_{max}\frac{mC}{1+mC} \qquad [5]$$

where S_{max} is the maximum adsorption (moles per kilogram) and m is an empirical constant (cubic meters per mole). Often, however, for simplicity, or in an absence of better information, a linear isotherm is assumed:

$$S = K_D C \qquad [6]$$

where K_D is called the distribution coefficient (cubic meters per kilogram).

Equations [3] and [4] provide a mathematical framework that is commonly used to interpret and predict the leaching processes that determine the quantity of leachate (D) and the leachate load of chemical (M).

Preferential Leaching Processes

The macroscopic theory on which eqns [3] and [4] is based requires that the soil be a uniform medium. Unfortunately, soil is a highly structured medium, such that the three phases of solid, liquid, and air are not isotropic. Indeed there is often a high degree of connectedness between these phases. Soil structure is typically characterized by aggregates, fractures, and biogenic macropores (**Figure 2**).

The spatial topology of the macroporous networks that typify structured soil determines the speed and extent of preferential flow processes. As a result of local nonequilibrium processes, there can be more rapid and deeper-than-expected leaching of surface-applied chemicals, relative to that predicted by eqns [3] and [4].

Not all of the soil's pore water would therefore seem to be actively involved in water transport and chemical exchange. Aggregates and macropores can

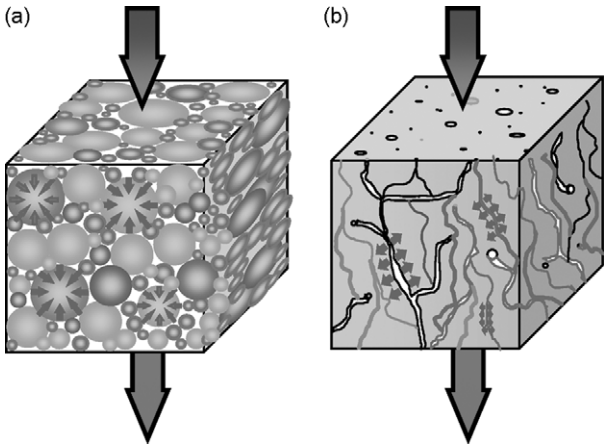


Figure 2 Soil structure is characterized either by aggregates (a) or macropores and fractures (b). Such structure results in rapid and far-reaching transport processes that lead to preferential leaching of applied chemicals. In such structured soils, there is meanwhile likely to be reduced leaching of chemicals produced from within the soil's matrix.

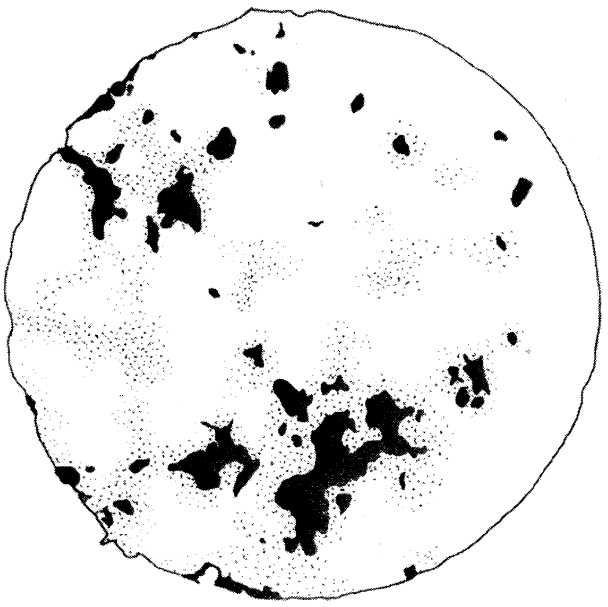


Figure 3 Contrasting patterns of the leaching of two dyes through a 120-mm-diameter core of undisturbed soil. The dark area was stained by a dye tracer contained in saturated flow from a free-water pond at the surface ($h_o = 0$), and the stippled area was from dye leached by unsaturated water applied at the slight negative head of $h_o = -20$ mm. (Reproduced from Scotter DR and Kanchanasut P (1981) Anion movement in a soil under pasture. *Australian Journal of Soil Research* 19: 299–307, with permission of CSIRO Publishing.)

create a wide spectrum of pore-water velocities associated with a solution that is either invading the soil or leaching from it. In an aggregated soil, the more-rapid water-flow paths will be those that are connected throughout the interaggregate domain. So interaggregate water will preferentially wend its way around the aggregates, which tend to possess smaller, albeit connected intra-aggregate pores. These would not seem to be as effective at leaching chemicals from root-zone soil, because the flow velocities are lower and the connected path-lengths shorter.

Alternative, nonmechanistic schemes have been developed for modeling the preferential leaching of chemicals through structured soil. Process-based transport schemes are difficult to parameterize, and so transfer-function models have been considered, because they are less data-intensive and they can support either a deterministic or a stochastic framework.

Pedogenic and biogenic processes can impart upon soil a macroporous structure characterized by a longitudinal connectedness. The enhanced mobility of flow down these variously sized macropores, as predicted by Poiseuille's law of flow through a pipe, is abetted by macropore connectivity to produce preferential transport of the invading solution or enhanced leaching of some part of the resident solution.

In an illustrative study of preferential leaching in the laboratory with undisturbed cores of soil that was growing pasture, two dye tracers were used to mimic leaching. Methylene blue dye, applied in free water ponded on the soil's surface ($h_o = 0$), was found to stain only a small volumetric fraction of the soil's volume (Figure 3). When water containing rhodamine-B dye was applied to the soil in water at the slight unsaturated surface potential of -20 mm, the pattern of leaching was found to involve a larger volumetric fraction of the soil's porosity. Obviously, leaching in structured soils can, depending on hydraulic conditions, involve differing volumes of the soil's wetted porosity.

In one of the most popular mechanistic models used currently to model the preferential leaching that results from the presence of either aggregates or macropores, Occam's razor is simply used to cut the soil's volumetric water content θ (cubic meters per cubic meter) in two. This binary assumption does not account for changes in the leaching processes as a function of the soil's degree of unsaturation (cf. Figure 3), but it does better-represent the preferential leaching processes that prevail in field soils. The great simplicity of this model is that one fraction of the soil's water, θ_{im} (cubic meters per cubic meter), is considered, for leaching purposes, to be stagnant. The complementary portion of the soil's water, θ_m (cubic meters per cubic meter), is deemed to be mobile, as might be imagined of the water that is between the aggregates or flowing in the macropores.

A two-domain approximation for the leaching of a conservative solute under steady flow at pore-water velocity ν_m, (millimeters per second) is given by:

$$\theta_m \frac{\partial C_m}{\partial t} + \theta_{im} \frac{\partial C_{im}}{\partial t} = \theta_m D_s \frac{\partial^2 C_m}{\partial z^2} - \nu_m \theta_m \frac{\partial C_m}{\partial z} \quad [7]$$

where C_m and C_{im} are the concentrations of inert chemical in the mobile and immobile domains, and $\nu_m = q_w/\theta_m$. It is assumed that molecular diffusion in the direction of leaching in the mobile domain is negligible, and that the only mechanism to transfer solute between domains is, for analytical convenience, limited to a diffusion-like process that can be described by:

$$\theta_{im} \frac{\partial C_{im}}{\partial t} = \alpha(C_m - C_{im}) \quad [8]$$

where α (per second) is a first-order mass transfer coefficient.

In order to measure θ_{im} independently, a tracer-filled disc permeameter is used to allow leaching into the soil of a solution at a concentration C_m. Ignoring any effects that might accrue from interdomain diffusion (eqn [8]), the flux concentration in the mobile domain in the soil just under the disc is C_m. Thus by removing soil samples from under the disc and determining the average resident concentration of solute C^*, the mobile fraction can be found using the partitioning formula:

$$\theta C^* = \theta_m C_m + \theta_{im} C_{im} \quad [9]$$

If a tracer not initially present in the soil is used and if diffusive exchange (eqn [8]) is considered sufficiently weak, then C_{im} remains essentially zero over the time scale of the experiment. So:

$$\theta_m = \theta\left(\frac{C^*}{C_m}\right) \quad [10]$$

via which θ_m can be obtained independently from measurements immediately under the disc of just C^* and θ. The results for a silt loam suggest that, for leaching, θ_m/θ is approximately 0.7–0.75 at a pressure head of $h_o = -40$ mm. Only about three-quarters of the soil's water-filled porosity appears to be actively involved in leaching. This qualitatively corroborates the cross-sectional staining areas in **Figure 3**, where it can be presumed that approximately 50% of the area is pore space, and of that only about 75% is stained by dye.

A novel aspect has been added to this technique through the use of the sequential invasion of *i* inert tracers into the same soil under the permeameter. They show how differences in the respective resident concentrations of C_i^* reflect the rate of diffusion, α, into the immobile domain. The resident concentration in the soil of any tracer *i* reflects the diffusion that has taken place during the time t_i it has been present in the mobile domain. From a regression of the measurements using:

$$\ln\left[1 - \frac{C_i^*}{C_{m,i}}\right] = \frac{-\alpha}{\theta_{im}} t_i + \ln\left[\frac{\theta_{im}}{\theta}\right] \quad [11]$$

both θ_m and α can be deduced. From many experiments on various soils, α has been found to range between approximately 5×10^{-7} and 5×10^{-6} s^{-1}.

For aggregated soils, there is a numerical model, the so-called MACRO, which allows convection between the interaggregate mobile zone and the intra-aggregate pores. This convection is considered gravity-free and is driven simply by water content differences between the respective domains. Solute transfer between these domains is allowed by diffusion, as well as by direct convection with the transferring water. By simulation, "... at short times, macropore flow increases leaching of non-reactive solute. At longer times, leaching is reduced by macropore flow [and it is] critical to account for the effects of macropore flow in clay soils on the leaching of sorbing, degrading compounds."

Modeling and Risk Assessment

So, despite the complicated mathematical forms of the coupled flow and transport eqns [3] and [4], and even [7] with [8], modern computers and efficient software schemes are easily able to solve them to predict the impact of land management on leaching processes. The results from models can then be used in a risk-assessment framework to help us better understand the functioning of environmental systems and to guide decisions for sustainable land management.

A simple, lumped-parameter model of the leaching processes under short pasture that is grazed and a pine forest provides insight into the processes of leaching in terms of the controls on leachate quantity (D), and the contrast between the chemical concentration in the leachate (C) and total leachate load (M). The simulation considers pasture and pines growing on soil typical of the Nelson region of New Zealand. Actual weather data from 1972 to 1998 are used. The maximum rooting depth, z_R, of 0.5 m is assumed for pasture, whereas the pine trees are considered able to root effectively to any prescribed depth. Simulations using a lumped-parameter model have been carried out for $z_R = 0.5$, 1, 1.5, 2, 2.5, and 3 m (**Figure 4**).

When pasture and pines are grown on soils of a similar shallow depth of 0.5 m, the evapotranspiration (ET) for the pines is lower due to sparseness of the canopy in the early stages of tree growth and again following the thinnings. None the less, due to

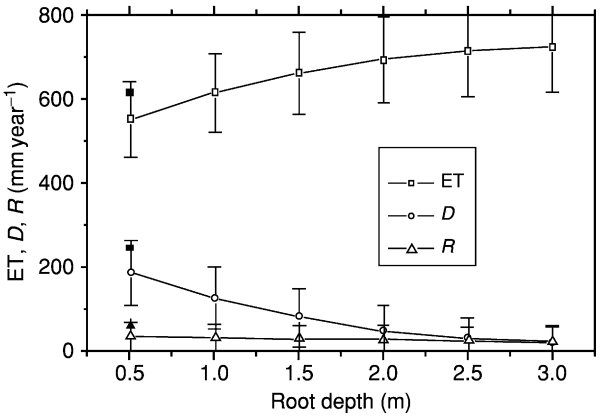


Figure 4 The influence of the depth of the root zone on the amount of leaching (D) 'lost' by pine trees (open symbols) in comparison with pasture (closed symbols). Pasture is assumed to root to a maximum depth of z_R of 0.5 m, whereas the pines can extend their roots beyond 3.0 m if soil depth permits. The mean annual rainfall over the 27 years was 974 mm. Reproduced with permission from Clothier BE, *et al. Modelling Rootzone Processes: a Tool for Risk Assessment*, Water 2000, NZWWA.

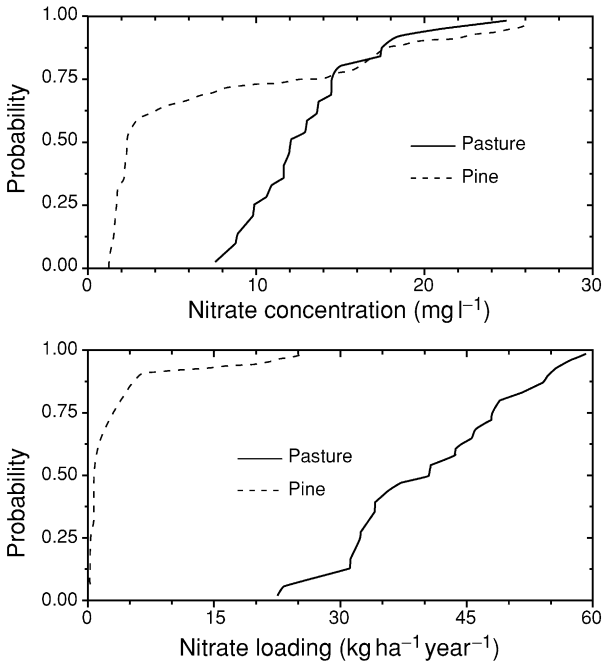


Figure 5 The predicted cumulative probability functions (CPF) of: (a) the concentration of nitrate in the leachate under pasture, and that under pine trees (soil depth 2 m; but the rooting depth for pasture is $z_R = 0.5$ m and pines, $z_R = 2.0$ m). This CPF relates to the annualized value that was derived by running the model in daily time-steps over 27 years, and summing the results to form annual totals; and (b) the total loading of nitrate on the receiving water bodies. Reproduced with permission from Clothier BE, *et al. Modelling Rootzone Processes: a Tool for Risk Assessment*, Water 2000, NZWWA.

interception losses by the trees reducing the effective rainfall, there is still a lower leachate quantity (D) under pines for such shallow soils. As the soil depth increases, the rise in ET constricts leachate quantity (D) under the pines. For a shallow soil of $z_R = 0.5$ m, the annual leachate load under pines is approximately 195 mm, or 20% of RF. When the soil is deep, for example $z_R = 3$ m, $D = 23$ mm or just 2.5% of RF. Thus the amount of water leaching through soil is critically dependent on the nature of the vegetation cover and soil's properties, and in particular the depth of effective rooting z_R.

Pasture production and pine tree growth are modeled through a nitrogen-balance scheme. Consequently, this model can be used to assess the risk posed to groundwater quality by the leaching of nitrate from the root zone of these crops. Figure 5 shows the average annual nitrate-leaching concentrations, C, and the total annual loadings, M, for the pasture ($z_R = 0.5$ m) and the pines with $z_R = 2$ m.

There are two distinctive characteristics: the difference between the two land uses, and the distinctly skewed and non-Gaussian nature of the cumulative probability functions (CPFs). The average concentration (probability 0.5) is much lower for pines (2.3 mg l^{-1}) than for pasture (12.1 mg l^{-1}), since a lot more of the nitrogen ends up stored in the plant tissue. Through repeated grazings, and additions of dung and urine, there are greater inputs of N into the pastoral system. The CPF for the trees is the most non-Gaussian, and this is due to the infrequent high loadings of nitrate that occur when, occasionally, a wet year follows a dry year and there is a surge in nitrate leaching. The simple, lumped-parameter model provides the amount of leaching (D) so that this can be multiplied by the nitrate concentration (C) to obtain a loading of nitrate of the receiving waters (M). The model results in Figure 5 show that the loading on groundwater from grazed pasture is predicted to be 41 kg · N ha^{-1} per year. However for the pine trees growing on this 2-m-deep soil, the low concentrations of nitrate combined with the low amount of drainage mean that there is simulated to be a very low leachate loading of just 0.6 kg · N ha^{-1} per year. So whereas pine trees are predicted to have a negative impact of the quantity of leaching (D), they are considered, by virtue of a lower total loading (M), to have a beneficial role in maintaining the quality of leachates moving to receiving bodies. The interplay between land management and soil character determines the quantity and quality of leaching.

A more-sophisticated modeling framework, the Soil–Plant–Atmosphere System Model (SPASMO), predicts the risks posed by the leaching of the active ingredient (AI) of three pesticides applied, in a simulation, to a vineyard on Twyford silt loam located above an unconfined aquifer in Hawkes Bay, New Zealand. This framework can account, in a simple

way, for the mobile–immobile partitioning of fast- and slow-flow pathways of leaching through soil. The SPASMO simulations consider actively growing grapevines and a grassed interrow. The calculations are run on daily time-steps using a 27-year sequence of weather data (1972–99) recorded from the local airport at Napier. The soil's physical and hydraulic properties are taken from the New Zealand Soils Database. Three herbicides, simazine, linuron, and amitrole, are 'applied' at the label rate on 15 September each year. This modeling of leaching predicts if and how often the soil-solution concentration at 3 m might exceed the maximum allowable value (MAV) of 2 ppb of active ingredient.

Simazine is quite mobile because of its low exchange with the soil's organic carbon (eqn 6). Also it is persistent, with a long degradation half-life of $t_{1/2} = 50$ days. Linuron is less mobile because of greater exchange with the soil's organic matter, and it is less persistent, with $t_{1/2} = 33$ days. Amitrole is also less mobile because of low exchange, but it is much less persistent, with $t_{1/2} = 3$ days. Therefore, concerning the likelihood of leaching to groundwater, the risks are expected to be ranked simazine > linuron > amitrole. This expectation is borne out in the simulated leaching results, presented for a depth z of 3 m (Figure 6). The average simazine concentration in the soil solution leaching through Twyford silt loam, over a 30-year period, is predicted to be just 0.03 $\mu g\,l^{-1}$, although 10% of the time it will exceed 0.12 $\mu g\,l^{-1}$, or 6% of MAV. Simazine should be used with great caution, for it poses a greater leaching risk than either of the other two pesticides.

The risk of pesticide leaching also depends on the depth to groundwater. Figure 7 shows the model results for linuron, as a function of depth. Here the data are presented as a cumulative probability function of exceedence. During the time taken to leach deeper into the soil profile, there is degradation of the pesticide. For linuron, this means that there is predicted to be, on average (50% exceedence), about an order-of-magnitude drop in the concentration of active ingredient with every meter of soil depth. Even in the worst case, $z = 1$ m, the concentration of linuron will still be approximately an order of magnitude less than the MAV. None the less, this analysis shows that shallow, unconfined groundwaters are at the greatest risk of contamination from the leaching of surface-applied chemicals.

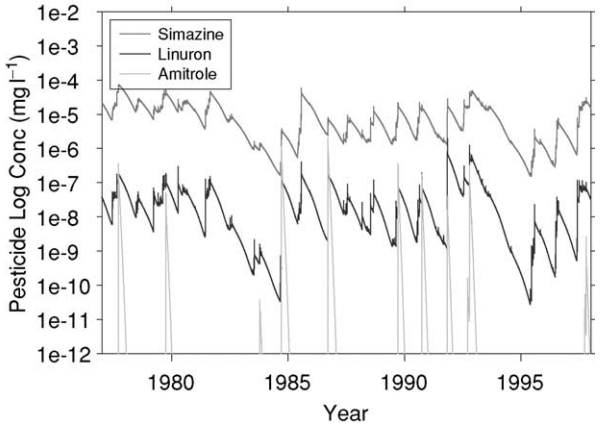


Figure 6 Simulated herbicide concentrations in the soil solution at 3 m depth under grapevines growing on a Twyford silt loam. Each year, over 30 years, the herbicides were considered to be applied at the label rate, to 100% of the soil surface, in mid-September.

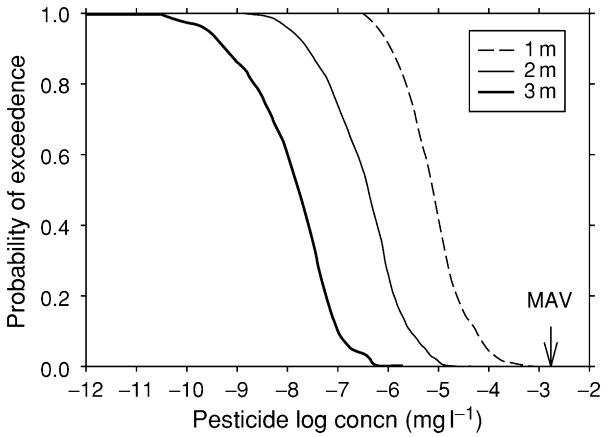


Figure 7 The predicted probability distribution for the concentration of linuron in leachate at three depths under grapevines growing on a Twyford silt loam. Linuron was considered to be applied once each year, in mid-September, at the label rate of 1.0 kg ha^{-1} of active ingredient. MAV, maximum allowable value.

Summary

The leaching of chemicals through soil by percolating pore-water is the result of a complex of interconnected biophysical and biochemical processes. Firstly, biophysical mechanisms determine the quantity of water that quits the root zone, destined for either groundwater or surface-receiving waters. The prevailing weather, the nature of the vegetative cover, and the properties of the soil conspire to determine the amount of leachate. Secondly, the biochemical processes of exchange, production, and degradation control the concentration of chemical in the leaching solution. Risk assessment of leaching needs to consider both the chemical concentration of the leachate solution and the total loading of chemical, which is the product of leachate quantity and leachate quality.

See also: **Diffusion**; **Evapotranspiration**; **Infiltration**; **Irrigation:** Environmental Effects; **Macropores and Macropore Flow, Kinematic Wave Approach**; **Pesticides**; **Plant–Soil–Water Relations**; **Pollution:** Groundwater; **Rhizosphere**; **Root Architecture and Growth**; **Soil–Plant–Atmosphere Continuum**; **Solute Transport**; **Sustainable Soil and Land Management**;

Vadose Zone: Hydrologic Processes; **Water Potential**; **Water-Use Efficiency**

Further Reading

Clothier BE and Green SR (1997) Roots: the big movers of water and chemical in soil. *Soil Science* 162: 534–543.

Clothier BE, Kirkham MB, and MacLean JE (1992) *In situ* measurement of the effective transport volume for solute moving through soil. *Soil Science Society of America Journal* 56: 733–736.

Clothier BE, Green SR, Roygard JFK, Vogeler I, and Duwig C (2000) *Modeling Rootzone Processes: a Tool for Environmental Risk Assessment*. Proceedings of the *Water 2000 Conference*, NZ Water and Wastes Association, Auckland, 19–23 March.

Feddes RA, Kowalik PJ, Malinka KK, and Zradny H (1976) Simulation of field water uptake by plants using a soil water dependent root extraction function. *Journal of Hydrology* 31: 13–26.

Gardner WR (1960) Dynamic aspects of water availability to plants. *Soil Science* 89: 63–73.

Green SR, Clothier BE, Mills TM, and Millar A (1999) Risk assessment of the irrigation requirements of field crops in a maritime climate. *Journal of Crop Production* 2: 353–377.

Green SR, Clothier BE, Caspari HW, and Neal SM (2002) Rootzone processes, tree water use and the equitable allocation of water to olives. In: Raats PAC, Smiles DE, and Warrick AW (eds) *Heat and Mass Transfer in the Natural Environment*. American Geophysical Union Monograph. Washington, DC: American Geophysical Union Books.

Jarvis NJ (1994) *The MACRO Model (version 3.1) – Technical Description and Sample Simulations*. Reports & Dissertations 19. Uppsala: Swedish University of Agricultural Science.

Jaynes DB and Horton R (1997) Field parameterization of the mobile/immobile domain model. In: Selim M and Ma L (eds) *Physical Nonequilibrium in Soils: Modeling and Application*. Ann Arbor, MI: Ann Arbor Press.

Jury WA and Roth K (1990) *Transfer functions and solute movement through soil: theory and applications*. Basel: Birkhauser.

Scotter DR and Kanchanasut P (1981) Anion movement in a soil under pasture. *Australian Journal of Soil Research* 19: 299–307.

Van Genuchten MT and Wierenga PJ (1976) Mass transfer studies in sorbing porous media. I. Analytical solutions. *Soil Science Society of America Journal* 40: 473–480.

LIEBIG, JUSTUS VON

R R van der Ploeg, University of Hannover, Hannover, Germany
W Böhm, Georg-August-University, Göttingen, Germany
M B Kirkham, Kansas State University, Manhattan, KS, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Justus von Liebig (1803–73) was one of the most productive, significant, and versatile German chemists of all times. Because his contributions to the various fields of organic, analytical, agricultural, physiological, industrial, pharmaceutical, and educational chemistry are so basic and numerous, his influence is recognized even today. He is particularly celebrated for his chemicophysiological work on the metabolism of organisms and on the chemistry of agricultural soils. He is often considered as one of the founders, if not as the founder, of agricultural chemistry and as the father of modern agronomy. He is one of the most decorated scientists of all times and continues to be commemorated throughout the world. In 1845, for example, at the age of only 42, he was ennobled (which explains the 'von' in his name after 1845). The front of the beautiful Main Library of Iowa State University at Ames, Iowa, USA, shows the name of Liebig amidst those of other all-time celebrities of science and literature such as Newton, Pasteur, Faraday, Darwin, Shakespeare, and Emerson. The International Union of Soil Sciences (IUSS) in 2002 decided at its Congress in Bangkok to call one of its major awards after Liebig. In Germany, in almost every sizeable town, there is a street that is named after Liebig. The Justus-Liebig-Universität in Giessen, where he was a professor from 1824 to 1852, has borne his name since 1946.

Liebig's engagement as an agricultural chemist started in 1840, when he published his classic work *Organic Chemistry in its Application to Agriculture and Physiology*. Based on an analysis of the literature, Liebig rejected in this publication the so-called humus theory, with which hitherto the nutrition of plants was explained. Instead, he presented the theory on the mineral nutrition of plants, which implies that plants need, besides CO_2 from the air, mineral salts from the soil to develop. At the same time, Liebig

recommended the use of mineral fertilizers in crop production.

In recent years, however, it has been shown that the agronomist and chemist Carl Sprengel (1787–1859) was the first to disprove the humus theory (1826), to formulate the theory of the mineral nutrition of plants (1828), and to recommend the use of mineral fertilizers (1828). Justus von Liebig was familiar with Sprengel's work, but did not cite his countryman's pioneering achievements and thus has to be blamed for plagiarism. Nevertheless, as an indefatigable leader of agronomy in the struggle for the acceptance of the theory of mineral nutrition of plants and the use of mineral fertilizers in crop production, as a reformer of agricultural education, and as a founder of the experimental and research station system, Liebig deserves to be commemorated as one of the most outstanding scholars of agriculture.

Life History

Justus Freiherr (Baron) von Liebig was born on 12 May 1803 in Darmstadt (Hesse) as Johann Justus Liebig, and he died on 18 April 1873 in Munich. He was the second son of Johann Georg Liebich (Liebig) and his wife Marie Caroline Fuchs. His father owned a drugstore in Darmstadt and sold paints and varnishes that he produced himself. The young Liebig gained his first chemical experience in his father's laboratory. At the age of 14, Justus Liebig dropped out of high school and worked for almost 2 years with his father. During this time, he started an apprenticeship in a pharmacy, which he did not complete. Although he had not finished high school, he registered at the age of only 16 as a university student, first at Bonn (Figure 1) and, from 1821, at Erlangen. There he worked, as he had in Bonn, under the supervision of the chemist Karl Wilhelm Gottlieb Kastner. In 1822 Liebig left for Paris to become a coworker of the famous chemist L.J. Gay-Lussac. In 1823 he earned, in his absence (because of political turmoils in Germany), his doctorate from the University of Erlangen with a thesis, 'The relation between mineral chemistry and plant chemistry.' One year later, at the age of only 21, Liebig accepted a professorship at the University of Giessen (Figure 1), and in 1825 he became a full professor of chemistry at the same university. In 1826 he married Henriette Moldenhauer. The Liebigs had five children, two boys and three girls.

At Giessen, Liebig established a laboratory to teach chemistry through research and experimentation. Within a few years, this unique laboratory developed into the most significant teaching and research facility in Germany. Because of his exceptional talent as a teacher and his epoch-making research, Liebig soon became a world-famous chemist. He attracted students from all over the world to Giessen. Among his German students were such chemists as Hoffmann, Kekulé, Pettenkofer, Fresenius, and Erlenmeyer. One of his British students was Joseph Henry Gilbert,

Figure 1 Germany, showing present and previous borders.

later coworker of John Bennet Lawes at Rothamsted. Of Liebig's American students, E.N. Horsford, J.A. Porter, and S.W. Johnson deserve special mention. After completing their studies in Liebig's laboratory at Giessen, Eben N. Horsford accepted a professorship at the Lawrence Scientific School at Harvard, and John A. Porter became a Professor of Analytical and Agricultural Chemistry at Yale University. Samuel W. Johnson also became a Professor of Agricultural Chemistry at Yale and, in addition, established the Connecticut Agricultural Experiment Station, the first institution of this kind in the USA. Later Johnson's influence did much to secure the passage of the so-called Hatch Act in 1887, which enabled the establishment of at least one agricultural experiment station in every state of the USA.

After 28 years at the University of Giessen, in 1852 Liebig accepted an invitation from King Maximilian II of Bavaria to go to the University of Munich. Here a new research institute was established, completely in accordance with Liebig's ideas and aspirations. In contrast to Giessen, this research laboratory was exclusively for Liebig and his assistants, not students. Liebig's picture of 1853 is shown in **Figure 2.** In 1859 Liebig was elected as President of the Royal Academy of Sciences, in Munich, an honorary position he held until his death in 1873. One of his sons, Hermann von Liebig (1831–94), became a respected agricultural chemist himself. Hermann worked for many years at the Agricultural Experiment and Research Station in Munich and published significant work. The other son, Georg von Liebig (1827–1903), who was a medical clinician, published after his father's death hitherto unpublished material, such as speeches by his father and correspondence with renowned colleagues.

Figure 2 The chemist Justus von Liebig (1803–73), who entered the field of agronomy in 1840 with his classic *Agrikulturchemie*. Engraved by A. Weger (Leipzig) from a photograph taken in 1853.

The Work of Liebig

From 1822 on, Liebig published extensively, at first mainly on topics of inorganic chemistry, and, after 1830, increasingly on organic chemistry and pharmacy. By the end of the 1830s, Liebig was a distinguished and world-famous scientist, with nearly 300 much-regarded, high-impact publications in the leading journals of chemistry and pharmacy. It was at this point that Liebig entered the field of agriculture almost by accident. In 1837, during a visit to Britain, Liebig was asked by the President of the British Association for the Advancement of Science to give a report on the state of organic chemistry and organic analysis. Liebig wrote this report, which was published in French in three parts in 1840 in Paris. The introduction to this report appeared in the same year as a book in German under the title *Die organische Chemie in ihrer Anwendung auf Agricultur und Physiologie* (*Organic Chemistry in its Application to Agriculture and Physiology*).

In this book (with the short title *Agrikulturchemie*), Liebig reviewed and analyzed current knowledge of plant physiology and plant nutrition. He especially criticized the so-called humus theory, which stated that plants need humus to develop. In studying the literature, Liebig concluded that plants did not need humus to grow, but they required CO_2 from the air and, in particular, mineral salts from the soil. For that reason Liebig recommended that farmers needed to replace all the mineral nutrients that were removed from their fields at harvest, to maintain the soil's fertility.

Because of the boldness with which Liebig demanded a change in fertilization, his book received much attention, from both farmers and agronomists. In a short time, five further editions appeared. From the fifth edition on, the title was slightly modified (*Die Chemie in ihrer Anwendung auf Agricultur und Physiologie*). The doctrines posed by Liebig gained ardent consent as well as sharp rejection. His claim to abandon organic fertilizing was questioned; also, the mineral fertilizers that he recommended were looked at with skepticism. In particular, his 'patent' fertilizer did not show the predicted yield increases. It probably failed because Liebig used highly insoluble mineral compounds; he assumed that soluble compounds would be leached from the soil too easily.

Another matter of skepticism, which hindered a quick breakthrough of the theory of the mineral nutrition of plants, was the unsolved question of the need to fertilize field crops with nitrogen. In the fifth edition of his *Agrikulturchemie* of 1843, Liebig said that nitrogen fertilization was not necessary, because crops would receive ample ammonia from the atmosphere. However, this statement contradicted farm experience; most crops had increased yields when fertilized with nitrogen. The strongest opposition to Liebig's view came from England, where John Bennet Lawes and his coworker Joseph Henry Gilbert had been carrying out field experiments since 1843 at Rothamsted. Their main objective was to disprove Liebig's theory of nitrogen fertilization.

Because his patent fertilizer failed, for the next 10 years Liebig did not take part in public discussions concerning unsolved problems of agricultural chemistry. In 1855 he reentered the ongoing debate. In a basic contribution in the *Zeitschrift für deutsche Landwirthe* (*Journal for German Farmers*), Liebig sharply criticized the work of Lawes and Gilbert and simultaneously formulated, in 50 precise theses, his own doctrines of plant nutrition. This work was published, in somewhat modified form, as a book with the title *Die Grundsätze der Agrikultur-Chemie mit Rücksicht auf die in England angestellten Untersuchungen*. In addition, it was published in English, also in 1855, under the title *The Relations of Chemistry to Agriculture and the Agricultural Experiments of Mr J.B Lawes*. With this publication Liebig tried to deflect the remaining resistance against the theory of the mineral nutrition of plants. He also reluctantly slightly changed his views on the usefulness of nitrogen fertilization of field crops.

The second edition of this book (1855) also contains Liebig's first exact formulation of the so-called Law of the Minimum. It states that the yield of a crop in a field (of a given condition and composition) is directly related to a particular nutrient, which is indispensable for complete development of the plant and which is available in the soil (in proper form and quality) in the smallest quantity (in 'minimo'). For many years it was believed that the Law of the Minimum was first stated by Liebig. A study of the literature has shown, however, that Liebig's countryman Carl Sprengel had conceived this law as early as 1828. Today it is generally accepted that Liebig derived the law, as well as much of his theory on the mineral nutrition of plants, from the work of Sprengel, which was published in the 1820s and 1830s.

Liebig's *Agrikulturchemie* initiated a public dispute of many years, because of his thesis that field crops did not need to be fertilized with nitrogen. In his book of 1856 *Über Theorie und Praxis in der Landwirthschaft* (*About Theory and Practice in Agriculture*), published in Braunschweig, Liebig again sharply attacked his opponents (the 'nitrogenists'). Main adversaries in Germany were Carl Sprengel, Julius Stöckhardt, and Emil Wolff, and, in England, Lawes and Gilbert. Carl Sprengel discussed the use of mineral nitrogen as a fertilizer as early as 1828.

During this period of ample and vocal agricultural controversies, Liebig published in 1858 his paper *Über das Verhalten der Ackerkrume zu den in Wasser löslichen Nahrungsstoffen der Pflanzen* (*About the Behavior of the Mold with Respect to the Water-Soluble Nutrients of Plants*). In this publication Liebig discussed the (cation) exchange capacity of arable soils, as observed around 1850 by John Thomas Way. From Way's work Liebig concluded that effective mineral fertilizers should be partly water-soluble. Based on this new insight, he was now able to explain the failure of his earlier patent fertilizer, which was highly insoluble. In 1859, with this new knowledge, he tried, by means of another book called *Naturwissenschaftliche Briefe über die moderne Landwirthschaft* (*Letters of Natural Sciences About Modern Agriculture*), to convince land users to practice farming solely on the basis of information derived from natural sciences.

Motivated by accumulated knowledge, Liebig revised in 1862 his *Agrikulturchemie*. The book now consisted of two volumes with different subtitles: volume 1 was called *Der chemische Prozess der Ernährung der Vegetabilien*, and volume 2, *Die Naturgesetze des Feldbaus* (*The Chemical Process of the Nutrition of Plants,* and *The Natural Laws of Crop Production*). By comparing previous editions, it is apparent that Liebig stuck to his earlier agricultural chemical principles. However, he did add new pieces of knowledge garnered since the appearance of his preceding editions. He also accepted some farming practices, gained by long-term experience rather than by scientific reasoning or modern agricultural experimentation. He made a small concession to his opponents by admitting that, under certain conditions, nitrogen fertilization might be necessary.

Liebig's colleagues generally gave positive evaluations of the seventh edition of his *Agrikulturchemie*. In a review for the *Agronomische Zeitung* of 1865, Wilhelm Hamm wrote, for example, that "with this book we have received an agricultural codex for all times." This appraisal seemed to be correct, because the seventh edition indeed listed almost all the facts of modern plant nutrition and fertilizer science. The seventh edition of *Agrikulturchemie* was not the last one. In 1865 the eighth edition appeared, and, in 1876, three years after Liebig's death, Philipp Zöller published a ninth edition. *Agrikulturchemie*, a classic

work in the history of science, was translated into many foreign languages, especially in the early years after its first appearance. By 1848, the book had passed through more than 20 editions and reprints: six in Germany; five in England; three in the USA; two in France; two in Italy; and one each in Denmark, the Netherlands, Sweden, Poland, and Russia.

As an agricultural chemist, Liebig also dealt with questions regarding the nutrition of animals. His most significant publication in this area is the book *Die organische Chemie in ihrer Anwendung auf Physiologie und Pathologie* (*Organic Chemistry in Its Application to Physiology and Pathology*), which was published in Braunschweig in 1842, with a second edition in 1843 and a third one in 1847.

Liebig also considered problems of agricultural chemistry in his *Chemische Briefe* (*Chemical Letters*), published in Heidelberg in 1844. This book had a number of editions; the sixth edition was published in both Heidelberg and Leipzig. Also Liebig had contributions dealing with agricultural chemistry in the journal *Annalen der Chemie und Pharmacie*, of which he was an editor. Additionally, Liebig published frequently in agricultural journals about fertilizers and plant nutrition. Of much interest for the history of agronomic sciences is the correspondence between Justus von Liebig and Theodor Reuning, a highly influential agricultural administrator in Saxony (Germany). This correspondence was published by Reinhold Echtermeyer and Georg von Liebig in 1884 in Dresden and was titled *Briefwechsel zwischen Justus von Liebig und Theodor Reuning über landwirthschaftliche Fragen aus den Jahren 1854–1873*.

Liebig's merits as a reformer of agricultural education are important. From 1860 discussions had been taking place in Germany as to how to improve agricultural education. Up to then, advanced education in agriculture was almost exclusively offered by isolated agricultural academies. In two public addresses during meetings of the Royal Academy of Sciences in 1861 in Munich, Liebig called for the closing of existing agricultural academies and, instead, for the establishment of educational programs in agriculture at universities. He felt that only at universities was it possible to teach adequately the basic sciences that agriculture needed. His addresses were published by his son Georg von Liebig and M. Carrière in 1874, in Leipzig and Heidelberg, as *Justus von Liebig: Reden und Abhandlungen*. Liebig's efforts to implement the study of agriculture at universities resulted in the creation of agricultural departments at a number of universities and colleges.

Last, but not least, Liebig's influence on the founding of agricultural extension programs in Germany and abroad needs to be mentioned. With much success, he convinced governmental authorities at home and in other European countries of the necessity of providing funds and facilities for the establishment of agricultural experiment stations. The first state agricultural experiment station in Germany was established at Möckern (now a suburb of Leipzig) in Saxony in 1851–52. Twenty-five years later there were 74 such stations in Germany, 16 in Austria, and 10 in Italy. At each station a chemist with one or more assistants was engaged in the analysis of food, feedstuff, fertilizers, and other agricultural products, and in conducting experiments on the fertilization of crops and the feeding of animals. Liebig and his school based agriculture so overwhelmingly on chemistry that, well after the beginning of the twentieth century, agricultural chemistry was considered synonymous with agricultural science.

Liebig and Sprengel

A discussion about the significance of Justus von Liebig as an agricultural chemist is not complete without a reference to Liebig's countryman Carl Sprengel. Philipp Carl Sprengel was born in 1787 in Schillerslage (today a suburb of Burgdorf) near Hannover in Northern Germany (**Figure 1**), and he died in 1859 in Regenwalde, now Resko (Poland). As a student of Albrecht Daniel Thaer (the founder of modern agriculture) and as a farm manager for many years, he was one of the most experienced and knowledgeable agronomists in Germany, when he registered in 1821, at the age of 34, as a student of natural sciences at the University of Göttingen. After earning his doctorate in 1823 (Justus von Liebig was awarded his doctorate in the same year), Sprengel worked eight more years at the University of Göttingen, before he transferred to Braunschweig in 1831 to become a college professor. From there he moved to Regenwalde (**Figure 1**), where he founded his private academy of agriculture. Sprengel's undated picture is shown in **Figure 3.**

Sprengel carried out pioneering research in Göttingen between 1823 and 1831. His first publications dealt with the humus theory and the role of soil organic matter (humus) as the only source of plant nutrients. He analyzed the water-soluble constituents in the humus extract and found a variety of salts such as alkali nitrates, sulfates, chlorides, and phosphates. Because of the salts that he measured in the ash of plants, Sprengel concluded that these soluble salts in the humus extract were the real plant nutrients. Sprengel published his results (with which he essentially disproved the humus theory) in 1826 in the *Archiv für die gesammte Naturlehre*. The editor of this journal was K.W.G. Kastner, under whose

Figure 3 The agronomist and chemist Carl Sprengel (1787–1859), who disproved, in 1826, the humus theory and who published in 1828 the theory of the mineral nutrition of plants, as well as the Law of the Minimum. (Undated picture from the Göttingen State and University Library, Göttingen, Germany.)

guidance Justus von Liebig had earned his doctorate in 1823. Initially, the work of Liebig was also published in this journal.

Sprengel's most important publication was probably Von den Substanzen der Ackerkrume und des Untergrundes (About the substances in the mold and the subsoil), which appeared in the *Journal für Technische und Oekonomische Chemie in 1828*. In this lengthy article, Sprengel explained in detail and precisely his theory on the mineral nutrients of plants. In this article he also stated the Law of the Minimum. In later publications, for example in Sprengel's books on agricultural chemistry of 1831 and 1832, the theory of the mineral nutrition of plants and the Law of the Minimum are presented again. Sprengel was a most productive scholar in the first half of the nineteenth century. Besides a large number of journal articles, he published fundamental books on agricultural chemistry, soil science, and crop production. As early as 1828, he had firm ideas about fertilizing field crops with mineral nitrogen. In his article of 1828, he discussed in detail the usefulness of $(NH_4)NO_3$, $(NH_4)_2SO_4$, $(NH_4)Cl$, and $(NH_4)_2CO_3$, as well as KNO_3, $NaNO_3$, and $Ca(NO_3)_2$ as mineral fertilizers.

Liebig was very familiar with the work of Sprengel; but he refused to acknowledge Sprengel as the scholar who had first published the theory of the mineral nutrition of plants and the Law of the Minimum, and who had first recommended the use of mineral fertilizers in crop production. There was a passionate dispute between Sprengel and Liebig about parts of Liebig's *Agrikulturchemie* of 1840, which were taken from Sprengel's work without proper acknowledgement. Liebig, however, who is also remembered as an ambitious, highly argumentative, and intolerant man, ignored Sprengel's justified complaints and continued to pass off the theory of the mineral nutrition of plants as his own. Because of his inappropriate behavior toward Sprengel, and also because of innumerable and harsh disputes with other colleagues, Liebig was not particularly liked by his peers. Some of his colleagues, like the Dutch chemist and physiologist G.J. Mulder from the University of Utrecht, showed their antagonism publicly. (In a widely circulated brochure, Mulder wrote, in 1846: "Freedom of scientific opinion has never been understood by Liebig. For years past a tribunal has been established in Giessen, before which Liebig is at the same time accuser, witness, public prosecutor, advocate, and judge. Before this tribunal a case is rapidly terminated, but grace or justice can never, never be obtained there. From this tribunal, even the purest innocence is never dismissed without being whipped and branded; and for this purpose Liebig holds also the office of executioner.")

Around the middle of the twentieth century, Günter Wendt, from the University of Göttingen, made a thorough study of Carl Sprengel and his pioneering work in soil chemistry and agronomy. Wendt's work about Sprengel and Liebig caused a good deal of confusion in Germany. In 1955, the influential Association of German Agricultural Experimental and Research Stations (VDLUFA) decided to take proper action. It created the Sprengel–Liebig Medal in gold (**Figure 4**), with which persons are regularly honored for outstanding achievements in, or services to, agriculture. By doing so, a dispute about priorities and impacts was avoided, and both pioneering scholars, Sprengel and Liebig, were equally recognized and commemorated.

Summary

Liebig, who spent more than 30 years in intensive engagement in agricultural chemistry, has influenced, since 1840, the course of agronomic developments in Germany and many other countries, particularly in the USA. Because he expounded on the doctrine of the mineral nutrition of plants, it became generally accepted. He promoted the production and use of mineral fertilizers in crop production. The establishment of numerous agricultural experiment stations during the second half of the nineteenth century in Germany and abroad was, in large part, a result of his efforts. More than anybody else, Liebig called for the use of scientific principles in agriculture. He initiated

Figure 4 The gold Sprengel–Liebig medal, established in 1955 by the Association of German Agricultural Experimental and Research Stations (VDLUFA), to commemorate and honor both pioneering agronomists.

a reform in the educational system of agronomists and made agriculture a university discipline. Occasionally, as in the case of nitrogen fertilization, he advocated incorrect views, but this does not affect his outstanding position in the history of the agricultural sciences. However, he should be commemorated in agriculture as a most prominent leader and reformer, and as an untiring propagandist of using mineral fertilizers in crop production, rather than as a discoverer of new knowledge. To Carl Sprengel, on the other hand, belongs the honor of having disproved the humus theory and of having conceived both the theory of mineral nutrition of plants and the Law of the Minimum.

Further Reading

Böhm W (1997) *Biographisches Handbuch zur Geschichte des Pflanzenbaus (Biographical Handbook on the History of Agronomy)*. Munich, Germany: Saur-Verlag.

Browne CA (1944) A source book of agricultural chemistry. In: Verdoorn F (ed.) *Chronica Botanica*, vol. 8, no. 1. Waltham, MA: Chronica Botanica.

Dyke GV (1993) *John Lawes of Rothamsted – Pioneer of Science, Farming and Industry*. Harpenden, UK: Hoos Press.

Liebig J (1840) *Die organische Chemie in ihrer Anwendung auf Agricultur und Physiologie (Organic Chemistry in its Applications to Agriculture and Physiology)*. Braunschweig, Germany: Friedrich Vieweg.

Liebig J von (1855) *Die Grundsätze der Agricultur-Chemie mit Rücksicht auf die in England angestellten Untersuchungen (The Relations of Chemistry to Agriculture and the Agricultural Experiments of Mr JB Lawes)*, 1st and 2nd edn. Braunschweig, Germany: Friedrich Vieweg.

Moulton FR (1942) *Liebig and after Liebig*. Publication No. 16. Washington, DC: American Association for the Advancement of Science.

Paoloni C (1968) *Justus von Liebig: Eine Bibliographie sämtlicher Veröffentlichungen*. Heidelberg, Germany: Carl Winter University Press.

Rossiter MW (1975) *The emergence of agricultural science: Justus von Liebig and the Americans, 1840–1880*. New Haven, CN: Yale University Press.

Sprengel C (1826) Ueber Pflanzenhumus, Humussäure und humussaure Salze (About plant humus, humic acids and salts of humic acid). *Archiv für die Gesammte Naturlehre* 8: 145–220.

Sprengel C (1828) Von den Substanzen der Ackerkrume und des Untergrundes (About the substances in the mold and the subsoil). *Journal für Technische und Ökonomische Chemie* 2: 423–474; 3: 42–99, 313–352, 397–421.

Sprengel C (1831) *Chemie für Landwirthe, Forstmänner und Cameralisten (Chemistry for Agronomists, Foresters, and Agricultural Economists)*, vol. 1. Göttingen, Germany: Vandenhoeck und Ruprecht-Verlag.

Sprengel C (1832) *Chemie für Landwirthe, Forstmänner und Cameralisten (Chemistry for Agronomists, Foresters, and Agricultural Economists)*, vol. 2. Göttingen, Germany: Vandenhoeck und Ruprecht-Verlag.

Thaer A (1809–1812) *Grundsätze der rationellen Landwirthschaft (The Principles of Agriculture)*, vols 1–4. Berlin, Germany: G Reimer. (Translated by Shaw W and Johnson CW.) New York: Greeley and McElrath.

van der Ploeg RR, Böhm W, and Kirkham MB (1999) On the origin of the theory of mineral nutrition of plants and the law of the minimum. *Soil Science Society of America Journal* 63(5): 1055–1062.

Wendt G (1950) *Carl Sprengel und die von ihm geschaffene Mineraltheorie als Fundament der neuen Pflanzenernährungslehre (Carl Sprengel and his Mineral Theory as Foundation of the Modern Science of Plant Nutrition)*. Wolfenbüttel, Germany: Ernst Fischer-Verlag.

LIMING

E J Kamprath and T J Smyth, North Carolina State University, Raleigh, NC, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Approximately 40% of the arable land in the world is naturally acid. The naturally acid soils have developed where rainfall exceeds evapotranspiration. Acid soils have also developed with intensive agriculture. The acidity of soils can be divided into that associated with the soil solution, exchangeable, and nonexchangeable forms. The poor growth of plants on acid soils is due to toxicities of Al, Mn, and H, and deficiencies of Ca, Mg, and Mo. Liming of acid soils increases plant growth due to neutralization of toxic elements, supplying Ca and/or Mg and increasing availability of Mo and P. Determination of the lime requirement for a soil can be based on the amount needed to adjust pH to a desired level, neutralization of exchangeable Al, or the amount estimated by the buffer pH method. Selection of a liming material has to take into account the neutralization value of the lime, fineness of the material, and the need for Mg.

Development of Acid Soils

In nature acid soils are found where rainfall exceeds evapotranspiration. Several processes contribute to soil acidification under these conditions. The excess water moving through the soil carries Ca^{2+} and Mg^{2+} which balance the negative charge of soluble anions like NO_3^- and Cl^-. Leaching of Ca and Mg in conjunction with soluble anions decreases the percentage of exchange sites occupied by basic cations. The sites vacated by Ca and Mg are initially replaced with H^+ supplied by decomposition of organic matter and residues. The H^+-clays then decompose to form Al-clays.

With intensive agriculture H^+ is also produced by the conversion of ammonium to nitrate:

$$NH_4^+ + 2O_2 \rightarrow NO_3^- + H_2O + 2H^+ \qquad [1]$$

The application of 100 kg N ha^{-1} in the ammonium form can require a maximum of 360 kg $CaCO_3$ ha^{-1} to neutralize the acidity produced by the conversion of NH_4^+ to NO_3^-. Plants also release H^+ from roots when their uptake of basic cations exceeds that of anions. Since legumes fix atmospheric N and have relatively high concentrations of Ca, cation uptake exceeds anion uptake. To maintain internal plant electroneutrality legume roots excrete H^+ which acidifies the rhizosphere. Neutralization of acidity produced in one growing season of N-fixing legumes can correspond to 30–400 kg $CaCO_3$ ha^{-1}.

Another source of H^+ is acid precipitation produced upon burning fossil fuels containing sulfur compounds which give rise to SO_2. The SO_2 reacts with water in the atmosphere to produce sulfuric acid.

Nature of Soil Acidity

Soil Solution Acidity

The solutions of acid soils contain two ions that are detrimental to plant growth, hydrogen and aluminum. The concentration of Al^{3+} in the soil solution increases with the proportion of cation exchange sites which are occupied with Al (i.e., the percentage Al saturation of the cation exchange capacity (CEC)). There is a marked increase in soil solution Al^{3+} concentration when the saturation is greater than 50%. The soil solution H^+ concentration is controlled by the hydrolysis of Al^{3+}, as shown by eqn [2].

$$Al^{3+} + H_2O \rightleftarrows AlOH^{2+} + H^+ \qquad [2]$$

Between pH values of 5.6 and 5.8 the soil solution Al^{3+} concentration is nil in mineral soils since the Al saturation is essentially zero. Organic soils with appreciable mineral matter have very low concentrations of soil solution Al^{3+} at pH values of 4.8–5.0, because Al is strongly bound to the carboxyl groups. High concentrations of soil solution Al^{3+} only occur in organic soils at pH values of 4.5 or lower.

Exchangeable Acidity

Early concepts were that exchangeable H^+ was the source of soil acidity. Titration studies showed that acid clays behaved as weakly ionized acids, which led to the conclusion that acid soils contained exchangeable H^+. Soils were assumed to be 100% base-saturated at pH values of 7 and the difference between the CEC at pH 7 and the sum of basic cations was the amount of exchangeable H^+. Later studies revealed that the weakly ionized acid characteristic of clays was due to exchangeable Al and the hydrolysis of Al^{3+} was the source of soil solution H^+. When saturated with H^+ the clays decomposed, releasing Al, which occupied the exchange sites.

Exchangeable acidity is defined as that which is extracted with a neutral unbuffered salt solution such as KCl. Neutral unbuffered salt solutions must be used to avoid changing the soil pH and altering the

Table 1 The exchangeable and nonexchangeable acidity of acid soils

Soil	pH	Organic matter ($g\,kg^{-1}$)	Exchange Al ($cmol\,kg^{-1}$)	Al saturation (%)	Acidity	
					Titratable[a] ($cmol\,kg^{-1}$)	Nonexchangeable $cmol\,kg^{-1}$
Typic Kandiudult	4.5	20	0.91	82	1.70	0.57
Typic Paleaquult	4.7	27	1.04	78	1.72	0.68
Aeric Paleaquult	4.5	40	2.33	73	3.77	1.44
Typic Haplustox	4.5	–	1.55	68	3.85	2.70

[a]Acidity titrated to pH 6.

Data from Evans CE and Kamprath EJ (1970) Lime response as related to percent Al saturation, solution Al, and organic matter content. *Soil Science Society of America Proceedings* 34: 895–896. Kamprath EJ (1970) Exchangeable aluminum as a criterion for liming leached mineral soils. *Soil Science Society of America Proceedings* 34: 252–254. Gonzalez-Erico E, Kamprath EJ, Naderman GC, and Soares WV (1979) Effect of depth of lime incorporation on the growth of corn on an Oxisol of Central Brazil. *Soil Science Society of America Journal* 43: 1155–1158.

exchangeable acidity. In mineral soils the exchangeable acidity is Al^{3+}. At pH values of 5 and lower, over half of the active exchange sites are countered by Al^{3+} (Table 1). As previously discussed, exchangeable Al is essentially zero in mineral soils at pH 5.6 and the active cation exchange sites are 100% base-saturated.

Nonexchangeable Acidity

Soils also contain H which is not extractable with neutral unbuffered salt solutions, but is titratable. The nonexchangeable H is associated with carboxyl groups of organic matter, hydroxy Al, and hydrated oxides of Fe and Al. The nonexchangeable acidity of the three Ultisols in Table 1 was primarily due to organic matter. However, with the Oxisol, the nonexchangeable acidity was associated with the hydrated oxides of Fe and Al. Once the exchangeable Al is neutralized the acid buffering capacity of the soil is a function of the nonexchangeable acidity. Neutralization of the nonexchangeable acidity has little effect *per se* on plant growth, but is the source of the pH-dependent CEC. With increasing neutralization of the nonexchangeable acidity, there is an increase in soil pH, CEC, and the retention of the basic cations Ca, Mg, and K.

Acid Soil Constraints to Plant Growth

Hydrogen Toxicity

Poor plant growth in acid soils is usually associated with low soil pH, but the direct effects of H^+ are often confounded with changes in solubility of various elements affecting plant growth. Experiments in hydroponics, therefore, are often used to assess the direct effects of H^+. Roots growing in solutions at pH 4.6 or lower have visual symptoms of injury, namely stunted growth, brownish color, and little lateral root development. Increasing solution Ca concentration alleviates root injury due to H^+ (Table 2). At low pH and Ca concentrations root membranes are damaged, leading to loss of organic substrates and absorbed cations. Under soil conditions with pH below 4.6, root injury due to H^+ can be expected along with reduced plant uptake and possible deficiencies of Ca and Mg.

Table 2 Soybean tap and lateral root length after 12 days of exposure to solutions with different pH and Ca concentrations

Solution		Root length	
pH	Ca ($mmol\,l^{-1}$)	Tap (cm per plant)	Lateral (cm per plant)
4.0	0.2	1.1	0.0
	2.0	3.2	1.0
4.6	0.2	10.2	3.4
	2.0	22.2	36.3
5.5	0.2	27.8	40.1
	2.0	44.6	85.6

Data from Sanzonowicz C, Smyth TJ, and Israel DW (1998) Calcium alleviation of hydrogen and aluminum inhibition of soybean root extension from limed soil into acid subsurface solutions. *Journal of Plant Nutrition* 21: 785–804.

Aluminum Toxicity

Aluminum toxicity is a major cause of poor plant growth in acid soils. Current evidence indicates that Al interferes with various root growth processes: disruption of regulatory signals in root cap cells and interference with cell division in root apices, enzyme activities, DNA replication, and P availability at membranes. The immediate and visible evidence of Al toxicity is a reduction in root length which, in turn, limits plant access to soil water and nutrients. Consequently, plants growing in soils with toxic levels of Al are underdeveloped, sensitive to drought stress, and often present symptoms of multiple nutrient deficiencies.

Concentrations of soil solution Al that reduce crop growth are in the micromolar range. As illustrated in

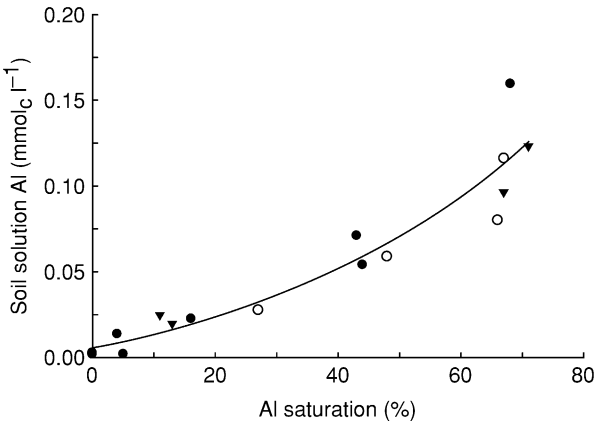


Figure 1 Relations between Al saturation of cation exchange sites and concentration of Al in the soil solution of Oxisols and Ultisols. Filled circles, Haplustox; open circles, Umbraquult; inverted triangles, Paleudult. Data source: Gonzalez-Erico E, Kamprath EJ, Naderman GC, and Soares WV (1979) Effect of depth of lime incorporation on the growth of corn on an Oxisol of Central Brazil. *Soil Science Society of America Journal* 43: 1155–1158; and Jallah JK (1994) *Assessment of Some Chemical Constraints to Root Growth in Four Acid North Carolina Soils*. PhD thesis. Raleigh, NC: North Carolina State University.

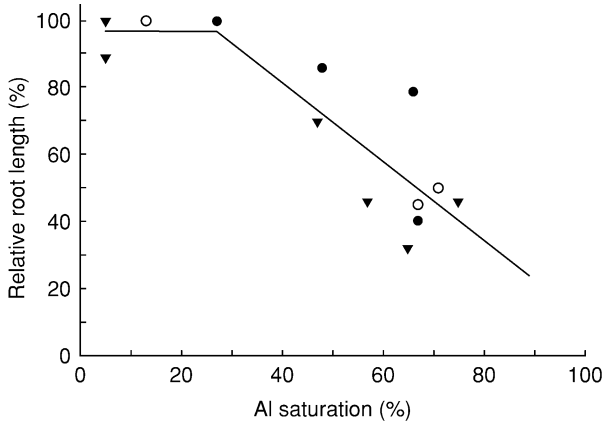


Figure 2 Effect of Al saturation on relative root length of field-grown corn in Oxisols and Ultisols. Filled circles, Umbraquult; open circles, Paleudult; inverted triangles, Haplustox. Data source: Gonzalez-Erico E, Kamprath EJ, Naderman GC, and Soares WV (1979) Effect of depth of lime incorporation on the growth of corn on an Oxisol of Central Brazil. *Soil Science Society of America Journal* 43: 1155–1158; and Jallah JK (1994) *Assessment of Some Chemical Constraints to Root Growth in Four Acid North Carolina Soils*. PhD thesis. Raleigh, NC: North Carolina State University.

Figure 1, the concentration of Al in soil solutions is related to the percentage of Al saturation of the cation exchange sites, with a marked increase in the Al concentration when the Al saturation exceeds 50%. Most field assessments of crop response to toxic levels of soil solution Al have been based on percentage of Al saturation as a proxy variable, due to its greater ease of measurement. Root length of corn, for example, is markedly reduced at percentage of Al saturation levels corresponding to high concentrations of Al in the soil solution (Figure 2).

Tolerance to Al varies widely among species and even varieties of the same species (Table 3). Yields for most varieties of cotton, mung bean, pearl millet, and wheat are reduced by low levels of percentage of Al saturation, whereas cassava and cowpea tolerate high levels of soil Al. However, considerable differences in Al tolerance/susceptibility have been observed among varieties within a given species. For example, critical percentage of Al saturation levels among field trials with different varieties range from 0 to 40 with corn, 0 to 44 with soybean, and 0 to 60 with upland rice.

Table 3 Threshold values of percentage Al saturation of the cation exchange capacity above which crop yield for various species is normally reduced by Al in mineral soils

Crop	*Critical % Al saturation*
Cassava	75
Corn	30
Cotton	0
Cowpea	60
Mung bean	0
Peanut	40
Pearl millet	0
Phaseolus bean	15
Potato	20
Sorghum	15
Soybean	15
Upland rice	40
Wheat	0

Data from Osmond DL, Smyth TJ, Reid WS *et al.* (2002) *Nutrient Management Support System, NuMaSS* (version 2.0). Raleigh, NC: Soil Management Collaborative Research Support System, North Carolina State University.

Manganese Toxicity

Acid soils with high contents of Fe and Al oxides often contain large amounts of soluble Mn. Plant growth in such soils is reduced when large quantities of the soluble soil Mn accumulate in their tissues. Although Mn is an essential nutrient, excessive plant uptake is toxic to plants. Liming these soils to pH 5.5 or greater decreases the solubility of soil Mn, which reduces the uptake of excess Mn and increases plant growth. In many instances, improved plant growth upon liming soils with percentage of Al saturation below the critical level is associated with the alleviation of Mn toxicities.

Calcium and Mg Deficiencies

Acid soils with high percentage of Al saturation have very low amounts of Ca and Mg on both the exchange sites and in the soil solution. Soils with very

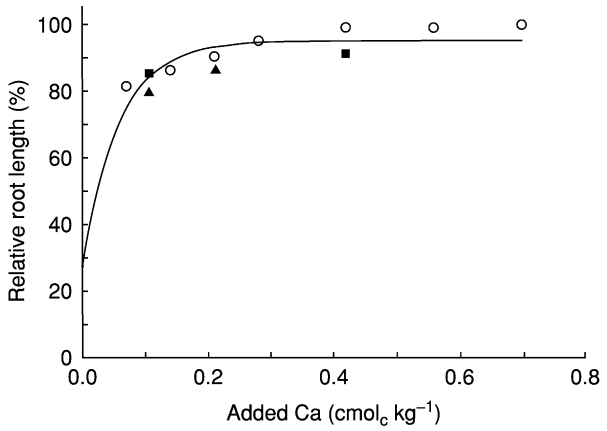


Figure 3 Root elongation of wheat as a function of Ca levels and sources in Ca-deficient Oxisols. Squares, lime; circles, $CaCl_2$; triangles, $Ca(H_2PO_4)_2$. Data source: Ritchey KD, Silva JE, and Sousa DMG (1983) Relação entre teor de calcio no solo e desenvolvimento de raizes avaliado por um metodo biologico. *Revista Brasileira Ciencia Solo* 7: 269–275.

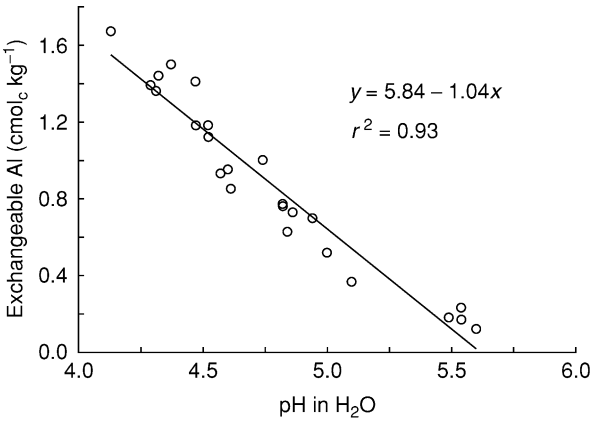


Figure 4 Relationship between soil pH and exchangeable Al for Coastal Plain soil. Data source: unpublished data of Soil Science Department North Carolina State University.

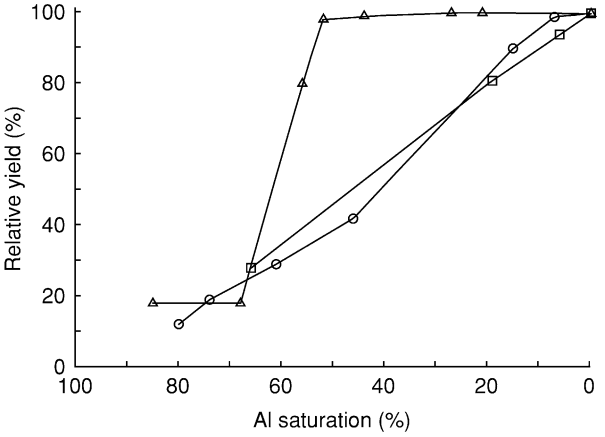


Figure 5 Relative yield of soybean (circles), corn (triangles), and cotton (squares) as related to Al saturation of the effective cation exchange capacity. Data source: E.J. Kamprath (unpublished data); and McCart GD and Kamprath EJ (1965) Supplying calcium and magnesium for cotton on sandy, low cation exchange soils. *Agronomy Journal* 57: 404–406.

low CEC have a limited capacity to retain Ca and Mg and supply of this nutrient may not be sufficient for normal plant growth. The immediate supply of Ca for continued growth of root tips is dependent on the soil solution concentration, because this nutrient is relatively immobile in plants. Root length of wheat seedlings, for example, was increased by over 50% when 0.1 cmol Ca kg^{-1} was added to an Oxisol which initially contained 0.02 cmol Ca kg^{-1} (Figure 3). Although it is often difficult to separate the extent to which a plant growth response to liming is due to correction of a Ca deficiency versus alleviation of an Al toxicity, the improved root growth of wheat was similar among various sources of Ca. Plant growth responses to additions of Mg are often associated with soil conditions where exchangeable Mg^{2+} and/or the proportion of cation exchange sites occupied by Mg^{2+} is very low.

Response to Liming

A major response of plants to liming is due to the neutralization of H, Al, and Mn, which if present in relatively high concentrations in the soil solution are toxic to plant growth. Toxic concentrations of H generally do not occur in soils except in certain organic soils when pH values are less than 4.5. In acid mineral soils with pH values less than 5.5 liming removes the negative effect of H on absorption of Ca, Mg, and K.

The response of plants to liming is mainly due to the neutralization of exchangeable Al and its replacement on the exchange sites with Ca and/or Mg. Liming of acid mineral soils to pH 5.6 reduces exchangeable Al to essentially zero (Figure 4). This essentially eliminates Al from the soil solution and removes it as a factor limiting plant growth. As previously discussed, the concentration of Al in the soil solution is a function of the Al saturation of the active cation exchange sites. Since the Al saturation of a soil is much easier to determine than soil solution Al, it is an excellent criterion for determining the need to lime a given soil. Since crops and genotypes vary in their tolerance to Al, knowledge about the Al saturation of a given soil provides information whether a particular crop or genotype will require liming. An example of this is the differential Al tolerance of corn as compared to soybean and cotton in a greenhouse study (Figure 5). Corn showed little effect to Al until

the Al saturation was greater than 50%, where there is a marked increase in soil solution Al. Relative grain yields of corn on three soils decreased sharply when Al saturation was greater than 40% (Figure 6), a value similar to that found for dry-matter production in the greenhouse. Organic soils and soils high in organic matter have a lower critical pH than most mineral soils. Maximum plant growth on acid organic soils is generally obtained at pH 5. On these soils the exchangeable Al is held on carboxyl sites and at pH 5 the Al is held quite strongly so there is very little Al in the soil solution. Liming of these soils to pH 5 provides adequate Ca to overcome the H effects on plant growth.

Response to liming can also be obtained on acid soils with low Al saturation but that have high amounts of Mn. Soil with high contents of sesquioxides often contain appreciable amounts of Mn minerals and in these soils Mn toxicity is a problem when pH values drop to 5 and less. Liming to pH 5.6 and above eliminates Mn toxicity on these soils by converting the soluble Mn^{2+} to the relatively insoluble oxides, Mn_2O_2 and MnO_2.

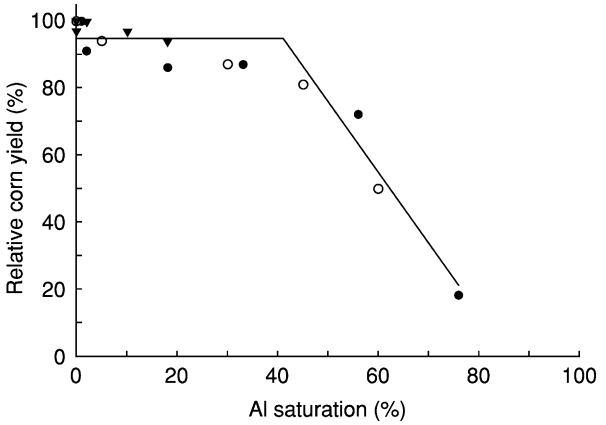


Figure 6 Relative grain yields of corn on three soils (filled circles, Goldsboro; open circles, Wharton; inverted triangles, Morrill) as related to exchangeable Al saturation of the effective cation exchange capacity. Data source: Alley MM (1981) Short-term chemical and crop yield responses to limestone applications. *Agronomy Journal* 73: 687–689; and Fox RH (1979) Soil pH, aluminum saturation and corn grain yield. *Soil Science* 127: 330–334.

Response to liming due to supplying of Ca *per se* is not often the case. In most instances Ca saturations of 20–25% will supply adequate Ca for plant growth except where toxicities to Al, H, or Mn limit plant growth. There is a good relationship between Ca or Ca + Mg saturation and plant growth, because these are the mirror image of Al saturation. On soils with very low CECs a minimum of 1 cmol Ca kg^{-1} soil is needed for optimum growth of most plants. When toxicities are not a factor the percentage of Ca saturation of the CEC is a good indicator of Ca availability. Calcium concentrations of a given plant species are similar when grown on soils with the same Ca saturation even though the soils contain a wide range of exchangeable Ca (Table 4). This is because the Ca concentration of the soil solution is determined by the percentage of Ca saturation rather than the amount of exchangeable Ca. Calcium is transported to the root by mass flow of water and thus the amount arriving at the root surface is a function of the Ca concentration. Magnesium deficiencies on sandy soils are generally associated with pH values of less than 5. The concentration of Al^{3+} and H^+ at pH 5 in mineral soils is sufficient to inhibit Mg uptake by plants. Liming to neutralize Al and H is necessary for optimum uptake of Mg where soil reserves are adequate. On acid sandy soils response to liming is also related to the supply of Mg with the application of dolomitic lime. Availability of soil Mg is related more to the Mg saturation of the CEC rather than the amount of exchangeable Mg. Various studies have shown that Mg saturation needs to be in the range of 5–10% for optimum plant growth.

Soybean and other legumes often show a response to increase in pH beyond that at which Al is neutralized. Raising the pH of mineral soils to 6 has

Table 4 Effect of Ca saturation of the effective cation exchange capacity (ECEC) on Ca availability

		Soil Ca			
Soil	*pH*	*ECEC saturation (%)*	*Exchangeable ($cmol_c\ kg^{-1}$)*	*Solution ($cmol_c\ l^{-1}$)*	*Soybean Ca (%)*
Norfolk	4.9	4	0.06	0.75	0.39
	5.8	72	1.29	4.00	0.86
Portsmouth	4.8	26	3.00	2.92	0.61
	5.3	77	8.31	5.00	0.99
Organic	4.2	31	7.08	3.75	0.47
	4.7	70	31.75	5.12	0.98

Data from Evans CE and Kamprath EJ (1970) Lime response as related to percent Al saturation, solution Al, and organic matter content. *Soil Science Society of America Proceedings* 34: 895–896. Evans CE (1968) *Ion Exchange Relationships of Aluminum and Calcium in Soils as Influenced by Organic Matter.* PhD thesis. Raleigh, NC: North Carolina State University.

increased yield because of increased solubility of soil Mo, which is required by bacteria for N_2 fixation. Increasing the OH^- concentration with liming brings about the replacement of MoO_4^{-2} absorbed by the Fe and Al hydrated oxides. Raising the pH to 6.0–6.2 makes sufficient Mo available for optimum N_2 fixation and optimum yields of legumes. Thus the higher critical pH for legumes as compared to nonlegumes on mineral soils is related to the Mo requirement for N_2 fixation and the effect of pH on Mo availability.

Availability of soil P in acid soils has been increased by liming. In acid soils Al at the root surface precipitates P and decreases the amount of P transported to plant tops. Removal of Al from the soil solution by liming prevents this from happening. Neutralization of exchangeable Al also results in more root growth and exploration of a greater soil volume. Phosphorus gets to roots by diffusion and when root growth in a given volume is increased, the uptake of P is increased.

The increased root growth in a given volume of soil when Al is neutralized by liming has a number of general beneficial effects on plant growth. Nutrient and soil water availability are increased. Microbial activity is enhanced; this affects N_2 fixation, decomposition of organic matter and plant residues, making available organically bound nutrients.

Cation Exchange Properties Affecting Liming of Acid Soils

The CEC of a soil is largely determined by the mineral composition of the clay-sized fraction. The source of the negative charge in clay minerals can be either permanent or variable. Permanent charge results from isomorphous substitution among cations with differing valences within the crystalline clay structure, and is a permanent characteristic of the mineral. Smectite, vermiculite, and illite are commonly occurring clay minerals with permanent charge. Variable charge results from the protonation and deprotonation of OH groups on the surfaces of Al and Fe oxides, crystalline or amorphous, and exposed edges of layer-silicate minerals. Kaolinite, hydroxy-interlayered vermiculite, goethite, gibbsite, quartz, and allophane are some of the clay minerals with variable charge. Carboxyl and OH groups in soil organic matter also have a variable-charge behavior, wherein the amount of negative charge increases with soil pH.

Most acid soils have a mixed composition of both permanent- and variable-charge clay minerals. However, the dominant type of clay mineral charge can be estimated through the sum of basic cations and exchangeable acidity (unbuffered salt extraction) relative to the amount of soil clay. Soils dominated by minerals with permanent charge have CEC values $>12\,\text{cmol}\,\text{kg}^{-1}$ of clay, whereas soils dominated by variable-charge clays have CEC values $<12\,\text{cmol}\,\text{kg}^{-1}$ of clay.

When lime is applied to soils with variable-charge clay minerals, part of the potential base reaction is consumed in the neutralization of H^+ released from OH groups. Consequently, the lime required to neutralize a targeted amount of exchangeable acidity is increased.

Lime Requirement Determinations

Soil pH

Soil pH was one of the first methods for determining the need for lime. At pH 7 soils were considered to be 100% base-saturated and had zero base saturation at pH 4. The relationship between soil pH and base saturation was considered to be linear between pH 4 and 7. Measurement of the soil pH provided an estimate of the base saturation. To determine the amount of lime to raise the soil to the desired soil pH requires information about the CEC of the soil which can be estimated based on soil texture and organic matter content. As an example, the amount of lime to bring the pH to 6 (90% base saturation) for a soil which had an initial pH of 5 (50% base saturation) and a CEC of $10\,\text{cmol}\,\text{kg}^{-1}$ soil would be $10\,\text{cmol}\,\text{kg}^{-1}$ soil $\times\, 40\% = 4\,\text{cmol lime}\,\text{kg}^{-1}$ soil, which is equal to $4000\,\text{kg lime}\,\text{ha}^{-1}$. One of the limitations of this approach is in estimating the CEC of the soil. If the estimate of the CEC is off, this will result in recommendations which are either too much or too little of the required amount of lime.

Titration of Soil with a Base

The lime requirement of soils can be determined by titrating acid soils to the desired pH with a base such as KOH or $Ca(OH)_2$. Titration of mineral soils to pH 5.6–5.8 will give the amount of lime required to neutralize exchangeable Al. The amount of lime on an equivalent basis is more than the exchangeable Al because, as the pH increases, nonexchangeable acidity ionizes and reacts with the lime. Titration to pH values above 5.6–5.8 results in the reactions of the base with nonexchangeable acidity which is H associated with hydroxy Al, carboxyl groups of organic matter, and hydrated oxides of Fe and Al.

Reaction of lime with exchangeable and nonexchangeable acidity is shown by the following equations:

1. neutralization of exchangeable acidity:

$$2\text{Al-soil} + 3CaCO_3 + 6H_2O \rightarrow 4\text{Ca-soil} + 2Al(OH)_3 + 3H_2CO_3$$
$$3H_2CO_3 \rightarrow 3H_2O + 3CO_2\uparrow \quad [3]$$

2. neutralization of nonexchangeable acidity:

$$2(\text{R-C-OOH}) + CaCO_3 \rightarrow Ca(\text{R-C-OO})_2 + 2H_2CO_3$$
$$2H_2CO_3 \rightarrow 2H_2O + 2CO_2\uparrow \quad [4]$$

This method is rather time-consuming and does not lend itself to use where large number of samples need to be analyzed.

Buffer pH Method

Many soil-testing laboratories use the soil-buffer pH method to determine the lime requirement. This method is essentially a titration of an acid, the soil, with a base, the buffer solution. The molarity of the buffer solution is known and the amount of base required to neutralize the acidity of the soil can be determined by the change in the pH of the buffer solutions. With these methods a given volume of the buffer solution is added to a given weight or volume of soil and the pH is measured. The decrease in pH of the buffer solution is related to the amount of acidity in the soil.

Four buffer pH methods are being used in the USA to determine lime requirements. The amount of lime needed for each 0.1 decrease of buffer pH to bring the soil to the indicated pH is given in Table 5 for each of the methods. The Woodruff buffer was developed for determining the amount of lime to adjust the pH of Mollisols to a range of 6.5–7.0. The SMP buffer is used for soils with a large amount of three-layer clays and high-organic-matter content, such as Alfisols. The Adams and Evans buffer was developed for soils with low CEC and kaolinitic clay mineralogy. The lime recommendation by this method is to adjust the pH to 6.5. The Mehlich buffer was developed to predict the amount of lime needed to neutralize exchangeable Al. For mineral soils lime rates based on the acidity extracted with the Mehlich buffer raise the pH in the range of 5.8–6.0, which neutralizes the exchangeable Al.

Exchangeable Al

Liming mineral soils on an equivalent basis to the amount of exchangeable Al targets achievement of pH 5.6–5.8 and a soil solution Al^{3+} concentration approaching zero. As previously discussed, lime will neutralize exchangeable Al and also react with protons on variable-charge mineral surfaces. Consequently, the lime equivalence factor to achieve 0% Al saturation of the soil CEC is greater than 1.0 and is usually in the order of 1.5–3.

Yields for many crops are not reduced until soil Al saturation values exceed 40% or even 60% (Table 3 and Figure 5). For these acid-tolerant crops and, especially, in regions where lime materials are expensive, lime requirements can be adjusted to achieve a targeted percentage of Al saturation instead of neutralizing all of the exchangeable Al. An example of this approach is the following:

$$CaCO_3 \text{ equivalent } (\text{t ha}^{-1}) = LF[Al - TAS(Al + Ca + Mg)/100] \quad [5]$$

where LF = the product of 1.5 equivalents of $CaCO_3$/equivalent of exchangeable Al and the conversion to a 20-cm layer for a hectare of soil, Al, Ca, and Mg = exchangeable cations extracted with a neutral unbuffered salt, such as KCl, and TAS = the targeted percentage of Al saturation after liming.

Liming soils to achieve a given percentage of Al saturation requires more information than when liming to neutralize all of the exchangeable Al. In addition to exchangeable Al, analytical data are needed for exchangeable Ca and Mg. Furthermore,

Table 5 Lime requirement as determined with buffer solutions

Method	*Buffer pH*	*Target pH*	*Intended use*	*Lime/0.1 pH decrease in buffer (t $CaCO_3$ ha^{-1})*
SMP	6.8	6.8	High exchangeable Al	0.6
		6.4	Alfisols	0.5
		6.0		0.4
Woodruff	7.0	6.5–7.0	Mollisols	0.5
Adams and Evans	8.0	6.5	Low cation exchange capacity Ultisols	0.1
Mehlich	6.6	6.0	Neutralize exchangeable Al, Ultisols	0.16

Data from van Lierop W (1990) Soil pH and lime requirement determination. In: Westerman RL (ed.) *Soil Testing and Plant Analysis*, 3rd edn, pp. 73–126. Madison, WI: Soil Science Society of America.
CEC, cation exchange capacity.

the percentage of Al saturation tolerated by the intended crop must also be known.

Factors to Consider in Lime Application

Liming Materials

There are a number of different materials which can be used for liming. The most common ones are calcium carbonate (calcitic limestone) and calcium-magnesium carbonate (dolomitic limestone). These are crystalline compounds which must be ground. Marls are soft, unconsolidated calcium carbonates which may contain some clay. Calcium oxide (burnt lime or quick lime) is a white powder and is caustic when it absorbs water. It is difficult to get uniform mixing with the soil because immediately after application absorption of water causes granules to form. Calcium hydroxide (hydrated lime) is a white powder and when wet is caustic. It is also difficult to apply. Basic slag (calcium silicate) is a by-product of the steel industry.

Neutralizing Value

The neutralizing value ($CaCO_3$ equivalent) of liming materials is related to the amount of acid that a unit weight of lime will neutralize. The standard for evaluating materials is pure calcium carbonate which is given a value of 100%. The neutralizing value for pure liming materials is given in Table 6.

The purity of the liming material must also be taken into account when determining the neutralizing value. Most liming materials are not 100% pure. In many instances the purity of the material will have been determined so that the buyer is given this information. Lime recommendations are often based on 100% purity of the material so that the application rate will have to be increased when the purity is less than 100%.

Fineness of Liming Materials

The most commonly used liming materials, calcitic and dolomitic limestone, are crystalline compounds. These materials must be ground to a fineness which will provide a large number of particles per unit volume of soil. The lime particles dissolve when coming into contact with acid soils and will release Ca^{2+} and HCO_3^-: this results in neutralization of H^+ and Al^{3+}. The rate at which the pH of the soil mass is affected depends on how close the individual lime particles are to one another. Since the diffusion of Ca is relatively slow the particles need to be close together so that the zones of neutralization around each particle overlap in a relatively short time.

Calcitic and dolomitic limestones are generally ground so that all of the particles pass a 20-mesh sieve and a certain percentage passes a 100-mesh sieve depending on the regulations of the state or province. Calcitic limestones are often softer than dolomitic limestones and therefore dissolve at a faster rate. The relative efficiency of different particle sizes of these two materials is given in Figure 7. Particles held on a 10-mesh sieve have little value for neutralizing acid soils.

Table 6 Neutralizing value of pure liming materials

Material	*Neutralizing value (%)*
Calcium carbonate, $CaCO_3$	100
Calcium-magnesium carbonate, $CaMg(CO_3)_2$	109
Calcium oxide, CaO	179
Calcium hydroxide, $Ca(OH)_2$	135
Calcium silicate, $CaSiO_3$	86

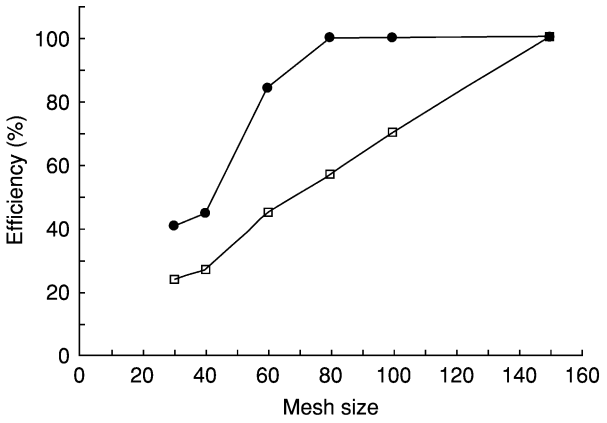


Figure 7 Relative efficiency of various mesh sizes of calcitic (circles) and dolomitic (squares) lime in neutralizing acidity. Data source: unpublished data of E.J. Kamprath.

Incorporation of Lime in the Soil

Since Ca^{+2} and HCO_3^- ions have a slow rate of diffusion, the lime must be well mixed with the soil to neutralize soil acidity throughout the zone which is to be amended. One approach is to apply half of the lime and incorporate it by disking, and then apply the other half and disk to incorporate. However this results in high costs of application and incorporation rather than doing it only once.

Conclusion

Poor plant growth on acid soils is due to toxicities of Al, H, and Mn and deficiencies of Ca, Mg, and Mo. Oxisols, Ultisols, and some Alfisols in their native state have pH values less than 5 and exchangeable Al of the effective CEC greater than 50%. Plant response to liming is due to neutralization of the toxic elements

and the supply of Ca and Mg. The lime requirement for a soil can be based on the amount of lime required to: (1) adjust the pH to a desired value; (2) neutralize exchangeable Al; or (3) neutralize the acidity measured by buffer pH methods. Factors to consider in selecting a lime material are fineness, purity, and the need for Mg.

See also: **Acidity**; **Calcium and Magnesium in Soils**; **pH**

Further Reading

Adams F (ed.) (1984) *Soil Acidity and Liming*, 2nd edn. Madison, WI: American Society of Agronomy, Crop Science Society of America, Soil Science Society of America.

Alley MM (1981) Short-term chemical and crop yield responses to limestone applications. *Agronomy Journal* 73: 687–689.

Cochrane TT, Salinas JG, and Sanchez PA (1980) An equation for liming acid soil mineral soils to compensate crop aluminium tolerance. *Tropical Agriculture* 57: 133–140.

Evans CE (1968) *Ion Exchange Relationships of Aluminum and Calcium in Soils as Influenced by Organic Matter*. PhD thesis. Raleigh, NC: North Carolina State University.

Evans CE and Kamprath EJ (1970) Lime response as related to percent Al saturation, solution Al, and organic matter content. *Soil Science Society of America Proceedings* 34: 895–896.

Fox RH (1979) Soil pH, aluminum saturation and corn grain yield. *Soil Science* 127: 330–334.

Gonzalez-Erico E, Kamprath EJ, Naderman GC, and Soares WV (1979) Effect of depth of lime incorporation on the growth of corn on an Oxisol of Central Brazil. *Soil Science Society of America Journal* 43: 1155–1158.

Jallah JK (1994) *Assessment of Some Chemical Constraints to Root Growth in Four Acid North Carolina Soils*. PhD thesis. Raleigh, NC: North Carolina State University.

Kamprath EJ (1970) Exchangeable aluminum as a criterion for liming leached mineral soils. *Soil Science Society of America Proceedings* 34: 252–254.

McCart GD and Kamprath EJ (1965) Supplying calcium and magnesium for cotton on sandy, low cation exchange soils. *Agronomy Journal* 57: 404–406.

Osmond DL, Smyth TJ, Reid WS *et al.* (2002) *Nutrient Management Support System, NuMaSS*. (version 2.0). Raleigh, NC: Soil Management Collaborative Research Support System, North Carolina State University.

Ritchey KD, Silva JE, and Sousa DMG (1983) Relação entre teor de calcio no solo e desenvolvimento de raizes avaliado por um metodo biologico. *Revista Brasileira Ciencia Solo* 7: 269–275.

Sanzonowicz C, Smyth TJ, and Israel DW (1998) Calcium alleviation of hydrogen and aluminum inhibition of soybean root extension from limed soil into acid subsurface solutions. *Journal of Plant Nutrition* 21: 785–804.

van Lierop W (1990) Soil pH and lime requirement determination. In: Westerman RL (ed.) *Soil Testing and Plant Analysis*, 3rd edn, pp. 73–126. Madison, WI: Soil Science Society of America.

LIPMAN, JACOB G.

J C F Tedrow, Formerly with Rutgers University, New Brunswick, NJ, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Jacob G. Lipman was one of the most prominent international leaders in soil science. In 1901 he formed the Department of Soil Chemistry and Bacteriology at Rutgers University, the first of its kind in the USA, and probably the world. He was the driving force in internationalizing soil science. He also initiated and edited Soil Science, the first international journal dealing with soil chemistry, soil fertility, soil physics, and soil microbiology.

Lipman (Figure 1) played a major role in the development of agriculture in the USA and beyond during the 1901–1939 period. His main contributions were in the area of soil microbiology and soil fertility, particularly in relation to increased crop production. During later years, he was also involved in worldwide agricultural problems. Not only did he advance science on several fronts but he was also a gifted administrator. He addressed the broader aspects of agriculture, particularly in increasing the world food supply, controlling soil erosion, and improving the general welfare of humanity.

The Early Years

Jacob Goodale Lipman was born in Friedrichstat, Russia, on November 18, 1874. He received his early academic training under private tutors in Moscow and in the classical gymnasium in Orenberg. In 1888, with his parents, he settled at the Baron de Hirsch colony in Woodbine, New Jersey. After graduating from farm school there he entered Rutgers

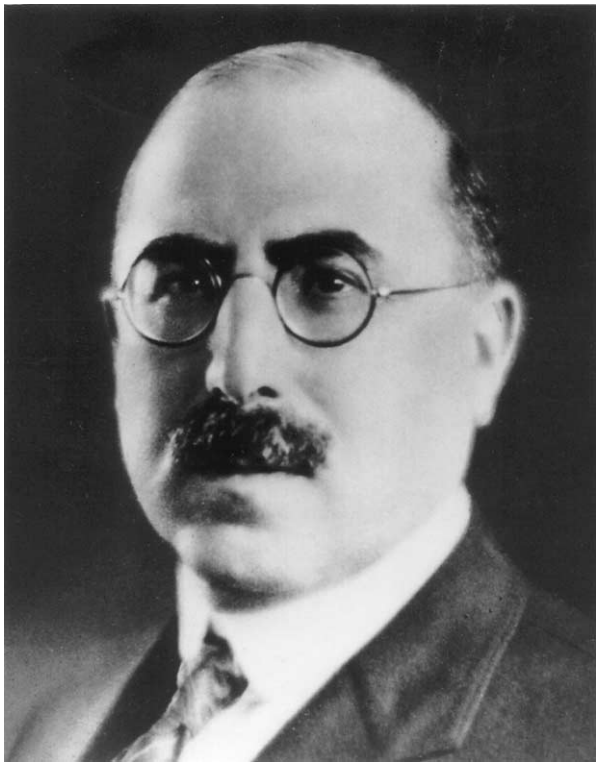


Figure 1 Jacob G. Lipman, 1874–1939.

College (University) in 1894. Next was graduate study at Cornell University where he studied under professors G.A. Caldwell and V.A. Moore. His master of arts degree was completed in 1900 and the doctor of philosophy degree in 1903. As his graduate studies were nearing completion he returned to the New Jersey Agricultural Experiment Station to organize a department of soil chemistry and bacteriology, the first of its kind in the USA. In 1902 he was appointed instructor in agriculture and by 1910 he had attained his (full) professorship. In 1911 he was appointed director of the Agricultural Experiment Station, but also retained his position as chairman of the Department of Soil Chemistry and Bacteriology. Additionally, in 1916 he was appointed as dean of the College of Agriculture, the first time that the office existed.

Lipman the Scientist

Early in his career people were just beginning to gain knowledge of the role of microbes and their applications to soil processes and to agriculture in general. During this time microbes were being recognized as agents responsible for numerous reactions in the soil that led to the liberation of chemical nutrients essential for plant growth.

Immediately after taking office, Lipman outlined a policy of increased service to the state. He saw the need for strengthening and enlarging the staff and adding new departments. This, however, could not be done without increased state appropriations.

During his first address to the Board of Managers after becoming director of the New Jersey Agricultural Experiment Station, Lipman announced the beginning of a policy of increased service to the state. He asserted that there should be a sound basis for development of the land. He also urged the expansion of extension activities. Through his close relationship with leading farmers and prominent businessmen, he convinced the legislature that there was a real need for greater financial support for agriculture. Accordingly, during his tenure, state appropriations were increased 16-fold. Legislators in Trenton looked forward to Lipman's appearance before the Appropriations Committee, not only for the humorous stories he told so well, but also because they thought he was sincere and they liked him. Pointing out that, for many years, the station had been a creative force in developing agriculture in New Jersey and beyond, he asserted that it was time to undertake a comprehensive campaign in the public interest. He proposed studies in farm management, irrigation, soil drainage, fertilization of crops, and product marketing. He also pointed out the need for upgrading the laws governing seed control, fertilizer manufacture, and animal feeds.

Dr. Lipman's own scientific activities began at the turn of the century when the importance of microorganisms in soil processes was beginning to receive increased attention, particularly relating to nitrogen studies. In the history of soil microbiology, the name of Lipman will forever be connected with the study of the cycle of nitrogen in nature, especially the problem of nitrogen fixation by nonsymbiotic bacteria, the liberation of nitrogen as ammonia in the process of decomposition of nitrogenous materials in the soil, and the general relations of nitrogen liberation in soil to the growth of vascular plants.

One of Dr. Lipman's most important scientific contributions dealt with the transformation and fixation of nitrogen by bacteria. In studying some of the organisms concerned in the process he broached two subjects that later developed greatly: the importance of joint action of the soil organisms, and the relation between denitrification and composition of the organic matter concerned, now commonly referred to as the carbon–nitrogen ratio.

During his studies Lipman isolated a newly discovered nonsymbiotic nitrogen-fixing organism, *Azotobacter vinelandii*; he also described several other newly discovered organisms. The genus Azotobacter was recognized as one of the important members of the soil population. Further, Lipman

was one of the first to prove that Azotobacter are capable of fixing large quantities of nitrogen from the atmosphere. Later, he identified two more important species of Azotobacter: *A. beijerinckii* and *A. woodstownii*. Lipman showed that the same soil type under different farming systems had the ability to produce ammonia from various organic compounds. Further, Lipman's work can be considered the forerunner of the Azotobacter method for determining the presence and abundance of available minerals in the soil.

The next signal accomplishment of Dr. Lipman in soil research concerned oxidation of sulfur in the soil. The underlying task was to make rock phosphate more available to plants. Lipman believed that if certain bacteria in the soil were capable of oxidizing the sulfur, the phosphate would be converted into a more soluble form. Together with his assistants, Selman A. Waksman and Jacob S. Joffe, he isolated the bacterium *Thiobacillus thiooxidans* and applied it to the soil. The concept had considerable agricultural appeal and patents were granted covering the process. Subsequent investigations by Lipman's colleagues, however, showed that in New Jersey soils the rate of sulfur oxidation in the uninoculated samples was about the same as that of the inoculated ones.

Dr. Lipman's scientific accomplishments soon achieved worldwide recognition. Students came from many foreign lands to carry out their graduate studies at Rutgers University. In all, students from some 24 foreign countries studied under Lipman and colleagues. Among the future luminaries from the USA and beyond were P.E. Brown, R.T. Dubos, J.S. Joffe, W. Rudolfs, R.L. Starkey, and S.A. Waksman. Further, young investigators who had already completed their doctor of philosophy degrees, as exemplified by Hans Jenny and W.L. Kubiëna, worked in the department for shorter periods.

It was often said that Dr. Lipman kept his associates happy and smiling. He believed in delegating responsibility and in so doing inspired his associates to supreme efforts. But once he delegated an assignment he would hold his staff member responsible. One staff member commented that Lipman could outline in one-half hour more than could be done in 6 months. His colleagues also said of him that, while demanding of students, he was also patient and understanding. He was never known to punish anyone. He felt that everyone should have another chance if they wanted it and he was never known to be angry with any subordinate.

As a writer and speaker Dr. Lipman was at his best. He delivered a series of lectures at other institutions such as University of Illinois, Cornell University, University of Tennessee, University of Nebraska, and Iowa State University. He was also called upon to speak at many agricultural meetings in New Jersey and beyond.

After he became director of the New Jersey Agricultural Experiment Station visitors came to learn from him, not only from nearby but also from all over the world and from all walks of life. Lipman had an encyclopedic mind. His conversations were not limited to agriculture but included general science, economics, and religion. Very few investigators have the skills to master the field of research and also have the ability to convey messages to the ordinary farmer. Lipman, however, bridged the gap admirably. He wrote many reports on the technical phases of soil science and agriculture and contributed articles to various farm journals such as Pennsylvania Farmer, Rural New Yorker, New Jersey Agriculture, and the John Wiley Agricultural Science series. He wrote technical reports for C.E. Marshall's Microbiology and E.O. Jordan and I.S. Falk's book The Newer Knowledge of Bacteriology and Immunology. Lipman was also an advisory editor of the Journal of Agricultural Research, the Internationale Mitteilungen für Bodenkunde, and Annales Agronomiques. Dr. Lipman also edited a series of textbooks in agricultural science. His Bacteria in Relation to Country Life gave him nationwide recognition. This was followed by his Laboratory Guide of Soil Bacteriology, coauthored with his first assistant, P.E. Brown.

Dr. Lipman held membership in a large number of professional societies throughout the USA. The list included the American Health Association, American Association for the Advancement of Science, American Chemical Society, Society for the Promotion of Agricultural Science, American Society of Agronomy, Association of Official Agricultural Chemists, Society of American Bacteriologists, Association of Land Grant Colleges, New Jersey State Sanitary Association, Washington Academy of Science, American Academy of Political and Social Science, National Research Council, American Peat Society, New Jersey Chemical Society, and others. The list of his membership in foreign societies included the International Commission of Agriculture Ecology, Reale Accademia del Georgofili de Firenze, French Academy of Agriculture, Academy of Agriculture of Czechoslovakia, and the Swedish Royal Academy of Agriculture. In 1922 he was awarded a silver medal by the French Academy of Agriculture. Through Lipman's efforts the New Jersey Agricultural Experiment Station received the Bacteria Award at the Panama-Pacific International Exposition of 1915.

In 1923 Rutgers University awarded Dr. Lipman an honorary doctor of science degree and in 1930 he

received an honorary doctor degree from the Catholic University of Santiago, Chile.

Dr. Lipman's research subjects during the 1920s and 1930s included soil microbiological populations, the influence of environmental conditions upon biological changes in the soil, the correlation of soil fertility and microbiological processes, enzymes of microorganisms, anaerobic fermentation of hemicelluloses, and transformation of organic materials.

Dr. Lipman's investigations were carried out during a period of soil microbiology when both scientists and practical microbiologists were seeking applications of science to agriculture. His work was carried out during the time when there was a transition between science developed in the laboratory and application of the results to field conditions. He contributed to both and thus carried the service through one of the most complex periods of development. He not only helped to advance science of the soil, but he also interpreted the application of science to practical agriculture.

At the height of his career Dr. Lipman turned to the broader issues of society. He was one of the pioneers in advancing rural settlement, rural rehabilitation, soil conservation, flood control, afforestation, wildlife sanctuaries, and the general good use of land.

Up until the World War I era there were virtually no important technical journals dealing specifically with soil science published in North America. Americans had to rely mainly on technical papers and books dealing with soil science from Germany and England. The great body of soils information from Russia had not, at that time, penetrated the western world to any extent. In 1916 Lipman founded and edited Soil Science, a journal he personally continued editing for the remainder of his lifetime. In all he edited 46 volumes (two volumes per year). Contributions dealing with soil science came from investigators from all over the globe. Lipman, at an early date, appointed some 35 consulting editors of the journal, over half of whom were from foreign lands. Soil Science was truly an international journal without any society or government affiliation. This venerable journal, published monthly since its inception, has had, over the years, 2000–3000 institutional and private subscribers, with nearly half of them from foreign lands. Lipman's labors with the journal alone served as a vehicle for internationalizing the field of soil science.

At the time when Dr. Lipman became dean of the College of Agriculture and director of the Agricultural Experiment Station, the soils of New Jersey were being surveyed on a scale of 1 in per mile (2.5 cm per 1.6 km). The surveys were being carried out as a joint project with the US Department of Agriculture. By 1927 surveys for the entire state were completed. New Jersey was one of the first states to have such a survey. Soils were also collected and analyzed for their chemical, physical, and mineralogical properties. Also, there were recommendations for cropping systems, land management, soil drainage, and related subjects.

During the latter part of his life Dr. Lipman made a number of trips to Europe, particularly England, the Netherlands, Germany, France, and Italy. Conversing with non-English-speaking colleagues presented few problems because he was fluent in Russian, German, and French. European investigators returned many visits to Lipman. He also made trips to Central and South America. His main purpose was to study various agricultural practices.

Lipman the Administrator

When Lipman became director of the New Jersey Agricultural Experiment Station the physical facilities at the station were rather limited. Office and laboratory space was limited to two buildings on the main campus of the university. There were also a few farm buildings at College Farm on the edge of New Brunswick. Lipman increased the acreage of College Farm from 142 to 800. Over a period of 20 years the entire operation of the Agricultural Experiment Station plus the teaching facilities in agriculture were moved to College Farm where they remain to this day.

During the 1911–1930 period many new, major structures plus ancillary buildings, including animal barns, were constructed at the College of Agriculture. By 1928, with the growth of research and the increase in the number of students, the state legislature approved the construction of an additional building to be devoted primarily to research and the teaching of soil science and in 1931 funds were provided for the new building. But just as the architectural plans were completed, the state, because of the world depression, rescinded all appropriations for capital construction. In 1950, however, the new building was completed and it bears the name Lipman Hall (**Figure 2**).

At the time when Dr. Lipman became director of the New Jersey Agricultural Experiment Station there were only a few departments. In addition to enhancing the physical plant he created departments of seed analysis, plant pathology, agronomy, and farm management. During the later years he also added departments of plant physiology, agricultural biochemistry, agricultural engineering, sewage research, and ornamental horticulture.

Dr. Lipman greatly increased the size and scope of the New Jersey Agricultural Experiment Station but there were times when it was difficult to obtain sufficient funds to carry out his plans, particularly during the World War I era. He was heard to say: "I wake up

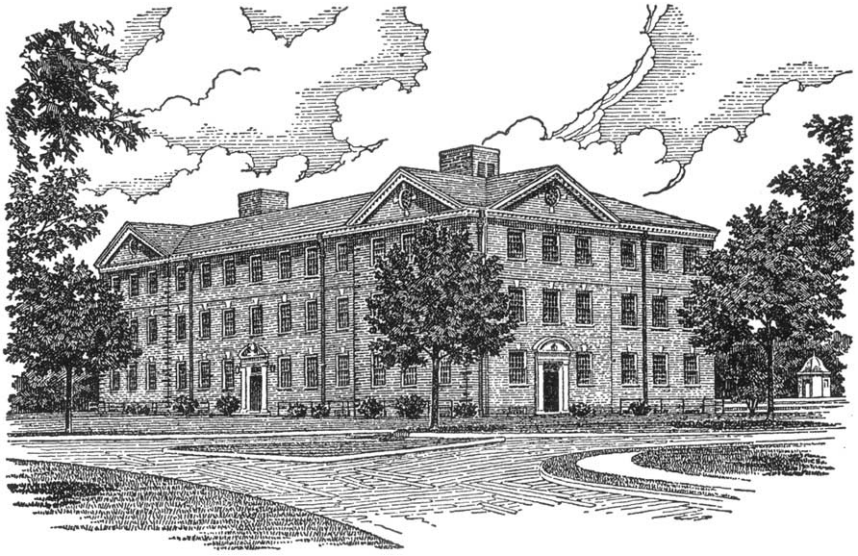


Figure 2 Lipman Hall on the campus of Rutgers University, New Brunswick.

Figure 3 Experimental soil cylinders at the New Jersey Agricultural Experiment Station, Rutgers University, New Brunswick.

in the morning with a vision of a pile of mail waiting for me, requests I cannot fill, complaints I cannot help, and work piled without end. I have an idea I'd rather prefer the trenches."

Dr. Lipman was well aware of the many varieties of soils in New Jersey. Accordingly, he set up management demonstrations on many farms. Branch stations were established for specialized crops such as blueberries and cranberries. Also, a shellfish research station was established, as was a dairy research station.

Experimental work dealing with growing crops on different kinds of soil was initiated by Dr. Lipman's predecessor, Dr. E.B. Voorhees. Initially there were 60 openend, steel cylinders, 4 ft (1.2 m) high with a diameter of 23 1/2 in. (60 cm) installed in a cluster at ground level. Various soils were placed in the cylinders and crops such as tomatoes, corn, and oats were grown (Figure 3). Later, under Lipman's directorship, the number of cylinders was increased to 320 and additional experiments included various soils and fertilizer mixtures. Also, chemical and bacteriological studies were made regarding the pathways of nitrogen and organic matter transformation. In addition to gaining general information on nutrient balance the cylinder experiments demonstrated that enormous

losses of nitrogen may take place in the soil. Also, it was found that soluble nitrogen in liquid manure possesses high value as a source of nitrogen. The cylinder experiments settled many worldwide problems dealing with the availability and fate of many kinds of fertilizers.

Lipman's predecessor at the State Agricultural Extension Service had been operating in New Jersey since about 1890 but the service was greatly expanded under Lipman. Early in his tenure as director there was a major consolidation of the various agricultural organizations in New Jersey. In 1912 extension activities were formally initiated. When the state legislature convened in 1913, New Jersey governor Woodrow Wilson had a bill drafted for an extension system of the governor's own design. Farm demonstrations were to be operated by county agents who would be responsible to a state superintendent appointed by, and responsible to, the governor. Wilson's idea was to have the extension service completely independent of the Agricultural Experiment Station. As soon as the draft bill was made public, Lipman called on the governor and urged that the extension service be incorporated into the college and station. But the governor said, "I want it my way." The bill passed the assembly but the governor was later due for a surprise. Lipman turned to his long-time friend Harry Wheaton, an elderly blacksmith who was chairman of the State Committee on Agriculture and urged him to "sit on the bill." March came and Wilson was inaugurated president. The controversy between Lipman and Woodrow Wilson continued but Lipman had convinced Wheaton to continue to withhold the bill from a final vote. Lipman said to him: "Woodrow Wilson is no longer Governor of New Jersey, now he is only President of the United States." Lipman persisted and, finally, had his own bill passed providing for the extension service to administer activities through Rutgers University.

Dr. Lipman realized the need for international collaboration particularly between investigators from Europe and those from the USA. During the Third International Conference on Soil Science in Prague in 1922 the USA delegated Lipman and C.F. Marbut (Head of the US Soil Survey) to attend. Lipman's mind had developed so keenly and rapidly as to extend well beyond the immediate problems with which soil science was struggling. At the Fourth Conference on Soil Science in Rome in 1924 Lipman again attended the meetings and continued to make very favorable impressions on European colleagues. It was during this meeting that the idea of forming the international society of soil science was finalized. Washington, DC, was selected as the meeting place and Lipman, because of his great organizational talent, was selected as president. Selection of the USA was based on several factors: (1) European investigators were anxious to make observations on US soils in general; (2) there was particular interest in seeing the highly weathered, reddish soils in the south-eastern USA; and (3) investigators wished to make studies and observe the similarity of the dark-colored soils of the prairies and steppes of the central USA to their counterparts in Eurasia.

The First International Society of Soil Science meeting was held during June 1927. President Calvin Coolidge gave the keynote address, which was followed by an address by Lipman. During the meeting Dr. Lipman said:

> Truly in man's upward climb he has accomplished the seemingly impossible. But much as the student of soils may be tempted to ponder on man's past, he must leave this to the philosopher and the poet. It is his task, rather, to inquire into the relations of climates, soils, plants, animals, and man. For all that, soil science must build a foundation large enough and strong enough to support the study of plant food resources, their mobilization, and the interrelations of soils and plants and the characteristics and peculiarities as reflected in the make-up of plants, animals, and man.

Immediately following the First International Soil Science Congress, Lipman was instrumental in arranging a 30-day, 12 000-mile (19 320-km) trip to the west coast of the USA and western Canada. A special train was engaged and it carried over 200 people in 17 railroad cars. The excursion gave delegates the opportunity to make firsthand observations concerning North American soils and make comparisons to European conditions. One individual went so far as to say that the transcontinental excursion was as important to his career as the voyage of the Beagle had been to Charles Darwin.

The End of an Era

In February 1939 Dr. Lipman developed a severe cold and was confined to his home. He was hospitalized several times during the following weeks. Local physicians found his heart completely gone. He died on April 19, 1939. His services to humanity, however, were not forgotten, as exemplified by the following tributes made by various speakers during the dedication of Lipman Hall.

Nobel Laureate Selman A. Waksman said that, with the death of Dr. Lipman, Rutgers completed a period of nearly a century of investigation in the field of soil science, a science which, even if not touched by this group and their coworkers in all its phases, has reached some of the highest pinnacles of its attainment. The investigations of the soil, its origin and

nature, its fertility and its relation to plant growth, its microscopic population, and its microbiological processes are closely interwoven with the other investigations carried out not only in the various institutions in this country but in the world at large. They have served to raise the science of the soil from its beginnings, as a child of geology and as a handmaiden of agronomy to an independent, natural science. We shall provide then for the intensification of production and the support in greater comfort of the growing population of the world. It is the duty of soil sciences to establish more clearly the relation of crops to soils in order that cropping methods may best conserve plant food, solar energy, and human labor. Soil science should teach the farmers how best to modify food quality by soil treatment. It should plan a more far-reaching program of research and education in order that it might serve in a constantly larger way the needs of a human society that from year to year is growing more complex in its organization and more exacting in its demand for such of the products of the land as are essential for maintaining the moral and spiritual values that are the final measure of human activities.

Sir E.J. Russell of England said that there are two ways in which a man may advance science: by his own personal investigations and by organizing investigations of others. And, since nothing succeeds like success, he must ensure success by keeping the whole program well-balanced, favoring none, repressing none. Dr. Lipman followed the above usual sequence and he was successful in all parts he had been called upon to play. His long and faithful service to the New Jersey Agricultural Experiment Station can never be forgotten, nor can his pioneer work in opening up new branches of knowledge.

Further, Rutgers president R.L. Clothier said of Lipman that he saw him as a man endowed with uncompromising intellectual honesty, relentless mental discipline, inspired reasoning, dogged perseverance, and a rare vision of the future of pure and applied research. Clothier also found Lipman to be endowed with the qualities of character and personality that made him a sympathetic friend, a rare raconteur, and a delightful companion.

See also: **Jenny, Hans**; **Marbut, Curtis Fletcher**; **Waksman, Selman A.**

Further Reading

Allison RV (1935) Jacob G Lipman as teacher and Director of Research. *Soil Science* 40: 31–36.

Blair AW (1935) Jacob G Lipman and New Jersey Agriculture. *Soil Science* 40: 25–29.

Martin WH (1952) *Lipman Hall*. New Brunswick: Rutgers University.

Russell EJ (1935) Jacob G Lipman and Soil Science. *Soil Science* 40: 3–7.

Waksman SA (1935) Jacob G Lipman as an investigator. *Soil Science* 40: 11–25.

Waksman SA (1966) *Jacob G Lipman: Agricultural Scientist and Humanitarian*. New Brunswick: Rutgers University Press.

Woodward CR and Waller IN (1952) *New Jersey's Agricultural Experiment Station 1880–1930*. New Brunswick: Rutgers University.

LOESS

A J Busacca and M R Sweeney, Washington State University, Pullman, WA, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Loess is a terrestrial deposit of eolian (wind-blown) dust composed predominantly of silt-sized particles. Most loess has been altered weakly to strongly by soil-forming processes as it has accumulated and its unique properties create highly fertile agricultural soils that are some of the most productive in the world. Loess covers large land areas, most often downwind of the meltwater streams of large Pleistocene glaciers, although loess is also found downwind of some large and small desert areas. Erosion of loess soils via wind and water can be severe and reduces crop productivity; however, the global redistribution of small amounts of fine dust by wind to areas of highly weathered and leached soils has recently been recognized as adding nutrients that allow these soils to sustain complex ecosystems throughout the world. Buried soil sequences within thick deposits of loess can be used to reconstruct the history of climate and ecological changes during the Pleistocene and Holocene Epochs, covering approximately the last 2.4 Ma of earth history.

Distribution

Loess covers approximately 10% of the earth's land surface (Figure 1) and is generally associated with semiarid to semihumid climate regimes. The thickest deposits are found in China, where approximately 500 000 km^2 is blanketed by loess that is up to 300 m thick. The Great Plains has the largest area of loess in North America. Loess there can be more than 60 m thick. Other loess deposits include the Pampas of Argentina, parts of Europe, the Ukraine, large areas in central Asia, the Palouse and Snake River Plain of the US Pacific Northwest (PNW US), parts of Alaska, and parts of northern Africa. Thinner and less extensive deposits are found worldwide.

Origin of Silt and Eolian Transport of Loess

There are three main processes responsible for the formation of loess: (1) production of silt, (2) wind sufficient to entrain the silt into suspension in the atmosphere, and (3) suitable trap areas where loess can accumulate owing to surface soil moisture or vegetative cover.

Glacial grinding is a dominant producer of fine-grained sediment; however, silt is also produced by fluvial and eolian abrasion, frost-shattering, salt-weathering, chemical-weathering, biological processes, and tectonic processes (i.e., overburden release). Since loess is sometimes found downwind of dune fields, it has been argued that grain-to-grain collision during saltation of sand creates silt particles, but this is only a minor contributor to dust production. Silt, produced by one of the above mechanisms, is commonly transported from its place of origin and redeposited before it is entrained by wind to form loess. Models for different types of loess help to explain its genesis. High-latitude loess is the most extensive and includes periglacial and perimontane loess. Periglacial loess is derived from silts produced by glacial grinding that are transported by water on to outwash or braidplains that are subject to redistribution by wind erosion. Perimontane loess is associated with high rates of erosion in mountainous areas that supply silt for adjacent loess accumulations. More recently recognized, peridesert or 'desert' loess is derived from the eolian reworking of sediments from alluvial fans, dry lake beds, or fluvial channels in deserts. Loess accumulates at the semiarid margins of the desert, where an increase in vegetative cover acts as a trap for the silt. Peridesert loess deposits tend to be thinner, sandier, and less continuous compared to high-latitude loess.

Mineral particles can be entrained by the wind by several processes. Particles can be entrained when a threshold velocity is exceeded, resulting in aerodynamic lift of the particle, or by fluid drag exerted on

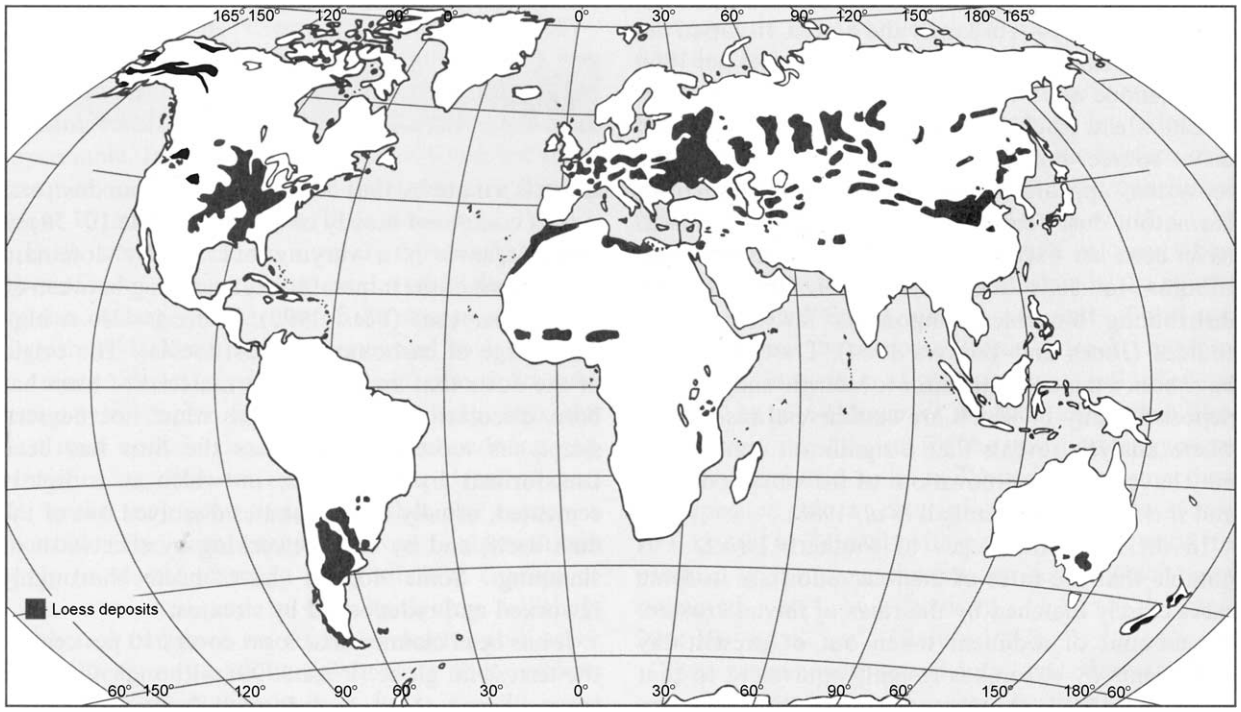


Figure 1 World distribution of major loess deposits. Adapted from AEOLIAN GEOMORPHOLOGY by Livingstone/Warren. © Reprinted by permission of Pearson Education, Inc., Upper Saddle River, NJ.

the particle by air. Particles may also be ejected into the air when they are struck by sand-sized grains. Silt- and clay-sized particles have electrostatic binding forces that can cause soil crusting and aggregation and inhibit their entrainment by the wind. (*See* **Erosion:** Wind-Induced.) For this reason, silt and clay are generally entrained as sand-sized aggregates or ejected by grain impacts. Entrainment is hindered by surface crusting, surface armoring (i.e., gravel lag), and surface soil moisture. Disturbance by animals and humans may aid in the deflation of dust.

Once silt and clay are released into the air column, transport is dominantly by suspension. While coarser sediments such as sand are transported close to the bed via creep (rolling) or saltation (jumping), the smaller silt and clay particles are carried vertically higher into the air column by turbulent eddies. Coarse silt and very fine sand are carried in short-term suspension and are deposited within tens of kilometers of the source. Finer silts and clays are carried in long-term suspension for tens to hundreds of kilometers or farther.

Deposition of dust particles out of suspension to form loess occurs when: (1) wind velocity and turbulence decreases; (2) particles are trapped by vegetation, moist soil surfaces, or uneven topography; or (3) particles are washed out of the atmosphere in precipitation events. If dust is deposited on a bare, dry surface, it is likely to be reentrained by the wind, because there is no means to hold it in place.

Sedimentology

Loess is composed predominantly of particles that range in size from 20 to 60 μm. Loess usually has a unimodal grain size distribution, is poorly sorted, and is finely skewed. Loess with more than 20% sand-sized particles can be called 'sandy loess' and that with more than 20% clay 'clayey loess.' Grain size of any loess is a function of the proximity to the dust source and the degree of weathering during soil formation. Loess near its source contains higher percentages of sand and coarse silt. Mean grain size typically decreases downwind, with a concomitant increase in fine silt and clay. This trend is a result of sorting of particles by the wind and decrease in wind velocities downwind. In some loess areas such as the Palouse of the PNW US, downwind sorting and fining are accompanied by increases in density of native vegetation and increased precipitation and surface soil moisture.

The color of loess ranges widely, depending on source mineralogy and weathering; it can be light to dark gray, yellow, pale to dark brown, red, or black. (*See* **Morphology.**) Mineralogy of loess reflects the source from which it is derived. Generally, loess is composed predominantly of quartz grains (50–70%) with feldspars, micas, carbonates, clay minerals, heavy minerals, and volcanic tephra. (*See* **Minerals, Primary.**) Clay minerals may be detrital (inherited from the source rocks) or formed *in situ* by weathering during soil formation. Carbonate minerals are often present in dust and may comprise more than 30% of loess. Inherited carbonate grains may be dissolved and precipitated to form pedogenic calcic horizons in semiarid to arid climates.

On a macroscopic scale, loess generally lacks sedimentary structures such as bedding and thus has a homogeneous appearance. Soil horizons can give exposures of loess a pseudobedded appearance. Frequent wetting and drying may result in formation of prismatic structure in loess. Closer detail of loess sedimentology may reveal fluctuations in the percentage of sand, silt, and clay with depth in the deposit that may reflect changes in wind intensity or in source areas through time. Rarely, laminations of coarse and fine dust have been preserved in loess that are interpreted to represent seasonal fluctuations in wind strength during times of very high dust accumulation rates that outpaced the mixing processes of soil organisms.

Stratigraphy

Loess accumulates in layers or 'strata' and soils commonly form at the top of a loess stratum. The soils become buried with continued deposition of dust, creating 'stratigraphic sequences' of alternating loess and buried soils (paleosols). Loess sequences such as those in China, the US, and in many other areas have been utilized to reconstruct the history of the Quaternary Period. The common occurrence of loess-paleosol sequences has been interpreted to represent episodes of high rates of dust accumulation during which a loess layer is deposited, alternating with episodes of low rates of dust deposition during which soils develop. Micromorphological techniques applied to loess and paleosols (microscope studies of slides of undisturbed materials) have shown that the idea of sequences recording either periods of dust accumulation or soil formation is misleading. In fact, soil-forming processes occur throughout the deposition of dust, and accumulation rates and climate affect the intensity of soil-forming processes (**Figure 2**).

Correlation of loess sequences across regions has been hampered by the lack of reliable dating techniques. Luminescence dating is a relatively new radiometric technique that dates the last exposure of the dust grains to sunlight. It can provide ages up to approximately 100 000 years or older with analytical

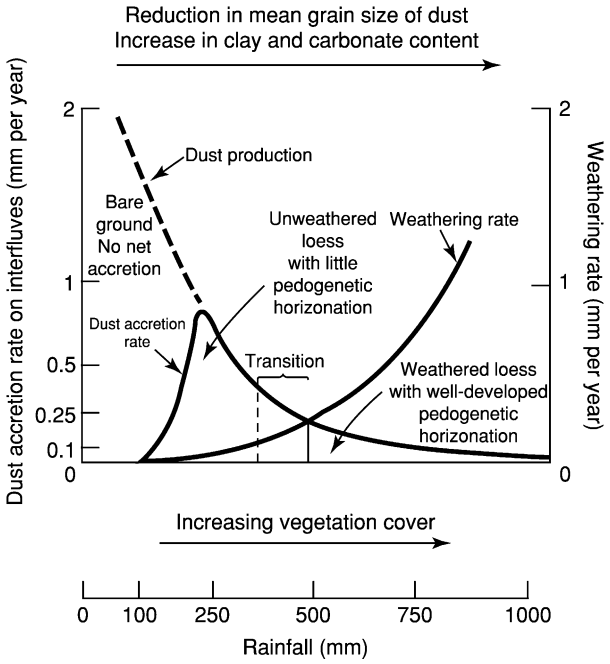


Figure 2 Schematic model showing increase in pedogenesis of loess with an increase in moisture and vegetation and decrease in dust accumulation rate and grain size. Reproduced with permission from Pye K and Tsoar H (1987) The mechanics and geological implications of dust transport and deposition in deserts with particular reference to loess formation and dune sand diagenesis in the northern Negev, Israel. In: Frostick L and Reid I (eds) *Desert Sediments: Ancient and Modern*, pp. 139–156. Special Publication No. 35. Geological Society.

errors as low as 5%. In some cases, prominent paleosols can be used to correlate sequences. Volcanic tephras in loess can be traced over large areas and can have unique chemical and mineralogical fingerprints that aid in correlation.

Chinese loess is as old as 7 million years, but the bulk of the loess there is less than 2.4 million years old. Loess deposition in China has been linked to the uplift of the Tibetan Plateau that may have resulted in climates more favorable to loess formation. Cycles of accumulation of loess followed by soil formation have been attributed in China to variations in the strength of the summer (rain bearing) compared with winter (dust-storm bearing) monsoons, and these have been correlated in turn to Milankovitch cycles (cycles in Earth's orbital parameters) that are thought by some to be the driver of the ice ages. Loess in the Great Plains and Europe show cycles that follow Pleistocene glaciations, with major loess accumulation during the dry and windy glacial periods, followed by soil formation during the more humid and less dusty interglacials.

Early scientists who studied loess were emphatic that loess was purely a glacial phenomenon. Some more recent studies show that there are cases where the opposite has occurred. Study of the Palouse loess indicates that maximum rates of loess deposition occurred during interglacials, while soil formation was dominant during glacials. The majority of the silt is glaciogenic, but it was transported to basins via catastrophic outburst floods triggered by ice dam failures that occurred toward the end of full glacial periods. The flood sediment was reworked by the wind and deposited as loess during interglacials. When full glacial conditions returned to the area, dust-transporting winds appear to have been suppressed by changes in atmospheric circulation, landscapes ceased aggrading, and soil formation ensued.

Geomorphology

Loess is a blanket deposit that accumulates from suspension fallout of dust particles. Regional patterns of thickness are controlled by a number of factors. Loess strata are thickest closest to the source area, as large volumes of the coarser sediment falls from suspension, leaving smaller volumes of finer sediment to be blown farther downwind. An exponential decrease in loess layer thickness with distance along downwind transects has been documented in many places such as the North American Great Plains (Figure 3). Thickness may be increased locally by the presence of trapping mechanisms such as topography or vegetation that separate the saltation load (sand that keeps the dust entrained) from the suspension load. A persistent source of dust-sized particles for eolian entrainment will also result in thick deposits and continuously aggrading land surfaces. The topographic form of loess landscapes can be quite different from place to place. The topography of the Palouse in the PNW US is a classic example. Loess in the drier, western Palouse (mean annual

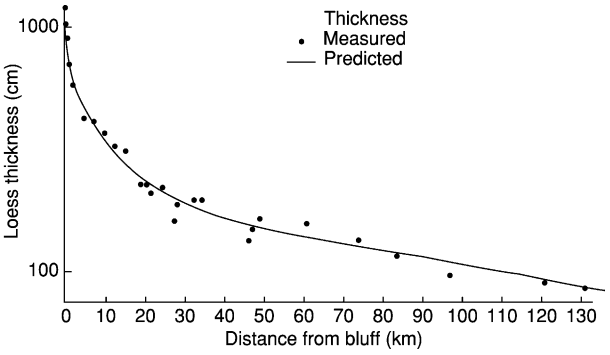


Figure 3 Exponential decrease in loess thickness downwind of the Missouri River source in Illinois, USA. Reproduced with permission from Frazee CJ, Fehrenbacher JB, and Krumbein WC (1970) Loess distribution from a source. *Soil Science Society of America Proceedings* 34: 296–301.

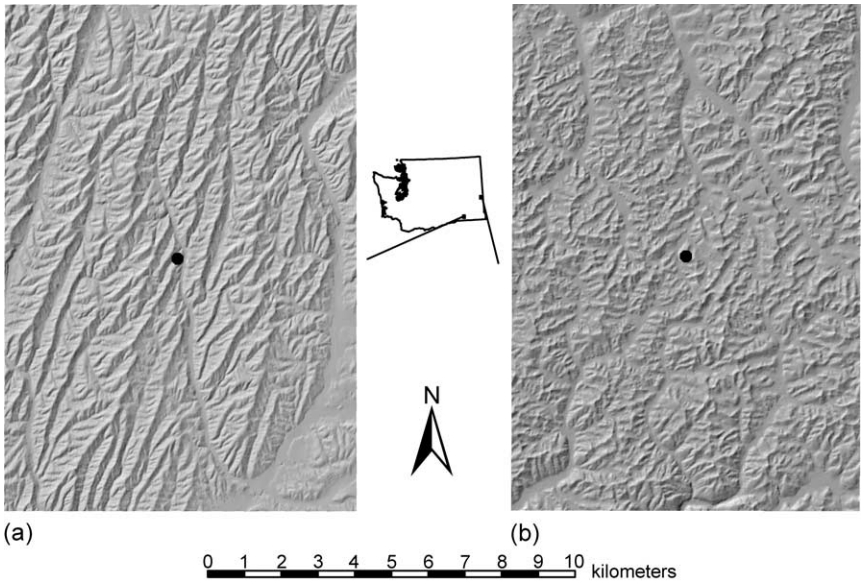


Figure 4 Hillshade models of Palouse topography, in Washington state (silhouette) in figure, based on 30-m digital elevation model resolution. × 2 vertical exaggeration. (a) Distinct, wind-aligned NNE linear ridges, formed during the deposition of sandy loess. Image of Hadley 7.5′ quadrangle; centroid is 118.44° west, 46.19° north; (b) rolling terrain of loess hills shaped predominantly by water erosion in the higher precipitation zone of the Palouse. Image of Garfield 7.5′ quadrangle; centroid is 117.19° west, 47.06° north.

rainfall less than 450 mm) forms a distinctive landscape of NNE-trending linear ridges and valleys (Figure 4a). This linear topography, called 'greda' or 'paha,' has also been observed in the loess of Hungary and the North American Great Plains. Linear ridges commonly form where loess has a more sandy texture. Local topographic highs may influence the deposition of loess in the lee of any preexisting topographic obstacles, resulting in streamlined ridges aligned with the prevailing wind direction. Other workers have argued that loess will accumulate on the windward side of topographic highs owing to deceleration of the wind, but this only may occur in the absence of sand particles. Loess in the wetter eastern Palouse (mean annual rainfall up to 800 mm) forms a dissected terrain of rolling hills with 'cirque-like' amphitheaters and a dendritic drainage network (Figure 4b). There, the combination of greater soil moisture, greater incidence and severity of soil freezing, and existence of buried soils with water-restricting, clayey argillic horizons create a landscape shaped by water erosion on frozen soils and rotational slumps and mass movements.

Loess can also blanket preexisting landforms. In eastern Europe, loess commonly blankets flights of fluvial terraces. Pseudoanticlines are described where loess has accumulated on undulating topography (Figure 5). Where dust accumulates in an area of higher relief, loess thickness can be greater in the valleys due to colluvial or alluvial reworking of loess off the hillslopes.

Figure 5 Thick accumulation of Palouse loess representing approximately 100 ky of deposition. Note the 'pseudoanticline' and the presence of buried soils represented by more resistant, white petrocalcic horizons. Photo by M. Sweeney.

Paleosols in Loess and Studies of Paleoclimate

Loess is considered to be one of the better terrestrial proxies for reconstructing paleoclimate for the Quaternary Period (the last 2.4 million years). This is at least in part because loess deposits aggrade through time, so that the land surface rises as dust accumulates and aggradation results in the preservation of former surface soils, or paleosols.

When soil formation keeps pace with dust deposition, soil horizons such as calcic, cambic, or argillic horizons form upward over time within the

accumulating loess. In situations where loess deposition and soil formation are more or less continuous, buried soils may be indistinct and hard to recognize, because subsoil (B horizon) features are superimposed on topsoil (A horizon) features. 'Welded soils' occur where increments of loess are insufficiently thick to separate episodes of soil development completely, resulting in partial overlap of soil horizons and modification of the properties of buried soils. Only where episodes of dust deposition are episodic and new loess layers are thicker than typical soil profiles are buried soils with distinct A–B–C horizon sequences likely to be preserved.

There are several properties of paleosols that can provide clues to past climates including soil morphology, stable isotopes in pedogenic carbonates and soil organic matter, magnetic susceptibility, preservation of plant opal phytoliths, and plant pollen (Figure 6).

The formation of a soil in loess itself is very informative, as it implies a period of relative landscape stability and decreased dust deposition. Specific soil features can be key indicators of past climates. For example, calcic (Bk), petrocalcic (Bkm), or duripan (Bqm) horizons suggest arid climates, whereas argillic (Bt) horizons may indicate higher precipitation that drove clay illuviation processes, and redoximorphic features suggest changing soil moisture conditions with periodic anaerobism. (*See* **Morphology**.) Platy soil structure may be indicative of freezing conditions. Faunal fabrics formed by burrowing organisms have also provided clues to past ecological conditions. Palouse paleosols often contain horizons rich in cylindrical peds that have been interpreted as burrows formed by cicada nymphs, whereas surface soils at the same sites lack burrows. Recent studies have shown that cicadas reached maximum numbers in association with a periglacial *Artemisia* (sagebrush) shrub steppe. The cicada-burrowed horizons, along with evidence from phytolith assemblages, suggest that these paleosols formed under cold, arid, full-glacial climates. Cicada-burrowed, late-glacial paleosols abruptly grade upward into worm-burrowed horizons of the modern surface soil associated with perennial grass-steppe vegetation and moister, warmer interglacial climates.

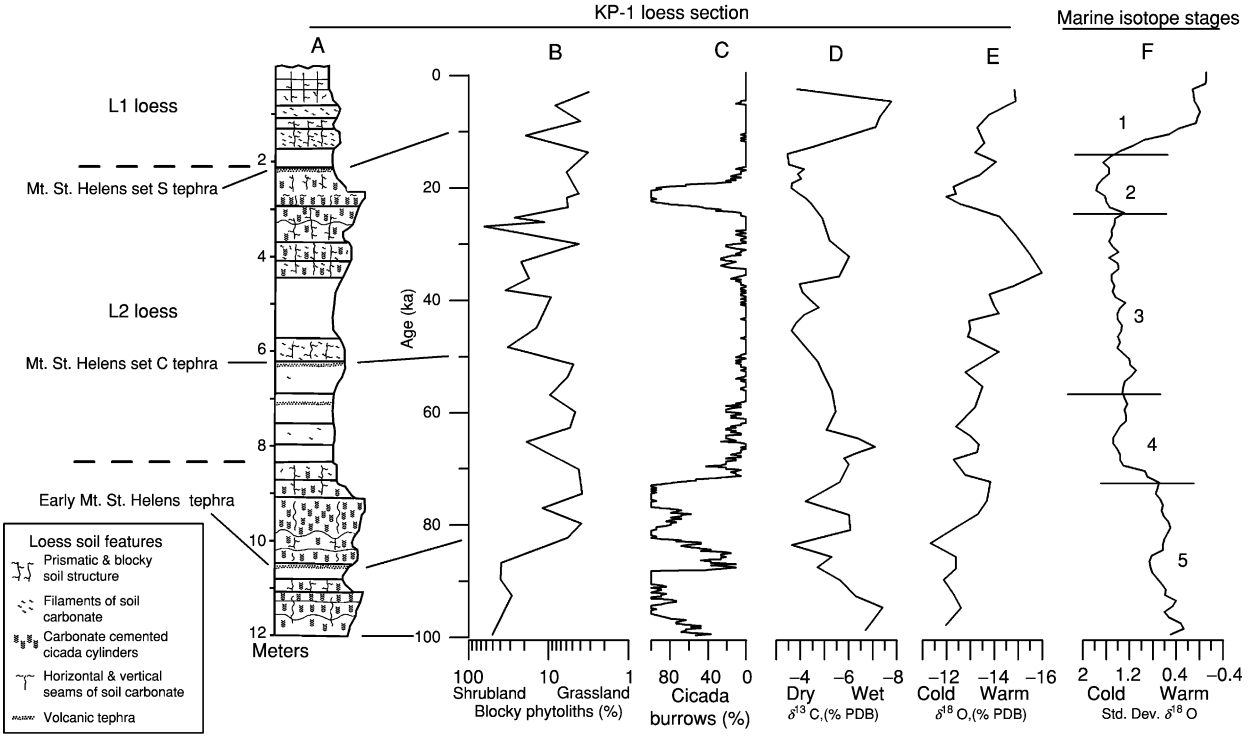


Figure 6 KP-1 loess (Palouse loess, US PNW) stratigraphy and paleoclimate proxies. A, loess stratigraphy and features of buried soils. Age–depth model of loess was derived from luminescence age dating. B, relative percentage of shrub (arid *Artemisia*) versus grass (moist perennial grasses) cover through time based on phytoliths preserved in the loess. D and E, change of carbon and oxygen isotopes (respectively) extracted from pedogenic carbonate. Palouse vegetation is approx. 100% C3 type. Carbon isotope signature is a proxy for paleowetness. Oxygen isotope signature is a proxy for paleotemperature. F, oxygen isotope curve obtained from marine sediments, separated into major isotope stages, with odd numbers representing interglacials and even numbers representing glacials. Cold and arid periglacial conditions in the Palouse indicated by relatively high abundance of shrubland phytoliths, cicada burrows, and positive isotopic signatures. Adapted with permission from Blinnikov M, Busacca A, and Whitlock C (2001) A new 100 000-year phytolith record from the Columbia Basin, Washington, USA. In: Meunier JD and Colin F (eds) *Phytoliths: Applications in Earth Sciences and Human History*, pp. 27–55. Lisse, The Netherlands: AA Balkema. © A.A.Balkema.

Carbon isotopes extracted from soil carbonate have been used as a proxy for temperature and moisture conditions. The isotopic signature can indicate the proportion of plants using the C_3 or C_4 metabolic pathway or, as in the example in **Figure 6**, signal changes in soil respiration rate in a C_3 plant community resulting from changes in precipitation. Oxygen isotopes also retrieved from soil carbonates and soil organic matter may provide information on paleotemperatures. Caution must be advised when interpreting carbon and oxygen isotope data from soil carbonates unless one can assess the possibility of atmospheric, groundwater, or other sources of contamination.

Magnetic susceptibility (MS) has been an important tool in correlating loess-paleosol sequences to deep-sea oxygen isotope records of paleoclimate. During soil formation, magnetic particles can concentrate via dust fall or due to the production of magnetite and maghemite via pedogenic processes, resulting in higher MS signals in paleosols than in unaltered loess. The depth function of MS in long loess-paleosol sequences in China in some cases matches strikingly well with paleoclimate records from deep-sea sediment cores.

Loess is generally too oxidizing for the preservation of plant matter, including pollen, but in some cases pollen analysis has been utilized. Phytoliths composed of opaline silica are formed in plant cells and are preserved in loess paleosols after the plants decay, providing a local, accurate assessment of past plant assemblages. Distinctive shapes can in some cases allow the identification of different genera of grasses, unlike pollen analysis. Phytolith analysis has been demonstrated as an important tool to reconstruct paleoclimate in the Palouse loess (**Figure 6**), New Zealand, and the Great Plains. Local assemblages can be compared with other local or regional proxy records of plant communities.

Surface Soils in Loess

European workers have repeatedly emphasized that loess is more than a sedimentary deposit; that its properties are a result of pedogenic weathering processes that impart porosity, unique structure, tawny color, and that reorganize calcium carbonates.

Large areas of loess soils such as the US Great Plains, Kazakstan, and the Ukraine, are in semiarid climates that supported grass-steppe vegetation prior to agricultural development. As a result, loess soils are dominated by the Mollisol soil order (US Soil Taxonomy System). (*See* **Civilization, Role of Soils**.) Where loess has accumulated in a forest setting or where hardwood or mixed forests succeeded on to loess areas after dust deposition ceased, gray moderately leached Alfisols with a clay-enriched argillic subsoil horizon are common. Soils formed in loess of true desert climates produce Aridisols with pale topsoil horizons and subsoil calcic horizons at shallow depth.

Historically it has been observed that loessial soils are very productive compared with soils formed in other parent materials. This is because silty soils provide high porosity and high water-holding capacity, which affords good aeration and high levels of plant-available water, relative ease of root penetration, and a stone-free seedbed. In comparison, sandier soils retain less moisture, and clayey soils retain more total moisture but have less plant-available water than loessial soils. Clayey soils pose additional problems, including limited depth of rooting and excessive wetness. High native fertility of loess soils is most notable in Mollisols, where nitrogen (N) is supplied by the mineralization of humus in the early years of cultivation; however, this must be supplemented by fertilizer N for continued high yields of cereal crops. Soil organic matter in loessial agricultural soils in the PNW US is less than half of native levels because of increased mineralization rates due to annual tillage and because of tillage erosion and water erosion. Loss of organic matter can lead to poor soil structure and increased soil crusting, which in turn can reduce seedling emergence and increase water erosion.

Environmental Concerns

Wind and water erosion have significantly degraded loessial soils worldwide. Erosion results in the loss of topsoil and soil nutrients, which have negative effects on productivity. Eroded soil materials can seriously degrade off-farm surface water, groundwater, and air quality. The North American Great Plains suffered severe wind erosion during the 'dust bowl' drought years of the mid-1930s that resulted in huge economic losses. In the Palouse area of the PNW US, wind erosion of agricultural fields has resulted in a 10-fold increase in atmospheric dust since the late 1800s. As of 2003, wind erosion from agricultural fields is a major producer of particulate matter less than 10 μm in size (PM_{10}), which poses significant heath risks to people of the region. Farmers have attempted to decrease wind erosion on fallow fields by leaving stubble, which increases surface roughness, and by the implementation of annual cropping. (*See* **Windbreaks and Shelterbelts**.) Loessial soils that are most prone to wind erosion are those with less than 1% soil organic matter content and less than 10% clay content.

Water erosion is a serious problem in loessial soils. Soil loss of up to 450 tonnes ha^{-1} per year has been

recorded in autumn-planted winter wheat, where surface protection from rainsplash is minimal over winter. The erosion hazard can be increased by surface crusting and by frozen soils. Surface crusting occurs during rainfall, when raindrop impact detaches particles from soil aggregates and orients platy minerals such as micas. Crusting inhibits infiltration and encourages overland flow of water, which enhances erosion. Frozen soil similarly inhibits or prevents infiltration of rainwater. Snowmelt on frozen loessial soils may cause liquefaction of the thawed surface layer. (*See* **Crusts:** Structural.)

Unweathered loess has high porosity of 40–55% that decreases during pedogenesis due to translocation of particles, precipitation of minerals, and collapse of pores. Dry loess can form vertical walls several meters high, because it has open microfabric, cohesion that may be attributed to cementation of particles by carbonate, clay bridges between silt particles, preferential grain orientation, or attractive forces between small particles. Loading may result in the collapse of open pores, and frequent wetting may wash clay bridges from the silt particles, allowing failure of loess. Subsurface water flow results in piping, and the excavation of these pipes through time may result in collapse. Thick loess in humid regions is typically deeply gullied with high, steep walls. Tensile fractures may be common and can result in failure. Large landslides triggered by seismic activity have been documented in the thick Chinese loess.

Dust Additions to Other Soils

In places where dust accumulation rates are not large enough to result in loess deposits, dust additions still may be of great importance to soil fertility. Dust additions to soils were originally recognized by the presence of minerals such as quartz in places where the bedrock such as basalt contained none. Terra Rossa soils on limestones have been argued to form by very slow dust accretion over eons rather than by residual enrichment of noncarbonate minerals in the limestone. Large dust plumes from Asia and Africa have been well documented to travel in the upper atmosphere around the globe and deposit thousands of kilometers from their source areas. Far-traveled Asian dust is a major contributor to soils in Japan, Korea, and Hawaii and has traveled as far as North America. The same can be said of Saharan dust, which contributes to soils of the Canary Islands and southern Europe. Saharan dust plumes also make their way to South and North America.

Biogeochemical studies of ecosystem development have shown that, in areas of highly weathered soils, productivity is sustained by the deposition of exotic mineral dust, which can supply critically limiting nutrients such as phosphorus and potassium. Rainforest soils of the Amazon in South America include highly weathered Ultisols and Oxisols. These old soils can no longer provide nutrients from weathering of parent material, yet they sustain a diverse ecosystem. It has been shown that African dust is a significant source of nutrients to the Amazon Basin. A similar situation exists for old, weathered tropical soils of Hawaii, which receive an influx of nutrients from Asian dust. Dust has been shown to be a major source of N and P and organic matter in soils of Australia and Israel. Soluble salts in dust may have an impact on soil pH.

On a geologic time scale, additions of calcium carbonate and silica from dust greatly aid the formation of duripans and petrocalcic horizons (cemented horizons of silica and calcium carbonate, respectively). (*See* **Dryland Farming; Productivity; Remote Sensing:** Organic Matter; **Sand Dunes**.)

List of Technical Nomenclature

Alfisol	A soil order in the US system of soil classification; a mineral soil, generally forested, that has a light-colored surface horizon and clay-enriched subsurface horizon that is not highly depleted in plant nutrients
Argillic horizon	A subsurface mineral soil horizon characterized by the illuvial accumulation of clays; a subsurface horizon of clay enrichment by the downward movement of clays
Aridisol	A soil order of the US system of soil classification; a mineral soil that formed in a desertic climate and that has some development of soil horizons, often a calcic or duripan horizon
Biogeochemical	In reference to systems influenced by a combination of biological, chemical, and geological processes
Calcic horizon	A subsurface mineral soil horizon characterized by enrichment in calcium carbonate, often by the downward movement of ions with precipitation of the carbonate in a lower horizon
Cambic horizon	A subsurface mineral soil horizon characterized by the alteration or removal of mineral materials as indicated by mottling or gray colors, or by more intense or redder colors than the parent material from which the soil formed, or by the removal of carbonates

Duripan A subsurface mineral soil horizon cemented by illuvial silica; a type of hardpan

Eolian In reference to entrainment, transport, and deposition of sedimentary particles by the wind

Greda A European term for a linear topographic ridge in a loess-covered landscape

Holocene An epoch of the Quaternary period beginning at the end of the ice ages approximately 10 000 years ago and extending to the present day

Horizon A layer of soil, usually parallel to the land surface, with a distinctive set of physical, chemical, and biological properties and differentiable from horizons above or below it by differences in appearance and properties

Illuviation A soil-forming process involving the movement of soluble or particulate soil materials from one horizon downward into another horizon (cf. argillic horizon, calcic horizon)

Luminescence dating A radiometric age-dating technique that measures the time elapsed since a mineral grain was last exposed to sunlight. Most often applied to age-dating a sample of eolian sediment

Magnetic susceptibility The ratio of induced magnetization to the strength of the magnetic field causing the magnetization

Micromorphology The study of undisturbed soil samples by light microscopy from specially prepared microscope slides, called thin sections

Mollisol A soil order of the US system of soil classification; a mineral soil that has a dark-colored humus-rich surface horizon and one of several kinds of subsurface horizons

Oxisol A soil order of the US system of soil classification; a mineral soil that has an oxic horizon, which is highly weathered, infertile, and is highly enriched in kaolin clays or in iron and aluminum oxide minerals

Paleoclimate Referring to climate conditions of times prior to today

Paleosol A buried soil profile formed in an environment of the past or a surface soil that has undergone weathering during several climatic episodes

Pedogenesis The physical, chemical, and biological processes of soil formation

Petrocalcic horizon A subsurface mineral soil horizon characterized by cementation by calcium carbonate

Phytolith Microscopic structures of opaline silica formed within the tissues of living plant. Phytoliths extracted from soils or sediments are sometimes used like fossil pollen to infer vegetation or climates of the past

Pleistocene An epoch of the Quaternary period that began approximately 2.4 million years ago and ended approximately 12 000 years ago. A time period corresponding to the ice ages

Porosity The ratio of the volume of void or pore space to the total volume of a rock or sediment

Quaternary A period within the Cenozoic era consisting of the Pleisocene and Holocene epochs

Saltation The transport of sand-sized sediment particles in which grains are moved by wind in a series of short leaps or bounces

Stable isotopes One of two or more species of the same nonradiogenic chemical element having the same number of protons but differing in the number of neutrons. The isotopes of an element have slightly different physical and chemical properties owing to their mass differences

Strata Layers of sedimentary rock

Suspension The transport of generally silt-sized and finer sediment particles in air whose velocity or turbulence allows the particle to remain for a time aloft

Tephra Any rock materials ejected from a volcano and transported through the air, including pumice and ash

Ultisol A soil order of the US system of soil classification; a mineral soil that has a clay-enriched subsurface horizon that is highly depleted in plant nutrients

Welded soil Two or more soils of different ages in a depositional sedimentary system that are superimposed on one another and partially overlap

See also: **Civilization, Role of Soils**; **Dryland Farming**; **Erosion:** Wind-Induced; **Minerals, Primary**;

Morphology; **Productivity**; **Remote Sensing:** Organic Matter; **Sand Dunes**; **Windbreaks and Shelterbelts**

Further Reading

Blinnikov M, Busacca A, and Whitlock C (2001) A new 100 000-year phytolith record from the Columbia Basin, Washington, USA. In: Meunier JD and Colin F (eds) *Phytoliths: Applications in Earth Sciences and Human History*, pp. 27–55. Lisse, The Netherlands: AA Balkema.

Busacca AJ, Beget J, Muhs DR *et al.* (2003) Eolian sediments. In: Gillespie AR, Porter SC, and Atwater BF (eds) *The Quaternary Period in the United States. Developments in Quaternary Science 1*, pp. 275–310. Amsterdam: Elsevier.

Catt JA (2001) The agricultural importance of loess. *Earth Science Reviews* 54: 213–229.

Chadwick OA, Derry LA, Vitousek PM, Huebert BJ, and Hedin LO (1999) Changing sources of nutrients during four million years of ecosystem development. *Nature* 397: 491–497.

Frazee CJ, Fehrenbacher JB, and Krumbein WC (1970) Loess distribution from a source. *Soil Science Society of America Proceedings* 34: 296–301.

Kemp RA, Derbyshire E, Xingmin M, Fahu C, and Baotian P (1995) Pedosedimentary reconstruction of a thick loess-paleosol sequence near Lanzhou in north-central China. *Quaternary Research* 43: 30–45.

Livingstone I and Warren A (1996) *Aeolian Geomorphology*. Singapore: Addison Wesley Longman Ltd.

Muhs D and Zárate M (2001) Late Quaternary eolian records of the Americas and their paleoclimatic significance. In: Markgraf V (ed.) *Interhemispheric Climatic Linkages*, pp. 183–216. London: Academic Press.

Pye K (1987) *Aeolian Dust and Dust Deposits*. London: Academic Press.

Pye K and Sherwin D (1999) Loess. In: Goudie AS, Livingstone I, and Stokes S (eds) *Aeolian Environments, Sediments and Landforms*, pp. 213–238. Chichester, UK: Wiley & Sons.

Wright JS (2001) 'Desert' loess versus 'glacial' loess: quartz silt formation, source areas and sediment pathways in the formation of loess deposits. *Geomorphology* 36: 231–256.

LOWDERMILK, WALTER CLAY

J D Helms, USDA Natural Resources Conservation Service, Washington, DC, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Walter Clay Lowdermilk (Figure 1) often described his profession as "reading the records which farmers, nations, and civilizations have written in the land." While others have belonged to this profession, certainly few had the inclination, ability, and opportunity to indulge in it, as did Lowdermilk. The profession required expertise in many fields of study, but as practiced by Lowdermilk it was not a purely academic exercise. Rather, he sought an ambitious objective – a permanent agriculture for the world. He had begun his career researching the basic processes of erosion and runoff, especially as influenced by humans. Lowdermilk became a member of the early-twentieth-century conservation movement in the USA, a movement with a strong scientific bent. Its scientists held that treatment of natural resources should be in accordance with scientific principles, not propelled by emotionalism or untested theories. A leader in organizing government conservation agencies in the USA and abroad, Lowdermilk put research findings into action. He enunciated plans for agricultural development in *Palestine, Land of Promise*. Later, he worked in Israel to write the plans on to the face of the countryside. As an author he spoke for the conservation

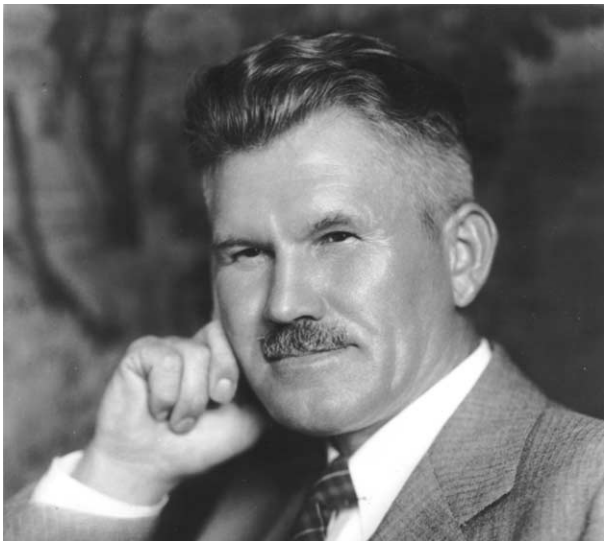


Figure 1 Walter Clay Lowdermilk.

movement, and in his international work served as a humanitarian by furthering conservation and the proper use of land to make and keep it productive.

Lowdermilk, Walter Clay (1 July 1888 to 6 May 1974), soil scientist, geologist, soil conservation leader, and author, was born in Liberty, North Carolina, the son of Henry Clay Lowdermilk, a businessman, lumberman, and rancher, and Helen Vashti Lawrence Lowdermilk. The family moved westward to Missouri, to Oklahoma, and finally to Arizona. Walter Lowdermilk graduated from the Park College Academy in Parkville, Missouri, in 1906 and then attended Park College (1908–10). In 1910 he enrolled at the University of Arizona; after 2 years there he won a Rhodes Scholarship to Oxford University, where he earned a BS degree in forestry (1914); a BA degree in geology (1915); and an MA degree, granted *in absentia* (1922). While at Oxford he had an opportunity to study forestry in Germany. He also served on Herbert Hoover's Commission for Relief in Belgium (1914–15). After the scholarship years, he returned to the USA as a ranger in the Southwest for the Forest Service. During World War I, he returned to Europe as timber acquisitions officer and then measured timber for assessing war-time damages. Returning from World War I, he became the Forest Service's regional research officer in Montana.

A man who enjoyed research work, he had found a position that offered satisfaction. Given his ability, there was opportunity for advancement. But he was not to remain on that career ladder. Soon he would be in China, where, he later recalled, the significance of soil erosion was burned into his consciousness. Through the years in England and afterward, the young forester had corresponded with Miss Inez Marks, a friend from Arizona. On leave from her missionary work with the Methodist Church in China, she agreed to meet him at the Rose Bowl, New Year's Day, 1922. Marriage plans quickly followed. Her entreaties that China desperately needed talented scientists led to his applying for a position with the University of Nanking's school of agriculture and forestry. The couple married in August and departed for China in September 1922. Lowdermilk's charge, for a small salary, was to assist in solving the flooding problems and resulting famines. Exactly how a forester was to help with food production remained a mystery as he attended university classes to learn Mandarin during the first year.

An expedition to the Yellow River solved the mystery. There he stood atop a section of the 650-km-long dike that held the river 12–15 m above the flood plain. This marvel was a result of Chinese labor necessitated by silting of the river's channel – 'aggradation' in the terms of earth scientists. Lowdermilk set out to find the source of the silt. In spring 1924, O.J. Todd, engineer of the International Famine Relief Commission, accompanied Lowdermilk on a 3000-km trip to the watersheds of the Yellow and Wei rivers. Todd's mission was to study the Wei-Peh irrigation project. Few foreigners had visited northwest China, where the pair completed a third of the journey afoot or on mulecart or muleback. In Shensi province, they found a plateau consisting of deep, undulated deposits of loessial soils. Depth, fertility, and erodibility made these fine, wind-deposited soils prime locations for man-induced erosion. In the deforesting activities of the people, Lowdermilk found the reason for the gigantic 200-m-deep gullies. The Chinese shaved the hills clean of all herbaceous shrub and tree growth for firewood. Paradoxes abounded on the trip. Temple forests, reproduced naturally and protected by Buddhist priests, provided evidence of the denuded hills' capability for sustaining vegetation. Bench terraces festooned some slopes. Yet some of the best agricultural land on the level, alluvial plains was used for timber production under irrigation. Surrounding hills were little used for timber.

The pair visited Sianfu, the capital city of China during its Golden Age, where Todd wanted to inspect the irrigation works. The area retained little of its former prosperity, which Lowdermilk conjectured had flowed from a great irrigation project, long since silted up. The forester returned to his post at the University of Nanking with an impression of colossal erosion contrasted with evidences of former grandeur. Already he had decided to expand his study of the sciences involved with natural resources to include the actions of people as well as man's relationship to nature.

Historical research revealed that the Yellow River had changed course eight times since AD 11. Several times the river had been restrained by dikes, only to break free. Once it emerged 600 km from its former outlet. Dikes, therefore, were essential to using the plain for agriculture. But building higher dikes, Lowdermilk concluded, was not a lasting solution unless the aggradation of the river was reduced by checking the supply of silt. Lowdermilk's supposition that erosion caused frequent and severe flooding had been recognized in the USA, but only on the small water courses, not on the lower reaches of major rivers. The China experience – siltation of a major river channel as a cause of flooding and channel relocation – was on a scale unknown in the USA.

Lowdermilk's recommendation for flood control gave some indication of the breadth of his training in sciences, especially geology, and his ability to assimilate the findings into a solution. The Yellow River and its tributaries had excavated a deep channel into

the plateau created by the wind-deposited soils. Recognizing that removal of vegetation allowed runoff to carve gullies in the loessial plain and that gully wash accounted for most of the silt, he proposed attacking erosion by planting trees on the talus slopes at the foot of the gullies. The forested gullies would be guarded and managed by villages as community forests to provide wood. Undissected portions of the loessial plateau could be used for agriculture. Where and when possible, check dams should be used to raise the base level of streams and prevent incision by the gullies farther into the plateau. Treatment of the watershed was directly tied to famine prevention. He concluded that soil and water conservation were urgently necessary to increase the productivity of this region of China.

Lowdermilk was not content to base his recommendations exclusively on empirical evidence. Certainly the scientific forestry school, whence he came, demanded another explanation. Using the runoff and erosion plot method devised by F.L. Duley and M.F. Miller at the University of Missouri, he and his Chinese associates set up plots on 20 temple forests and on denuded areas for comparison. After 3 years of study, he presented the findings. Runoff from denuded areas greatly exceeded that of temple forests or areas reclaimed through reforestation. The main reason for the excess runoff, he believed, was that particles of soil on the denuded areas clogged the pores of the soil surface. Forest litter arrested this action.

Further study convinced Lowdermilk that 40–60% of the uplands in northern China had little cover to retain runoff. So great had been the rapid runoff that it had reduced evaporation and brought on a period of decreased precipitation in the area. With this argument, Lowdermilk projected a hypothesis that he would later apply to other lands. Scholars had long been presented with anomalies of twentieth-century poverty contrasted with evidences of former civilizations which possessed a high degree of culture and prosperity. Some scholars, notably Ellsworth Huntington and Baron Von Richthofen, found the answer in climatic change. In the case of north China, Lowdermilk not only saw soil erosion and flooding as the reason for decline, but also claimed their effects as the reason for a climatic change.

The communist uprising of 24 March 1927 in Nanking ended the Lowdermilks' stay in China. Leaving behind all possessions, they barely escaped. At the University of California, Lowdermilk combined study for a PhD from the School of Forestry (minors in soil science and geology) with research at the California Forest Experiment Station. Here he reentered the fray over the effects of vegetative cover on runoff, erosion, and flooding. On one of his treks in China, Lowdermilk had heard the proverb "Mountains empty – rivers gorged." He judged the application of timber management in that locale to be superior to any system he had observed in Germany. The Chinese and other civilizations had recognized the value of forest cover and acted upon their observations. Scientists in the conservation movement, however, demanded more than proverbs for proof, and the influence of forest cover on soil erosion and streamflow continued to be debated by hydrologists, engineers, and foresters.

In his study of the influence of forest litter on runoff and erosion, he used rain-making machines, soil profiles transferred to tanks, and measuring instruments of his design. In 1929, he presented the confirmation for what he and others had observed. On bared soil the raindrops splashed up muddy. As muddy water percolated into the soil profiles, suspended particles were filtered out at the soil surface. The thin layer thus formed reduced percolation and increased runoff. The water-absorbing capacity of forest litter had little influence on runoff. However, the litter filtered the water and kept the soil profile open to percolation. The experiments confirmed a hypothesis that Lowdermilk had first presented at the Third Pan-Pacific Science Congress in 1926 in Tokyo. He also proposed terms for classifying erosion. 'Accelerated erosion' arose from the 'artificial disturbance of factors which controlled the development of soil profiles.' In the absence of such disturbances, one could view any erosion as the 'geologic norm of erosion.'

As Lowdermilk was completing his dissertation, another research opportunity developed with the Forest Service in California. Partly in response to the prodding of Hugh Hammond Bennett, a US Department of Agriculture (USDA) soil scientist, Congress, in the Agricultural Appropriations Act of 1930, authorized a group of experiment stations to study erosion and runoff. Bennett supervised most of the stations from his post in the Bureau of Chemistry and Soils, but the Forest Service also received funds for stations. Lowdermilk designed the research station and program on a large, isolated watershed in southern California. The San Dimas watershed provided an excellent opportunity to test the effects of watershed management on water yield. Expanding towns and citrus orchardists at the foot of the watershed had to dig increasingly deeper wells to reach underground aquifers. Whether the vegetative mantle should be burned to reduce transpiration or protected from fire for maximum ground water supplies was a matter of controversy. To demonstrate and measure the relationship of percolation to aquifer levels, Lowdermilk

had Civilian Conservation Corps (CCC) enrollees build water-spreading structures, which led to a gravelly basin where the silt settled out and water percolated to the aquifers.

Experiments at San Dimas also contributed to a better understanding of soil genesis processes such as horizonization and the anthropogenic influence on soils in a densely populated region. CCC enrollees sifted and mixed soil, which was then placed in lysimeters. They also kept soil samples in glass containers. Soil scientist Robert C. Graham and colleagues examined the uncontaminated, glass-enclosed samples and compared them with soils in the lysimeters. Lowdermilk had read and cited the major soil scientist treatises of the day by Harry O. Buckman, Jacob S. Joffe, Jacob Lipman, Curtis F. Marbut, G.W. Robinson, and Nathaniel Southgate Shaler; but he did not specialize in topics such as soil genesis, soil mineralogy, or soil chemistry, all topics which were being pursued in the new soil science departments at land-grant universities. Rather his studies addressed the factors that prevented erosion and other soil degradation, and thus preserved the upper soil environment for agriculture.

While working on the San Dimas project, Lowdermilk was recruited to be the vice-director of a temporary New Deal agency in the Department of the Interior, the Soil Erosion Service, in October 1933. The aforementioned Hugh Hammond Bennett had been selected as Director. The news service commenced the campaign for adopting conservation practices on American farms through demonstration projects based in small watersheds. Young soil conservationists worked with farmers to develop conservation plans for the farms. The plans had to achieve both the conservation objectives as well as satisfy the farmers' needs. Enrollees in CCC camps located on the projects and Works Progress Administration workers supplied labor. Later the service expanded operations nationwide by assigning trained conservationists to work with locally organized conservation districts. An act of 27 April 1935 created the Soil Conservation Service (SCS) in the USDA and gave it continuing authorities and duties. Lowdermilk served as associate chief (Dec. 1935 to Mar. 1937), chief of the research division (Mar. 1937 to Nov. 1939), and then head soil conservationist until his retirement.

Lowdermilk contributed to organizing all facets of the new service, but he especially concentrated on the research program. USDA transferred the experiment stations established under the 1930 Agricultural Appropriations Act to SCS in 1935. The stations were already engaged in research on terracing, crop rotations, strip-cropping, tillage methods, and the value of all these to soil conservation. Lowdermilk added runoff and erosion studies. Furthermore, the demonstration projects afforded an opportunity to observe the effects of farmland conservation practices on the entire watershed. Lowdermilk was thus instrumental in the new agency's contributions to understanding small-watershed hydrology.

In 1938 chance again intervened in Lowdermilk's life; and as usual he seized the opportunity. Representative Clarence Cannon suggested that a survey of the Old World could be useful in the United States' efforts toward a permanent agriculture. The trip, August 1938 to November 1939, involved more than 25 000 miles of automobile travel in Europe, the Mediterranean area, and the Middle East. Here he perfected his art of reading the land for evidence of past use and misuse. Before undertaking surveys in each country, Lowdermilk consulted agriculturalists, scientists, and officials. Geologists and archaeologists were especially interested, and valuable to Lowdermilk in explaining the cultural and physical factors involved in land use. In addition to searching for soil conservation and flood prevention measures that might be imported to the USA, Lowdermilk was engaged in what he called 'agricultural archaeology.' Ruins of some preindustrial civilizations indicated a prosperous agriculture, although these areas now had serious resource problems. What events brought about such conditions? What were the lessons for contemporary civilizations?

Lowdermilk's land-read records of past civilizations appeared in numerous articles. Indeed, there were 'Lessons from the Old World to the Americas in land use,' as Lowdermilk titled an article in the annual report of the Smithsonian Institution. He gladly noted the cases of wise land use through centuries, but was usually obliged to find a story of deterioration. The Soil Conservation Service published a summary, *Conquest of the Land Through 7000 Years*, in 1953 and followed it with several reprintings until more than one million copies were distributed. Early on, Lowdermilk had learned the value of written communication and practiced his writing skills. Readers who know Lowdermilk only through this publication have perhaps a truncated view – that of the globe-trotting chronicler of calamities awaiting civilizations that abuse their resources. He realized that a civilization's decline could not be interpreted solely on the basis of soil erosion. However, in writing the pamphlet, he embarked on a didactic mission aimed at all Americans, not just farmers. Soil fertility was a matter of concern for the farmer. Maintaining the medium for fertility – the physical body of soil resources – concerned the

entire nation. Without it, 'liberty of choice and action' was gone.

World War II terminated the trip in Europe but it opened a new opportunity, a return to China. At the behest of the Chinese government, Lowdermilk undertook the dangerous journey to advise the Chinese about increasing their food supply. The Communist triumph in China meant the Lowdermilk's and other advisors' plans for postwar-China would not be acted upon. But the trip afforded Lowdermilk an opportunity to pursue his interest in agricultural archaeology. During the intervening years in the USA, he had continued to study the agricultural archaeology of China. While in China he bought gazetteers, local histories, which Dean R. Wickes, a Chinese-language specialist, then researched for evidences of erosion problems. This research showed that in northern China, an area with a small percentage of level land, the population had increased threefold since the mid eighteenth century. This rapid population increase sent people to the hills for firewood and arable land, without any orderly installation of engineering measures for soil conservation. Unlike areas of central and southern China, they had no elaborate bench terraces to protect farmland. The gazetteers provided accounts of clearing the slopes, removing farmland from the tax rolls as wasteland, and abandoning homes along streams due to frequent flooding.

Lowdermilk retired from the SCS on 30 June 1947, but he continued his international work, which had been fairly continuous since the 1938 trip to the Middle East. He consulted in French North Africa (1948) and in the British colonies in Africa (1949–50). In the USA he contributed to the President's Water Resources Policy Commission. Controlling erosion on the upper reaches of watersheds became a passion for Lowdermilk's generation of conservationists. They favored land cover for increased absorption and engineering works for the controlled disposal of water without erosion. The upstream reservoir on the small watersheds was an integral part of the river development – an assertion that was often contested. Proponents of the control and use of headwaters had stated their case in the publication *Little Waters and Headwaters: Control and Use*. In the late 1940s, conservationists had another opportunity when Morris Cooke, a force behind *Little Waters*, became chairman of the President's Water Resources Policy Commission. Lowdermilk assumed chairmanship of the Committee on Standards for Basic Data. The Cooke and Lowdermilk views held sway in the committee report that emphasized a comprehensive, interdisciplinary approach. The interdependence of land and water called for watershed management, which had been neglected due to "our natural endowment and relatively low population density." Furthermore, the small watershed, the unit of watershed management preferred by the authors, was a cultural unit. The watershed unit had to be small enough so that residents understood its influence on their lives. Then they would devote the time and money needed to bring it to fruition as a community watershed. Lowdermilk's experience in semiarid climates came through in the committee's attitude toward flood control. Where feasible, reservoirs should not be used solely to control floods, but also to store storm waters for later use. Following work with the US Water Resources Policy Commission (1950) and the Supreme Allied Command in Japan (1951), he worked in Israel under the auspices of the Food and Agriculture Organization of the United Nations to organize an SCS.

During the trip to the Middle East in 1938–39, Lowdermilk became inspired by the efforts of urban-born European Jews to reclaim land. Upon returning to the USA, he wrote *Palestine, Land of Promise* (1944), wherein he proposed a Jordan Valley Authority to provide power, water supply, and irrigation water. The book argued, as did his popular pamphlet *Palestine Can Take Millions* (1944), that the land there could support a much larger population. The message contravened the British and American foreign policy of the early 1940s that discouraged Jewish emigration to Palestine. After retirement from the SCS, he worked with the Israelis to implement some of the measures outlined in the book. Drawing upon his Western experience, he advocated diversion of the Jordan River water before it entered the Dead Sea. The water would be utilized for hydroelectric power and irrigation water. Many Israelis favored technical assistance for agricultural development over direct food assistance. That sentiment was concisely conveyed when Minister of Development, Mordecai Bentov, coined the saying, "We don't need powdered milk; we need Lowdermilk." While there, Lowdermilk helped establish at Haifa a school to train conservationists, a school which later bore his name. The Lowdermilk School of Agricultural Engineering emphasized the basic sciences as preparatory to agricultural studies. Students took 2 years of mathematics, chemistry, physics, geology, and biology before moving on to the agricultural sciences. A job-related project in the 5th year was necessary to earn the degree.

Lowdermilk became best known for his work in Israel, probably more so than even for his work in China, at San Dimas, or with the SCS. In a career

that has generally been lauded, recommendations for development in the future state of Israel occasioned the most criticism of his work. A progressive conservationist who came of age in the early twentieth century, he advocated what he saw as positive human intervention to correct past detrimental human actions. In the Middle East he usually discounted climatic causality for land degradation and emphasized the human element. He singled out the nomadic herders as a cause for soil erosion. In place of nomadic herding, he would substitute development of the Jordan River to supply electric power and irrigation water for intensive agriculture and regional economic development. The new state of Israel needed such development to support the immigrants, he argued. In the USA he and Mrs. Lowdermilk spoke for Israeli statehood and were counted as 'Christian Zionists.' Lowdermilk cited an area he knew well, southern California, as an analogy of economic development brought on by the combination of hydroelectric power and irrigation water. While he was careful to say that Arabs would also benefit from economic development, none the less such arguments could be seen as antithetical to Arab interests. During recent decades environmentalism and the study of environmental history have generally looked with some disfavor on the progressive conservationists' merging of conservation with economic development. Some favor indigenous, less-intensive systems such as nomadic grazing as more appropriate.

Lowdermilk received many professional honors. He was a fellow and president of the American Geophysical Union (1941–44) and a fellow of the Society of American Foresters and of the Soil Conservation Society of America. The Lowdermilks had two children, William Francis and Winifred Esther. In retirement the Lowdermilks had a home in the Berkeley hills, California. Walter Lowdermilk died in Palo Alto, California in 1974, having spent his last years at the Veterans Hospital there.

See also: **Archeology in Relation to Soils**; **Climate Change Impacts**; **Degradation**; **Desertification**; **Erosion:** Water-Induced; **Jenny, Hans**; **Loess**; **Marbut, Curtis Fletcher**; **Sustainable Soil and Land Management**; **Terraces and Terracing**; **Water Harvesting**; **Watershed Management**

Further Reading

Chall M (ed.) (1969) *Walter Clay Lowdermilk, Soil, Forest, And Water Conservation and Reclamation in China, Israel, Africa, and the United States*. Oral History Interview. Bancroft Library, University of California–Berkeley, CA.

Davis RCB (1999) *Development as a Tool of Diplomacy: The Domestic Models for United States Policy in the Jordan River Valley, 1939–1956*. PhD dissertation. Georgetown University, Washington, DC.

Egerton-Warburton LM, Graham RC, Allen EB, and Allen MF (2001) Reconstruction of historical changes in mycorrhizal fungal communities under anthropogenic nitrogen deposition. *Proceedings of the Royal Society of London. Series B: Biological Sciences* 268: 2479–2484.

Graham RC and Wood HB (1991) Morphologic development and clay redistribution in lysimeter soils under chaparral and pine. *Soil Science Society of America Journal* 55: 1638–1646.

Helms JD (1984) Walter Lowdermilk's journey: forester to land conservationist. *Environmental Review* 8: 132–145.

Lowdermilk IM (1985) *All in a Lifetime: An Autobiography*. Berkeley, CA: Lowdermilk Trust.

Lowdermilk WC (1944a) *Lessons From the Old World to the Americas in Land Use*. Annual Report of the Board of Regents of the Smithsonian Institution, 1943, pp. 413–427. Washington, DC: Government Printing Office.

Lowdermilk WC (1944b) *Palestine, Land of Promise*. London, UK: Victor Gollancz.

Lowdermilk WC (1953) *Conquest of the Land Through 7000 Years*. Agriculture Information Bulletin No. 99. Washington, DC: US Department of Agriculture.

Lowdermilk WC and Wickes DR (1942) Ancient irrigation brought up to date. *Scientific Monthly* 55: 209–225.

Lowdermilk WC and Wickes DR (1943) China and America against soil erosion. *Scientific Monthly* 56: 393–413; 505–520.

Stross RE (1986) *The Stubborn Earth: American Agriculturalists on Chinese Soil, 1898–1937*. Berkeley, CA: University of California.

LYSIMETRY

T A Howell, USDA Agricultural Research Service, Bushland, TX, USA

Published by Elsevier Ltd.

Introduction

Lysimeters are standard instruments used in hydrology and water-quality research. In particular, they are used to measure percolation beneath the vegetation root zone and water use through evaporative processes from vegetation. (*See* **Evaporation of Water from Bare Soil; Evapotranspiration**.) They can also measure 'net' additions to the soil from precipitation or irrigation. When used to measure percolation, they become *in situ* water-quality sampling instruments. (*See* **Drainage, Surface and Subsurface; Leaching Processes; Macropores and Macropore Flow, Kinematic Wave Approach; Salination Processes; Salinity:** Management; **Solute Transport; Vadose Zone:** Hydrologic Processes.)

The word 'lysimeter' is derived from the Greek roots 'lysis,' which means dissolution or movement, and 'metron,' which means to measure. The standard definition of a lysimeter is: a device for measuring the percolation of water through soils and for determining the soluble constituents removed in drainage; so clearly the *lysimeter* is a soil physics and hydrologic measurement tool. It defines the water flux at some arbitrary boundary beneath the soil surface so that water balance can be measured or a water sample extracted that represents the naturally percolated chemical constituents. As a water-balance measurement device, lysimeters are used to determine evaporation from the soil–plant system, commonly called evapotranspiration (ET), and/or the latent heat flux in a surface energy balance.

History

Lysimeters are reported to have been used in the Netherlands in the early seventeenth century and in France later in the seventeenth century. Most of these reported studies were focused on measuring crop water use. Lysimetry for determining water use from vegetation has advanced through better knowledge of soil physics of water flow, soil hydraulic and thermal processes, plant root water uptake, and micrometeorologic influences from advection and aerodynamic processes (fetch related).

Lysimeters can be used in a laboratory, greenhouse, environment chamber, or situated in the natural outdoor environment. When used to simulate natural outdoor environments, design details and site selection are critical, together with management, so the vegetation within the lysimeter duplicates the surrounding vegetation. In practical terms, a well-managed lysimeter is indistinguishable from its surroundings. When used in greenhouse, laboratories, or environmental chambers, lysimeters are more 'relative' than 'absolute' measurement devices. The physical and chemical characteristics of the lysimeter soil should duplicate the natural states under study.

Construction materials and drainage methods are critical in lysimetry used to determine soil-water chemical transport or solute fluxes. Care is exercised to select design materials that will not react with or absorb any of the chemicals that are of interest.

There is not a standard for lysimeters that can be followed to construct either a lysimeter for measuring chemical transport or for measuring water evaporation from soils or vegetation. However, one principle has emerged to guide design considerations, which is that the lysimeter must be representative of the conditions desired to be measured. If the lysimeter design, construction, and its operation cannot permit the vegetation or soil condition within the lysimeter to duplicate or accurately represent the environment under study, then reliable measurements of neither the evaporative processes nor the soil-water chemical transport processes can be expected.

Lysimeter Types

Lysimeters are foremost devices, typically tanks or containers, that define a specific boundary to contain soil water and permit measurement of either the soil-water balance or the volume of water percolating vertically and/or its quality. Lysimeters can be broadly defined as percolation lysimeters (sometimes called drainage lysimeters) or weighing lysimeters (Figure 1). Both can serve the purposes of determining the soil-water balance, the vertical percolation flux

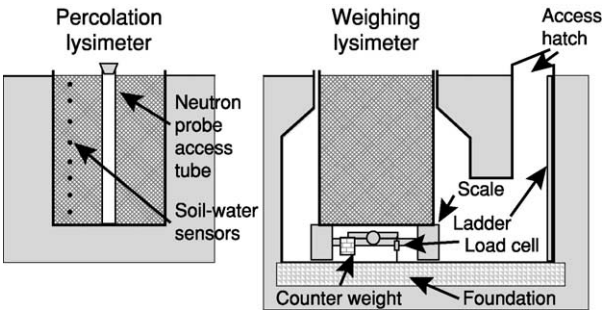


Figure 1 Schematic drawings of a percolation (drainage) lysimeter and a weighing lysimeter (not to scale).

(drainage), and the chemistry of the percolating water.

Percolation lysimeters may be installed to determine simply the vertical soil-water flux (drainage) or the chemical movement within the soil at a defined boundary. Soil-water solution sampling lysimeters can be simple porous devices that can extract soil water under a vacuum or access wells to permit water sampling from beneath a free water surface (i.e., water table). Percolation and weighing lysimeters differ in their measurement methods to determine the vegetation water use and/or soil-water evaporation. Percolation lysimeters must be used with a soil-water profile measurement method (*See* **Neutron Scattering; Tensiometry; Time-Domain Reflectometry**) to estimate 'indirectly' the water use in evaporative processes. Weighing lysimeters permit the mass or volumetric soil-water content change to be determined by weighing the lysimeter and determining its mass change over time. Weighing lysimeters, thus, can determine the 'net' infiltration from rainfall or irrigation and the amount of evaporation between wetting events. If properly designed, weighing lysimeters can coincidently measure the drainage rate as well as the evaporation rate. Percolation-lysimeter accuracy of the evaporative water balance is directly related to the precision of the soil-water measurement and its integration through the vegetation root zone. Weighing-lysimeter precision depends on many factors, the main ones being scale resolution, counterbalancing, and area-to-volume ratio.

Lysimeter Water Balance

The water balance of a lysimeter is given as:

$$\theta_i = \theta_{i-1} + P_{i_i} + I_i - D_i - ET_i \qquad [1]$$

where: θ_i is the lysimeter water volume per unit lysimeter area, in millimeters, at time i; P_i is precipitation, in millimeters; in I_i is irrigation, in millimeters; D_i is drainage from the lysimeter, in millimeters; and ET_i is evaporative loss (evapotranspiration) from the lysimeter, in millimeters, all during the time interval $i-1$ to i. Typically, lysimeters have a rim height of 10–40 mm above the soil surface to prevent runoff and runon. Some lysimeters used in hydrology are equipped with tanks to capture volumetrically runoff above a specified depth of surface storage. If runoff does not occur from the lysimeter, the lysimeter 'freeboard' may increase rainfall or irrigation capture, thereby increasing water available for evaporation or adding to lysimeter percolation volumes. Generally, P and I are measured separately using various methods (i.e., rain gauges, flowmeters, etc.), while D can be measured volumetrically as drainage occurs from the lysimeter.

Although water can move upward, most lysimeters are not equipped to measure or add water to simulate upward water fluxes. Some lysimeters use a static or dynamic water table to mimic field water conditions, and these lysimeters are usually equipped to measure the water additions needed to maintain the water table depth, or raise the water table, or measure the drainage water volume when the water table is lowered. Most percolation lysimeters use manual vacuum pumping or volumetric collection to measure drainage. Some weighing lysimeters that have greater belowground access have used tipping-bucket devices or tanks suspended from load cells from the lysimeter itself to measure drainage rates or volumes.

The main difference between percolation and weighing lysimeters is that weighing lysimeters measure the mass of the lysimeter (in megagrams or kilograms), and this mass change is then converted, taking into account the lysimeter area (meters squared) and the density of water (typically assumed to be $1\,Mg\,m^{-3}$; note: $1\,kg\,m^{-2} = 1\,mm$ assuming a water density of $1\,Mg\,m^{-3}$), to determine θ_i. Assuming no drainage, precipitation, or irrigation during the measurement time period ($ET_i = \theta_i - \theta_{i-1}$), the lysimeter mass measurement precision is a direct ratio of the lysimeter area (i.e., $1\,kg\,m^2$ per millimeter), while the soil volumetric measurement precision is much greater (i.e., $0.001\,m^3\,m^{-3}$ per millimeter). Therefore extremely high soil-water measurement precision is required to determine ET accurately for time periods of a single to a few days. Weighing lysimeters with proper design and electronic measurements can achieve equivalent accuracies for a day without counterbalancing. With counterbalancing of the soil 'dead' mass, excellent ET measurement precision has been obtained for time periods from 15 min to 1 h with typical weighing lysimeters. Weighing lysimeter mass measurement precisions have been noted to decline as wind speeds exceed approximately $3\,m\,s^{-1}$ due to wind load imposed on the lysimeter. Under very windy conditions, longer mass integration intervals may be necessary or numerical smoothing can be used.

Lysimeter Design Considerations

Experimental purpose is the principal consideration in lysimeter design. If the lysimeter facility is intended to represent natural outdoor environmental conditions, then siting becomes a most critical consideration. The site should represent the soil and plant conditions desired. Surrounding conditions, for perhaps up to a kilometer or more, can certainly impact

the use of water by a crop, especially for taller crops. It is important to be aware that in many cases these surrounding areas may not be directly controlled by the investigators using the lysimeters, and features and land use of surrounding land areas can change rapidly, compromising the lysimeter facility. Siting needs to include considerations for adequate surrounding fetch (similar cropped distance usually expressed in the direction of the predominant wind direction), utility access (telephone, electrical power, and sometimes water), and foot-traffic access for maintenance. Fetch needs vary with the research purpose and site conditions, but easily extend from 100 to 200 m. Utility access needs vary if alternate current (AC) power is required or if a telephone modem is used for data transfer. Most current lysimeter data loggers are battery-powered and some use an AC power backup for the batteries. Although analog telephone lines and modems are commonly used for data transfer, cellular telephones are becoming more common, as well as other wireless network technology. All lysimeters require repeated, periodic site visits for maintenance and measurements. Care must be taken to avoid crop damage and 'path' creation that might affect the lysimeter microclimate.

Lysimeter Shape and Area

Percolation and weighing lysimeter shapes are typically circular, square, or rectangular. Cylindrical tanks are inherently stronger per unit wall thickness than square or rectangular tanks, owing to their greater hoop strength. However, circular lysimeters can cause inaccuracies, depending on the vegetation type, especially for row crops with wide row spacing.

Figure 2 shows several typical lysimeter tank surface configurations. An important consideration is the lysimeter surface area. Obviously, larger sizes are desired, but they are constrained by capital resources as well as physical ones. Lysimeter areas in use vary considerably from less than 1 to 30 m^2 or greater. Two principles guide the selection of the appropriate area: (1) soil pedological considerations (size needed to duplicate saturated hydraulic conductivity of natural soil; i.e., macropores and soil cracks); and (2) the unit area of vegetation that might be considered as a minimum representative sample. The soil-pedology factor is particularly important for more structured, heavy vertisols and mollisols that have a high level of shrink–swell and structural cracking. The vegetation-area factor is particularly important in terms of matching the row spacing of crops or the size of the area needed to capture the natural variation in native species. This area is linked to the balance of the evaporative processes for soil water (E) and plant transpiration (T). For closely seeded crops (alfalfa, small grains, grasses, etc.) or selected forestry or native brush species, circular shapes are adequate; but for agronomic crops or horticultural crops (trees, vines, or vegetables) grown in rows, a square or rectangular shape is desirable. Figure 2 shows the 'equivalent' surface area of small, circular lysimeters, which introduces an uncertainty in knowing exactly the 'effective area' of the lysimeter. The square area of the row or multiple rows through the lysimeter (see the dashed lines in Figure 2) represents the sunlight-absorption area of the crop. The E and T components of a circular shape are constrained by the lysimeter surface area. For larger circular lysimeters with many rows, the lysimeter diameter can depart from an even multiple of the row spacing without a large bias on the balance of E and T and not affect the measured ET accuracy. However, for smaller circular lysimeters with a single row or even two rows, it is difficult to avoid biasing the ET accuracy by unbalancing the E and T components of crop water use if the lysimeter diameter is not larger than the row spacing (i.e., for a single row, row spacing is less than lysimeter diameter, which is less than 1.8 × row spacing; for a single row, this is $D_l = \pi^{-1/2} W$, where W is row spacing (in meters) and D_l is lysimeter diameter (in meters)).

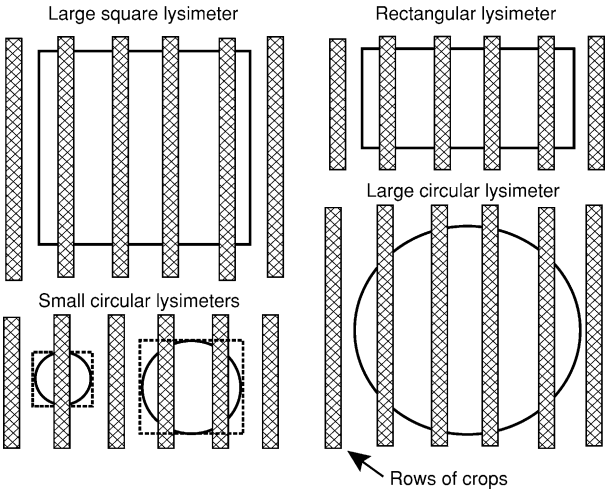


Figure 2 Lysimeter shapes for large and small lysimeters. The dashed lines represent the 'equivalent' area of sunlight interception for the smaller circular lysimeters.

An often unreported factor for interpreting lysimeter water balances has been the lysimeter effective area. For percolation lysimeters and some weighing lysimeters, the area considerations discussed above are important. Most percolation lysimeters have a single wall (Figure 1), so the effective area is assumed to match the inside tank area (i.e., π $(D_l^2/4)$ for a circular lysimeter and L_l^2 or $L_l \times W_l$ for a square or rectangular lysimeter tank, where L_l and W_l are the

lysimeter length and width, respectively, for a square and rectangular design). For a weighing lysimeter, which has two walls (the inner tank and outer tank walls) and an air gap (see Figure 1), the effective area for evaporation is computed for the 'midgap' dimensions. This midgap and wall area might represent approximately 5–10%, or even greater in some cases, of the lysimeter tank inner area. So the measured change in mass is adjusted downward for this area over which the outside and inside vegetation are assumed to be equal. Some hydrologic weighing lysimeters have used precisely beveled upper gap covers to capture the precipitation on the lysimeter correctly.

Lysimeter Depth

The depth of a lysimeter is an important design consideration; greater depths increase construction and installation problems. The basic consideration is to duplicate adequately the field rooting depth for the species being studied. Additionally, lysimeter depth has been shown to affect the lysimeter water-holding capacity and the soil-water flux through the profile. If the lysimeter depth is less than 1.5–2 m, vacuum drainage is required to mimic natural soil-water pressure potentials. If a water table is used to match field conditions and/or to simulate high water tables, care must be exercised in selecting the desired lysimeter depth. Few lysimeter designs have used depths greater than 2.5–2.7 m; but several newer lysimeters (one in China) are 5 m in depth, while some in Arizona in the western USA are 4 m in depth. For weighing lysimeters, greater depth (volume per unit area) necessitates counterbalancing a large fraction of the soil and tank mass to achieve high precision in the ET measurements.

Soil Profile Characteristics

The lysimeter soil profile should match or duplicate the site characteristics so that vegetative growth and water use will be representative of the site and experimental purposes. Both soil monoliths and repacked (reconstructed) soil profiles are used. Under 'well-irrigated' cropping situations, a repacked soil lysimeter column is often adequate. Exact physical, chemical, and/or vegetation characteristics of a soil can only be preserved in many cases by using a soil monolith. Naturally occurring soil spatial heterogeneity (both vertically and horizontally) can affect soil hydraulics, chemical composition, and thermal processes. In many lysimeters, repacked soil profiles have been successfully used for coarsetextured soils and finetextured soils that do not have significant profile stratification. Larger soil monoliths are more difficult to obtain and often require heavy industrial-type equipment (cranes, backhoes, coring equipment, etc.), but the time required to obtain a representative-sized monolith (approximately 5–15 m^3) must not exceed the time or labor needed to repack the same soil volume. Also, an intact soil monolith is ready for research applications almost instantly following collection, while repacked soil columns might take several years to duplicate the natural soil state for research with restricted irrigation (i.e., rain-fed or dryland applications). Figure 3 shows a lysimeter being repacked.

Figure 3 Technician installing soil-water sensors in a repacked lysimeter near Five Points, California. The lysimeter is a 2-m square (4 m^2). (Photo courtesy of USDA–ARS, Water Management Research Laboratory, Parlier, California, USA.)

Soil monolith collection Large soil monoliths are usually collected by coring, encasing, or compositing several smaller monoliths together. In the coring method, an open-ended tank (usually steel) is pushed into the soil using static weights or pulled into the soil using an anchor secured to the earth and various hydraulic or manual jacking methods. The tank's leading edge may be beveled to facilitate soil shearing while reducing internal compression on the monolith. In most cases, coring requires outside excavation as the monolith tank is lowered to reduce outside soil friction and to reduce the shearing forces. Soil monolith tanks that are square or rectangular are supported on the outside of the tank walls, while similar tanks that are repacked can be structurally supported by the floor and internal gussets as well as horizontal wall supports. Circular monolith tanks require simpler hoop-bracing. Figure 4 shows a square soil monolith prior to being pressed into the soil using four anchors to earth piers using a manually operated, hydraulic jacking assembly (shown on top of the lysimeter tank). This hydraulic lowering process allows precise control of installation speed in contrast to adding static 'dead' weights. The hydraulic installation

Figure 4 A monolith soil tank (1.5 m square and 2.3 m in depth) ready for insertion at Bushland, Texas. The hydraulic rams are located on the top corners and offset to permit a backhoe to excavate along the outside of the tank as it is pressed into the soil against the anchors secured into concrete piers. (Photo courtesy of USDA–ARS, Conservation and Production Research Laboratory, Bushland, Texas.)

procedure using rams or jacks avoids any erratic, jerky loading such as those that can occur when static weights are added.

Soil monoliths have been obtained in structured soils by carefully trimming a soil pedestal and gradually pressing a container tank over the soil pedestal while trimming the leading edge or using expandable foam (care has to be used in material selection as some materials react with or leach chemicals into the soil) to seal the soil pedestal against the tank walls. In encasing, a trimmed soil pedestal is carefully fitted with sections that can later be sealed to form the tank. In some encasings, pylon-like sheets are driven into the soil and welded or sealed together after excavating around the soil monolith.

In both coring and encasing methods, the soil monolith is undercut and lifted before a drainage system and tank bottom is installed. Undercutting methods are varied, too, including forcing shear plates under the monolith, boring holes and using 75- to 100-mm pipes, and driving steel rods or structural steel (angle iron, channel iron-type cross sections) to support the monolith during lifting. The boring method works best with more structured subsoils, while plates may be required for coarse sand or gravel subsoils. Monoliths may be desirable in certain situations to preserve natural soil-profile layering that cannot be reconstructed; but they are primarily soil samples with unique heterogeneities inherent in the spatial soil variability at the site. If soil monoliths have to be transported to the lysimeter site, care must be used to minimize the vibrations that inevitably occur during transport which can physically damage the monolith.

Lysimeter tanks The lysimeter tank, whether for a repacked column or for a monolith, should include measures to reduce water flow down the inside walls and even perhaps a division in the tank floor to separate the wall drainage from the inner core drainage. In repacked lysimeter tanks, the inside wall flow can be buffered by the inside horizontal bracing. In monolith tanks, an antiseep collar may be required at a short distance (approximately 200–300 mm) below the soil surface to reduce direct wall flow, particularly for a soil that may shrink away from the inner tank walls. Larger lysimeters require an adequately designed floor-bracing system to avoid tank-floor warping and for scale supports for weighing lysimeters.

Lysimeter drainage systems Lysimeter drainage system designs are quite varied, from those used to maintain static or falling water tables, to those with gravity drainage or manually pumped vacuum drainage, to those with automated vacuum systems. Gravity drainage can be measured volumetrically or by rate (tipping-bucket types). Manual vacuum drainage is often measured volumetrically while pumping. Automated vacuum systems have been used to maintain a constant vacuum on drainage lines, with drain outflows into tanks attached to a weighing lysimeter or suspended with load cells. Drain materials vary widely also. Gravity drains are made from simple, perforated polyvinyl chloride (PVC) pipes or perforated polyethylene (PE) pipes. Vacuum drains are made from ceramic cups, porous blocks, or sintered stainless tubes; and it is important that the 'bubbling pressure' of the soil in contact with the drain media is not exceeded. Some lysimeter drain systems have used outside soil matric potential sensors to match the drainage vacuum and even to add water to the lower soil profile if upward fluxes are detected. The drainage lines are usually embedded in fine sand or diatomaceous earth to apply the vacuum to the soil column uniformly.

Weighing-Lysimeter Scale Systems

Lysimeter scale systems are highly diverse and often uniquely conceived by the lysimeter design team. Scale systems include commercial lever load-cell types, flexure load cells, hydraulic floating-cell types, and hydraulic load-cell types. Mechanical lever-type scales permit a large part of the soil 'dead' (dry) mass and the tank mass to be tared statically to permit a small- to medium-sized electronic load cell to be used to measure the 'live' mass changes (water). Some mechanical scales have movable tare weights to adjust the offset counterbalance mass to maintain

a small electronic measurement range. The greatest advantages in weighing lysimetry are probably widespread availability and low-cost, highly precise analog-to-digital data loggers now commonly used to measure load cells. Typical ET mass accuracy can easily surpass 0.05- to 0.02-mm ranges. Modern data loggers also permit intensive data collection to monitor load-cell drift and other possible electronic problems.

Floating lysimeters were popular before the more expensive electronic–mechanical scales became popular with the improved data-recording precision. Floating lysimeters were designed with lightweight ballast tanks beneath the lysimeter tank, so highly accurate recordings (usually manual liquid depths that were later measured electronically) could be made of the lysimeter displacement to track the lysimeter mass changes. Often the more sensitive of these floating lysimeters were even more accurate than the mechanical scale types, with accuracies as small as 0.025 mm.

Various hydraulic weighing lysimeters have been built with a diverse range of hydraulic reading mechanisms. Some have used simple, homemade hydraulic load cells (often referred to as 'pillows') to measure the lysimeter mass change as the load (pressure) on the hydraulic load cells changes. Hydraulic weighing lysimeters are subject to a greater uncertainty owing to the density changes of the fluid (usually water or a glycol–water solution) and/or a Hg manometer due to temperature changes or flexure changes in the 'pillow' materials. These hydraulic weighing lysimeters achieved accuracies approaching 0.05–0.1 mm.

Lysimeters have been used with only load cells or load-beam scales without counterbalances. Placement of the load cells or scale platform under the lysimeter tank (at approximately 1.5- to 2.5-m depths) greatly reduces temperature influences on the electronic load cells. With highly precise data loggers and larger area-to-mass lysimeters, accuracies within the range 0.1–0.05 mm have been reported. These accuracies are acceptable for good short-term ET research, while greatly reducing the lysimeter scale complexity and construction costs.

Measuring ET or Latent Heat Fluxes with Weighing Lysimeters

The ET rate or mass ET accumulation can be measured easily with precise lysimetry provided no water is added or drained from the lysimeter during the measurement period. If water additions or drainage occur, the water balance can be calculated, but the additions or drainage must be computed separately to determine ET (see eqn [1]). This is a common problem with lysimetry in hydrology. Seldom will either the 'gross' precipitation or irrigation volumes be known in accurate comparison with the measured net soil mass water balance determined via lysimetry. During periods without water additions or drainage, the ET rate is computed as:

$$\mathrm{ET_i} = \{A_\mathrm{l}[(M_\mathrm{i} - M_{\mathrm{i}-1})/A_\mathrm{l}]/A_\mathrm{f}\}/T_\mathrm{i} \qquad [2]$$

where: ET is the evapotranspiration rate, in millimeters per hour (kilograms per square meter per hour), for time interval i; M is lysimeter mass, in kilograms; A_l is the lysimeter inner tank surface area, in square meters; A_f is the lysimeter foliage area (mid wall–air gap area), in square meters; and T is the time in period i in hours. The ratio A_f to A_l^{-1} is the correction factor for the lysimeter effective area. This correction factor assumes the outside and inside vegetation foliage overlap evenly on all of the sides or edges. If there is no overlap, as might occur with a shorter grass, then the correction factor becomes 1.0. Various techniques have been used to obtain reliable short-term ET rates. The lysimeter depletion over a short interval (approximately 30 min to 1 h) might be plotted and a linear line fit to the data so the derivative (ET rate) can be computed for that period. Alternatively a 'numerical' difference can be computed from eqn [2]. If the M terms are 'noisy' (common with high winds), a numerical smoothing technique can be used (equal or weighted averaging running averages).

The latent heat flux is λET, where λ is the latent heat of vaporization in megajoules per kilogram. When the time interval is in seconds, then λET will be in watts per meter squared. **Figure 5** shows the ET rate of irrigated alfalfa on 1998 June 13–14 (day of year (DOY) 164 and 165) at Bushland, Texas, measured with two precision weighing lysimeters. On DOY 164, advection took the form of exceptionally strong winds and a large vapor pressure deficit (VPD), while the second day had calmer winds and a lower VPD, with almost identical net radiation. The increased wind speeds on the 2nd day during the 18.00–19.00 time period was noticeable together with the greater postmidnight, predawn evaporation on that day; while winds and VPD were still greater than those typically experienced in that time period. On DOY 164 peak λET exceeded 1000 $\mathrm{W\,m^{-2}}$ (greater than peak incoming solar irradiance), but, because the crop had a fetch of almost 300 m (NE lysimeter; 100 m fetch for the SE lysimeter) and was 'well watered,' it could maintain the extremely large transpiration fluxes with no discernible 'stomatal control' above typical evaporative conditions (i.e., the next day).

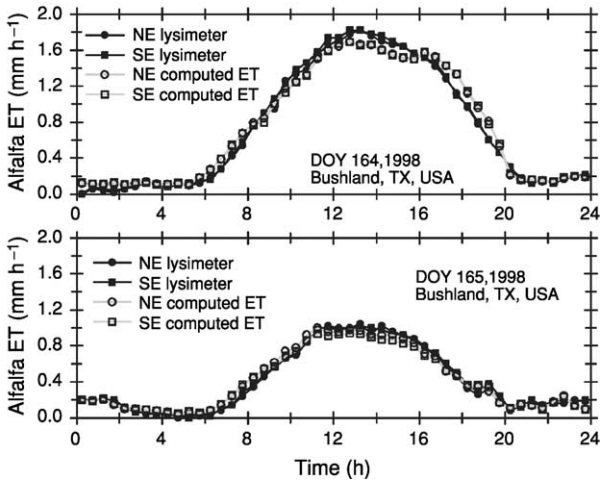


Figure 5 Example alfalfa evapotranspiration (ET) measured with two precision lysimeters at Bushland, Texas, under advective conditions (DOY 164) and on a more typical day (DOY 165) in 1998 contrasted with reference alfalfa ET computed with a Penman–Monteith equation and standardized parameters based on the ASCE/EWRI (American Society of Civil Engineers/ Environmental Water Resources Institute) procedures.

Summary

Lysimeters are hydrologic instruments to measure the percolating water quality and/or to measure the evaporation of water from vegetated surfaces. They can be percolation or weighing types. The water balance is measured in percolation types by soil water balance while the mass change in the weighing types measures the water use (or gain from irrigation or precipitation). The lysimeter design must carefully consider its purpose. Then its size, depth, type, and drainage are key considerations. With well-designed lysimeters that are carefully maintained, high-quality data can be obtained on vegetation water use and percolation water quality.

See also: **Drainage, Surface and Subsurface**; **Evaporation of Water from Bare Soil**; **Evapotranspiration**; **Leaching Processes**; **Macropores and Macropore Flow, Kinematic Wave Approach**; **Neutron Scattering**; **Salination Processes**; **Salinity:** Management; **Soil–Plant–Atmosphere Continuum**; **Solute Transport**; **Tensiometry**; **Time-Domain Reflectometry**; **Vadose Zone:** Hydrologic Processes; **Water Content and Potential, Measurement**; **Water Cycle**; **Water-Use Efficiency**

Further Reading

Aboukhaled A, Alfaro A, and Smith M (1982) *Lysimeters*. FAO Irrigation and Drain Paper No. 39. Rome, Italy: Food and Agricultural Organization of the United Nations.

Allen RG and Fisher DK (1990) Low-cost electronic weighing lysimeters. *Transactions of the American Society of Agricultural Engineers* 33: 1823–1833.

Allen RG, Howell TA, Pruitt WO, Walter IA, and Jensen ME (eds) (1991) *Lysimeters for Evapotranspiration and Environmental Measurements*. Proceedings of the International Symposium on Lysimetry. New York: American Society of Civil Engineers.

Bergström L (1990) Use of lysimeters to estimate leaching of pesticides in agricultural soils. *Environmental Pollution* 67: 325–347.

Brutsaert W (1982) *Evaporation into the Atmosphere*. Dordrecht, The Netherlands: D. Reidel.

Duke HR and Haise HR (1973) Vacuum extractors to access deep percolation loses and chemical constituents of soil water. *Soil Science Society of America Proceedings* 37: 963–964.

Harrold LL and Dreibelbis FR (1951) *Agricultural Hydrology as Evaluated by Monolith Lysimeters*. Technical Bulletin No. 1050. Washington, DC: US Department of Agriculture.

Hatfield JL (1990) Methods of estimating evapotranspiration. In: Stewart BA and Nielsen DR (eds) *Irrigation of Agricultural Crops*, pp. 435–474. Monograph No. 31. Madison, WI: American Society of Agronomy.

Howell TA, Schneider AD, and Jensen ME (1991) History of lysimeter design and use for evapotranspiration measurements. In: Allen RG, Howell TA, Pruitt WO, Walter IA, and Jensen ME (eds) *Lysimeters for Evapotranspiration and Environmental Measurements*, pp. 1–9. Proceedings of the International Symposium on Lysimetry. New York: American Society of Civil Engineers.

Kirkham RR, Gee GW, and Jones TL (1984) Weighing lysimeters for long-term water balance investigations at remote sites. *Soil Science Society of America Journal* 48: 1203–1205.

Marek TH, Schneider AD, Howell TA, and Ebeling LL (1988) Design and construction of large weighing lysimeters. *Transactions of the American Society of Agricultural Engineers* 31: 477–484.

Pruitt WO and Angus DE (1960) Large weighing lysimeter for measuring evapotranspiration. *Transactions of the American Society of Agricultural Engineers* 3: 13–15, 18.

Ritchie JT and Burnett E (1968) A precision weighing lysimeter for row crop water use studies. *Agronomy Journal* 60: 545–549.

Schneider AD and Howell TA (1991) Large, monolithic, weighing lysimeters. In: Allen RG, Howell TA, Pruitt WO, Walter IA, and Jensen ME (eds) *Lysimeters for Evapotranspiration and Environmental Measurements*, pp. 37–45. Proceedings of the International Symposium on Lysimetry. New York: American Society of Civil Engineers.

Schneider AD, Howell TA, and Steiner JL (1993) An evapotranspiration research facility using monolithic lysimeters from three soils. *Applied Engineering in Agriculture* 9: 227–232.

Schneider AD, Steiner JL, and Howell TA (1994) Pulldown forces for collecting large soil monoliths. *Transactions*

of the American Society of Agricultural Engineers 37: 831–835.

Schneider AD, Howell TA, Moustafa ATA, Evett SR, and Abou-Zeid WS (1998) A simplified weighing lysimeter for monolithic or reconstructed soils. *Applied Engineering in Agriculture* 14: 267–273.

Soileau JM and Hauck RD (1987) A historical review of US lysimetry with emphasis on fertilizer percolation losses. In: *Proceedings of the International Conference on Infiltration Development and Application*. St. Joseph, MI: American Society of Agricultural Engineers.

Tanner CB (1967) Measurement of evapotranspiration. In: Hagan HM, Haise HR, and Edminister TW (eds) *Irrigation of Irrigated Lands*, pp. 534–574. Monograph No. 11. Madison, WI: American Society of Agronomy.

Yang J, Li B, and Liu S (2000) A large weighing lysimeter for evapotranspiration and soil-water-groundwater exchange studies. *Hydrologic Processes* 14: 1887–1897.

Young MH, Wierenga PJ, and Mancino CF (1996) Large weighing lysimeters for water use and deep percolation studies. *Soil Science* 161: 491–501.

M

MACRONUTRIENTS

C W Wood, J F Adams, and B H Wood,
Auburn University, AL, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Macronutrients are defined as plant nutrients that typically attain a concentration $>500\,\mathrm{mg\,kg^{-1}}$ in mature plants. They include nitrogen (N), phosphorus (P), potassium (K), calcium (Ca), magnesium (Mg), and sulfur (S). Proper management of these nutrients is essential to ensure that farmers get maximum economic yield while protecting environmental quality. Sensible management of N and P by producers is especially important for environmental protection, because these elements ordinarily limit plant growth in natural terrestrial and aquatic ecosystems. Eutrophication (enrichment) of natural ecosystems with N and P can result in a myriad of undesirable environmental consequences, such as loss of species diversity, fish kills, increased cost of water treatment, impaired navigational and recreational use of waterways, noxious odors, and increased populations of insect pests.

Conservation management of macronutrients implies that these elements will be handled in a manner promoting incorporation into harvestable plant biomass and retardation of movement away from the site of application. To accomplish this goal, management decisions should be based on their cycling patterns, which vary among macronutrients. Two macronutrients (N and S) have gaseous pathways in their cycles that complicate management, while the others have simpler cycles with no atmospheric pathway.

Nitrogen

Nitrogen is the most limiting nutrient in biomass production systems, due to its large requirement by plants and the finite ability of soils to supply available N. Plants take up N as nitrate (NO_3), ammonium (NH_4), or urea ($CO(NH_2)_2$). In moist, warm, well-aerated soils the NO_3 form is dominant. Once inside plants NO_3 is reduced to NH_4 using energy provided by photosynthesis. Plants normally contain between 10 and $50\,\mathrm{g\,N\,kg^{-1}}$ on a dry basis.

Conservation of N in biomass production systems involves preventing losses of N that enters soils through various N-cycle pathways. These pathways include: (1) addition by precipitation; (2) fixation by lightning; (3) biological N fixation, both symbiotic and asymbiotic; (4) addition of organic fertilizers and crop residues; and (5) addition of industrially fixed N as fertilizers. Nitrogen added to soil systems in precipitation is generally in the form of NO_3 and NH_4, and usually amounts to 5–$10\,\mathrm{kg\,N\,ha^{-1}\,year^{-1}}$, with greatest amounts being deposited in regions where acid rain is prevalent. Fixation by lightning occurs when it oxidizes atmospheric dinitrogen (N_2) to $NO + NO_2$, which in turn is converted to NO_3, but this only amounts to 10–20% of the N deposited via precipitation. Amounts of N added in precipitation are not enough to support agricultural biomass production, but may be enough to sustain mature natural ecosystems.

N Fixation

Atmospheric N_2 can be fixed asymbiotically or symbiotically by specialized organisms into organic forms that can subsequently be mineralized into plant available forms. Asymbiotic fixation is performed by free-living organisms such as blue-green algae that fix up to $25\,\mathrm{kg\,N\,ha^{-1}\,year^{-1}}$, and aerobic (e.g., *Azobacter* spp.) and anaerobic (e.g., *Clostridium* spp.) bacteria that fix up to $0.5\,\mathrm{kg\,N\,ha^{-1}\,year^{-1}}$. These organisms do not require a specific host, unlike those which bring about symbiotic N fixation. The latter include rhizobia bacteria spp. in legumes and actinomycete spp. in some trees, that fix atmospheric N_2 in symbiotic association with the higher plants. These microorganisms reside in nodules on plant roots and derive their energy from carbohydrates produced photosynthetically by their higher plant symbiont. Symbiotic fixation's contribution of N to terrestrial biomass production systems can be substantial. For example soybean, cowpea, clover, lupin and alfalfa have been shown to fix up to 75, 85, 130, 160, and

360 kg N ha^{-1} $year^{-1}$, respectively. The alder, a small tree, has been shown to fix up to 170 kg N ha^{-1} $year^{-1}$. Symbiotic N fixation has been recognized as being second in importance among biological processes only to photosynthesis in regard to sustaining life on earth.

Fertilizers and Manures

Atmospheric N_2 is fixed industrially as NH_3 via the Haber-Bosch process, which requires high temperature (1200°C) and pressure (350 atm). Subsequently the NH_3 gas can be converted to various liquid and solid NO_3, NH_4, or $CO(NH_2)_2$ fertilizers.

The major types of organic fertilizers include barnyard manures, liquid manures, processed organic materials, and crop residues returned directly to the soil. Barnyard manures are relatively dry, bulky, solid materials that derive their nutrient value from animal excreta. Liquid manures consist of animal excreta that have been deposited on solid or slatted floors, without bedding material, and then washed into lagoons or storage tanks. Processed organic materials are produced on- or off-farm and include composts, sewage sludge, food-processing wastes, forestry by-products, industrial wastes, and municipal refuse. Crop residues include plant parts (straw, stover, roots, etc.) that remain on the land after crop harvest and green manure or cover crops grown for incorporation into the soil.

Organic fertilizers vary widely in N concentration depending on their origin, and consequently, some materials have greater N fertilizer value than others. Of the barnyard manures, those generated by birds tend to have the greatest N concentration owing to their relatively low water content. Due to low dry matter concentrations, liquid manures typically have low N concentrations. Sewage sludges tend to have relatively high N concentrations. N concentration generally declines when organic wastes are composted, but co-composted dead poultry and chicken manure may have a greater N concentration than barnyard manures. Crop residues (e.g., maize and cotton stover, and wheat and rice straw) usually have low N concentrations in comparison to leguminous green manures (e.g., clover and alfalfa) which, due to their typically greater N concentration, can provide substantial amounts of residual N for succeeding crops.

Typically, for net release of plant available N to occur in the soil, i.e., net mineralization, organic fertilizers should have a carbon (C):N ratio <20. Application of organic materials with C:N >30 will result in immobilization of N by soil microbes. Typically, for immediate release of plant available N, organic fertilizers should have a N concentration >17 g kg^{-1} on a dry-weight basis. The end product of the mineralization of organic forms of N is NH_4, which can be further transformed to NO_3 via the process of nitrification.

Nitrogen-cycle pathways that lead to soil system losses include crop removal, leaching, runoff, NH_3 volatilization, biological denitrification, chemical denitrification, and urea hydrolysis. From a biomass production standpoint, crop removal is the desired outcome, while the others result in economic loss to producers and pollution of the environment.

Leaching results in movement of NO_3 below the rooting zone. Nitrate that escapes the rooting zone can eventually enter groundwater and present a health hazard to humans and livestock. Management to prevent nitrate escape into groundwater includes use of appropriate N application rates; losses of NO_3 via leaching will be minimized if N is not added beyond that needed for maximum yield. Synchronization of N application with crop needs will also decrease risk of NO_3 leaching; plants need little N at planting when the potential for leaching loss can be high. Split or multiple applications in irrigation water allow plants to be fertilized when they need N most, lowering potential for leaching losses. Slow-release N fertilizers may also reduce NO_3 leaching losses, but are expensive, and generally limited to use on high-value crops such as vegetables and turf. Lastly, winter cover crops can help reduce NO_3 leaching by capturing NO_3 that otherwise would be lost below the rooting zone.

Losses of N in runoff are a function of rainfall amount and intensity, infiltration of water into soil, and vegetative cover. Management to prevent losses of N in runoff include maintaining vigorous ground covers on production fields when possible, and maintaining healthy filter strips and riparian areas on borders of production fields year-round. Any management that results in greater water infiltration into the soil will decrease runoff N losses, but may increase losses via leaching and/or denitrification.

Under alkaline conditions the NH_4 ion is unstable and subject to chemical decomposition and subsequent volatilization: NH_4:$NH_{3(solution)} + H$; $NH_{3(solution)}$:$NH_{3(gas)}$. Volatilization losses from surface-applied N can be substantial (>50% of N applied). Ammonia volatilization represents an economic loss to farmers, and atmospheric NH_3 can promote negative environmental impacts such as respiratory problems, acid rain production, soil acidification and ecosystem eutrophication following deposition, and toxicity to plant tissue. Many factors promote NH_3 volatilization as they increase: surface area of the soil solution–atmosphere interface, velocity of air moving across the soil surface, temperature, partial pressure of NH_3 in soil solution, and pH.

Ammonia is also volatilized when surface-applied urea, $CO(NH_2)_2$, comes in contact with the ubiquitous soil enzyme urease. Reduction of NH_3 volatilization is most easily accomplished via techniques such as tillage, injection (liquid manures), irrigation, or timing of application with natural rainfall, which move NH_3/NH_4 to depths where it can be adsorbed by soil colloids.

Biological denitrification is the facultative microbial reduction of NO_3 and nitrite (NO_2) with liberation of N_2 and nitrous oxide (N_2O). Common, widespread species of typically aerobic bacteria use NO_3 and NO_2 as an oxygen source in the absence of O_2. The potential for denitrification losses is high given the appropriate requisites including: anaerobiosis, carbohydrates for energy, $pH > 5.5$, NO_3/NO_2 supply, and warm temperatures. Besides being an economic loss to farmers, release of N_2O into the atmosphere via biological denitrification serves to exacerbate the 'greenhouse effect' by decreasing escape of infrared energy into space. Moreover, N_2O has been implicated in photo-oxidation of stratospheric ozone (O_3), which increases penetration of destructive ultraviolet rays to the earth's surface. Managing for inhibition of biological denitrification involves any processes that: (1) slow nitrification, such as chemical nitrification inhibitors; (2) enhance NO_3 removal such as plant uptake; and (3) promote aeration, such as drainage and evapotranspiration.

Denitrification can also occur chemically under aerobic conditions when NH_4 and NH_4-forming fertilizers are nitrified. This result can be realized through different chemical pathways, but they all involve destruction of the nitrification intermediate NO_2. Thus, from a management standpoint, chemical denitrification is not an important process since NO_2 is ephemeral in soil systems.

Phosphorus

Phosphorus is one of the key essential elements in modern agriculture, and it has many important functions in plants, the premier one being energy storage and transfer. The agronomic literature is replete with examples of grain, fiber and forage yield increases owing to proper maintenance of P fertility in arable soils. Although salutary effects of P in agricultural production are evident, P can also function as a pollutant if it moves from the site of its intended use. The main concern is P transport in surface or subsurface flow from soils to streams, rivers, lakes, and eventually the oceans. Phosphorus transported from agricultural soils can promote eutrophication, and algal growth in surface waters is usually limited by P instead of N availability.

Origins of P in soils include residual soil minerals, or inputs of P from commercial fertilizers and organic fertilizers/wastes. The most common mineral forms of P in soils are: (1) apatite (calcium phosphate), which is found in unweathered and moderately weathered soils; and, (2) iron and aluminum phosphates, which are found in highly weathered soils. The raw material for commercial fertilizer P is apatite mined from various deposits around the world and treated with acid to increase its solubility. Organic P fertilizers/wastes derive from a variety of sources including animal manures, crop residues, and municipal and industrial wastes.

Once a P atom enters the soil environment it is subject to several soil processes that control its fate. In soil solution, P is present as either a monovalent (H_2PO_4) or divalent (HPO_4) anion, the form depending on soil pH. Phosphorus enters the soil solution via either: (1) dissolution of primary minerals; (2) dissolution of secondary minerals; (3) desorption of P from clay and mineral surfaces; and (4), biological conversion of P in organic materials to inorganic forms (mineralization). In most soils, soil solution P ranges between <0.01 and $1\,mg\,l^{-1}$, and a value of $0.2\,mg\,P\,l^{-1}$ is commonly accepted as the solution P concentration desired to support agronomic crops.

Dissolution and erosion are the processes by which P is moved into runoff from agricultural fields. Dissolution of P from a thin zone of surface soil and vegetative material yields dissolved P that is immediately available for uptake by aquatic biota. Erosion via surface runoff transports particulate P in the form of soil and vegetative matter; only that portion of particulate P in equilibrium with dissolved P is available for aquatic biota. Thus, bioavailable P includes dissolved P and a portion of particulate P. Once bio-available P moves from agricultural fields into receiving waters it can contribute to eutrophication.

Although generally considered a less important mechanism than surface runoff, P leaching followed by shallow lateral subsurface flow can contribute dissolved P to surface waters under high water table conditions. This mechanism becomes more important in soils with large accumulations of P that saturate surface soil sorption capacity, leading to downward movement of P.

Escape of soil P to surface waters is affected by a host of soil management practices (e.g., tillage method, cover crops, riparian zones, terracing, etc.). However, the soil P concentration that correlates with P bioavailability is the greatest determinant of the balance between adequate soil P fertility and offsite P escape. In that regard, soil testing is currently the best management tool available to ensure that

crops are provided with adequate soil P. Moreover, soil testing is likely the best management tool available to ensure that soils do not become overloaded with P, which increases the likelihood of their contribution to pollution of downstream water.

Potassium

Potassium is the seventh most common element in the earth's crust, accounting for 2.6% of its mass. It is present in igneous, sedimentary, and metamorphic rocks, and deposits that are mined for fertilizer are evaporites, largely KCl, derived from drying of enclosed bodies of seawater. The quantity of K in soils is dependent on parent material and degree of weathering. Minimally weathered soils in arid regions may contain up to 2.5% K, while weathered soils in the tropics may contain little or no K.

Potassium is taken up by plants in amounts greater than those of N, and depending on species, plant tissue concentrations can range as high as 5%. It does not combine with C in plants to form organic compounds, but exists in a mobile ionic form (K^+). Potassium has many functions in plants, but owing to its high tissue concentration and ionic form, it is not surprising that its primary one is osmoregulation.

Potassium has no atmospheric pathway, and exists in the soil as primary minerals (feldspars and micas), secondary minerals (illite and vermiculite), exchangeable K, and soil solution K. Of the primary minerals, feldspars are more resistant to weathering than micas (biotite and muscovite), and exist mostly in the sand and silt size fractions. Weathered micas exist in the clay size fraction, and eventually release K from interlayer spaces as they hydrate to form secondary minerals. In a reverse process, K can be trapped when the interlayer spaces of secondary minerals collapse upon dehydration, reforming mica. When K^+ leaves the interlayer space of a secondary mineral, it will likely be adsorbed on a cation exchange site of a soil colloid. This exchangeable K maintains an equilibrium with the K^+ in soil solution, together constituting 'available K.'

Loss pathways for K include leaching and erosion/runoff. Leaching of K is increased under conditions that include a high degree of soil colloid saturation with K, sandy subsoils, poor vegetative cover and shallow rooting, and low soil-pH. Soil acidity impacts K leaching through the complementary ion principle: there will be less leaching of K if Ca^{++} rather than hydrogen (H^+) or aluminum (Al^{3+}) is the dominant complementary ion on soil-colloid exchange sites. Thus, management to reduce K leaching from soils includes following soil test recommendations to avoid saturation of soil colloids with K, maintaining vigorous vegetative cover, and liming to maintain soil pH according to soil test recommendations. In regard to K loss via erosion/runoff, management previously described to prevent surface soil losses of P will also prevent K loss.

Calcium

Calcium concentration in the earth's crust is approximately 3.6%, while soils contain only 1.37%. The primary source of soil Ca is from weathered limestone and the weathering of certain primary minerals. This indicates that parent material plays a major role in determining soil Ca concentration. The soil concentration of Ca can vary greatly and is generally related to soil pH. Soils with a pH of 6.0 or greater usually have adequate Ca for plant growth. The majority of plant-available Ca is exchangeable, is much greater than that found in soil solution, and is a divalent cation (Ca^{2+}). An average Ca concentration in soil solution is approximately $0.01\,mol\,l^{-1}$. The approximate range of exchangeable Ca is 20% to 65% of the cation exchange capacity (CEC) for most soils. Soils may have an average CEC of $15\,cmol_c\,kg^{-1}$ and 50% Ca saturation. This results in the exchangeable Ca being about 12 000 times greater than soil solution. Some soils contain mineral Ca in the form of precipitated $CaCO_3$ or $CaSO_4 \cdot 2H_2O$ within the soil profile, which results in Ca concentrations much greater than the exchangeable. In the western USA, these soils are usually found in areas of reduced rainfall.

Soil pH and available Ca are usually directly correlated. For example, as pH increases the base saturation increases, thus the amount of exchangeable Ca, Mg, and K increase. These are considered to be basic cations while Al and H are considered acidic cations. Often, liming soils to raise low soil pH will greatly increase Ca concentration in soils. Limestone is the primary liming material and there are different types of limestone depending on the ratio of Ca to Mg in the limestone. Calcite is $CaCO_3$ with essentially no $MgCO_3$ but as the amount of $MgCO_3$ begins to increase within the limestone, the limestones are referred to as calcitic, dolomitic, and dolomite. Dolomite contains about 50% $CaCO_3$ and 50% $MgCO_3$. Other sources of $CaCO_3$ are marl or chalk, which are soft sources of limestone. Choices of liming material may be determined by the nutrient status of Ca or Mg. Another available Ca source is gypsum ($CaSO_4 \cdot 2H_2O$), often used when soil pH is adequate, but additional Ca is required. Industrial byproducts that have neutralizing properties such as quick lime, burnt lime, basic slag, fly ash and some sludge materials may also contribute Ca.

Calcium has fewer pathways than other macronutrients and thus, the management of adequate soil Ca is simpler. The loss of Ca is generally by plant uptake, since it does not volatilize and leaching is minimal. For example, Ca is not easily leached because it is held to cation exchange sites more strongly than other nutrients, which is demonstrated by the lyotropic series:

$$Al^{3+} > Ca^{2+} > Mg^{2+} > K^{+} = NH_4^{+} > Na^{+}.$$

The strength of retention increases with the charge on the cations, and also depends on their hydrated size with smaller cations being held more tightly. Inputs into the pool of available Ca must be considered for proper management. Very little available Ca is derived from mineralization of soil organic matter. So the input of available Ca is from weathering of Ca minerals or from supplemental applications of Ca sources as those mentioned earlier. Plant uptake of Ca is another pathway and the uptake characteristic is important in the method of supplying adequate Ca.

The major role of Ca in plants is cell-wall integrity and function, and it plays a role in cell division and elongation. Calcium is bound to negatively charged phosphate groups of membrane lipids. This prevents inorganic and organic molecules from leaking out of the cells. When Ca deficiency continues there is a general disintegration of the membrane and cell-wall leaking. In the whole plant, the meristematic tissue shows the deficiency in the growing point and root tips. Most visual Ca deficiency symptoms occur in fruit and storage tissue, because translocation of Ca is by way of xylem tissue. Thus, low soil solution Ca or very low transpiration rate in humid conditions may cause Ca deficiency. Examples of common symptoms include: apples display 'bitter pit' where the surface is pitted with brown spots; tomatoes, peppers, and watermelons have blossom-end rot where there is complete tissue breakdown at the distal end of the fruit; celery has black heart and brussel sprouts have internal browning; and peanuts exhibit 'pops' where the hull is completely formed but the kernel did not develop.

Magnesium

The lithosphere contains about 2.1% Mg and soil generally contains 0.5%. Parent material plays a major role in determining available soil Mg concentration. Soil Mg is derived from the weathering of several minerals, much like Ca. Limestones (dolomitic and dolomite) supply Mg in many soils as well as some primary and secondary minerals. Mg, unlike Ca, is contained in several secondary minerals e.g., montmorillonite, vermiculite, illite, and chlorite. Soils rich in these minerals usually have sufficient available Mg and do not need supplements. Much like Ca, most of the available Mg is exchangeable but usually at a lower concentration. Generally Mg soil solution concentration is about 0.001 mol l^{-1}, which is lower than soil-solution Ca. The lower availability of Mg can be attributed to the lyotrophic series that indicates Mg can be leached more easily than Ca.

As mentioned earlier with Ca, availability of Mg is also often related to soil pH. Soil Mg is normally increased when applying dolomite or dolomitic limestone when correcting soil pH. When adding Mg without adjusting pH other sources are available; such as Epsom salt ($MgSO_4 \cdot 7H_2O$), magnesium ammonium phosphate ($MgNH_4PO_4 \cdot H_2O$), and a more common source Sul-Po-Mag ($K_2SO_4 \cdot 2MgSO_4$). Other organic sources may supply Mg, but mineralization of organic matter does contribute substantially to Mg availability.

The pathway for Mg is the same as Ca, except Mg is not held to exchange sites as strongly as Ca and can be subject to some leaching sooner than Ca. Uptake and translocation of Mg is different than Ca. Magnesium is very mobile in the phloem and is translocated from older to younger leaves and translocates to storage tissue. The deficiency symptom occurring first in the older tissue, depending on the type, is an intervenal yellowing or chlorosis, in severe cases becoming necrotic. Magnesium is the central component of chlorophyll and acts as a cofactor in most enzymes activating phosphorylation.

Sulfur

Management of soil S is more complicated than most other nutrients and may be considered similar to that of N. Sulfur content of the lithosphere is about 600 ppm and while there is a large range (30–10 000 ppm) in soil S, the average concentration is estimated at 700 ppm. The four major forms of S in soils are: (1) sulfides, (2) sulfates, (3) organic sulfur, and (4) elemental sulfur. Understanding the S cycle in soil is important for proper management of soil S. Sulfur mineralization is very similar to N mineralization with decay products being in a reduced form (sulfides) then oxidized to a sulfate form. Sulfate is subject to reduction, immobilization, fixation by clays, precipitation, anion exchange, leaching and plant uptake. In maintaining adequate soil S fertility all these processes must be considered. The reduction of sulfates to sulfides requires anaerobic conditions in combination with an available C source. Once reduced it may be lost to the

atmosphere or precipitated and then be subject to oxidation when conditions become aerobic. The oxidation of sulfides or elemental sulfur causes acidification of soils. This may be desirable when attempting to grow acid-loving plants on initially high pH soils. Major inputs into the pool of available soil sulfur are from soil organic matter, fertilizers, and deposition from the atmosphere (wet or dry). Other minor sources are from irrigation water, plant residues or manures, and sulfur-bearing minerals.

In many cases the available sulfur must be supplemented. The common sources of sulfur are gypsum ($CaSO_4 \cdot 2H_2O$; 19% S), elemental S (flowers of sulfur; 95% S), ammonium sulfate ($(NH_4)_2SO_4$; 24% S) and sulfur coated urea (13–16% S). Sulfur-deficient soils requiring these inputs are usually sandy-textured and the deficiency is usually more prominent in early cool spring months. This occurs mainly for two reasons: mineralization is minimal and earlier mineralized S has been taken up or lost due to previously mentioned pathways. Also, root growth is limited due to cool soil temperature. These reasons often explain why crops may grow out of S deficiencies when soil temperature increases, enhancing mineralization and root growth.

Sulfur is taken up mainly as SO_4, is considered nonmobile and is only translocated in the xylem. Thus, deficiencies usually appear in the younger tissue. Deficiency symptoms are often a general chlorosis in younger leaves. Sulfur is a component of many proteins in plants, but when plants become deficient in S there is an accumulation of non-S proteins and an associated decrease in sugars. The decrease in sugar inhibits chlorophyll production, resulting in a general chlorosis.

List of Technical Nomenclature

°C	Celsius
Al	Aluminum
atm	Atmosphere
Ca	Calcium
$CaCO_3$	Calcium carbonate
$CaSO_4 \cdot 2H_2O$	Gypsum
CEC	Cation exchange capacity
$CO(NH_2)_2$	Urea
$cmol_c\,kg^{-1}$	centimole charge per kilogram
$g\,kg^{-1}$	gram per kilogram
H	Hydrogen
HPO_4	Monohydrogen phosphate
H_2PO_4	Dihydrogen phosphate
K	Potassium
KCl	Potassium chloride
$K_2SO4 \cdot 2MgSO_4$	Sul-Po-Mag
$kg\,ha^{-1}\,yr^{-1}$	kilogram per hectare per year
Mg	Magnesium
$MgCO_3$	Magnesium carbonate
$MgNH_4PO_4 \cdot H_2O$	Magnesium ammonium phosphate
$MgSO_4 \cdot 7H_2O$	Epsom salt
$mg\,kg^{-1}$	milligram per kilogram
$mg\,l^{-1}$	milligram per liter
$mol\,l^{-1}$	moles per liter
N	Nitrogen
N_2	Dinitrogen
Na	Sodium
NH_3	Ammonia
NH_4^+	Ammonium
$(NH_4)_2SO_4$	Ammonium sulfate
NO_2	Nitrite
NO_3	Nitrate
N_2O	Nitrous oxide
O_3	Ozone
P	Phosphorus
ppm	part per million
S	Sulfur
SO_4^{2-}	Sulfate

See also: **Eutrophication**; **Nitrogen in Soils:** Plant Uptake; **Organic Residues, Decomposition**; **Phosphorus in Soils:** Biological Interactions; **Pollution:** Groundwater; **Sulfur in Soils:** Biological Transformations

Further Reading

Brady NC and Weil RR (2002) The nature and properties of soils, 13th edn. Upper Saddle River, NJ: Prentice Hall.

Havlin JL, Tisdale SL, Nelson WL, and Beaton JD (1999) *Soil Fertility and Fertilizers: An Introduction to Nutrient Management*, 6th edn. Upper Saddle River, NJ: Prentice Hall.

Hewitt EJ (1984) The effects of mineral deficiencies and excesses on growth and composition. In: Bould C, Hewitt EJ, Needham P, and Robinson JBD (eds) *Diagnosis of Mineral Disorders in Plants,* vol. 1, pp. 54–110. New York, NY: Chemical Publications.

Lindsay WL (1979) *Chemical Equilibria in Soils*. New York, NY: Wiley-Interscience Publications.

Mengel K and Kirkby EA (1987) *Principles of Plant Nutrition*, 4th edn. Bern, Switzerland: International Potash Institute.

Pierzinsky GM, Sims JT, and Vance GF (2000) *Soils and Environmental Quality*, 2nd edn. Boca Raton, FL: CRC Press.

MACROPORES AND MACROPORE FLOW, KINEMATIC WAVE APPROACH

P F Germann, University of Bern, Bern, Switzerland

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Discussion about the significance of structural soil voids carrying water flow dates back at least to the mid nineteenth century when Dupuit applied Darcy's law to two-dimensional flow towards drainage ditches. Many soil scientists at that time were convinced that water flows almost exclusively through wide structural soil pores, which are thought of as any cavity wider than expected from the mere arrangement of soil particles. Structural voids are due to processes like cryo- and bioturbation, stress and strain, shrinking and swelling, and internal transport of particulate and dissolved matter. They are called macropores and permit small water amounts to move quickly through soils.

The dispute intensified and subsided at times depending on the development of various approaches to water flow and transport of solutes and particles in water-saturated and nonsaturated soils. However, the Richards equation has evolved to become the predominant approach to calculating water flow in unsaturated soils since the 1930s, and solute transport has been treated as a convection–dispersion process since the 1960s. Darcy's law, Richards equation and the convection–dispersion approaches to flow and transport in unsaturated soils do not require information on soil structure. Therefore, any new approach to flow and transport in soils that requires structural information or that is designed to explain the effects of soil structure on transport tends to rekindle the controversy and perplexity.

The last three decades have seen intense development in field and laboratory instrumentation, electronic data acquisition and processing, and a generally increased demand for information about environmental processes. Thus, it was recognized that information on soil structure is required to explain better the observed transport of water, solutes, and particles in soils. Knowledge about the operational links between structure and transport processes in soils is still in its infancy despite the demand for suitable scientific and applied approaches. Numerous computer codes simulate the transport processes, although no commonly accepted theoretical framework of rapid flow and transport in structured soils has evolved so far.

Macropores are thought to be structural voids that are responsible for the observed deviations of transport from predictions with the Richards equation and the convection–dispersion approaches. Many empirical measures about the minimum widths and lengths of macropores have been proposed. Because of the nonexistence of alternatives to the Richards equation and the convection–dispersion approaches, there are, at present, few ways of functionally defining the dimensions of macropores with respect to transport processes.

These difficulties result in two consequences for this article: (1) terms like 'macropores' refer to arbitrarily defined soil structural voids. 'Macropore flow' and 'transport' infer processes strongly related to them, but without being properly based on independent and more general flow theory. Therefore, the process-oriented expressions 'preferential flow' and 'preferred' flow paths will be used; (2) because of the many evolving alternate approaches to transport in structured soils, this article cannot assume to be complete and final. Thus, reference to past achievements may illustrate the dynamics of the topic and hint at its evolution.

Early Considerations on Flow and Transport in Soils

Poiseuille, presumably the first researcher rigorously dealing with fluid flow in closed conduits, stated in

1844 that laminar flow through a cylindrical pipe is proportional to the pressure gradient and the fourth power of the pipe's diameter, and inversely proportional to the length of the pipe and the viscosity of the fluid.

Darcy, possibly inspired by the emerging laws of resistance-to-flow – Ohm's law dealing with electric currents, Poiseuille's law dealing with flow of liquids in closed conduits, Laplace's and Fourier's work on heat flow – found in 1856 that flow through a completely water-saturated porous medium is proportional to the medium's cross-sectional area and the gradient of the hydraulic head, and inversely proportional to the length of the macroscopic flow path. The factor of proportionality is called hydraulic conductivity of the water-saturated medium.

In the 1880s agricultural scientists in England reported an early breakthrough concerning drainage flow of water with a chemical composition similar to that of rain water, whereas the composition of drainage flow usually reflected the chemical soil environment. This phenomenon was later dubbed as 'by-pass flow'.

At the beginning of the twentieth century, agricultural plant physiologists discovered that the air-filled soil pores must exceed 5% by volume to guarantee proper aeration of pot-grown sunflowers. At about the same time, Swiss forest scientists, in their quest to justify vast reforestation projects by demonstrating the forests' superiority over many other types of canopies to mitigate peak runoff in steep upland catchments, developed a test that linked infiltration with soil porosity information. A steel cylinder 100 mm high and with a volume of 1 l was completely driven into the soil and covered with a wire mesh. A second cylinder of the same dimensions was tightly set on the first one and filled with water to its rim, and the time for the 100 mm of water to infiltrate completely was recorded. Soil cores were sampled nearby with the same cylinder type. They were completely saturated with water, weighed, drained on a gravel bed for 24 h and weighed again. The weight difference was divided by the density of water and the volume of the core, and its multiplication with 100% resulted in the core's air capacity. More than 50 pairs of infiltration times vs air capacities led to the results shown Figure 1.

The results support the 5%-threshold reported by the plant physiologists. In the 1970s, additional support of the 5%-threshold followed from measurements of oxygen conductivity vs air-filled porosity. The experiments indicate that pores quickly draining to about 5% by volume are hydraulically well-connected.

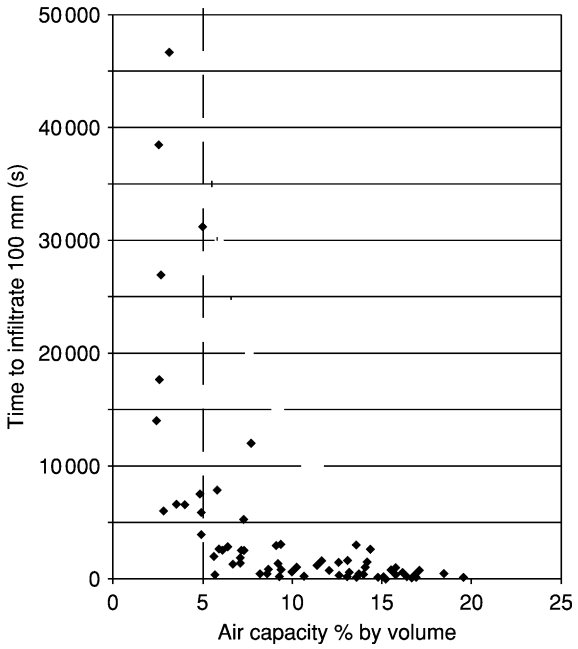


Figure 1 Time to infiltrate 100 mm of water vs air capacity. The empirical threshold at 5% by volume approximately separates the easily-drainable soil samples from those with reduced permeability.

Deviations from Richards Equation and Convection–Dispersion Approaches

Some of the early approaches to flow in unsaturated soils included information on porosity; however, many were not theoretically based and they had to yield to a more rigorous approach. In the late 1920s, Richards included the capillary potential Ψ of the water in unsaturated soils in the hydraulic gradient. He then combined Darcy's law with the volume balance and assumed instantaneous equilibration during transient flow between water content θ and Ψ according to the retention curve. The assumptions led to a quasi-hydrostatic approach to transient flow that expresses the diffusion of water potentials. He called the hydraulic conductivity function $K(\Psi$ or $\theta)$, an expression of proportionality.

Veihmeier's empirical concept of field capacity is close to the one of the 5% threshold of air-filled pore space. It hints at domains not predictable by the Richards equation. However, no theoretical explanation followed, presumably due to the lack of alternative approaches to the Richards equation.

In the 1960s, the concept of hydrodynamic dispersion was introduced to explain the spread of concentrations during solute transport through porous media. Ideally, half of the concentration difference is attained in constant drainage flow when one pore volume of soil water has been replaced. Likewise, at

some depth in the soil the normalized time variation of the solute concentration follows the S-shaped integral of the normal distribution function.

The increasing number of *in-situ* experiments and observations of flow at the scales of soil profiles, the vadose zone to unconfined groundwater tables, hillslopes, and first-order catchments, increasingly demonstrated that small-volume fractions of soil moisture may move rapidly and over long distances, and that the processes were not satisfactorily explained by the Richards equation. The deviating observations are summarized under the term of non-equilibrium conditions with respect to the Richards equation, and macropores of some length are thought to be the cause. In the late 1970s, the hunt started for concepts that supposedly link structure with transport.

Particular interpretations of water retention curves, Ψ vs θ, in terms of the equivalent pore diameters near saturation, were thought to transform nonequilibrium flow conditions to conform to the Richards equation. Also dual-porosity media have been proposed, with one fraction responding rapidly to infiltration and the other fraction showing more gradual responses, allowing for water to flow from one fraction to the other depending on the hydraulic gradient between them. In addition, observed breakthroughs of dye and particle tracers were frequently faster and distorted when compared with the predictions of the convection–dispersion approaches. A broad variety of transport concepts are offered in the literature based on dual-porosity, neural networks, fuzzy logic, and fractal dimensions.

However, the explanation of finger flow is the most subtle case of nonequilibrium flow and does not require structural pores. The hysteresis in the Ψ-θ relationship initiates finger flow that shows in heterogeneous water breakthroughs in soils in the shapes of vertical fingers. Various theoretical approaches to finger flow have been developed, all of which indicate intricate relations between soil and flow parameters.

So far, macropores and macropore flow, preferential flow and preferred flow paths as well as nonequilibrium flow and transport are explicitly or implicitly defined in contrast to 'ordinarily' sized pores and flows at equilibrium with respect to the Richards equation and the convection–dispersion concepts.

Flow paths are frequently visualized *in situ* with dye tracers that are applied to the infiltrating water with subsequent preparation and recording of the stained areas. The staining procedures are applied to sample sizes ranging from typical thin sections covering a few square centimeters to pedons and polypedons. The color patterns provide a visualization of infiltration and indicate the paths taken by the water; however, they do not provide information on the flow dynamics and reveal neither the quantity nor the duration of flow along them. Thus, they are of limited value when developing quantitative approaches to flow and transport in structured soils.

Alternative Approaches to Flow and Transport

General Requirements

Alternative approaches to the Richards equation dealing with transient water flow in unsaturated structured soils have to go back to the basic Navier–Stokes equations to avoid the assumption of equilibrium along the Ψ-θ relationship.

Flow and transport in structured soils and in their artificial physical counterparts are increasingly visualized with computer-aided tomography using X-ray, nuclear magnetic resonance (NMR), or various other radiation techniques. Information on the three-dimensional structure and transport is then used to solve the appropriate Navier–Stokes equations at the pore scale by either applying Boltzman transformations or lattice gas simulations. These process-oriented studies are laborious and require a combination of sophisticated tools and skills. They provide excellent insight into the processes, but up-scaling their results to larger soil volumes is still challenging, mainly because of the pronounced nonlinearity of the processes involved.

According to Newton, a solid body is accelerated when a constant force is acting on it, whereas the same constant force acting on a fluid results in its moving with a constant velocity after a short period of acceleration. Likewise, a solid body gains momentum, whereas a fluid achieves constant momentum when both are exposed to the same constant force. The difference is due to the fluid's viscosity which dissipates linear momentum $\rho \cdot \partial v/\partial f$ perpendicular to the direction of flow towards the solid. However, Richards equation applies to flow in soils as long as momentum completely dissipates within a representative elementary volume, i.e., when the Richards condition is fulfilled.

Preferential flow may expressly consider linear momentum and kinetic energy of transient flow in structured soils. Dissipation of linear momentum during laminar flow results in a shear force in the direction opposite to the flow driving force. Newton's law of shear states that:

$$\tau(f) = -f \cdot g \cdot \rho = \eta \cdot \rho \cdot \frac{\partial v}{\partial f} \qquad [1]$$

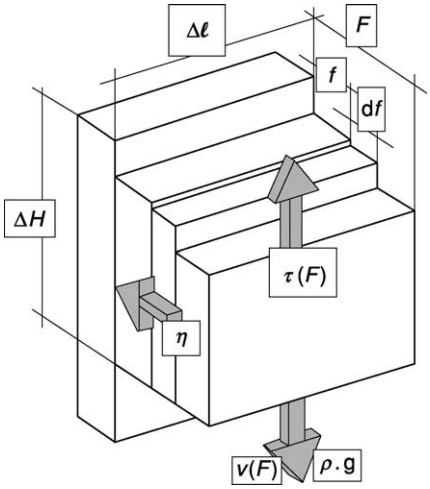


Figure 2 Schematic representation of Newton's law of shear. F(m) Film thickness with its variable f; g (m s^{-2}) acceleration due to gravity; ρ (kg m^{-3}) density of water; $\Delta\ell$ (m) contact length of the water film with the parts at rest; ΔH (m) macroscopic length of coherent water films; η (m^2 s^{-1}) kinematic viscosity dissipating momentum towards the resting part of the soil-water system; $\tau(F)$ (Pa) and $v(F)$ (m s^{-1}) are the shear force and the velocity of the lamina at F.

where $\tau(f)$ (Pa) is the shear force as a function of the film thickness f(m) acting in the opposite direction of gravity (gravity is assumed to be the only flow driving force), g ($=9.81$ m s^{-2}) is acceleration due to gravity pointing vertically downwards, ρ ($=1000$ kg m^{-3}) is the density of water, η (approx. 10^{-6} m^2 s^{-1}) is the kinematic viscosity of water (i.e., the coefficient of momentum dissipation), and $v(f)$ (m s^{-1}) is the velocity of a water lamina at distance f from the parts at rest. Figure 2 illustrates the definitions of the parameters.

Integration of eqn [1] leads to the velocity profile:

$$v(f) = \frac{1}{2\eta} \cdot f^2 \cdot g \qquad [2]$$

Mass and momentum of a layer at f are:

$$dM = \Delta H \cdot \Delta\ell \cdot df \cdot \rho \qquad [3]$$

and:

$$dI(f) = dM \cdot v(f) = \frac{1}{2\eta} \Delta\ell \cdot g \cdot \rho \cdot \Delta H \cdot f^2 \cdot df \,(\text{kg m}^{-1}\,\text{s}^{-1}) \qquad [4]$$

where ΔH is the macroscopic length of well-connected water films. Integration of eqn [4] within $0 \le f \le F$ yields horizontal momentum dissipation due to ρ, g and η from the mobile liquid towards the resting parts, like the solid matrix or stagnant water:

$$I_h(F) = -\frac{1}{6\eta} F^3 \cdot \rho \cdot g \cdot \Delta H \cdot \Delta\ell \,(\text{kg m}^{-1}\,\text{s}^{-1}) \qquad [5]$$

The combination of eqns [1] and [2] yields the vertical momentum of the water film:

$$I_v(F) = \frac{1}{6\eta} F^3 \cdot \rho \cdot g \cdot \Delta H \cdot \Delta\ell \,(\text{kg m}^{-1}\,\text{s}^{-1}) \qquad [6]$$

Steady flow occurs when $I_v = I_h$, acceleration when $I_v > I_h$ and deceleration when $I_v < I_h$.

Mobile volumetric soil moisture is:

$$w = \frac{\Delta H \cdot F \cdot \ell}{\Delta H \cdot A} (\text{m}^3\,\text{m}^{-3}) \qquad [7]$$

where A (m^2) is the cross-sectional area of the soil and $\ell = \Sigma\Delta\ell$(m) is the total contact length of w in A. Integration of eqn [2] under the consideration of eqn [7] leads to the volume flux density of momentum-dominated flow as:

$$q = b \cdot w^a (\text{m s}^{-1}) \qquad [8]$$

where the conductance is:

$$b = \frac{g}{2 \cdot a \cdot \eta} \cdot \frac{A^2}{\ell \cdot \Delta H} (\text{m s}^{-1}) \qquad [9]$$

The dimensionless exponent a is related to the channel geometry. Flow in a cylinder, as in Hagen-Poiseuille's law, results in $a = 2$, whereas flow along a plane or in a planar fissure, also referred to as the cubic law, results in $a = 3$.

Kinematic Wave Theory

The application of eqns [1] to [9] to transient laminar flow leads to the kinematic wave theory as developed by Lighthill and Witham in the early 1950s. It combines Eulerian and Lagrangean concepts of flow where the celerity c (m s^{-1}) describes the velocity of a particular wave property.

The water balance of a kinematic wave is:

$$\frac{\partial q}{\partial t} + c \cdot \frac{\partial q}{\partial z} = 0 \qquad [10]$$

where t and z are time and depth coordinates, the latter is positive vertically down.

The initial and boundary conditions of a rectangular input pulse to the soil surface of duration t_s and volume flux density q_s are:

$$0 \le z \le \infty \quad t \le 0 \quad w(z,t) = 0 \quad q(z,t) = 0 \qquad [11]$$

$$z = 0 \quad t \le 0 \quad \text{and} \quad t \ge t_S \quad w(z,t) = 0 \quad q(z,t) = 0 \qquad [12]$$

$$0 \le t \le t_S \quad w(z,t) = w_S; \quad q(z,t) = q_S = b \,.\, w_S^a \qquad [13]$$

A sharp wetting shock front starts moving vertically down at $t = 0$ with a celerity of:

$$c_W = \frac{q_S}{w_S} = b \cdot w_S^{a-1}\ (\mathrm{m\,s^{-1}}) \qquad [14]$$

The position of the wetting shock front as function of time is:

$$z_W(t) = t \cdot b \cdot w_S^{(a-1)}\ (\mathrm{m}) \qquad [15]$$

and the time lapsed for the wetting front to move to depth z is:

$$t_W(z) = \frac{z}{b \cdot w_S^{(a-1)}} \qquad [16]$$

At $t = t_S$, water input and mobile soil moisture at the soil surface drop to zero. However, the sharp draining shock front at the surface decays to an infinite number of layers (i.e., laminae) in the range of $0 \le w \le w_S$. Each lamina moves with a celerity:

$$c(w) = \frac{\mathrm{d}q}{\mathrm{d}w} = a \cdot b \cdot w^{(a-1)} = \frac{z(w)}{(t - t_S)}\ (\mathrm{m\,s^{-1}}) \qquad [17]$$

corresponding to the position f within the water film. The fastest lamina is the draining shock front whose celerity is:

$$c_D = \frac{\mathrm{d}q_S}{\mathrm{d}w_S} = a \cdot b \cdot w_S^{(a-1)}\ (\mathrm{m\,s^{-1}}) \qquad [18]$$

The depth of the draining front as function of time is:

$$z_D(t) = (t - t_S) \cdot a \cdot b \cdot w_S^{(a-1)}(\mathrm{m}) \qquad [19]$$

and the time lapsed to its arrival at z is:

$$t_D(z) = t_S + \frac{z}{a \cdot b \cdot w_S^{(a-1)}} \qquad [20]$$

The position in depth and time of any mobile water content in the range of $0 \le w \le w_S$ and at time $t \ge t_S$ follows from solving eqn [17] for:

$$w(z,t) = \left[\frac{z}{(t - t_S) \cdot a \cdot b}\right]^{\frac{1}{(a-1)}} \qquad [21]$$

Because $c_D = a \cdot c_w$ the draining front intercepts the wetting front at time:

$$t_I = t_S \frac{a}{(a-1)}\ (\mathrm{s}) \qquad [22]$$

and depth:

$$z_I = t_S \cdot \frac{a}{(a-1)} \cdot b \cdot w_S^{(a-1)}\ (\mathrm{m}) \qquad [23]$$

The total water volume that has infiltrated into the soil and that is contained in the kinematic wave is:

$$V = t_S \cdot q_S = t_S \cdot b \cdot w_S^{a}(\mathrm{m}) \qquad [24]$$

Integration of eqn [21] i.e., $\int_0^{zF(t)} w(z,t)\mathrm{d}_z = V$, yields the position of the wetting shock front for times $t \ge t_I$ and depths $z_F(t) \ge z_I$:

$$z_F(t) = a \cdot b \cdot w_S^{(a-1)} \cdot \left[\frac{t_S}{(a-1)}\right]^{\frac{(a-1)}{a}} \cdot (t - t_S)^{1/a} \qquad [25]$$

Inserting eqn [25] into eqn [21] yields mobile soil moisture at the wetting front at $z_F(t)$:

$$w_F(t) = w_S \cdot \left[\frac{t_S}{(a-1) \cdot (t - t_S)}\right]^{1/a} \qquad [26]$$

and the celerity of the wetting front is:

$$c_F(t) = b \cdot w_S^{(a-1)} \cdot \left[\frac{t_S}{(a-1) \cdot (t - t_S)}\right]^{\frac{(a-1)}{a}} \qquad [27]$$

Equation [21] applies to the ranges of time $t > t_D$ and depth $0 < z < z_D$, z_F. It describes the trailing wave of a kinematic wave. The combination of eqn [19] with eqn [21] yields the decline of mobile soil moisture at a particular depth after the passing of the draining front ($t_S \le t \le t_I$) or the wetting front ($t \ge t_I$):

$$w(z_D,t) = w_S \cdot \left[\frac{(t_D - t_S)}{(t - t_S)}\right]^{\frac{1}{(a-1)}} \quad \text{or}$$
$$w(z_F,t) = w_S \cdot \left[\frac{(t_F - t_S)}{(t - t_S)}\right]^{\frac{1}{(a-1)}} \qquad [28]$$

A complete kinematic wave is shown in **Figure 3**. Application of eqn [8] leads to the volume flux densities, and a combination of eqns [7] and [9] provides the film thickness and the contact length of mobile soil moisture with the resting parts.

Kinetic energy of flow per unit volume of mobile soil water is:

$$E_{kin,w} = \rho \cdot \frac{q^2}{2 \cdot w^2} = \rho \cdot b^2 \cdot w^{2(a-1)}\ (\mathrm{Pa}) \qquad [29]$$

and momentum of flow per unit volume of mobile soil water is:

$$i_w = \rho \cdot \frac{q}{w} = \rho \cdot \mathrm{b} \cdot \mathrm{w}^{(a-1)}(\mathrm{kg\,m^{-2}\,s^{-1}}) \qquad [30]$$

Application of the Kinematic Wave Model

Sprinkling at the soil surface during $0 < t < t_S$ with an intensity of q_S provides for the boundary conditions, eqns [11] to [13]. Because t_I depends only on

t_S, eqn [22], the duration of the latter is selected long enough for the maximum depth of observation to be within $0 < z < z_I$. The kinematic wave approach is applied either to drainage flow at some depth or to soil moisture variations at several depths. Temporal soil moisture variations are typically recorded with time-domain reflectometry (TDR) measurements using horizontally installed waveguides and frequencies of data recording higher than 1/300 Hz. (The system takes repeated water-content readings at each waveguide at intervals of 300 s, i.e., every 5 min.) A time series of drainage flow or soil moisture in the depth range of $0 < z < z_I$ follows three stages:

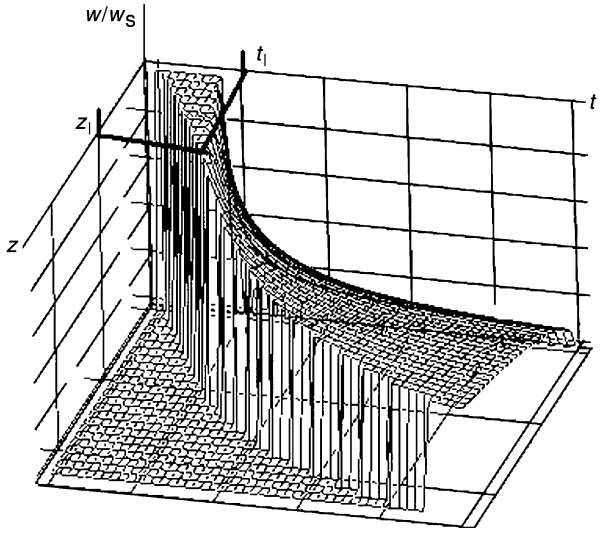


Figure 3 Schematic representation of a kinematic wave normalized with respect to the wave's amplitude w_S.

$$\text{Stage 1: } 0 < 1 < t_W(z) \quad w = 0 \qquad q = 0 \qquad [31]$$

$$\text{Stage 2: } t_w(z) < t < t_D(z) \quad w = w_S \qquad q = q_S \qquad [32]$$

$$\text{Stage 3: } t_D(z) < t < \infty \quad w = w_S \cdot \left(\frac{t_D(z) - t_S}{t - t_S}\right)^{1/(a-1)}$$
$$q = q_S \cdot \left(\frac{t_D(z) - t_S}{t - t_S}\right)^{a/(a-1)} \qquad [33]$$

An example of a model application to data is shown in Figure 4, where the overall soil moisture maximum w_S is replaced by its local amplitude w_{max}.

Soil layers of lower hydraulic conductivity, mainly when they are lacking paths of preferential flow, may backlog water with subsequent distortion of the wave's shape. Water backlogging may also temporarily produce positive pressures, thus requiring adaptation of eqn [1]. In addition, capillary attraction may adsorb parts of the rapidly flowing water at low antecedent degrees of saturation, thus markedly reducing the effects of preferential flow at depths. These more local effects are thought to distort a kinematic wave such that no single set of parameters a and b has so far been found capable of quantifying flow over a considerable depth range. Thus, the kinematic wave model needs calibration to data at the various depths of measurement. However, once calibrated, the model allows for the estimation of flow parameters at particular depths, as their approximate ranges are compiled in Table 1.

Moreover, the local quantification of flow with the kinematic wave model at various depths provides information on the water sorption processes during

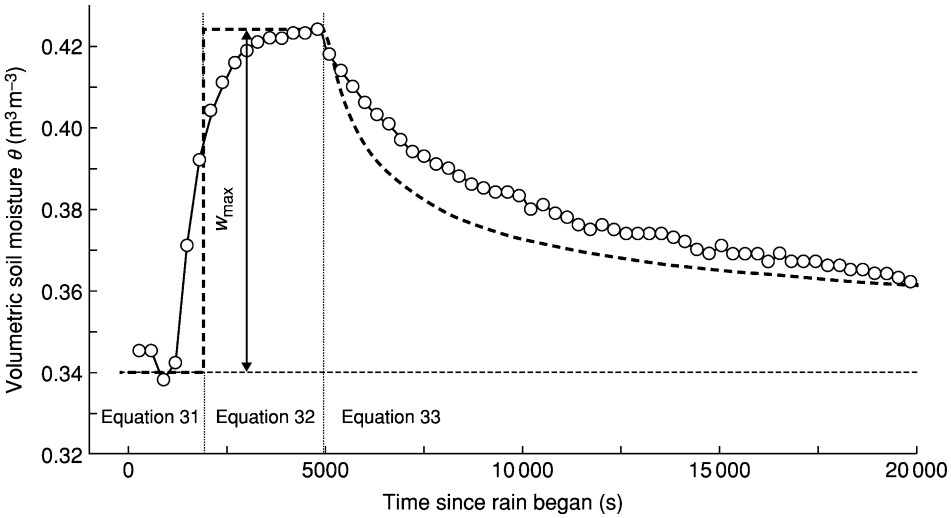


Figure 4 Soil moisture variation at depth 0.35 m due to sprinkling 125 mm during 4500 s (run 1, site Rütti). The local amplitude w_{max} replaces w_S; t_W and t_D are the arrival times of the wetting and draining fronts. The ranges of equation applicability are also shown. (See also Figure 5.) --- model; -o- data.

Table 1 Ranges of flow parameters

Parameter	*Symbol*	*Unit*	*Minimum*	*Typical value*	*Maximum*
Input rate	q_s	$m\,s^{-1}$	3.10^{-5}	8.10^{-6}	2.10^{-6}
Mobile soil moisture	w_S	$m^3\,m^{-3}$	0.003	0.03	0.12
Exponent	a	—	2.0	3.0	6.0
Conductance	b	$m\,s^{-1}$	0.001	0.3	5.0
Contact length of w_S per cross-sectional area	ℓ	$km\,m^{-2}$	0.1	5.0	30
Film thickness	F	μm	2.0	15	50
Momentum per w_S	i_w	$kg\,m^{-2}\,s^{-1}$	0.01	0.5	1.0
Kinetic energy per w_S	$E_{kin,w}$	Pa	10^{-5}	10^{-4}	10^{-3}

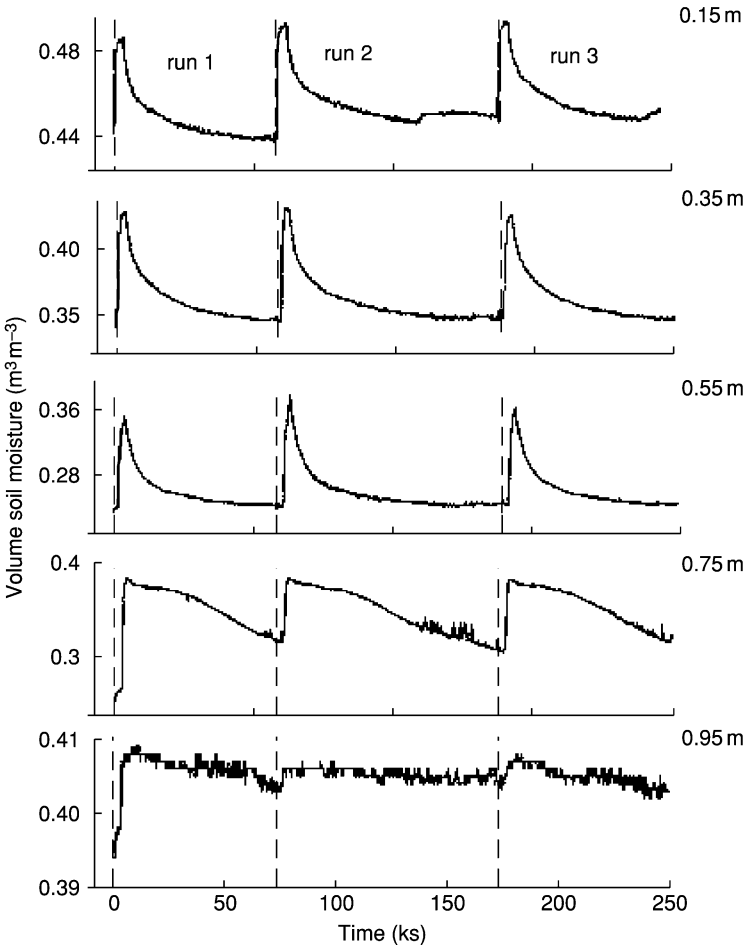


Figure 5 Soil moisture variation at depths 0.15, 0.35, 0.55, 0.75, and 0.95 m due to sprinkling 125 mm three times during 4500 s (site Rütti). The kinematic waves are easily discernible at the upper three depths of measurements. The convex reactions at depth 0.75 m indicate water backlogging above an impermeable layer between this and the 0.95-m depth. Note the different scales on the ordinates.

preferential flow between the two particular depths of soil moisture measurements. Examples of $\theta\,(Z, t)$, variations indicating deviations from the kinematic wave approach are shown in Figure 5.

The contact lengths and film thickness indicate that preferential flow is through a large number of small pores and not exclusively through 'macropores' of not-so-well-defined geometries. The range of the amplitudes includes the empirical threshold of 5% by volume of air capacity and air-filled porosity, as shown in Figure 1. On one hand, the kinetic energy per unit of flowing water is orders of magnitude smaller than the usual range of capillary potential in unsaturated soils, indicating their mutual independence. On the other hand, momentum per unit of flowing soil water is about 100 times greater than

its counterpart of the infiltration rate, indicating channeling of flow in soils. Regarding the fact that capillarity was observed in a pipe of 12 mm inner diameter, preferential flow may concur with Richards-type flow in an approximate pore size range of $6\,\text{mm} > R > 30\,\mu\text{m}$, which represents a relative range of about 1:200. Momentum of infiltration in combination with a minimum degree of antecedent saturation seem to be the two decisive flow parameters that may trigger preferential flow if the geometry of the soil structural voids permit a wave's coherence over a considerable length of flow paths.

Soil hydrologists deal mainly with vertical preferential flow, whereas catchment hydrologists relate preferential flow with lateral subsurface storm flow. The link between the two aspects requires bending of flow from the vertical to the lateral direction at some soil depth. Research on momentum of flow includes the investigation of its scalar and its vector. Early experiments in a layered soil on a hillslope using three-dimensional arrangements of TDR-waveguides indeed revealed bending of preferential flow.

Further Developments of Kinematic Wave Approaches to Preferential Flow

The kinematic wave approach, eqns [1] to [30], explains local soil moisture variations reasonably well. However, no single wave has been found so far that would explain infiltration at the profile scale. Water losses from the preferred flow paths into the soil matrix were assumed to be responsible for the deviations. Therefore, a sink function was introduced in the mid-1980s that rides on the kinematic wave. The water balance of a kinematic wave with a sink-function extension is:

$$\frac{\partial q}{\partial t} + c \cdot \frac{\partial q}{\partial z} = c \cdot w \cdot r \qquad [34]$$

with the sink strength $r\,(\text{s}^{-1})$ defined as:

$$\frac{\partial w}{\partial t} = r \cdot w\,(t) \qquad [35]$$

Due to the sink the wetting front may not penetrate beyond the depth of:

$$z^* = \frac{a \cdot b \cdot w_S^{(a-1)}}{r \cdot (a-1)} \qquad [36]$$

However, no improvement was found when calibrating the extended kinematic wave model, eqns [34] to [36], against data. It seems that water losses from the kinematic wave to the surrounding matrix pores are more complicated than assumed with eqn [35].

An approach based on a traveling diffusive wave was introduced by Di Pietro in 2002. Its water balance is:

$$\frac{\partial q}{\partial t} + c \cdot \frac{\partial q}{\partial z} = u \frac{\partial^2 q}{\partial z^2} \qquad [37]$$

where u $(\text{m}^2\,\text{s}^{-1})$ treats the wave's spreading during flow. The approach based on eqn [37] was successfully calibrated against observed drainage flow from a macroporous soil.

Applications

The concepts of preferential flow may apply to areas outside the hydrology of soils and hillslopes.

The transport of particulate matter through soils is related to the kinetic energy of flow, eqn [29], which results in a drag force acting on suspended particles. Recent investigations demonstrated that breakthrough of latex beads with diameters ranging from 0.5 to 1.75 μm and having electrically neutral surface charges was more strongly related with the kinetic energy of flow than with the variable particle diameters. In contrast, breakthrough of bacteriophages showed very strongly the effects of their variably charged surfaces.

In many areas of the industrialized world, soil compaction is the result of the employment of increasingly more efficient and thus increasingly heavier agricultural machinery. Rotating wheels may not only compact a soil in the vertical direction but knead it back and forth to some depth. The pores allowing preferential flow, rapid drainage, and efficient aeration may be squeezed at times below the plow layer. Soil strength decreases with increasing soil moisture. Thus a vicious cycle starts in that some squeezed preferred flow paths delay drainage and prolong the periods of reduced soil strength. Compaction potential thus increases if the successive soil operation happens under relatively wet soil conditions. Infiltration experiments with subsequent soil moisture recording with TDR-equipment at several soil depths allows an assessment of whether the structural pores are still adequate for drainage and aeration in the 5%-volume fraction range shown in **Figure 1**.

Summary

Poiseuille's early application of momentum dissipation to flow in closed conduits was not considered suitable to deal with flow in porous media. Darcy's law and the Richards equation were developed to deal with such flow. They do not require *a priori* information on soil structure, and the latter is based on the diffusion of potential.

Momentum dissipation approaches were applied to laminar flow in structured and nonsaturated soils with some success, still leaving open many areas for research. The major breakthrough in understanding flow in nonsaturated and structured porous media is expected to come from diffusion wave approaches that seem to lead from strict momentum dissipation and to include diffusive processes. Solute and particle transport in structured and nonsaturated soils will possibly follow similar tracks.

The relative range of pore diameters that are able to carry Richards-type flow as well as preferential flow is about 1:200. Momentum of infiltration is thought to trigger preferential flow when it is exceeding some minimum. The importance of the uppermost soil layer as being able either to transmit or mitigate momentum to deeper layers for the initiation of or prevention of preferential flow is also implicated.

List of Technical Nomenclature

η	Kinematic viscosity of water ($m^2 s^{-1}$)
θ	Volumetric soil moisture ($m^3 m^{-3}$)
ρ	Density of water ($kg m^{-3}$)
τ	Shear force per cross-sectional area (Pa)
ψ	Capillary potential (Pa)
A	Cross-sectional area of soil (m^2)
a	Exponent of kinematic wave (–)
b	Conductance of kinematic wave ($m s^{-1}$)
c	Celerity of kinematic wave ($m s^{-1}$)
c_D	Celerity of the draining front ($m s^{-1}$)
c_F	Celerity of the wetting front after interception ($m s^{-1}$)
c_W	Celerity of the wetting front ($m s^{-1}$)
$E_{kin,w}$	Kinetic energy per unit volume of mobile water (Pa)
F	Thickness of water film (m)
f	Film thickness variable (m)
g	Acceleration due to gravity ($m s^{-2}$)
H	Vertical macroscopic length of connected water films (m)
I	Momentum of water film ($kg m s^{-1}$)
I_h	Horizontal momentum loss of water film ($kg m s^{-1}$)
I_v	Vertical momentum of water film ($kg m s^{-1}$)
i_W	Momentum per unit volume of mobile water ($kg m^{-2} s^{-1}$)
K	Hydraulic conductivity ($m s^{-1}$)
ℓ	Contact length of mobile water (m)
M	Mass of soil water (kg)
q	Volume flux density ($m s^{-1}$)
q_s	Volume flux density of input ($m s^{-1}$)
R	Radius (m)
r	Sink strength ($m s^{-1}$)
t	Time (s)
t_D	Arrival time of the draining front (s)
t_I	Time of front interception (s)
t_S	End of input to the soil surface (s)
t_W	Arrival time of the wetting front (s)
u	Coefficient of wave diffusion ($m^2 s^{-1}$)
V	Water volume of kinematic wave per cross-sectional area (m)
v	Velocity of water film ($m s^{-1}$)
w	Mobile soil moisture ($m^3 m^{-3}$)
w_F	Mobile soil moisture at the front after front interception ($m^3 m^{-3}$)
w_{max}	Local amplitude of the kinematic wave ($m^3 m^{-3}$)
w_S	Mobile soil moisture at the surface during infiltration ($m^3 m^{-3}$)
Z	Specified soil depth (m)
z	Depth vertically down (m)
z^*	Maximum depth of front penetration (m)
z_D	Depth of draining front (m)
z_F	Depth of wetting front after front interception (m)
z_I	Depth of front interception (m)
z_W	Depth of wetting front (m)

See also: **Aggregation:** Physical Aspects; **Fauna**; **Morphology**; **Structure**; **Subsoiling**; **Unstable Flow**; **Vadose Zone:** Microbial Ecology

Further Reading

Bosh D and King K (eds) (2001) *Preferential Flow: Water Movement and Chemical Transport in the Environment – Proceedings of the 2nd International Symposium 3–5*

January 2001. St Joseph, Michigan: American Society of Agricultural Engineers.
Di Pietro L, Ruy S, and Capowiez Y (2003) Predicting preferential water flow in soils by travelling diffusive waves. *Journal of Hydrology* 278: 64–75.
Germann PF (2001) A hydromechanical approach to preferential flow. In: Anderson MG and Bates PD (eds) *Model Validation – Perspectives in Hydrological Sciences*, pp. 233–260. Chichester: John Wiley.
Germann PF and Di Pietro L (1999) Scales and dimensions of momentum dissipation during preferential flow in soils. *Water Resources Research* 35(5): 1443–1454.
Gish TJ and Shirmohammadi A (eds) (1991) *Preferential Flow – Proceedings of the National Symposium 16–17 December 1991*. St Joseph, Michigan: American Society of Agricultural Engineers.
Selim HM and Ma L (eds) (1998) *Physical Nonequilibrium in Soils: Modelling and Application*. Ann Arbor: Ann Arbor Press.
Selim HM and Sparks DL (eds) (2001) *Physical and Chemical Processes of Water and Solute Transport/Retention in Soil*. Soil Science Society of America Special Publication No. 56. Madison: Soil Science Society of America.

Magnesium in Soils *See* **Calcium and Magnesium in Soils**

MANURE MANAGEMENT

J T Sims, University of Delaware, Newark, DE, USA
R O Maguire, North Carolina State University, Raleigh, NC, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Animal manures are the solid, semisolid, and liquid by-products generated by animals grown to produce meat, milk, eggs, and other agricultural products for human use and consumption. They are mixtures of animal feces, urine, bedding materials (e.g., straw, sawdust, rice hulls), and other materials associated with animal production, such as waste feed, soil, wash waters, and any chemical or physical amendments used during manure handling and storage. Manures have been used as beneficial soil amendments since the dawn of civilization and were the primary soil amendment used in agriculture until the advent of chemical fertilizers in the 1940s. Today, manures continue to be regarded as valuable agricultural resources, because they are important sources of plant nutrients and are well known to improve soil physical and biological properties through the addition of organic matter. However, the changing nature of animal production in some countries, the move away from small farms that have adequate land bases to recycle manure nutrients through crop production, and the move toward geographically concentrated confined animal feeding operations (CAFOs) have raised concerns about the environmental and human health impacts of modern animal production and manure utilization practices.

Manure Generation, Use, and Properties

Summary of Global Trends in Animal Production

Animal production has increased worldwide in the past 40 years, although it is not distributed evenly over the globe (Figure 1a and Table 1). Long-term trends in animal production are most easily recognized if developed and developing countries are viewed separately, as greater intensification of animal production initially occurred in developed countries. Today, however, many developing countries are rapidly increasing animal production, while in most developed countries the number of animals produced is decreasing or leveling off, as seen for poultry production in Figure 1b. Calories from animal protein cost more to produce than calories from crops. Therefore as a general rule, as poorer countries develop and disposable income increases, the population can more easily afford animal protein, and there is a corresponding increase in animal production. As a result, meat consumption in developing countries is projected to double in the next 20 years, although it will still be below that of developed countries.

Globalization of agriculture has also changed the nature of animal production over the last few decades, increasing the distances that animal feed and animal products travel between sites of production and consumption. For example, poultry and swine produced in the eastern USA may use feed grains grown in the Midwestern US, with the meat

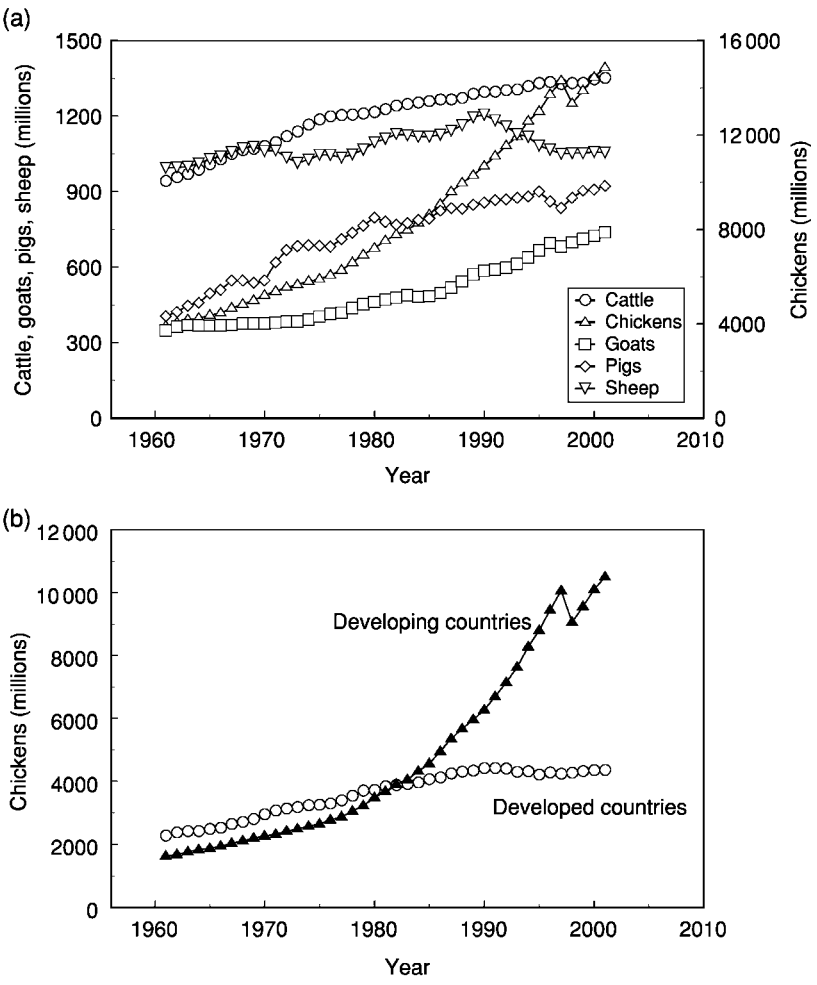


Figure 1 (a) Global trends in animal production, by major animal types (1961–2001); and (b) comparison of global poultry-production trends for developed and developing countries (1961–2002). Data from Food and Agriculture Organization of the United Nations (2003) Available online at: http://apps.fao.org/page/collections?subset-agriculture.

Table 1 Global animal production, by major animal species, for major geographic regions and for developed and developing countries, 2001

	Animal species				
Country	*Cattle (millions)*	*Chickens (millions)*	*Goats (millions)*	*Pigs (millions)*	*Sheep (millions)*
Asia	455	7180	462	551	371
Europe	101	1257	15	164	129
Africa	230	1276	219	18	250
North America	110	1988	1	72	8
South America	309	1765	22	55	75
Developed countries	326	4358	30	282	381
Developing countries	1026	10 501	708	641	675

Data from Food and Agriculture Organization of the United Nations (2003) Available online at: http://apps.fao.org/page/collections?subset-agriculture.

ultimately being consumed in Asia. In many countries these trends have also resulted in marked increases in the geographic intensity of animal production, with more animals being grown in confinement on farms that are located in close proximity to each other and that have the same, or declining amounts of cropland available for manure use. Figure 2 shows the change in the geographic distribution of confined animal production (all species) in the USA from 1982 to 1997.

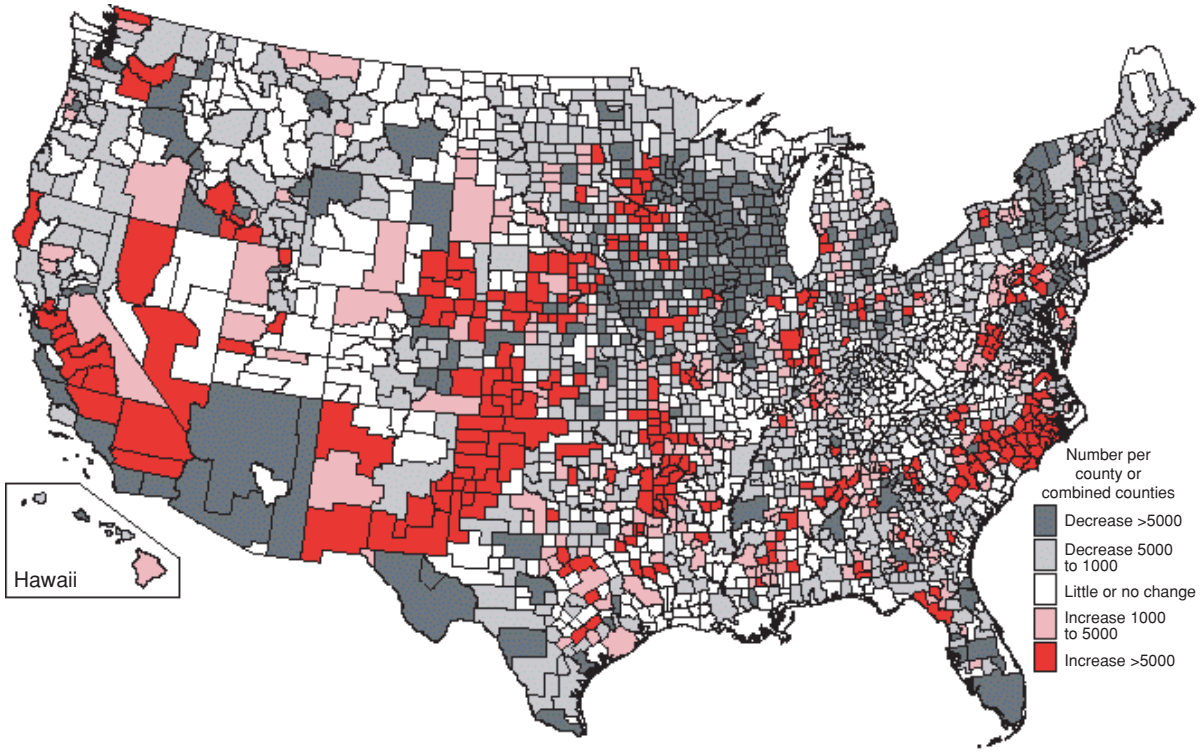


Figure 2 Change in the number of animal units in each county of the USA (1982–97). (Reproduced from Kellogg RL, Lander CH, Moffitt DC, and Gollehon N (eds) (2000) *Manure Nutrients Relative to the Capacity of Cropland and Pastureland to Assimilate Nutrients: Spatial and Temporal Trends for the United States*. Fort Worth, TX: GSA National Forms and Publication Center.)

Trends in Manure Production

The nature of the animal production system directly affects the amount of manure available for use as a soil amendment. For example, unlike the manure from animals grown in confinement (e.g., barns, feedlots), manure is not regularly collected from animals mainly grown on grassland or rangeland. Thus the amount of 'recoverable manure' that is available for land application as a soil amendment can vary widely between and within countries. However, as intensive cattle, swine, and poultry operations are projected to account for an increasing proportion of animal production in both developed and developing countries, collection and spreading of manure will probably become more and more common in the future. Manure production by species also differs widely by country; for example goats are very common in Asia and Africa, while cattle are more predominant in North and South America. In global terms, poultry production has grown the most in the last few decades, followed by pig and goat production, while the number of cattle produced has only increased slightly (Figure 1a). Indeed, in developed countries, poultry meat is seen as the only growth area toward 2010, in terms of per capita consumption. Despite these variations, the amount of manure produced by different animals is reasonably well known, as shown in Tables 2–4, which summarize manure production and characteristics for the major animal species raised today.

In general, the major trend today for animal, and hence manure, production in both developed and developing countries is for increasing intensification and consolidation of producers. This means that in developed countries, even though total manure production is leveling off, it is confined to fewer producers operating at fewer sites, with the result that the amount of manure generated in some regions is increasing markedly. For example, in the USA, the number of livestock operations decreased by 24% between 1982 and 1997, even though total animal production numbers were fairly stable. This concentration of animal production in specific areas in developed countries applies mainly to swine, cattle (beef, dairy, and veal), and poultry (chicken and turkey) production. In developing countries animal production is generally not as intensive but, as these countries develop, their production systems will inevitably become more intensive for the same economic reasons that promoted this trend in developed countries.

Table 2 Liquid-pit manure production and nutrient content for major animal species

Animal	Manure produced (kg year^{-1})	Nutrient concentration (kg 1000 l^{-1})			
		Total N	NH_3-N	P	K
Swine (farrow-finish)	17 007	3.4	1.9	1.3	2.3
Dairy (cow)	24 490	3.7	0.7	0.8	1.9
Dairy (heifer)	11 338	3.8	0.7	0.7	2.8
Beef (cow)	13 605	2.4	0.8	0.8	2.4
Poultry (layer)	59	6.8	4.4	2.7	3.3
Poultry (broiler)	38	7.6	1.6	2.1	2.9
Poultry (turkey)	128	6.4	1.9	2.1	2.9

Manure production values are expressed on a per-animal basis for beef, dairy, and swine and per bird space for poultry.
Adapted from Midwest Plan Service (2000) *Manure Characteristics. Midwest Plan Service*. Arnes, IA: Iowa State University Press.

Table 3 Lagoon-effluent manure production and nutrient content for major animal species

Animal	Manure produced (kg year^{-1})	Nutrient concentration (kg year^{-1})			
		Total N	NH_3-N	P	K
Swine (farrow-finish)	29 025	16	15	5	11
Dairy (cow)	41 270	21	19	4	12
Fattening cattle	19 955	12	11	4	10

Manure production values are expressed on a per-animal basis for beef, dairy, and swine and per bird space for poultry.
Adapted from Midwest Plan Service (2000) *Manure Characteristics. Midwest Plan Service*. Arnes, IA: Iowa State University Press.

Table 4 Solid-manure production and nutrient content for major animal species

Animal	Manure produced (kg year^{-1})	Nutrient concentration (kg Mt^{-1})			
		Total N	NH_3-N	P	K
Swine (farrow-finish)	7773	5.8	2.5	1.4	1.7
Dairy (cow)	12 698	4.1	0.8	0.5	2.0
Dairy (heifer)	5896	4.1	0.8	0.5	2.4
Beef (cow)	6077	2.9	1.2	0.7	2.4
Poultry (layer)	18	14.0	4.9	9.2	8.9
Poultry (broiler)	8	18.9	4.9	9.5	12.3
Poultry (turkey)	21	16.5	3.3	9.0	10.2

Manure production values are expressed on a per-animal basis for beef, dairy, and swine and per bird space for poultry.
Adapted from Midwest Plan Service (2000) *Manure Characteristics*. Arnes, IA: Iowa State University Press.

In some cases, the geographic intensification of animal production has led to manure generated within confined areas of production, containing nutrients in excess of local crop requirements, while other areas within the region or country have little or no local supply of manure nutrients for crop production. This is because the economic value of the manure cannot support the costs of its redistribution to areas where it may be needed as a soil amendment. As illustrated for the USA in Figure 3, this has resulted in large 'surpluses' of manure nutrients in some areas where animals are produced and increasing concerns about nonpoint source pollution of air, soil, and water in areas where animal production is concentrated. These concerns arise because farmers who have excess manures, relative to what is needed to produce crops, generally have no economically viable alternatives to land application. Because of this, manure nutrients are often overapplied or applied at inappropriate times of the year, leading to increased nutrient losses to surface water and groundwater by runoff and leaching, and to the atmosphere by volatilization. In some developed countries (e.g., the Netherlands), regulations have been enacted to decrease animal production intensity because of the environmental impacts associated with the disposal of excess manures. Efforts to control manure generation and use are expected to increase in developed countries in the future; however, in developing countries this may not occur until manure and nutrient excesses begin to create similar environmental problems.

Manure Characteristics by Species

The physical properties and nutrient content of animal manures are reasonably well known and tables summarizing manure characteristics have been developed for use by farmers and their advisors (Tables 2–4). Values such as these, however, are only guidelines, as it is recognized that manures can vary widely in composition both between and within species. Therefore, regular testing of manures for properties that influence their agricultural value and potential environmental impacts is often advised. Results from manure testing have consistently shown that differences in manure nutrient content between animal species are apparent (Tables 2–4). For example, both N and P concentrations in manure increase in the order cattle < swine < sheep and goats < poultry. Poultry litter contains two to four times the concentration of manure nutrients as other livestock manures, which leads to particular problems in areas where the poultry industry has intensified, compared with areas of other intensified livestock industries. For example, in the USA, of the 73 counties that

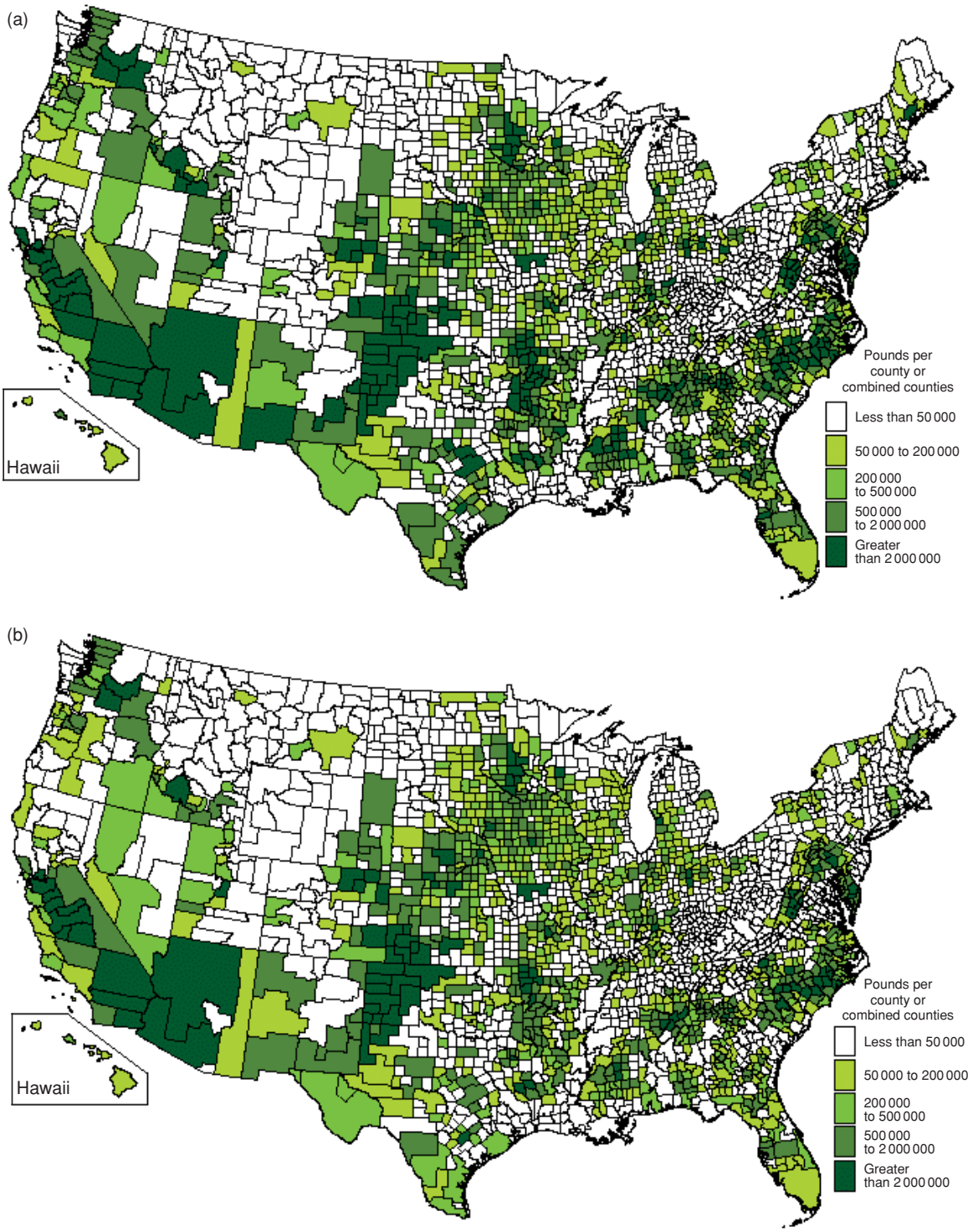


Figure 3 Potential excess of manure (a) nitrogen and (b) phosphorus for each county of the United States in 1997, assuming no export of manure from farms. (Reproduced from Kellogg RL, Lander CH, Moffitt DC, and Gollehon N (eds) (2000) *Manure Nutrients Relative to the Capacity of Cropland and Pastureland to Assimilate Nutrients: Spatial and Temporal Trends for the United States*. Fort Worth, TX: GSA National Forms and Publication Center.)

had insufficient land for the amount of manure N generated, 82% had poultry as the dominant livestock, while of the 160 counties with excess manure P, 64% were dominated by poultry production (Figure 3).

The variability in manure composition is due to a number of factors, including animal species and age, diet, digestibility of feed materials, environmental conditions in animal housing facilities, and the

means by which manures are handled and stored prior to use or disposal. For example, the amount of feed concentrates and other feed additives (e.g., enzymes) in a diet (and thus nutrients in manure) can vary between developed and developing countries; manure storage and drying can reduce nitrogen (N) content through ammonia volatilization; and manure collected from open-housing areas or stored in lagoons can be diluted by rainfall.

Major Uses of Animal Manures

Land application to meet crop N, P, and other nutrient requirements is by far the most widespread and long-standing use of animal manures. However, other uses do exist and are becoming more common, especially in areas where animal production has become so intensive that the amount of nutrients in the manure produced exceeds the nutritional requirements of crops grown on the local land base. Sometimes manure is fed back to animals as a cheap nutrient source (e.g., feeding chicken litter to beef cattle). However human and animal health concerns have been raised about this practice, in the wake of the 'mad cow disease' outbreak in Europe (caused by feeding contaminated animal products back to cattle), and this is not viewed today as a major alternative to land application of manures. Mushroom production in certain areas can use large amounts of manure, especially from horses and poultry. In China, manures are used extensively in fish production. Several other options exist for manure use, and although these uses are only carried out on a relatively small scale today, they represent a growth area, as increasing intensification of agriculture produces excess amounts of manure in localized areas of more countries. Examples of other manure uses are composting to produce high-value soil amendments, methane generation via anaerobic digestion, burning to generate energy, and pelletizing manures to make organic fertilizers.

Best Management Practices (BMPs) for Animal Manures as Soil Amendments

Overview

Years of research have identified the most effective approaches to manage animal manures as agricultural soil amendments and many references provide detailed discussions on this subject. In general, manures have historically been viewed as organic fertilizers and have been applied at rates that would provide adequate nutrients for crops to be grown on the farm, based on realistic yield goals and the results of soil and manure tests that quantify the amount of manure nutrients needed to attain economically optimum yields. In conjunction with this, soil and water conservation practices (e.g., buffer strips, grassed waterways, and cover crops) are usually recommended to prevent off-site movement of manures or manure nutrients. Other benefits of manure use (e.g., increasing soil organic matter content and improving soil physical and biological properties) are recognized but usually do not serve as the basis for determining manure application rates. The most effective BMPs for manure use vary with the type of agriculture being practiced. This can be illustrated by comparing two of the most predominant agricultural settings worldwide: grassland agriculture and arable crop production.

Manure management for grassland agriculture In general, there are two basic systems associated with grassland agriculture: (1) pasturing, and (2) confinement. Animals such as cattle or sheep graze in pastures and deposit manure as they produce it. Thus, manure 'application' is somewhat random and occurs without any form of incorporation into the soil. Where animals are confined indoors for at least part of the year, the manure they generate is collected before being land-applied. For confinement systems, manures are most commonly applied to the surface of pastures without tillage, although in some cases they may be injected into the soil using specialized equipment. With intensification of grassland agriculture, there has been a trend toward animal confinement and manure spreading, and away from solely pasturing animals. It is now estimated that about half of the manure produced in the USA can be collected, with much of the remainder being directly deposited by grazing animals. As more manure is collected, the opportunities to implement BMPs that better utilize the collected manures as soil amendments, or for other end uses, increase.

In general, manures are valuable sources of nutrients for grass and animal production in pasture systems. However, because they are surface-applied, several pathways for the loss of nutrients and other manure constituents (e.g., pathogens, hormones, and antibiotics) exist that are more pronounced than when manures are incorporated into arable soils. These nutrient losses are both economic (loss of plant nutrients) and environmental (deleterious impacts on air and water-quality) problems. For example, substantial gaseous losses of N by volatilization of ammonia occur directly from manures lying on the soil surface. This may increase the need to purchase N fertilizers to compensate for the lost manure N and also contribute to environmental concerns when the volatilized NH_3 is redeposited on to

soils (which may cause soil acidification) or into aquatic ecosystems (where the N may contribute to eutrophication).

As grassland is unlikely to be tilled regularly, there is also a tendency for nutrients to accumulate very near the soil surface as manures slowly decompose. This is particularly true for P, which is strongly adsorbed by soil clays and oxides. When runoff occurs, dissolved and colloidal P either in the manures lying on the surface or contained in the nutrient-rich uppermost layers of the grassland soil can be transported to surface waters, increasing the potential for eutrophication of surface waters.

Managing manures in grassland agriculture thus focuses on the use of BMPs that sustain optimum grass production and also prevent surface and gaseous losses of nutrients and other manure constituents. Examples include the use of soil testing and equipment calibration to prevent overapplication of manures; avoiding poorly timed and badly placed manure applications (e.g., immediately before rainfall events or snowmelts and in close proximity to surface waters or wellheads); use of chemical and physical manure amendments that decrease the solubility or volatility of N and P; and injection of manures below the grassland soil surface, although this is a practice that entails extra technology and costs and is not widely used at present.

Managing manures for arable crop production Unlike manures in grassland agriculture, animal manures used in arable crop production are commonly incorporated into soils by some form of tillage (e.g., plowing, disking, rototilling) or by injection. This is not always the case, however, as many agricultural crops today are produced using reduced or 'no-till' systems, where the soil is not plowed annually or is only tilled to a shallow depth to prepare a seedbed for annual crops. Manures incorporated into soils in arable cropping systems decompose as a function of soil microbial activity, which in turn is dependent upon soil temperature, moisture, aeration, and soil chemical properties (e.g., pH, salinity, nutrient status). As the manures decompose, they release nutrients into the soil solution that can be retained by chemical processes (e.g., ion exchange, adsorption, precipitation) or as part of the soil biota (e.g., incorporation into microbial biomass). Manure decomposition can also affect soil pH, soluble-salts concentration, and soil organic matter content, all of which affect plant growth and the potential for nutrient losses. Much research has been directed toward the development of predictive models that can estimate the rate of manure decomposition in soils and the total amount of nutrients released from different types of manures as a function of soil properties and environmental conditions.

As with grassland agriculture, manures incorporated into arable soils are well known to build soil fertility and improve soil physical properties such as soil structure and tilth and thus are important soil amendments for crop production. Nutrient losses can also occur from arable soils, with the greatest concerns being the erosion of soil particles (and nutrients that are bound to these particles, particularly P) and the leaching of soluble nutrients, especially nitrate-N, through the soil profile to groundwater. Soil erosion concerns are greater with arable crops than grasslands because of the tillage practices associated with manure incorporation and the lack of the permanent plant cover that can filter eroding soil particles from runoff. Gaseous losses of N are much reduced when manures are incorporated into arable soils, because manure ammonium-N is bound to soil colloids and is also not in direct contact with the atmosphere. Similarly, the potential for losses of soluble P is less in arable soils, because manures are not lying on the soil surface and because tillage prevents the development of the highly enriched zones of P accumulation often found in the upper few centimeters of untilled grassland soils.

Manure management for arable crops is similar to that described above for grasslands, with a few key differences. There is a much greater focus on controlling soil erosion and there are greater limitations on the times of year that manure can be applied to annual row crops. In many cases manures must be applied well before planting the crop, which then increases the likelihood for nutrient losses.

Contemporary Approaches to Manure Management

As animal agricultural operations have become increasingly intensive, many problems have emerged that point to the need for improvements in manure-management practices. While these problems are more serious in developed countries, where animal production is most intensified, they are emerging issues in developing countries, where information on how to avoid the environmental problems associated with intensive animal agriculture will be needed in the future.

One of the main problems with traditional approaches to manure management is the fact that, while manures provide most of the nutrients needed by crops, the relative proportions of plant nutrients present in manures are not in balance with the amounts needed by many crops. This can result in under- or overapplication of some nutrients. For

example, manures have usually been viewed as N sources and applied to meet crop N requirements. However, because the N:P ratio in manures is not in balance with the N:P ratio in most crops, applying manures to meet crop N needs, long the recommended practice, has often resulted in the overapplication of P and the buildup of P in soils to values well above those needed for optimum crop yields. Because of this, concerns now exist in many countries about losses of P to surface waters and shallow groundwaters from soils that have received long-term manure applications. Other emerging concerns for manure use as soil amendments include environmental impacts of other manure constituents such as pathogens, hormones, antibiotics, and nonessential trace metals on human health and environmental quality.

In some countries, these concerns have resulted in the development of more-intensive, structured, and even regulatory approaches for manure management, such as the comprehensive nutrient management plans (CNMPs) now recommended by the US Natural Resources Conservation Service for animal feeding operations (AFOs). The six basic components of a CNMP are described below and may apply, in principle, not only to the intensive animal agriculture common in the USA, but to animal production operations worldwide, including the less-intensive systems in many developing countries.

1. Manure and wastewater-handling and storage: This element of the CNMP focuses on the activities at the animal production facility (e.g, barn, feedlot) and any areas where manure is handled and stored prior to land application. The general goals are to minimize the amount of waste produced and to handle, store, treat, and transfer wastes to off-site uses in a manner that minimizes any losses, spills, or direct discharges to the environment. Areas of particular concern include providing adequate manure storage, preventing direct losses of pollutants from the farmstead and manure storage areas to nearby surface waters, minimizing odors and air-quality concerns, and safely disposing of dead animals, spoiled feed, and the like;

2. Land treatment practices: Because land application is the main use of most animal manures, well-designed land-treatment practices are vital components of a CNMP. Factors that must be addressed are: selection of the appropriate rate, method, and timing of application of manures, relative to crop needs; implementation of soil conservation practices that prevent soil erosion and the loss of soluble nutrients from manures and manure-amended soils; and establishment of monitoring programs to assess the effectiveness of the land-treatment practices over time;

3. Nutrient management: This refers to the development of a systematic approach to manage all nutrient sources on the farm, not just manures. Typically this begins with an effort to maintain nutrient balance in the operation, where nutrient inputs in feed, fertilizers, animals, and other sources do not exceed nutrient outputs in crops and animal products, except as needed to build soil fertility to optimum levels for crop growth. Key factors considered in this component of the CNMP include a farm-scale nutrient budget (i.e., is there a nutrient surplus or deficit on the farm?), crop nutrient requirements at realistic yield goals, soil-testing to assess the need for additional nutrients, manure-testing to determine the quantity of manure nutrients available and the need for off-farm nutrients, and the most efficient means to apply all nutrient sources to crops on the farm. Other considerations include how the approach used to manage nutrients might affect air quality, odors, pathogens, and the buildup of soluble salts and heavy metals in soils;

4. Record-keeping: Effective implementation of a CNMP requires good record-keeping. The following information should be recorded, by field, and reviewed to assess current practices and detect trends with time: basic farm characteristics (e.g., amount of land; weather conditions; number and type of animals or animal products produced; type and acreage of crops grown; amount of manure produced, used on the farm or transferred off-site); manure storage, handling, and treatment; soil, manure, and plant tests; crop-management practices (hybrid, planting date, seeding rate, irrigation, yields), and nutrient application practices (source, amount applied, when applied, method applied, equipment calibration);

5. Feed management: Animal-based agricultural operations are increasingly investigating the opportunity to improve nutrient use efficiency by animals, which, in turn, reduces the excretion of nutrients such as N and P. Research has shown that modifying animal diets by reducing the concentration of mineral supplements, or by including enzymes and other feed additives, can result in the same level of animal performance but reduce the amount of N and P excreted. This results in both an economic savings (fewer nutrients are needed for the same level of production) and lower nutrient concentrations in manures, which can be of benefit in areas where nonpoint nutrient pollution of surface and ground waters is of concern;

6. Alternative uses for manures: There is growing interest in developing alternatives to land application for some types of animal manures, particularly in areas where animal production has been geographically intensified and nutrient surpluses exist (see Figure 3). To sustain, or expand, operations in areas

such as these, it is necessary to find alternatives to land application, and this is a central part of an efficient CNMP. This can generate additional income and also help farmers redistribute surplus nutrients to off-site uses. The most common approaches investigated to date have been energy production (methane generation and incineration for electricity), production of value-added products such as composts and pelletized fertilizers, and transportation of manures from areas of nutrient surplus to nutrient deficit, usually mediated by 'manure brokers.'

Summary

The use of animal manures as soil amendments must be done in a comprehensive manner to maximize their agricultural value and minimize their potential impacts on environmental quality and human health. There is no doubt that animal manures are valuable natural resources, particularly in countries where the availability of inorganic fertilizers is limited. When managed properly, manures provide plant nutrients, build soil organic matter, and improve soil physical properties, all of which are vitally important for soil quality and crop production. At the same time, the growing global trend toward intensification of animal production is gradually concentrating manure generation on smaller amounts of land. Transportation costs and the lack of an organized, efficient manure processing and distribution infrastructure are the major limitations to relocation of manures from areas where they are in surplus to areas where they are badly needed. When an adequate land base is not available to recycle manure nutrients in a balanced manner, overapplication of nutrients such as N and P, which can negatively affect air, soil, and water quality, becomes a potential problem. Regional- and/or national-scale planning is required to develop and implement programs that can sustain modern animal production and use animal manures in an environmentally sound manner.

Further Reading

Beegle DB, Carton OT, and Bailey JS (2000) Nutrient management planning: justification, theory, practice. *Journal of Environmental Quality* 29: 72–79.

Food and Agriculture Organization (1995) *World Agriculture: Towards 2010*. New York: John Wiley.

Food and Agriculture Organization (FAO) (2002) http://www.fao.org.

Hatfield JL and Stewart BA (eds) (1998) *Animal Waste Utilization: Effective Use of Manure as a Soil Resource*. Chelsea, MI: Ann Arbor Press.

Kellogg RL, Lander CH, Moffitt DC, and Gollehon N (eds) (2000) *Manure Nutrients Relative to the Capacity of Cropland and Pastureland to Assimilate Nutrients: Spatial and Temporal Trends for the United States*. Fort Worth, TX: GSA National Forms and Publication Center.

Midwest Plan Service (2000) *Manure Characteristics. Midwest Plan Service*. Ames, IA: Iowa State University Press.

Sims JT (1994) Animal waste management. In: *Encyclopedia of Agricultural Sciences*, pp. 185–201. New York: Academic Press.

Sims JT (1995) Organic wastes as alternative nitrogen sources. In: Bacon PE (ed.) *Nitrogen Fertilization in the Environment*, pp. 487–535. New York: Marcel Dekker.

Steele K (1995) *Animal Wastes and the Land–Water Interface*. New York: Lewis Publishers/CRC Press.

MARBUT, CURTIS FLETCHER

J P Tandarich, Hey and Associates, Inc., Chicago, IL, USA

© 2005, Elsevier Ltd. All Rights Reserved.

The brilliance that characterized the life of Curtis Fletcher Marbut, dedicated to soil science in the service of humanity, ended on August 25, 1935 in what was then Harbin, Manchuria, China. He was the best-known soil scientist–geographer of his day, and his reputation was worldwide. Marbut was a pioneering soil scientist in the classroom, office, field, and laboratory. More than that, he was a disseminator of ideas; an internationalist who did not let politics, ideology or his personal health and welfare stand in the way of scientific advancement and friendship. The biography of Marbut is a remarkable story of a 'self-made' man who came from a progressive Ozarks family in rural southwest Missouri to become one of the foremost agricultural geologists, geographers, and soil scientists in the world during the formative period of modern soil science in the early twentieth century. Modern soil science has its origins in late eighteenth- and early nineteenth-century agricultural chemistry and agricultural geology. The term Bodenkunde, coined in 1837 by German agricultural chemist Carl

Sprengel, meant soil knowledge – a collection of facts. Friedrich Albert Fallou, a German agricultural geologist, attempted to elevate Bodenkunde to a formal science of soil that he called Pedologie. The German scientists never adopted the term, preferring to use Bodenkunde to mean soil science. Russian agricultural geologist Vasilii V. Dokuchaev used the word 'pochvovedenie' for soil science. In the twentieth century, pedology came to mean the study of the genesis, morphology (appearance), nature, classification, and geographic distribution of soil.

This short narrative summarizes the highlights of Marbut's life as a native of the Missouri Ozarks, an academic, a scientist with the US Soil Survey, an explorer and internationalist, and a scientist with a Unitarian philosophy of service to humanity for which he ultimately gave his life.

Missouri Ozarks Native and Academic

Marbut was born during the American Civil War, on July 19, 1963, to Nathan and Jane Browning Marbut, along Little Flat Creek near Verona in Lawrence County, Missouri. After surviving whooping cough in the first weeks of life, he grew up in an atmosphere of love, learning, and a concern for others, largely fostered by his Quaker mother.

His school attendance began in the one-room Marbut family school, and continued at a school in McDowell, Missouri, where he also taught when he was 17. Using his teacher's earnings, he enrolled in the Collegiate Institute of Professor Noah Lee Maiden at Cassville, Missouri (**Figure 1**), an experience that led him ultimately to enroll in the University of Missouri at Columbia (UMC) in 1885. He went on to receive a BS degree from UMC in 1889 and, in 1890, was appointed to the Missouri Geological Survey. In 1891, he married Florence L. Martin of Cassville, Missouri (**Figure 2**) and established residence in Jefferson City, Missouri. They had five children: Louise Marbut Moomaw (**Figure 3**), Thomas Fiske Marbut, William Martin Marbut, Helen Marbut, and Fredrick Browning Marbut. The lifelong correspondence between Louise and her father gives an insight into his life not possible through his published works alone.

The family moved to Cambridge, Massachusetts, in 1893 so that Marbut could enroll as a graduate student in geology at Harvard University. Although his major professor and adviser was the eminent pioneer geographer and geomorphologist William Morris Davis, Marbut revealed in his correspondence to Louise that Nathaniel Southgate Shaler, a pioneer geographer and agricultural geologist, was his academic and professional inspiration to study soils.

Figure 1 Professor Noah Lee Maiden and Curtis Marbut, 1880. Curtis Fletcher Marbut Papers, 1852–1963, Western Historical Manuscript Collection, Columbia.

Just prior to and during Marbut's tenure at Harvard, Shaler had become highly regarded for his publications on soils and natural resources conservation. The most notable of these were *Animal Agency in Soil-Making* (1888), *The Origin and Nature of Soils* (1891), *Nature and Man in America* (1891), *The Interpretation of Nature* (1893), and *The Economic Aspects of Soil Erosion* (1896). Shaler was a compelling lecturer who promoted the education of the whole student: mentally, physically, and spiritually. Marbut received an MA in 1894 under Davis. In 1895, he completed the residence requirements for the PhD and returned with his family to the UMC with a faculty appointment. Although he finished and published his PhD dissertation in 1896 with Davis'

Figure 2 Florence Martin Marbut, *c.* 1890. Western Historical Manuscript Collection, Columbia.

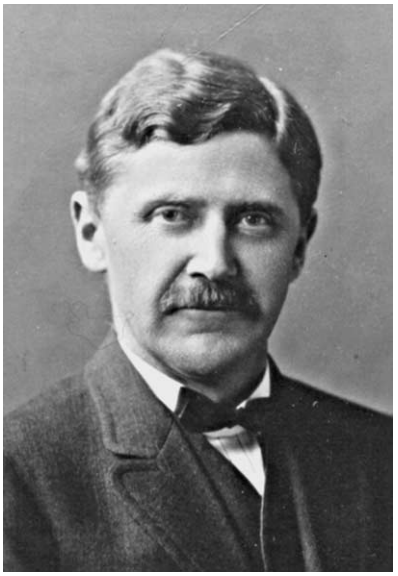


Figure 4 Merritt Finley 'M.F.' Miller, *c.* 1910. Personal collection of John P. Tandarich.

Figure 3 Louise Marbut Moomaw, *c.* 1920. Western Historical Manuscript Collection, Columbia.

approval, he never returned to Harvard for his oral defense exam and the PhD was never awarded.

Marbut served the UMC Geology Department as instructor (1895–1897), assistant professor (1897–1900), and professor (1900–1910), and was simultaneously appointed assistant geologist in the Missouri Geological Survey until 1904. His PhD dissertation *Physical Features of Missouri* published in 1896 by the Missouri Geological Survey was a pioneering study in physiography and surface hydrology. While at the UMC, Marbut published numerous studies, including: erosional remnants Crowley's Ridge, Cote Sans Dessein, and Grand Tower (1898); coal stratigraphy of areas of northern Missouri (1898); the lowlands of Southeastern Missouri (1902); the geology of Morgan County, Missouri (1908); and the soils of the Ozark region (1910).

In 1904, Marbut and Agriculture College Dean Henry J. Waters collaborated on a large-scale plaster physiographic map of Missouri that won a gold medal at the 1904 Louisiana Purchase Exposition in St Louis. In addition, Marbut and Waters coauthored the first attempt at a general soil map of Missouri that, together with two geological maps and a chapter on geology and physiography by Marbut, were published in the Exposition's volume, *The State of Missouri*. The following year, due to the influence of Waters and Cyril G. Hopkins of the University of Illinois, the University of Missouri Board of Curators organized the Missouri state soil survey and named Marbut as director. For the next 5 years, Marbut oversaw and contributed to the preparation and publication of individual county soil surveys. Thus began a lifelong interdisciplinary collaboration between Marbut and UMC faculty members Merritt Finley (known as M.F.) Miller of the Agronomy Department (**Figure 4**) and Homer L. Schantz of the Botany Department (**Figure 5**).

When, in 1909, USDA Bureau of Soils Chief Milton Whitney came to the University as keynote speaker at the annual Farmer's Week activities, he lured Marbut into the federal service as Special Agent in the national soil survey. Marbut's first task was to make a reconnaissance soil survey of the Ozark region of

Figure 5 Homer L. Schantz and Curtis Marbut taking soil samples in South Dakota, 1925. Western Historical Manuscript Collection, Columbia.

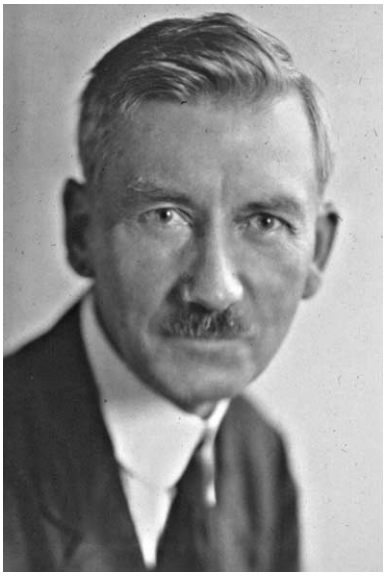


Figure 6 Curtis Marbut, *c.* 1920. Western Historical Manuscript Collection, Columbia.

Arkansas and Missouri. However, Marbut's wife Florence died from pneumonia that same year, leaving him in care of the family. When he was promoted to Scientist in the Soil Survey in 1910, he took a 2-year leave of absence from UMC for him and the family to take up residence in Washington, DC.

Scientist in Soil Survey

Marbut began his federal professional career by publishing the results of his Ozark soil survey in 1911, and leading the preparation of a monumental tome on US soils published in 1913. Marbut was dissatisfied with the latter work because the soil concepts and descriptions used in the soil survey program were unsystematically applied throughout the country. Nevertheless, he was appointed Scientist-in-Charge of the Soil Survey in 1913, but he and the family returned to the UMC. Suddenly, he decided to return to Washington and left the family at home in Columbia with a housekeeper and Louise to begin the most significant part of his professional life.

Upon his return to Washington, Marbut revisited the soil survey concept–description problem. He wrote to Louise about the situation that, "For four years I continued to grope in darkness without a ray of light." Then in 1914, Marbut discovered one of the two copies of *Die Typen der Bodenbildung* by Konstantin D. Glinka, a student of pioneer Russian soil scientist Vasilii V. Dokuchaev. The books were obtained by the USDA Departmental Library just prior to the outbreak of World War I. After his first reading of Glinka, Marbut wrote, "it comes to me suddenly – the soil profile." He translated the work in 3 years, and then unsuccessfully tried to publish it 1920 in during the height of the 'Red Scare' of anti-Bolshevik sentiment. Another obstacle was Whitney's disdain for foreign ideas and work. As with the departure of his predecessor Charles Nelson Coffey, Marbut feared a similar fate awaited him from Whitney who "had a reputation for getting rid of those who did not agree with him." When confronted by Whitney, Marbut told him that he would "follow his policies, but would think for myself." Marbut encountered anti-Bolshevik sentiment in the Bureau of Soils, writing: "I was called a Russian worshipper by my own men."

Marbut evaded this ethnocentric attitude, as he said, by having "to work in secret." By 1920, he had prepared a limited number of carbon copies of the Glinka translation for circulation to a few selected colleagues (Figure 6). In 1921, at the American Soil Survey Workers Association (ASSA) meeting in East Lansing, Michigan, he introduced the Russian concepts of classification and soil profile without identifying their origin. The ASSA, founded in 1920, was an organization started by scientists in soil survey for the promotion and publication of their ideas. Roy W. Simonson, however, has stated that the ASSA was formed "to counteract the growing influence of C.F. Marbut." Nevertheless, the ASSA was influential in disseminating ideas and it joined with the Soils Section of the American Society of Agronomy to create the Soil Science Society of America (SSSA) in 1936.

He worked with ASSA colleagues and soil survey parties to apply these 'new ideas' in the USA. Marbut's own notes in his personal papers clearly indicate his dedication to convince US soil scientists of the

efficacy of using the Russian profile concept as a basis for soil descriptions. To this end, he personally visited as many soil survey field parties as possible and participated in their yearly field reviews to exercise his influence and powers of persuasion – a practice that he continued his entire career.

During this time, Marbut's reputation grew in disciplines outside soil science, such as Pleistocene geology. An example of this is his interaction with University of Illinois faculty member and Illinois state geologist Morris M. Leighton. In the 1920s, Leighton was working to develop the weathering profile concept, originally conceived by Iowa state geologist George F. Kay. He stated that both Glinka's book and Marbut influenced him during the latter's visits to the University of Illinois between January and July 1923. Leighton credited Marbut for further insights obtained on the soil profile and its homologous application in geology as the weathering profile.

By the 1924 ASSA meeting, Jethro Otto Veatch of Michigan State University knew of the Russian origin of the ABC soil profile when he publicly referred to it as 'the Glinka scheme.' The foreign source of this soil concept was known among the soil scientists and seemed to be less of a problem.

Marbut's 1927 presentations at the First International Congress of Soil Science (see below; Figure 7) and his 1928 lectures in the Graduate School of the USDA represented further opportunities to promote the Russian ideas. The lectures, some of which were published in 1951, contained numerous references to the Russian work and influenced many soil scientists, particularly his own staff members, such as Mark Baldwin, Charles Kellogg, and James Thorp (Figure 8). The results of his lifelong work on the description and classification of the soils of the USA were published in the 1935 *Atlas of Agriculture* that included a monumental soil map.

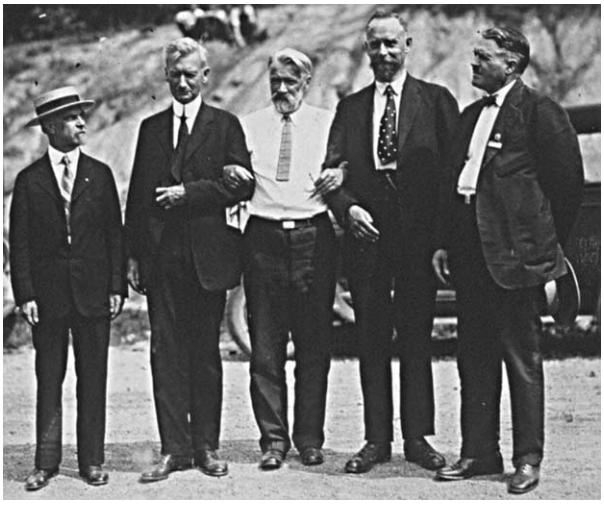


Figure 7 Oswald Schreiner, Curtis Marbut, Konstantin D. Glinka, D.J. Hissink, and Arthur G. McCall, 1927. Western Historical Manuscript Collection, Columbia.

Internationalist–Explorer

From 1920 until 1935, Marbut assumed the role of an internationalist in the field of soils. His extensive travels and explorations during this period carried him to many parts of the world and gained for him the well-earned reputation as an international authority on soil geography – as much at home in the study of the soils of Europe, Africa, and South America as those of his homeland (Figures 9 and 10).

Marbut acquired his first-hand knowledge of soils and landscapes of the world through his exhausting schedule of trips and expeditions. His first trip to Europe was on a sabbatical from UMC in 1899–1900 with colleague Isidor Loeb, during which they studied the Alps, the Hungarian plain, the Mediterranean region, the British Isles, and Scandinavia. In

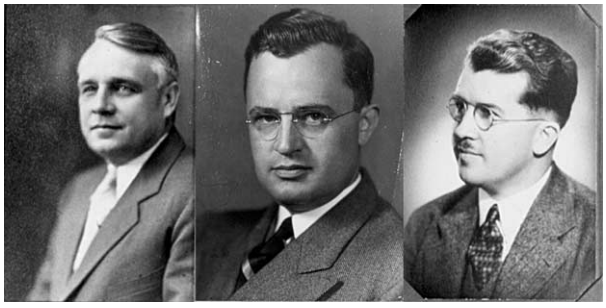


Figure 8 Mark Baldwin, Charles E. Kellogg, and James Thorp, 1937. Personal collection of John P. Tandarich. Baldwin and Thorp Portraits originally courtesy of Mrs Mark (Mary) Baldwin and Cyril Harvey family.

Figure 9 During the Amazon expedition, 1923. Western Historical Manuscript Collection, Columbia.

1918, Marbut and Schantz, who had gone from the UMC to the USDA Bureau of Plant Industry, studied the vegetation and soils of Africa for the American Commission to Negotiate Peace. In April 1919, Marbut and Bureau of Soils colleague Hugh Hammond Bennett went to Latin America to soil survey a disputed boundary between Guatemala and Honduras, sponsored by the American Geographical Society (AGS). He traveled to Prague, Czechoslovakia, in 1922 for the International Agrogeology Conference and subsequent field trip. From July 1923 to January 1924, he mapped the interior Amazon basin as a representative of the Brazilian-American Rubber Commission. Immediately afterward, until July, he studied the soils of Argentina under sponsorship of the AGS. Later that year, he traveled to Rome for the meeting of agro-geologists at which he assisted in the formation of the International Society of Soil Science (ISSS). At this meeting, Jacob G. Lipman of Rutgers University and he began preparations for holding the First International Congress of Soil Science in Washington, DC in 1927.

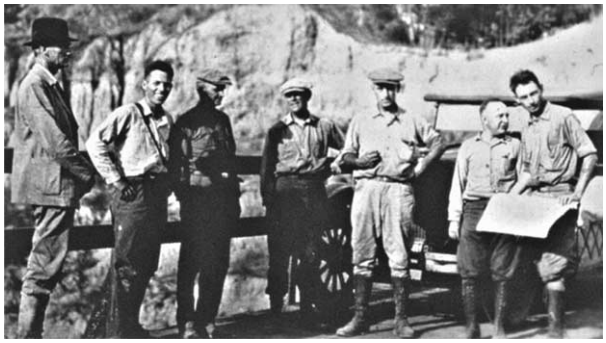


Figure 10 Curtis Marbut (far left) visiting soil survey party in North Dakota in early 1920s. Western Historical Manuscript Collection, Columbia.

Though Marbut made two presentations on soil classification during the Commission V session of the 10-day First International Congress that lasted from 13–22 June, 1927, his legacy of the meeting was the organization of the post-meeting Transcontinental Excursion (Figures 11 and 12). The Excursion was a 12 000-mile railroad train trip of 30 days (commencing on June 22) through 23 states in the

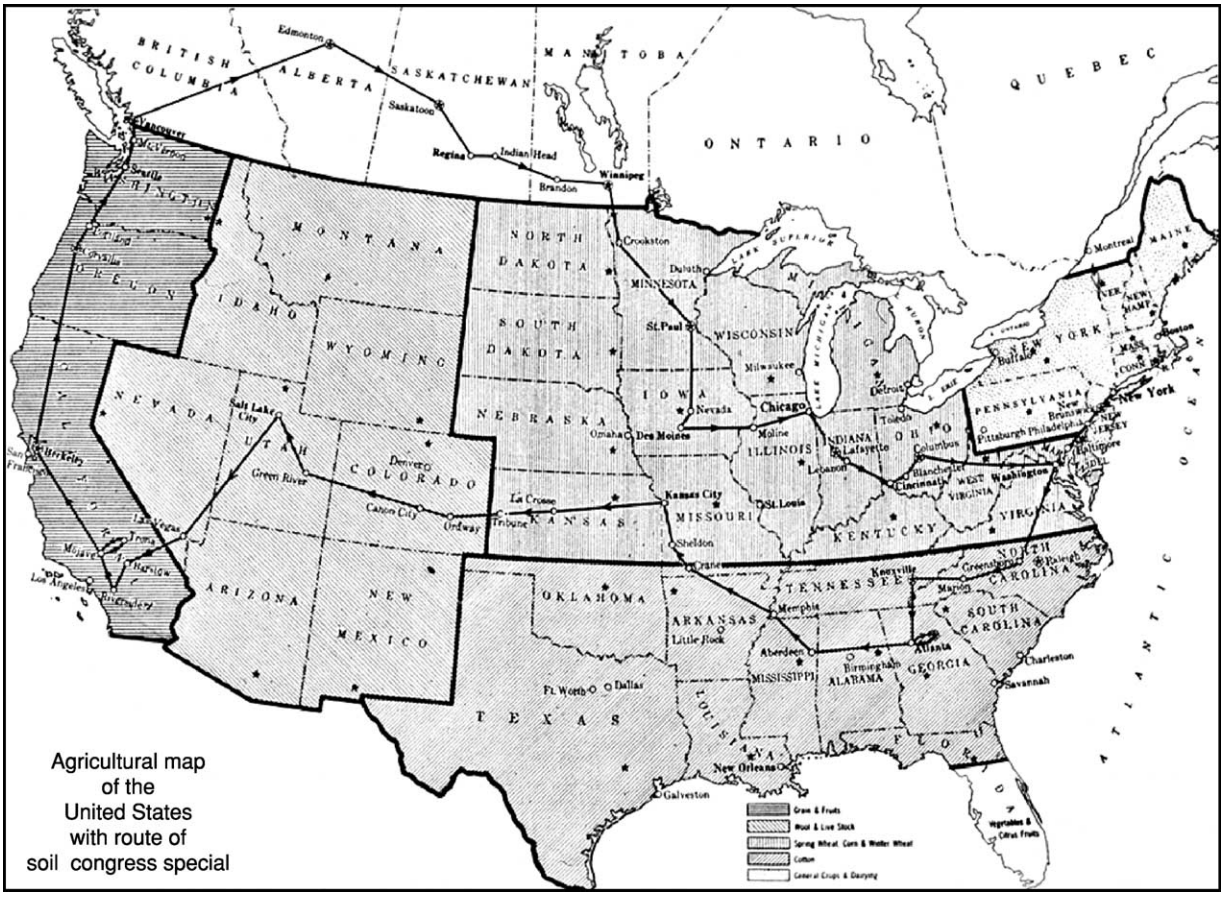


Figure 11 Route of the Transcontinental Excursion train 'The Soil Congress Special' through the USA and Canada, 1927. Archives of John P. Tandarich.

Figure 12 Typical Transcontinental Excursion stop to examine a soil pit, 1927. Marbut is standing third from right. Western Historical Manuscript Collection, Columbia.

Figure 13 Curtis Marbut (second from left) with soil scientists in Kansas, 1932. Western Historical Manuscript Collection, Columbia.

USA and four Canadian provinces to examine soils and landscapes at 34 stops. The Excursion advanced the sharing of ideas among soil scientists from around the world through the constant opportunities for interaction on the train, at the individual stops and through the daily Excursion newspaper, the *Boden Bull*.

In early 1928, Marbut visited Bennett's soil survey party in Cuba. In May 1929, he went to Danzig (modern Gdansk) for ISSS Commission V meetings and then to Soviet Russia in mid-1930 for the Second International Congress of Soil Science. The Excursion following the Second Congress meetings gave Marbut a view of the soils of western Russia from Moscow south to the region between the Black and Caspian Seas. He also visited Germany, particularly the Black Forest region, immediately afterward. James Thorp reported that Marbut visited his soil survey party in Puerto Rico early in 1932. This was immediately after attending the Fourth International Congress of Sugar Cane Technologists (Figures 13 and 14).

In 1933, Marbut accepted an assignment to advise the National Geological Survey of China in their organizing of a soil survey. To this end, Marbut sent James Thorp, his colleague from the Washington, DC soil survey office, to China in 1933 to train students in soil survey methods. By early 1935, enough progress in the soil survey of China had been made that it was necessary for Marbut to travel to China to render opinions on the descriptions and classification of the soils. In consultation with Thorp, Marbut decided to travel to China following his attendance at the Third International Congress of Soil Science in Oxford, UK (Figure 15).

In August 1935, Marbut traveled to England as a delegate to the Third Congress. Immediately after the

Figure 14 Participants in Second International Congress of Soil Science at Leningrad, 1930. Marbut is second from left in front. Western Historical Manuscript Collection, Columbia.

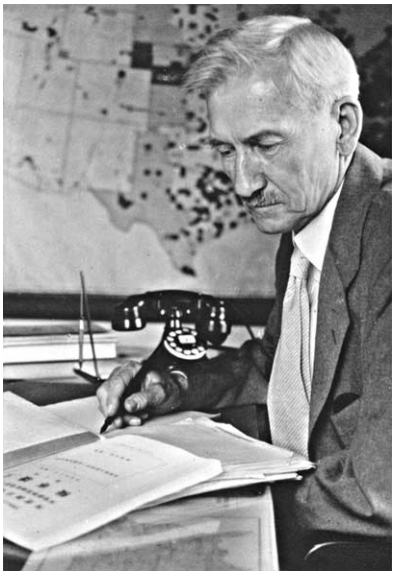


Figure 15 Curtis Marbut preparing for trip to China, 1935. Western Historical Manuscript Collection, Columbia.

meetings, he traveled through Europe and then, by the Trans-Siberian Railroad, through southern Russia. During the Siberian leg of the trip, Marbut was stricken with double pneumonia. He was taken from the train in Harbin, Manchuria (now Haerpin in Heilungkiang province) to the German-Russian hospital there. When summoned to Marbut's bedside, Thorp recalled Marbut telling him, as he neared death at age 72, "You've got to get me out of here; I have ten more years of work to do." Unfortunately, he died early the next morning. According to Thorp, the American Consulate held a beautiful memorial service at which a Russian refugee chorus sang. His ashes were sent back to southwestern Missouri.

In Service to Humanity

Marbut made his own independent place in the emerging discipline of soil science, but his concerns became more than the scientific facts of soils and their classification. His subjects at professional meetings and in his publications became more international in scope. His presidential address to the Association of American Geographers in December 1924 dealt with the Malthusian theory. He also examined Soviet soils and the capacity for that country's wheat production potentially to rival that of the USA.

The AGS, in awarding him the Cullum Medal in 1932, its highest distinction, described his achievements in the following words: "For geographical works on the study of soils – the foothold of all things." He also received awards from Berlin and Czechoslovakia, and honorary doctorates from UMC in 1916, and Rutgers in 1930. About this and other honors he received, Marbut was grateful but, as he indicated in his correspondence to Louise, they were but a means to an end. He wrote Louise that all his recognition and achievements were important only as they could be used to strengthen his position and influence to bring to the countries of the world knowledge of their soil resources and, thereby, self-sufficiency in agricultural production so that food would not be a reason to wage war. He wrote that he believed it was his duty and responsibility as a Unitarian scientist to do all one man could do to ensure that the countries of the world could feed themselves without having to rely on other countries for food. Marbut shared his unspoken goal of knowledge in the service of humanity, a point-of-view advocated by Unitarians such as Shaler and Davis, in his correspondence with Louise. However, she never mentioned his philosophy in the biographical sketch she wrote for the SSSA memorial volume; it remains only within the letters that she transcribed and incorporated into the manuscript biography of her father.

Epilog

Curtis Fletcher Marbut is mentioned along with Russian Vasilli V. Dokuchaev and American Eugene W. Hilgard when speaking of the 'founders' of modern soil science. He was a pivotal person – well-known as a teacher and promulgator of ideas. Marbut had to battle provincialism to achieve acceptance of the Russian ideas in the USA that he believed necessary to the advancement of the science. He was the world authority on soils in the 1920s and 1930s. Marbut became a scientist in demand worldwide for his knowledge, a circumstance that prompted US President Franklin D. Roosevelt to grant Marbut three yearly extensions beyond the mandatory retirement age of 70. As Thorp said in his eulogy: "[W]ho of us can fail to aspire to a long life such as his, with its earlier hardships, struggles and disappointment and its ultimate triumph; and who can but wish to follow our chosen work until we are old ... His life was full and rich and he passed on his flaming torch to younger hands while he was still running a strong race." Marbut was installed in the US National Agricultural Hall of Fame in 1989.

Further Reading

Chorley RJ, Beckinsdale RF, and Dunn AJ (1973) *The History of the Study of Landford and the Development of Geomorphology. Volume of the Life and Work of William Morris Davis*. London: Methuen.

Helms D, Effland ABW, and Duran PJ (eds) (2002) *Profiles in the History of the US Soil Survey*, pp. 43–46. Ames, Iowa: Iowa State Press.

Krusekopf HH (ed.) (1942) *Life and Work of C.F. Marbut/Soil Scientist/A Memorial Volume Published by the Soil Science Society of America*. Columbia, Missouri: The Artcraft Press.

Livingstone DN (1987) *Nathaniel Southgate Shaler and the Culture of American Science*. Tuscaloosa: University of Alabama Press.

Marbut CF (1896) Physical features of Missouri. *Missouri Geological Survey (1st series)* 8: 227–316.

Marbut CF (1910) Soils of the Ozark region. *Missouri Agricultural Experiment Station Bulletin*, no. 3.

Marbut CF, translator (1927) *The Great Groups of Soil and their Development by K.D. Glinka*. Ann Arbor, Michigan: Edwards Brothers.

Marbut CF (1928) *Soils, Their Genesis and Classification and Development/A Course of [30] Lectures given in the Graduate School of the United States Department of Agriculture/February to May, 1928*. Mimeograph.

Washington, DC: [USDA], The first 10 lectures of this series were published in 1951 (see below).
Marbut CF (1935) *Soils of the United States*. *USDA Atlas of American Agriculture*, Part 3. Washington, DC: US Government Printing Office.
Marbut CF (1951) *Soils: Their Genesis and Classification/A Memorial Volume of Lectures Given in the Graduate School of the United States Department of Agriculture in 1928*. Madison, Wisconsin: Soil Science Society of America.
Marbut CF and Schantz HL (1923) *The Vegetation and Soils of Africa*. Research Series No. 13. Washington, DC: American Geographical Society and National Research Council.
Morrow L (1981) An Ozarks landmark: the Curtis Fletcher Marbut house. *White River Valley Historical Quarterly* 7(9): 6–9.
Morrow L (1999) Curtis Fletcher Marbut. In: Christensen LO, Foley WO, Kremer GR, and Winn KH (eds) *Dictionary of Missouri Biography*, pp. 518–519. Columbia: University of Missouri Press.
Tandarich JP (1998) Pedology – disciplinary history. In: Good G (ed.) *Sciences of the Earth: An Encyclopedia of Events, People, and Phenomena*. vol. 2, H–Z, pp. 666–670. New York: Garland Publishing.
Tandarich JP (2002) The development of soil science in the western world to 1927. In: Lal R (ed.) *Encyclopedia of Soil Science*, pp. 659–662. New York: Marcel Dekker.
Tandarich JP and Johannsen CJ (1985) Introduction to the life and work of Dr Curtis Fletcher Marbut. *Soil Survey Horizons* 26(1).
Tandarich JP and Sprecher SW (1994) The intellectual background for the factors of soil formation. In: Amundson R, Hardin J, and Singer M (eds) *Factors of Soil Formation: A Retrospective*, pp. 1–13. Special Publication 33. Madison, Wisconsin: Soil Science Society of America.
Tandarich JP, Johannsen CJ, and Wildman WE (1985) James Thorp talks about soil survey, C.F. Marbut, and China. *Soil Survey Horizons* 26(2): 5–12.
Tandarich JP, Darmody RG, Follmer LR, and Johnson DL (2002) The historical development of soil and weathering profile concepts from Europe to the United States of America. *Soil Science Society of America Journal* 66(2): 335–347.
Thorp J (1935) *Curtis Fletcher Marbut passes on*. Peiping, China: privately printed.

Matric Potential *See* **Hydrodynamics in Soils**; **Water Potential**; **Water Retention and Characteristic Curve**

MEDITERRANEAN SOILS

J Torrent, Universidad de Córdoba, Córdoba, Spain

© 2005, Elsevier Ltd. All Rights Reserved.

The term 'Mediterranean' is applied to a specific climate and kind of vegetation that is typical of areas surrounding the Mediterranean Sea but also found in five other world regions, namely California, central Chile, Southwest and South Australia, and the Cape region of South Africa. Although there is no consensus on the precise delineation of the Mediterranean region, the total area with a true Mediterranean climate is probably some $2\times10^6\,km^2$, the Mediterranean region *sensu stricto* occupying approximately $1.2\times10^6\,km^2$. Because soil-forming factors vary widely among Mediterranean areas, the resulting soils share neither morphology nor genesis. Despite their high diversity, the body of properties of soils in the Mediterranean areas makes them significantly different from soils in subtropical and other temperate climatic regions. This justifies separate consideration of the genesis and properties of 'Mediterranean soils,' a term that is used here with no intended specific genetic meaning and encompasses all soils in areas with a Mediterranean climate.

Soil-Forming Factors

Climate

In Köppen's classification scheme, the Mediterranean climate is designated as a warm, temperate, rainy climate with dry summers (*Cs*), the wettest winter month receiving at least three times as much rain as the driest summer month. Areas with a Mediterranean climate lie mostly around the Mediterranean sea and at latitudes of 30–45° on the west coast of some continents and thus form transitional zones between

the influence of the westerlies of higher latitudes and subtropical high-pressure cells. The winter weather is dictated by the fronts of the westerlies, whereas the dry summer reflects the dominance of the midtropospheric anticyclone.

The Mediterranean climate encompasses an interesting variety of subclimates, as illustrated by the temperature and rainfall data from selected stations listed in Table 1. The mean annual precipitation ranges from 300 to more than 1500 mm, with one or two maxima in the rainy season. Summer drought lasts from less than 2 to approximately 5 months. The mean annual temperature ranges from about 11°C (e.g., in regions of Anatolia and central Spain) to approximately 19°C (e.g., in coastal areas of Cyprus), and absolute maxima above 45°C and minima below −20°C have been recorded at some stations. The difference between the mean temperatures of the warmest and coldest month is less than 8°C in some coastal areas affected by cold ocean currents (e.g., San Francisco Bay, California) but exceeds 20°C in inland areas with little oceanic influence. The difference at stations by the Mediterranean Sea is typically 13–16°C. From the standpoint of pedoclimate, the contrast in moisture content between winter and summer was used as the basis for defining the *xeric* moisture regime, an important diagnostic criterion in the soil taxonomy classification system. Under this regime, the soil zone affecting plant growth is dry for more than 45 days in summer and moist for more than 45 days in winter.

The influence of the Mediterranean climate on soil development can be gleaned by considering two simple parameters, namely: the degree of leaching (L) and the actual evapotranspiration (AET). The annual value of L is calculated as the sum of the excess precipitation (P) over potential evapotranspiration (PET) for those periods in which $P > PET$ less the soil water-holding capacity (WHC). The annual value of AET is equivalent to $P - L$. Table 2 shows the water balance for a station with a warm continental Mediterranean climate (Córdoba, Spain) and includes the mean monthly and annual values of P, PET, AET, and L. Winter leaching (37 mm) is low, consistent with other Mediterranean stations with mean annual P below 600–700 mm, if mean monthly P data are used in the calculations. As can be seen from Table 2, the summer drought leaves the soil with little water to meet plant needs and enable mineral weathering and biological activity. The intensity of these processes peaks in autumn and late spring when soil is moist and temperatures are mild. In summary, the relative lack of coincidence between high soil temperatures and high moisture contents precludes strong weathering and continuous biological activity.

Variability in annual rainfall is high in the Mediterranean region. For instance, the annual rainfall in Córdoba for the 1901–2000 period ranged from less than 250 mm to more than 1250 mm, i.e., by more than five times. The annual leaching for the same period ranged from 0 to more than 650 mm (at WHC 125 mm), with a moderately bimodal distribution (Figure 1). Because of the high variability in rainfall, L takes relatively large values in some years. Mean climatic data thus lead to gross underestimation of the leaching potential of the Mediterranean

Table 1 Climatic data from selected stations in the Mediterranean region

Month *Station*	*J*	*F*	*M*	*A*	*M*	*J*	*J*	*A*	*S*	*O*	*N*	*D*	*Year*
Athens													
Temperature (°C)	9	10	11	15	20	24	27	27	23	19	15	11	18
Precipitation (mm)	62	36	38	23	23	14	6	7	15	51	56	71	402
Rome													
Temperature (°C)	8	9	11	14	18	22	24	24	22	17	13	10	16
Precipitation (mm)	83	73	52	50	48	18	9	18	70	110	113	105	749
Cuenca (Spain)													
Temperature (°C)	3	4	7	10	13	18	22	21	18	12	7	4	12
Precipitation (mm)	43	41	70	48	72	50	19	26	43	55	49	55	571
San Francisco													
Temperature (°C)	9	11	12	13	15	16	17	17	18	16	13	10	14
Precipitation (mm)	102	88	68	33	12	3	0	1	5	19	40	104	475
Perth (Australia)													
Temperature (°C)	23	24	22	19	16	14	13	14	15	16	19	22	18
Precipitation (mm)	7	12	22	52	125	192	183	135	69	54	23	15	889

Mean monthly and annual temperature and precipitation values are rounded to the next unit.

Table 2 Soil water balance (mm) for Córdoba (southern Spain; 90 m, 38°51′ N, 4°50′ W), on the basis of the agricultural year, September–August (mean of 100 years: 1901–2000 period)

Month *Variable*	*S*	*O*	*N*	*D*	*J*	*F*	*M*	*A*	*M*	*J*	*J*	*A*	*Year*
P	27	74	91	87	73	73	79	61	45	19	3	3	635
PET	134	78	46	33	38	48	75	108	140	178	218	194	1288
P − *PET*	−106	−4	45	54	35	25	4	−47	−95	−159	−215	−191	−653
Soil storage[a]			45	99	125	125	125	78					
L					8	25	4						
AET	27	75	46	33	38	48	75	108	123	19	3	3	598

[a]For a soil with a water-holding capacity of 125 mm.

P, mean precipitation; *PET*, potential evapotranspiration according to Penman; *L*, leaching; *AET*, actual evapotranspiration.

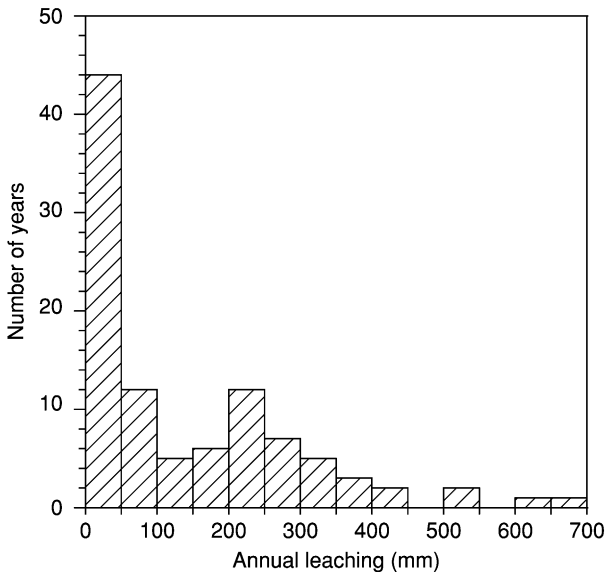


Figure 1 Frequency distribution of annual leaching in Córdoba, southern Spain, for the 1901–2000 period, as calculated from monthly rainfall data for the period, the mean monthly potential evapotranspiration (**Table 2**), and a soil water-holding capacity of 125 mm.

climate, which is generally significant if the mean annual *P* exceeds 500 mm.

Vegetation

Most Mediterranean areas were once covered with a sclerophyllous forest, which adapts itself readily to both summer drought and light winter frost. The exact nature of this forest in the areas bordering the Mediterranean Sea is difficult to ascertain, but the native vegetation probably included coniferous and broad-leaved evergreen trees. Among the latter, holm oak (*Quercus ilex* and *Q. rotundifolia*), cork oak (*Q. suber*), and Kermes oak (*Q. coccifera*) are widespread. Typically sclerophyllous species include wild olive (*Olea europaea*), carob (*Ceratonia siliqua*), and lentisk (*Pistacia lentiscus*). Isolated specimens of stone pine (*Pinus pinea*) and masses of Aleppo pine (*P. halepensis*) are also commonplace. Degradation of the original forest has often resulted in the establishment of scrub communities that are given local names (*garriga* in Catalonia, *maquis* in France, *macchia* in Italy, etc.) and include cistus (*Cistus* spp.), heathers (*Erica* spp.), gorse (*Ulex* spp.), broom (*Genista* spp.), and thyme (*Thymus* spp.). Some scrub communities probably constitute the original vegetation in areas where either low precipitation or pervious rocks and wind exposure gave rise to habitats that were too dry for the sclerophyllous forest to survive.

In North America, the sclerophyllous forests of western California are rich in species of *Quercus* and *Cupressus* similar to those of the Mediterranean basin (e.g., *Q. agrifolia* resembles *Q. ilex*). In areas with mean annual rainfall below 500 mm, a formation similar to the maquis, namely the *chaparral*, reaches 3 m in height and encompasses bush oaks in addition to species of *Ceanothus* and *Arctostaphylos*. In South America, the Chilean sclerophyllous forest exhibits a completely different floristic composition but is remarkably similar in appearance to analogous formations; *Quillaja saponaria*, *Rhus caustica*, and *Peumus boldus* are especially commonplace. Uncultivated land in the Cape region is now occupied by sclerophyllous scrub called *fynbos*, which varies between 1 and 4 m in height and exhibits a bluish-green hue; the only tree is *Leucodendron argenteum*, which is accompanied by scrub species pertaining to the Proteaceae, Brassicaceae, and Rosaceae families. Finally, in the Southwest and South Australia Mediterranean areas, the vegetation typically consists of *Eucalyptus* species with coriaceous leaves. The *jarrah* forest with *E. marginata* is typical of areas with a mean annual rainfall of 600–1200 mm; on the other hand, the drier (500–600 mm) *wandoo* zone has a less dense forest of *E. redunca*.

Sclerophyllous vegetation seems to be efficient in cycling of bases; however, relatively little information is available on the biogeochemical cycles for the

different areas. Transformation of fresh organic matter usually results in a mull humus that grades into a moder humus in the subhumid borders of the Mediterranean region, particularly when soils develop on coarse-grained, acidic parent materials.

Lithology and Relief

A number of Mediterranean areas are mountainous and exhibit a complex lithology that derives from the strong tectonic activity typical of zones involved in crustal plate collision (Mediterranean Sea, California, central Chile). Sedimentary rocks predominate, but plutonic, volcanic, and metamorphic rocks are well represented in most areas. In the Mediterranean region proper, limestones and dolomites constitute a fairly high proportion of the mountains. Frequent erosion and deposition episodes since the Upper Tertiary have led to the formation of large surfaces covered by fine-grained sediments at the foot of the mountains.

The abundance of steep slopes has contributed to make water erosion the principal process in slope development, which was particularly active in Quaternary periods with scarce or uneven rainfall and soils without a continuous vegetation cover. In periods of more-even rainfall, slope development under a continuous vegetation cover was probably more markedly influenced by soil creep and subsurface water erosion, similarly to slopes in humid temperate regions. The strong influence of water erosion is apparent in the slope profiles of many lithologically uniform areas, which, as predicted by geomorphic models, consist of a long, concave element at the foot, a straight middle segment, and a short, convex summit.

Structural reliefs with parallel-retreating slopes are commonplace in areas of sedimentary rocks (particularly in the drier regions). These slopes commonly merge into pediments. Depositional pediment surfaces, which are more common in Mediterranean areas than in humid temperate regions, generally have slopes of less than 1–2% and occasionally extend over tens of square kilometers. Some such surfaces are very old and covered by strongly weathered soils. Not rarely, opposite pediment surfaces form V-shaped, wide valleys dissected to a greater or less extent by a central stream and its tributaries.

Alluvial plains and their contiguous terraces dominate landscapes over stretches of tens of kilometers in the middle and lower parts of major river valleys, where they support irrigated agriculture. River terrace systems provide excellent soil chronosequences for the study of the major soil-forming processes as the systems formed over long periods during the Quaternary. For instance, the oldest river terraces reach heights of more than 200 m above the present river channel in the river Guadalquivir valley (southern Spain).

Time

Because the Mediterranean region underwent no glaciation, its landscape exhibits geomorphic surfaces of widely different age. Some surfaces (e.g., the *raña* surfaces in several parts of Spain) date back to the Pliocene. Lower Pleistocene surfaces are not uncommon, and are well represented by different pediments and river terraces. An intricate mosaic of soil formation times is therefore the rule rather than the exception in most areas. Unfortunately, absolute dating is impossible in most cases, so many pedologists rely on geomorphic data to establish relative soil age and its relationship with soil development.

Soil-Forming Processes

Mass Additions and Losses

The strong influence of water erosion results in marked catenary differentiation in mountainous and hilly areas. Thin, occasionally skeletal soils in the upper part of the slope give way to increasingly thick soils as the footslope is approached. Increasing thickness is usually associated with increasing textural fineness and with changes in the mineralogical assemblage of the clay fraction.

Some Mediterranean areas lie at the fringe of deserts (e.g., the Sahara and Negev deserts). This has resulted in more or less frequent episodes of eolian dust addition. Direct measurement of the dust carried by 'mud rains' and detailed mass balances in soils close to the Mediterranean have yielded or supported figures of a few grams per square meter and per year. Dust accretions, gradually assimilated by the soil during the Quaternary, may account for up to 50% of the mass of some soils. Unfortunately, balance calculations are difficult because the receiving soils and the dust contain similar minerals.

Leaching and Redistribution of Calcite and Silica

Weak-to-moderate winter leaching enables weathering of silicates and significant translocation of basic cations. Thus, mature soils that developed on base-poor parent materials exhibit acid pH values in the surface horizons, with pH and base saturation generally increasing with depth. If the parent material contains calcite, dolomite, or Ca-rich weatherable minerals and winter leaching is scarce to moderate, Ca precipitates as secondary calcium carbonate (in the form of calcite) in some soil horizons. Calcic horizons with carbonate nodules of different forms and sizes occur in soils in areas with $P < 700$–800 mm

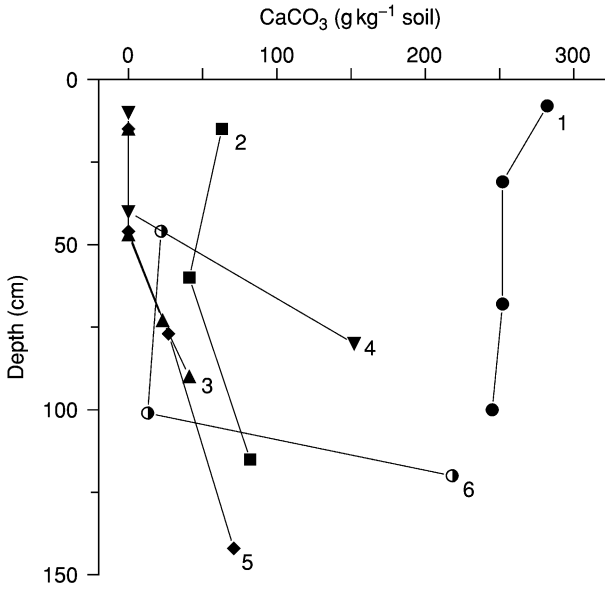


Figure 2 Depth distribution of $CaCO_3$ in soils of a river terrace sequence (Henares River, central Spain). Numbers denote relative soil age (1, recent alluvium; 6, highest terrace). Data from Daz MC (1986) *Estudio de los Procesos de Rubefacción de los Suelos de las Terrazas de la Cuenca del Ro Henares y del Ro Jarama*. PhD thesis, Universidad Politécnica de Madrid, Spain.

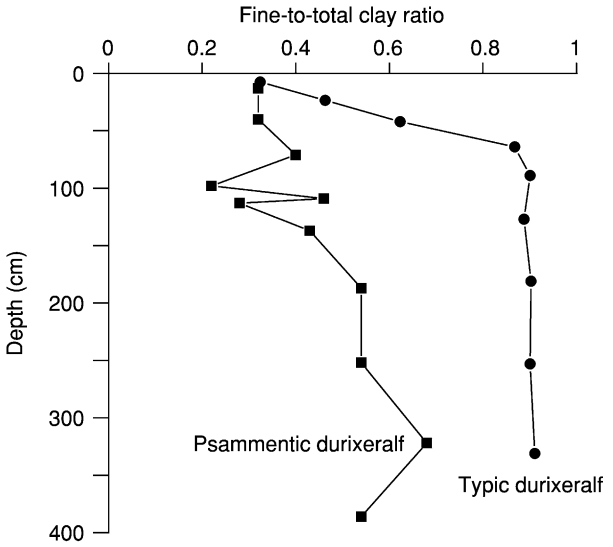


Figure 3 The increase with depth in the fine-to-total clay ratio in two deep soils of southern California is taken as evidence for clay translocation. Data from: Torrent J, Nettleton WD, and Borst G (1980a) Clay illuviation and lamellae formation in a Psammentic Haploxeralf of Southern California. *Soil Science Society of America Journal* 44: 363–369; and Torrent J, Nettleton WD, and Borst G (1980b) Genesis of a Typic Durixeralf of southern California. *Soil Science Society of America Journal* 44: 575–582.

and are replaced by indurated calcrete (petrocalcic) horizons where $P < 400$–500 mm. Long stretches of the dry Mediterranean areas (e.g., La Mancha region in central Spain) are covered by soils with calcrete. The depth at which carbonate accumulates is a function of not only the climate and parent material, but also relief and, particularly, time (Figure 2).

The weathering of silicates releases Si into solution; dissolved silica is translocated and can precipitate as silica (opal) in the form of nodules or crusts (duripans). This process is generally restricted to the drier areas, provided the parent material contains highly weatherable silicates or volcanic glass. Therefore, silica and mixed silica–carbonate are less abundant than carbonate accumulations in Mediterranean soils.

Clay Translocation

Clay translocation (illuviation) has been, and continues to be, a major process in the genesis of Mediterranean soils, particularly in the more humid areas and periods. Because of the strong flocculating effect of Ca ions, decalcification is required for clay particles to disperse and be carried by percolating water; however, translocation of clay has occasionally been observed in calcareous soils. Summer drought seems to favor clay translocation, probably because, when the soil dries, cracks develop and clay on the freshly exposed surfaces is easily dispersed and transported by the water percolating after the first autumn rains. Clay translocation is generally a slow process, but soils with horizons rich in illuviated clay (argillic horizons) have formed within only 2000–3000 years on the alluvium of some rivers of Spain. Summer drought generally contributes to disruption of the clay cutans resulting from clay translocation. Thus, evidence for clay translocation is provided by the presence of undisturbed cutans in deep soil horizons, or by the increase in the fine-to-total clay ratio with increasing depth (Figure 3), since the fine clay particles are assumed to be more mobile than the coarse ones.

Rubefaction

Rubefaction (i.e., reddening due to pedogenic hematite formation) is quite common in the Mediterranean region. It follows decalcification because the high pH of calcareous soils hinders weathering and Fe release from primary minerals. Appropriate environmental conditions for the formation of hematite from ferrihydrite (the initial product of the weathering of Fe minerals) include a high temperature, water activity of less than 1, low organic matter content, near-neutral pH, and high concentrations of Ca^{2+} and Mg^{2+}. In the absence of one or more of these conditions, only goethite (a yellow Fe oxide) is formed.

The vivid red hue of some Mediterranean soils has attracted the attention of pedologists, geologists, naturalists, and travelers. Terms such as the Italian 'Terra Rossa' or 'red Mediterranean soils' have been incorporated into some classification schemes and used (and abused) by pedologists to denote different red-colored Mediterranean soils. Indeed, pedogenic hematite possesses a high pigmenting power, so a concentration of more than than 10 g kg^{-1} soil often results in reddish brown or redder hues. Interestingly, redness ratings derived from the Munsell color notation and the soil hematite content are correlated (Figure 4).

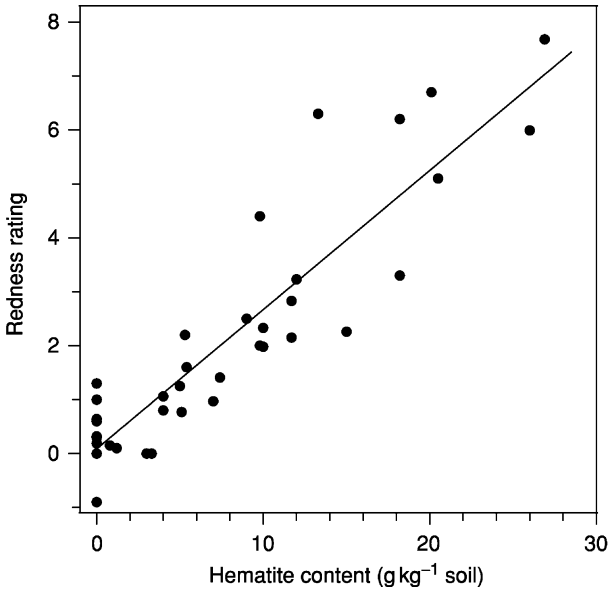


Figure 4 Relationship between a redness rating derived from the Munsell color notation and the hematite content in a group of Mediterranean soils. Data from: Peña F and Torrent J (1984) Relationships between phosphate sorption and iron oxides in Alfisols from a river terrace sequence of Mediterranean Spain. *Geoderma* 33: 283–296; Torrent J and Cabedo A (1986) Sources of iron oxides in reddish brown soil profiles from calcarenites in southern Spain. *Geoderma* 37: 57–66.

Soil chronosequences indicate that rubefaction and clay illuviation develop at similar rates once the upper soil horizons have been decalcified. These three processes play a crucial role in soil development on stable geomorphic surfaces. This is shown in Figure 5, which depicts an idealized development sequence on calcareous alluvium under an 'average' Mediterranean climate. Soils comparable with members of this sequence are seen on widely different parent materials that range from hard igneous rocks to unconsolidated sediments. Two related development sequences must be considered here. One has soils with no carbonate accumulation, so it must be assigned to base-poor rocks. In the other, the old soils exhibit a seasonally perched water table that develops as a result of pore clogging by illuviated clay. In this case, reductomorphic features appear and the red hue is lost due to reductive hematite dissolution.

Development of Vertic Features

The xeric regime favors development of the vertic features typical of Vertisols and other clayey soils that usually experience several major shrink/swell cycles each year. Vertic features seem to develop over short pedogenic times in clayey sediments as recent alluvium or Tertiary marls. Also, vertic features are enhanced by smectite formation in the soils

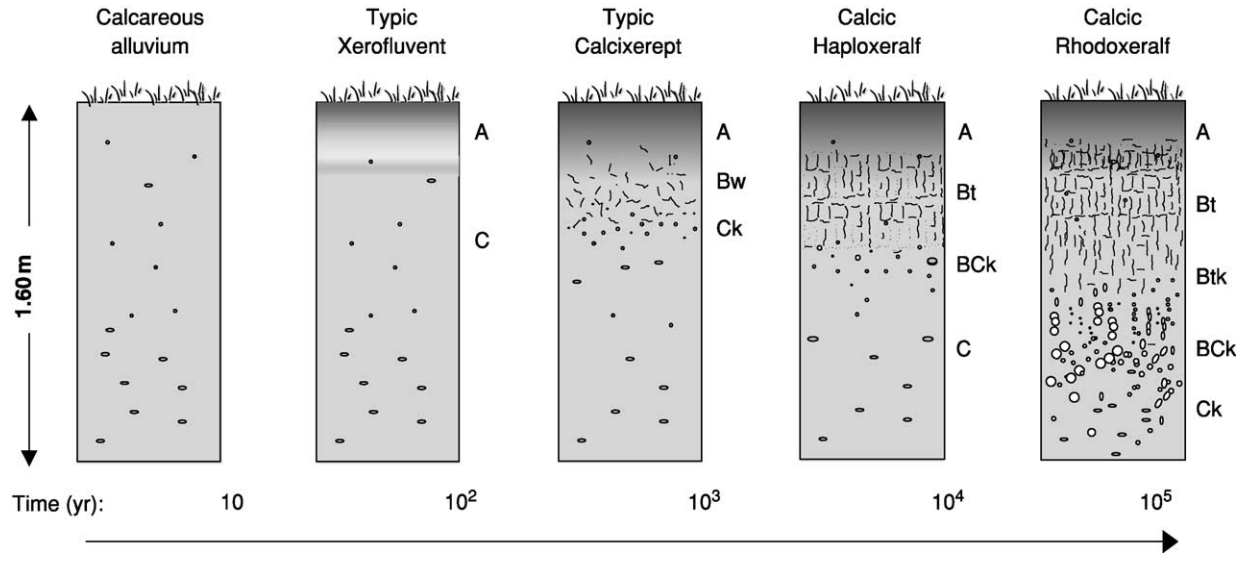


Figure 5 Idealized soil development sequence from a calcareous loamy alluvium in a typical Mediterranean environment.

of footslopes and closed basins as a result of lateral illuviation of silica and bases. For this reason, Vertisols occupy the lowest topographic positions of some Mediterranean soilscapes, even though the parent material may not be clayey.

Major Soil Types

Many pedologists would subscribe to the idea that the Mediterranean climate dictates the predominance of some specific soil-forming processes and thus justifies the term 'Mediterranean soils.' However, pedodiversity is probably higher in the Mediterranean region than in any other climatic zone. How can these two conflicting views be reconciled? Probably by remembering that: (1) the Mediterranean climate encompasses a rich variety of subclimates and (2) the effect of the Mediterranean climatic and biotic environments is modulated by the high lithological and geomorphic diversity resulting from the complex tectonic and climatic history of Mediterranean regions.

Various classification schemes have been used – some only at a national level – to encompass the Mediterranean pedodiversity. Some of the older schemes incorporated geographical and color terms (e.g., 'Black Andalusian Earths') with somewhat diffuse definitions. Most pedologists have by now adopted the UN Food and Agriculture Organization (FAO) legend (or the more recent International Society of Soil Science (ISSS)–International Soil Reference and Information Center (ISRIC)–FAO *World Reference Base for Soil Resources*) and the Soil Taxonomy System developed by the Soil Survey Staff of the United States Department of Agriculture (USDA). The latter is used here for a cursory description of the main soil types.

Eleven of the 12 orders in the soil taxonomy are represented in the Mediterranean region: Alfisols, Entisols, Inceptisols, Mollisols, and Vertisols are commonplace to abundant; Aridisols and Ultisols can be locally important; Andisols, Histosols, and Spodosols are rare; and Oxisols, as those found in some areas of California, are considered to be relict soils.

Alfisols are widespread and correspond to most of the development stages of the chronosequence in Figure 5 (and similar chronosequences). Alfisols are generally found on stable geomorphic surfaces over a wide variety of parent materials and are represented by the Aqualf and Xeralf suborders. The former are locally significant in depressions, closed basins, and plateaus, and they tend to exhibit a perched rather than a permanent water table. Among the Xeralfs, Durixeralfs (with duripan) are generally restricted to volcanic areas, and Fragixeralfs (with fragipan) and Plinthoxeralfs (with plinthite) are relatively rare. Rhodoxeralfs, which exhibit vivid red colors, are commonplace – but not so much so as the Haplo- and Palexeralfs – and cover a variety of well-drained parent materials (calcarenites, hard limestones, arkoses, sandstones, colluvia of acid rocks, basalts, etc.). Terra Rossa *sensu stricto* (i.e., a red, clayey soil on pure, hard limestone) generally belongs to the Rhodoxeralf group. Reddish-brown Xeralfs not meeting the stringent color requirements for Rhodoxeralfs are also commonplace and generally classed as Haploxeralfs. Palexeralfs ('old' Xeralfs) generally occupy old geomorphic surfaces such as high river or marine terraces and ancient pediments. Some possess a petrocalcic horizon; others tend to support a perched water table. In any case, Haploxeralfs, which are regarded by many as the ragbag of Xeralfs, constitute the most widespread great group.

Entisols, which exhibit minimal pedogenic alteration, are particularly abundant in the mountainous or hilly areas, where water erosion and other mass-wasting processes keep the soil permanently thin. These soils are usually classed as Xerorthents. Less abundant, but agriculturally important, are Xerofluvents, which occupy alluvial plains and some colluvial areas. Psamments (sandy Entisols) are confined to alluvia, coastal dunes, and sandy continental deposits, as are Aquents (wet Entisols) to some depressions, deltas, and alluvial flats. Not uncommonly, the soils of these two suborders underlie highly ecologically significant rather than agriculturally valuable areas.

Inceptisols in Mediterranean areas often represent steps in evolutionary sequences ranging from raw soils (Entisols) to Alfisols or Vertisols, but they are truly mature soils in other cases. Most are Xerepts, but Aquepts are also present. Xerepts include soils with calcic or petrocalcic horizons (Calcixerepts) and acidic soils (Dystroxerepts). They occur in all types of landscape and climatic area, and include agricultural and forest soils. Aquepts are restricted to alluvial and coastal plains, and closed basins; also, like Aquents, some are environmentally significant and can be turned into good agricultural soils after drainage.

The presence of a dark, well-structured (mollic) epipedon (a requirement for the Mollisol order) was probably common at some time under the typical sclerophyllous vegetation. Man's activities (deforestation, ploughing, grazing, etc.) have resulted in substantial losses of organic matter and deterioration of the good structural properties of the epipedon. Assessing the extent of such changes, however, is a difficult task. At present, less than 10% of the Mediterranean region is covered by Mollisols (mostly Xerolls), which tend to be concentrated in mountainous areas, man-made grasslands, and other uncultivated areas.

Table 3 Soil types most commonly found in Mediterranean areas

Order[a]	*Suborder*[a]	*Group*[a]	*WRB*[b]	*Probable extent (%)*
Entisols	Orthents	Xerorthents	Leptosols, Regosols	20–25
	Fluvents	Xerofluvents	Fluvisols	1–3
	Psamments	Quartzipsamments	Arenosols	<1
		Xeropsamments	Arenosols	1–3
Inceptisols	Xerepts	Calcixerepts	Calcisols	10–15
		Dystroxerepts	Cambisols	3–5
		Haploxerepts	Cambisols	10–15
	Aquepts		Gleysols	<1
Vertisols	Xererts	Calcixererts	Vertisols	1–2
		Haploxererts	Vertisols	1–2
	Aquerts		Vertisols	<1
Alfisols	Xeralfs	Haploxeralfs	Luvisols	15–25
		Palexeralfs	Luvisols	3–5
		Rhodoxeralfs	Luvisols	3–5
		Durixeralfs	Durisols	<1
	Aqualfs		Luvisols	<1
Mollisols	Xerolls	Calcixerolls	Kastanozems	2–4
		Haploxerolls	Phaeozems	2–4
Ultisols	Xerults	Haploxerults	Acrisols	1–2
		Palexerults	Acrisols	<1
Aridisols	Salids		Solonchaks	<1

[a]According to Soil Survey Staff (1999) *Soil Taxonomy*, 2nd edn. Agricultural Handbook 436. Washington, DC: USDA.
[b]ISSS–ISRIC–FAO (1998) *World Reference Base for Soil Resources*. Leuven, Belgium: Acco Press.

Vertisols form readily where the parent material is clayey or the local conditions favor the formation of smectite. Mediterranean Vertisols (Xererts) occupy relatively large areas on Tertiary marls and Quaternary clays (e.g., in the river Guadalquivir valley in Spain and some plains of western Morocco, where smectite is largely inherited from the parent material ('lithogenic' Vertisols)). In closed basins, footslopes, and some poorly drained areas, where smectite is neoformed ('topogenic' Vertisols), Xererts alternate with Aquerts and gradually merge into Alfisols at some higher topographic locations. Among Xererts, Calcixererts are probably more abundant than Haplo- and Durixererts, because the low hydraulic conductivity hinders decalcification in Vertisols.

Ultisols exhibit low base saturation in the subsoil, so they cannot be expected to form under the leaching regime imposed by the average Mediterranean climate. However, they do occur in areas where at least two of the following conditions are met, namely: (1) high (more than 200–300 mm) winter leaching; (2) base-poor parent materials (e.g., slates or quartzites); and (3) a stable geomorphic surface. These three conditions are fulfilled, for instance, by Xerults and Aquults in Pliocene *raña* formations, which occupy more than 10 000 km^2 in Spain.

The only Aridisols present in the Mediterranean regions are Salids. Although not abundant, these saline soils fairly frequently host ecologically unique areas in deltas, enclosed depressions, and coastal plains.

Finally, only the coincidence of unusual local conditions favors the occurrence of Spodosols, Andisols, and Histosols. For instance, Spodosols in western Portugal form on coastal dunes under a pine forest in a subhumid Mediterranean environment.

Table 3 shows the probable extent for the most common soils in the Mediterranean region. These figures can only be tentative, given the relative scarcity of detailed soil maps for some Mediterranean countries.

Physical, Chemical, and Mineralogical Properties

Many lands bordering the Mediterranean have been deforested and cultivated for centuries and even millennia. Man-induced water erosion and organic matter depletion have affected extensive areas, which have been worn down to unacceptable limits. Soil degradation in other Mediterranean areas has been less marked, but current risks derived from agriculture, forestry, industrial activities, or modern urban needs run high in most areas. Soil properties must be envisaged in any case from the standpoint of their significance and evolution in different man-influenced scenarios, which change – sometimes dramatically – with time.

Physical Properties

Generally, the quality of Mediterranean soils is more strongly influenced by physical than by chemical properties. Moreover, some physical properties cannot be significantly modified (e.g., texture) or changed without much effort (e.g., structural stability).

Mediterranean soils tend to be more clayey than soils from cool temperate areas. Textural contrasts such as those exhibited by Alfisols are common and generally result in water-holding capacity increasing with increase in depth, which favors growth of rainfed crops. Stoniness is widespread and generally reduces crop production; however, surface stoniness helps reduce surface evaporation, protects soil against splash erosion, and delays runoff.

The low structural stability of many cultivated soils has been attributed to their low content in organic matter. Various studies, however, have revealed other factors (Fe oxides, Ca saturation, clay type) to be equally important with a view to creating structure. Soil crusting affects many soils rich in silt and fine sand, where it hinders emergence of small seedlings and triggers runoff. Hardsetting A horizons, which combine structural instability and high strength when the soil is dry, have been reported in the Xeralfs of Southwest Australia and some countries bordering the Mediterranean.

Low water availability in summer limits the yield potential of Mediterranean soils, even though they possess, on average, a water-holding capacity exceeding that of soils in humid temperate regions. Soil crusting and low infiltrability may offset the latter advantage, however.

Mineralogical Properties

The lack of coincidence of high temperatures and high moisture contents prevents intensive weathering in the Mediterranean environment. Therefore, highly weathered soils occur only on old, stable geomorphic surfaces in subhumid regions. Generally, because of the low-to-moderate degree of weathering, much of the soil silicate clays is inherited from the parent material or produced by transformation of primary micas rather than neoformed from solution. Illite constitutes approximately 50% of the clay fraction on average, as it abounds in sedimentary rocks and is produced by transformation of muscovite.

The overall sequence of abundance of clay minerals is illite > smectite, illite–smectite > kaolinite > chlorite, vermiculite, hydroxy-interlayered vermiculite >> halloysite, palygorskite, pyrophyllite, talc. The composition of the soil solution in many soils falls near the lines bounding the kaolinite and smectite stability fields. There is indeed evidence that kaolinite in the more leached soil environments and smectite in solute-rich environments are neoformed in substantial amounts. Neoformation of illite has been documented in K-rich solutions, such as those in microcracks of potassium feldspar grains. The presence of different microenvironments, leaching gradients in the soil profile, and lateral eluviation combine to create differences in clay mineralogy with depth and along a catena. Figure 6 depicts an idealized spatial distribution of the minerals present in the clay fraction in a typical catena.

Iron oxides are the most abundant among nonsilicate clay minerals; they typically account for 5–10%

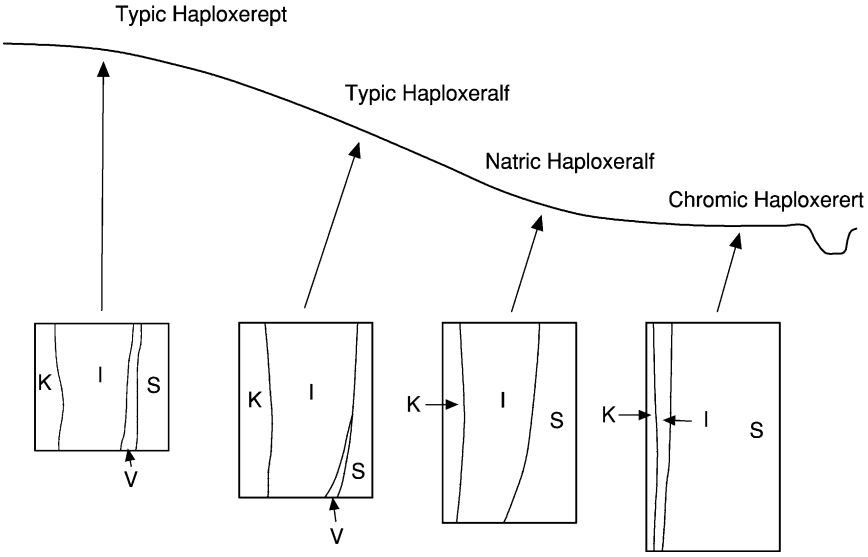


Figure 6 Relative abundances of silicate clays in the clay fraction of soils of a hypothetical catena on a granodiorite saprolite in a typical Mediterranean environment. K, kaolinite; I, illite; V, vermicullite; S, smectite.

of the clay fraction. The Mediterranean pedoenvironment favors crystallization of ferrihydrite to goethite and hematite, so that the ratio of poorly crystalline to crystalline Fe oxides is generally low. Aluminum oxides are rare, gibbsite being present in some excessively drained soils formed on coarse-grained, acidic plutonic rocks.

Among the abundant carbonates, calcite and magnesium calcite are either inherited or secondary, whereas dolomite is always inherited. Calcite crystals can be dispersed in the soil matrix or aggregated in nodules, tubules, or crusts. Crystals range in size from clay to fine sand, but the fine silt fraction is generally the most abundant.

Chemical Properties

The pH of Mediterranean soils generally lies within the slightly acidic to moderately alkaline range and, because of the occasionally steep vertical leaching gradient, can increase substantially with depth. Thus, the idealized mature soils of Figure 5 typically have a pH of approx. 6 on the surface and approx. 8 in the horizons where carbonate accumulates. However, ploughing and accretion of dust from nearby or distant sources often result in recalcification of surface horizons. Consequently, the highly acidic soils formed on old geomorphic surfaces on acidic rocks lie usually far from areas with calcareous soils and rocks.

The characteristic mineral assemblages of Mediterranean soils endow them with interesting quantitative adsorption features. Thus, the cation exchange capacity of the clay fraction generally ranges from 40 to 60 $cmol_c\,kg^{-1}$ as a result of the prevalence of 2:1 clay minerals. Specific ion adsorption by hydroxylated mineral surfaces largely arises from Fe oxides and, to a lesser extent, from the edges of clay minerals and aluminum oxides. Calcite surfaces also play a prominent role in adsorption and heterogeneous precipitation processes. Quantitative relationships between adsorption capacity and simple mineralogical or chemical properties can be established for most of the inorganic adsorbates of interest (e.g., phosphate).

Generally, the adsorptive properties of Mediterranean soils depend strongly on depth as a result of steep gradients in pH, particle size distribution, and mineralogical composition. Consequently, the appropriate 'filters' for the different nutrients and pollutants lie at different depths – and make morphological and mineralogical characterization of the soil profile necessary.

From the standpoint of chemical fertility, Mediterranean soils compare favorably with soils in other geographic areas. After N, P is the most limiting nutrient in native soils; fortunately, most soils are acceptably responsive to fertilization with P. Most are well supplied with K, because of the high illite content, and contain adequate levels of Ca and Mg. However, Fe, Cu, Mn, Zn, and B deficiencies frequently arise in some horticultural crops cultivated on calcareous soils.

Salinization, occasionally accompanied by specific ion toxicity problems, has made its way in both old and recently irrigated areas. Even though the principles for good management of irrigation have been known for decades, salinization continues to be a major issue; the scarcity of good irrigation water in most areas, and the use of water-saving irrigation systems, contribute to the problem.

Further Reading

Daz MC (1986) *Estudio de los Procesos de Rubefacción de los Suelos de las Terrazas de la Cuenca del Ro Henares y del Ro Jarama*. PhD thesis, Universidad Politécnica de Madrid, Spain.

ISSS–ISRIC–FAO (1998) *World Reference Base for Soil Resources*. Leuven, Belgium: Acco Press.

Peña F and Torrent J (1984) Relationships between phosphate sorption and iron oxides in Alfisols from a river terrace sequence of Mediterranean Spain. *Geoderma* 33: 283–296.

Roquero C (1979) The potential productivity of Mediterranean soils. In: *Soils in Mediterranean Type Climates and their Yield Potential* (Proceedings of the 14th Colloquium of the International Potash Institute held in Seville/Spain) pp. 21–42. Bern, Switzerland: International Potash Institute.

Soil Survey Staff (1999) *Soil Taxonomy*, 2nd edn. Agriculture Handbook 436. Washington, DC: US Department of Agriculture.

Torrent J (1995) *Genesis and Properties of the Soils of the Mediterranean Regions*. Napoli, Italy: Dipartimento di Scienze Chimico–Agrarie, Università degli Studi di Napoli Federico II.

Torrent J and Cabedo A (1986) Sources of iron oxides in reddish brown soil profiles from calcarenites in southern Spain. *Geoderma* 37: 57–66.

Torrent J, Nettleton WD, and Borst G (1980a) Clay illuviation and lamellae formation in a Psammentic Haploxeralf of southern California. *Soil Science Society of America Journal* 44: 363–369.

Torrent J, Nettleton WD, and Borst G (1980b) Genesis of a Typic Durixeralf of southern California. *Soil Science Society of America Journal* 44: 575–582.

Yaalon DH (1997) Soils in the Mediterranean region: what makes them different? *Catena* 28: 157–169.

METAL OXIDES

A C Scheinost, Swiss Federal Institute of Technology (ETH), Zurich, Switzerland

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

The oxide minerals typically present in soils comprise oxides, hydroxides, oxyhydroxides, and hydrated oxides of Si, Fe, Mn, Al, and Ti. With the exception of the Si oxide quartz and some Ti oxides, which are predominantly inherited from primary (rock) minerals, most oxides form in soil. The metal cations Fe^{2+}, Mn^{2+}, Ti^{4+}, Al^{3+}, and Si^{4+} are released from silicates by weathering. The divalent cations Mn and Fe oxidize and hydrolyze, and precipitate almost exclusively as oxide minerals. Al^{3+} and Si^{4+} have a strong tendency to form secondary aluminosilicate clay minerals; the formation of Al and Si oxides requires special conditions. All oxides have a very low solubility at common soil pH and are therefore enriched during pedogenesis. Highly weathered soils, which have lost a substantial part of their alkalis, alkaline earths, and Si, may contain as much as 50% wt metal oxides.

Oxides of Fe, Mn, and Al may exhibit a high surface area, with reactive surface sites which strongly bind oxyanions and metal cations, thereby impacting the availability and mobility of plant nutrients and toxic metals. Some Mn oxides have a strong oxidizing power and may degrade organic pesticides, decrease or increase the toxicity of heavy metals, and contribute to the formation of soil organic matter. In contrast, sand-sized crystals of the Si oxide quartz are chemically very inert.

The basic structural unit of Fe, Mn, Al, and Ti oxides are cationic metal centers bound to six oxygens. If one drastically reduces the size of the oxygens in relation to the metal cations, this arrangement can be visualized as an octahedron (Figure 1). The octahedra may be linked to each other in three ways, sharing oxygen corners (1 O), sharing edges (2 O), or faces (3 O). Many oxide minerals such as goethite contain only the first two types of connections, while the Fe oxide hematite contains all three. Silicon is bound to four oxygens, forming a tetrahedron. These tetrahedra are connected via corners only. We will use this polyhedral approach in the text and in figures to introduce the mineral species and highlight their major differences and similarities. For each oxide group, we will give an overview including general formation pathways and influence on soil properties. This overview is followed by a description of the minerals, their occurrence in soil, and specific conditions of their formation.

Fe Oxides

Iron is released by weathering of Fe(II)-containing silicates (biotite, pyroxene, amphibole, olivine). After oxidation to Fe^{3+} and hydrolysis, most Fe precipitates as Fe(III) oxides. (Only a small remaining amount of Fe^{2+} becomes a structural part of pedogenic phyllosilicates.) Under oxic conditions, Fe oxides are very insoluble, hence are enriched during pedogenesis. They readily dissolve, however, under reducing soil conditions:

$$2Fe_2O_3 + CH_2O + 8H^+ \Rightarrow 4Fe^{2+} + CO_2 + 5H_2O \qquad [1]$$

This reaction is driven by microorganisms, which metabolize biomass (in eqn [1] represented by CH_2O), and transfer freed electrons to Fe^{3+}, thereby

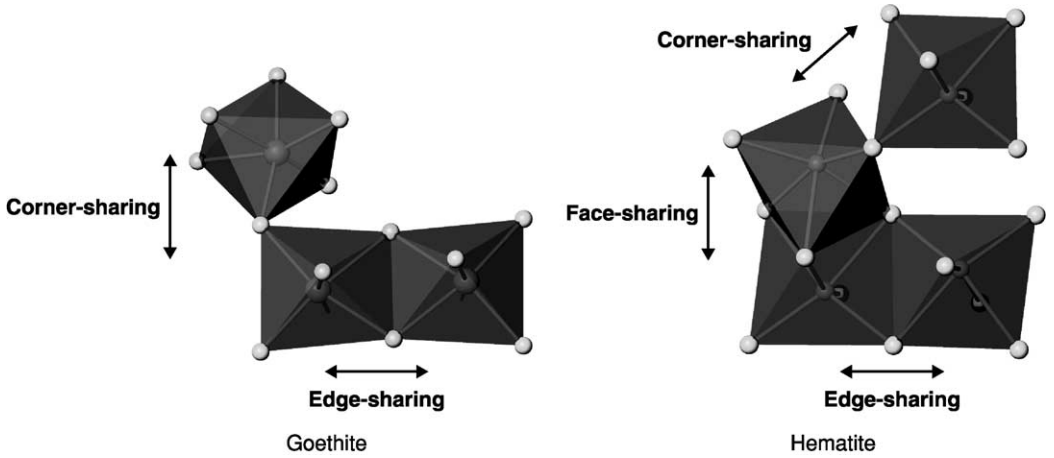


Figure 1 Linkage of metal octahedra by corners, edges, and faces, involving one, two, and three oxygen ligands, respectively. Shown are small structural subunits of the two Fe oxides goethite and hematite.

reducing it to Fe^{2+}. Dissolution of Fe oxides is therefore linked to microbial activity. Depending on parent material, degree of weathering, and soil redox conditions, the amount of Fe oxides in soils varies between less than 1 and more than 500 g kg^{-1}. The Fe oxides are commonly the strongest pigments in soil, and small quantities may impart vivid colors of red, orange, yellow, brown, and even blue-green (Table 1). Depending on pedogenesis, they occur either evenly distributed in the soil matrix, or locally enriched as skins, patches, and concretions. Soil horizons indurated by Fe oxides form ferricretes or laterites.

Soil Fe oxides commonly exist as very small crystals (2–50 nm), responsible for a large surface area (50–450 $m^2\,g^{-1}$). Due to the their circumneutral point of zero charge (PZC), they contribute to the anion exchange properties of soils. In addition, their surfaces have a strong affinity for oxyanions and many metal cations. Finally, other metal cations with an ionic diameter similar to Fe^{3+} (e.g., Al^{3+}, Ni^{2+}, Ti^{4+}, Mn^{4+}, Co^{3+}, Cr^{3+}, Cu^{2+}, Zn^{2+}, V^{3+}) may partially replace Fe in the crystal structure of Fe oxides. Due to these properties, Fe oxides play a significant role in the environmental cycling of plant nutrients such as N and P, of transition metals relevant for metabolic reactions (e.g., Mn, Zn, Co), and of toxic elements such as As and U. The 10 Fe oxide minerals observed in soils or sediments at the Earth's surface and some of their diagnostic criteria are listed in Table 1. Crystal structures are shown in Figure 2. Schematic formation and transformation pathways of the eight most common Fe oxides are shown in Figure 3.

Goethite (α-FeOOH)

Due to its high stability, goethite is the most ubiquitous Fe oxide mineral in soils. In the absence of hematite, its yellowish color is responsible for the yellowish-brown hue of many soils (7.4 YR–3.5 Y). Goethite formation is favored by slow hydrolysis of Fe^{3+} hydroxy cations at low temperature. The even distribution of goethite in a soil horizon indicates therefore that this soil has formed under aerated, temperate, humid conditions. The crystal structure

Table 1 Iron oxide minerals

Mineral name and formula	*Structural type*	*Median color*	*Range of hue*	*Typical crystal shape*	*Occurrence*	*Figure*
Goethite[a] α-FeOOH	Diaspore	Strong yellowish-brown	7.3 YR–1.5 Y	Needles, laths	Ubiquitous in soils of all climates	2a
Lepidocrocite[a] γ-FeOOH	Boehmite	Moderate orange	4.9 YR–7.9 YR	Laths	Seasonally anaerobic, noncalcareous soils in cool-temperate climate	2c
Akaganeite β-FeOOHCl	Hollandite	Strong brown	1.2 YR–6.8 YR	Spindle-shaped rods	Hot springs, volcanic deposits	2f
Feroxyhyte δ'-FeOOH		Brownish orange	2.8 YR–9.2 YR	Spheres		2e
Schwertmannite $Fe_8O_8(OH)_6SO_4$	Hollandite	Dark orange yellow	6.2 YR–0.3 Y	Hedgehog aggregates	Acid-sulfate springs and mine drainages	2f
Ferrihydrite[a] $Fe_5HO_8 \cdot 4H_2O$		Strong brown	3.7 YR–5.4 YR	Plates, needles	Frequent in seasonally wet and/or organic-matter rich soils in cool-temperate climate	–
Hematite[a] α-Fe_2O_3	Corundum	Moderate reddish brown	3.5 R–4.1 YR	Hexagonal plates	Frequent in soils of tropical, subtropical, and Mediterranean climate	2b
Magnetite[a] $Fe^{II}Fe^{III}_2O_4$	Inverse spinel	Dark yellowish brown	6 YR–9.5 YR	Laths, cubes	Inherited from magnetite-containing bedrock, microbial origin possible	2d
Maghemite[a] γ-Fe_2O_3	Defect spinel	Black		Octahedra	Common in highly weathered, tropical and subtropical soils	2d
Green rust $Fe^{II}Fe^{III}(OH)_4Cl$	Hydrotalcite (pyroaurite)	Bluish-green		Plates	Permanently wet subsoils	2g

[a]Minerals commonly occurring in soil.

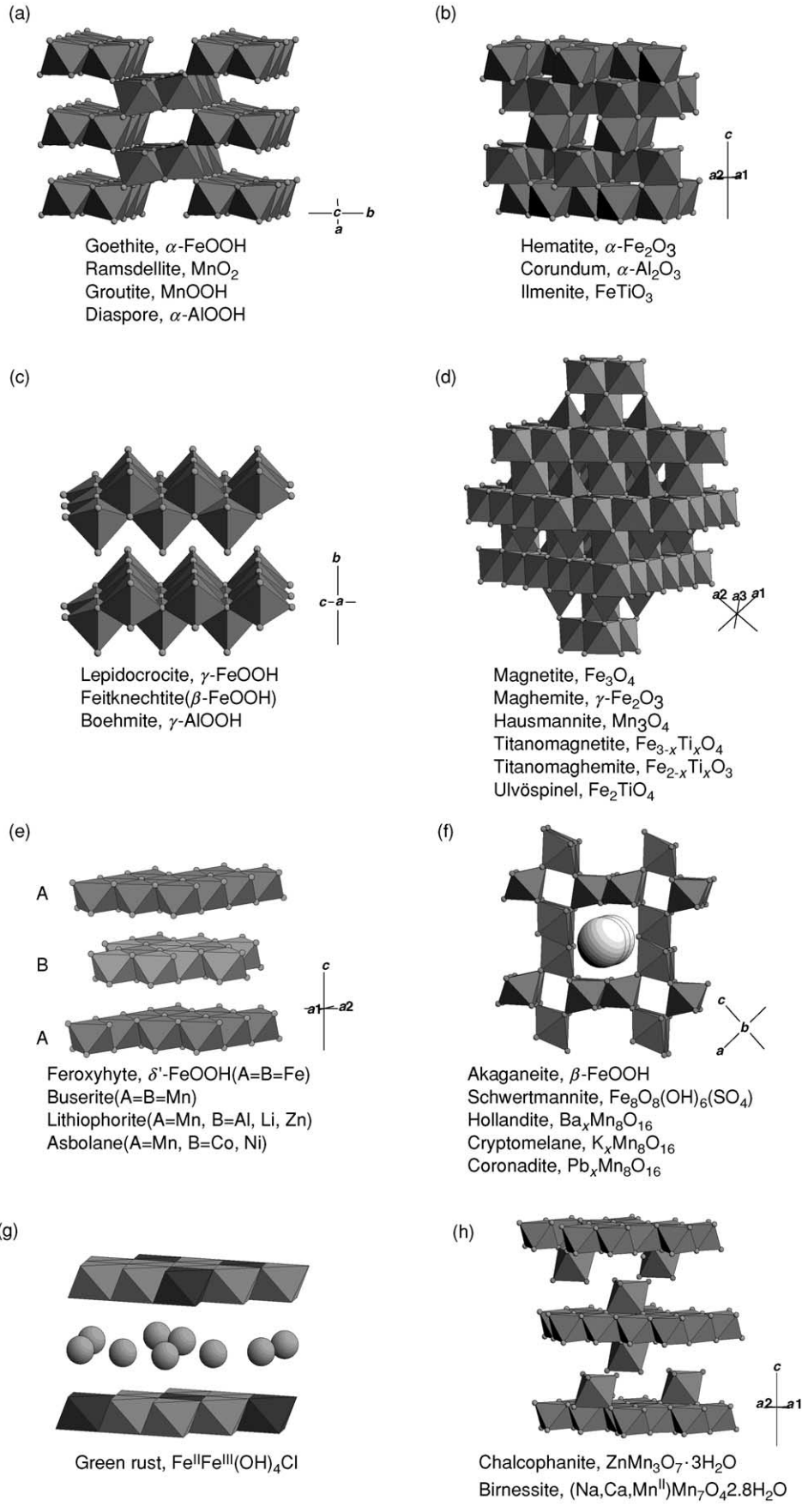


Figure 2 Crystal structures of common oxide minerals.

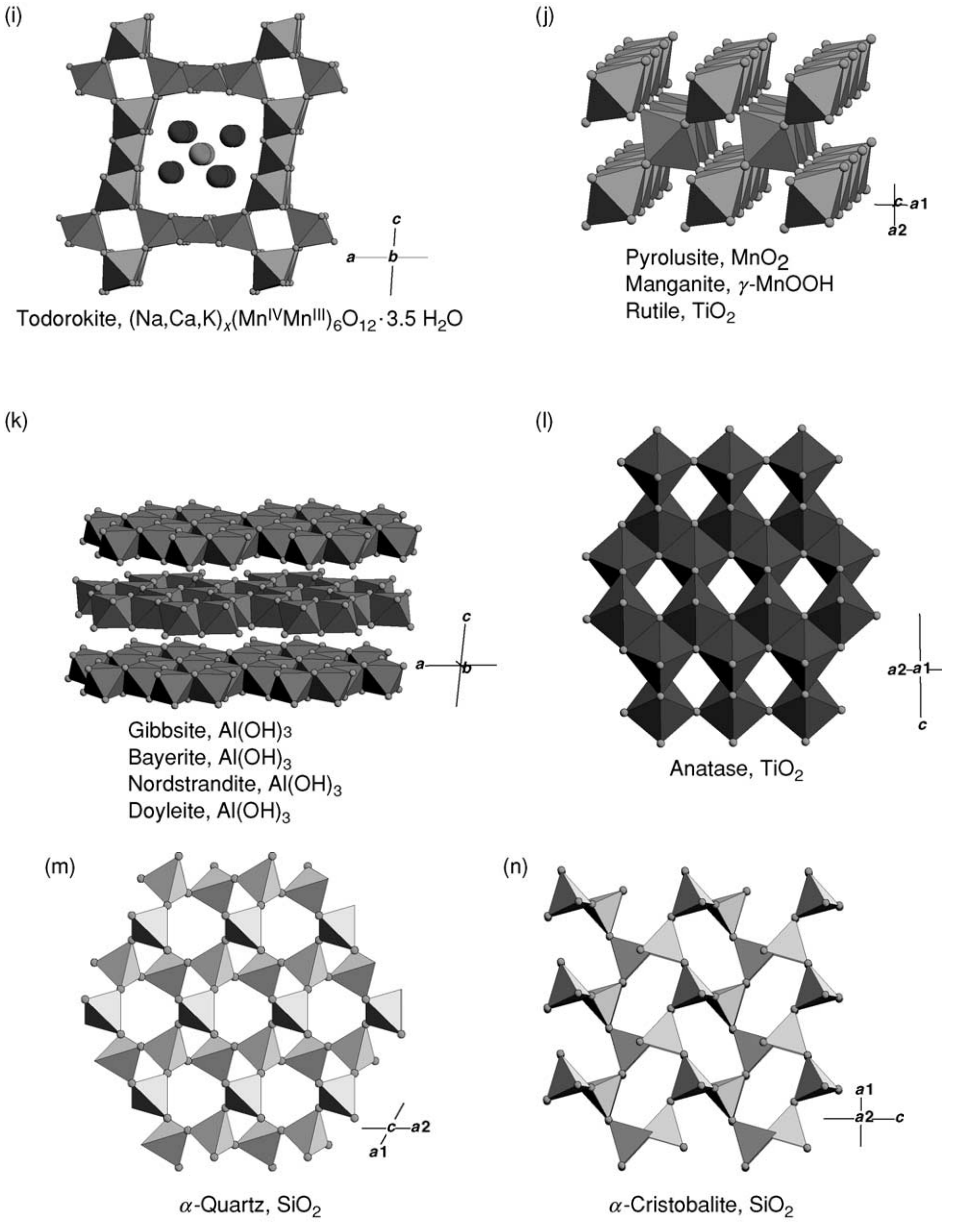


Figure 2 (*Continued*)

of goethite consists of double chains of edge-shared octahedra that are joined to other double chains by sharing corners and by hydrogen bonds (Figure 2a).

Hematite (α-Fe_2O_3)

Soils forming at subtropical or tropical climates often have reddish or even purplish hues (8.0 R–1.4 Y), derived from blood-red hematite. Hematite usually coexists with goethite, but, due to its greater tinting strength, masks the yellowish color of goethite even at low concentrations. Hematite forms by dehydroxylation of ferrihydrite (see below) via a solid-state mechanism which is not yet fully understood. In comparison with goethite, the formation of hematite is favored by neutral pH, increasing soil temperature, and decreasing water activity. The crystal structure of hematite consists of sheets of edge-sharing octahedra. The sheets are connected by edge- and face-sharing octahedra (Figure 2b). The unique face-sharing arrangement and the resulting proximity of neighboring Fe centers is responsible for the red color of this mineral.

Ferrihydrite ($Fe_5HO_8\cdot 4H_2O$)

This poorly crystalline, hydrous oxide commonly occurs in cooler soils with changing redox conditions. Its formation is favored by rapid oxidation and hydrolysis of Fe^{3+} from solution. The transformation

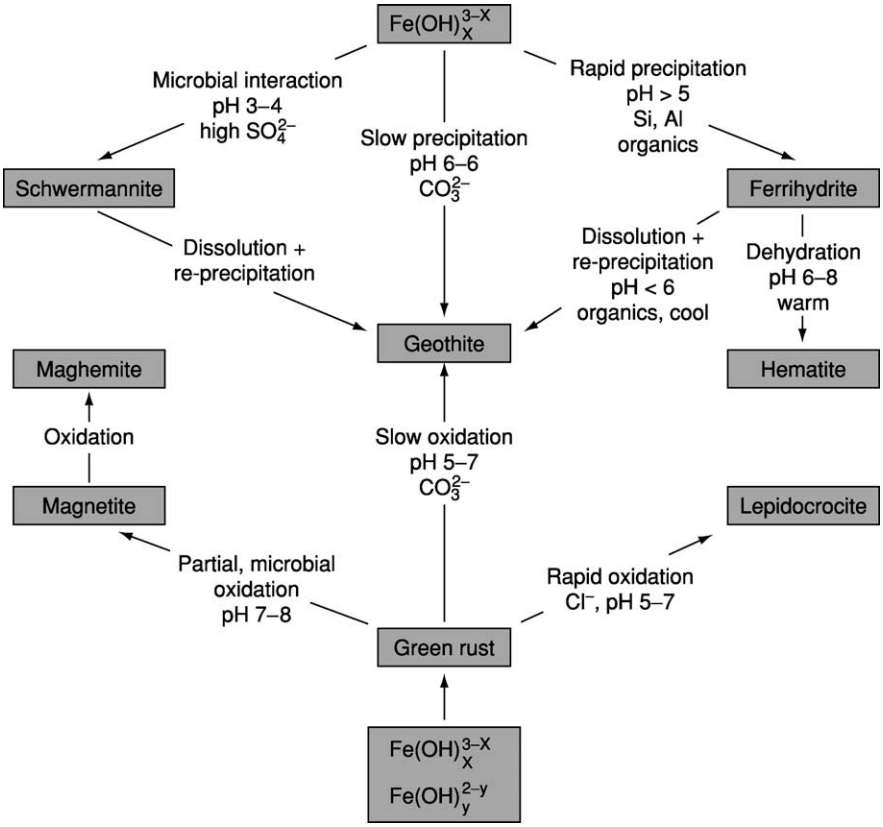


Figure 3 Schematic formation and transformation pathways of common Fe oxides in soils. The precipitation of Fe oxide minerals proceeds either from ferric hydrolysis species (top) or from mixed ferric-ferrous hydrolysis species (bottom). Figure modified after Bigham JM, Fitzpatrick RM, and Schulze DG (2002) Iron oxides. In: Dixon JB and Schulze DG (eds) Soil Mineralogy with Environmental Application, pp. 323–366. Madison, WI: Soils Science Society of America and Cornell RM and Schwertmann U (2003) The Iron Oxides: Structure, Properties, Reactions, Occurence and Uses, 2nd edn. Weinheim, Germany: VCH Verlagsgesellschaft.

into thermodynamically more stable goethite may be hindered by adsorption of silicate, phosphate, and organic anions, making ferrihydrite a common constituent of younger soils in association with goethite. The brownish-orange color of ferrihydrite often marks gradients between anoxic and oxic soil regions, e.g., at the transition between water-saturated and aerated soil horizons, and along root channels or other macropores. Ferrihydrite is also responsible for the brownish-orange color in podzol B horizons. The structure is still being investigated, but can be visualized as a defective hematite structure containing both edge- and face-sharing octahedra.

Lepidocrocite (γ-FeOOH)

This vividly orange, less abundant mineral occurs in seasonally anaerobic, clay-rich, carbonate-free soils in cool climates in association with the more stable goethite. Lepidocrocite forms through relatively slow oxidation of Fe^{2+} at low concentrations of carbonate. Its crystal structure contains double chains of octahedra which are joined by shared edges, resulting in corrugated sheets of octahedra. These corrugated sheets are stacked one on top of the other and are held together by hydrogen bonds (Figure 2c).

Magnetite (Fe_3O_4) and Maghemite (γ-Fe_2O_3)

Blackish magnetite is usually inherited from the parent material, but may also form in soil through biotic processes. The brownish-red maghemite is common in soils of tropical and subtropical climate, but occurs occasionally also in temperate soils that have been exposed to forest or heath fires. Maghemite forms either through oxidation of magnetite, or by heating (300–425°C) of other Fe oxides in the presence of organic compounds. Both magnetite and maghemite have the same inverse spinel structure, with octahedral and mixed tetrahedral/octahedral layers stacked along the direction [111] (Figure 2d). In magnetite, which contains both Fe^{2+} and Fe^{3+} ions, Fe^{3+} occupies both tetrahedral and octahedral sites, while the larger Fe^{2+} ion occupies octahedral sites only.

Rare Fe Oxide Minerals

Several minerals have rarely or never been observed in soils. They may form in extreme natural environments (akaganeite, schwertmannite), in polluted surface waters (schwertmannite, feroxyhite), or transform rapidly into more stable phases (green rust). Feroxyhyte, δ'-FeOOH, has been observed in rusty

precipitates of rapidly flowing, Fe^{2+}-containing water, which was quickly oxidized. It consists of sheets of edge-sharing octahedra (Figure 2e). Akaganeite (β-FeOOHCl) has been found in environments with high chloride concentrations (0.1 mol l^{-1}), low pH (3–4), and high temperature (60°C), e.g., in hot springs and volcanic deposits. Akaganeite consists of 2×2 channels (referring to the width of the channels in octahedral units) built by edge-sharing double chains (Figure 2f). These tunnels 0.5 nm in cross section contain Cl^- ions that stabilize the structure. Schwertmannite, $Fe_8O_8(OH)_6SO_4$, is isostructural with akaganeite, but SO_4^{2-} occupies the tunnels (Figure 2f). Schwertmannite has been found in strongly acid sulfate waters associated with mining activities (acid mine drainage) or pyritic rock outcrops. Green rust, $[Fe^{II}_{1-x}Fe^{III}_x(OH)_2]^{x+} \cdot (x/n)A^{n-} \cdot mH_2O$ (where A represents anions such as Cl^-, NO_3^-), has been observed rarely in reducing, weakly acid to weakly alkaline subsoils. Its greenish-blue color changes rapidly to yellowish brown on exposure to the air, indicative of its instability against oxidation. Green rust consists of sheets of edge-sharing octahedra, which are occupied by both Fe^{2+} and Fe^{3+}. The resulting positive layer charge is balanced by hydrated anions in the interlayer space (Figure 2g). Due to its anion exchange capacity, redox reactivity and high surface area, even small amounts of green rust may play an important role in aquifers and the vadose zone. Since the mineral forms as a metastable corrosion product of metallic iron at anaerobic conditions, green rust is also relevant for barriers of scrap iron engineered to clean up groundwater.

Mn Oxides

Mn is released by weathering of Mn(II)-containing silicates (biotite, pyroxene, amphibole). After oxidation of the soluble Mn^{2+} to Mn^{3+} and Mn^{4+}, brownish-black Mn oxides of low solubility precipitate, which redissolve only under reducing soil conditions. The reaction equation assuming a Mn(IV) oxide can be written in analogy to the dissolution of Fe oxides

$$2MnO_2 + CH_2O + 4H^+ \Rightarrow 2Mn^{2+} + CO_2 + 3H_2O$$

As for the Fe oxides, the dissolution depends on microbial activity. While oxidation of Mn^{2+} is catalyzed by mineral surfaces, microbial oxidation seems to be the more dominant process in soil environments. Therefore, both formation and dissolution of Mn oxides are intimately linked to microbial activity.

Mn oxides may occur evenly distributed in the soil matrix like the stable Fe oxides goethite and hematite. Due to their low concentration (about 1/50 of the concentration of Fe oxides) and their low crystallinity, however, they are extremely difficult to identify. Therefore, most soil Mn oxides have been identified in local enrichments, such as black coatings of peds, dendrites, or nodules, which form along redox gradients. The blackish colors of Mn oxides are often separated from the brownish-orange color of Fe oxides (ferrihydrite) by extending further toward the higher redox potential, indicative of a higher mobility of Mn^{2+} as compared to Fe^{2+}.

Mn oxides are able to accumulate a wide range of other elements such as Li, Ba, As, Pb, and almost all first-row transition metals. These elements are either part of the crystal structure or they are tightly sorbed to the large surface area of Mn oxides. Mn oxides have an oxidizing potential stronger than that of O_2. Therefore, they are able to oxidize inorganic ions such as Co^{2+}, Cr^{3+}, and As^{3+}, thereby either increasing (e.g., Cr^{6+}) or decreasing (e.g., Co^{3+}, As^{5+}) their mobility and toxicity. They are also able to oxidize organic molecules, thereby enhancing the degradation of anthropogenic compounds like pesticides. Furthermore, Mn oxides catalyze condensation reactions such as the Maillard reaction. Hence, Mn oxides may play an important role in the abiotic formation of humic substances.

In spite of their importance for the geochemistry of metals and carbon, knowledge of their mineralogy and formation in soils is still relatively poor. This is due to their low crystallinity and low concentration, but also due to their complex mineral chemistry. Since microspectroscopic methods became available recently which are suited to study Mn oxides in soils, this may change within the next years.

Based on the arrangement and linkage of their metal octahedra, the Mn oxides have been categorized into three major groups: (1) phyllomanganates or layer structures, (2) tectomanganates or tunnel structures, and (3) chain structures. The most common Mn oxide minerals are listed in Table 2, and their crystal structures are shown in Figure 2.

Phyllomanganates (Layer Structures)

Phyllomanganates are commonly observed after microbially catalyzed oxidation of Mn^{2+} and they are ubiquitous in soils.

Birnessite forms in a wide variety of soils. It consists of sheets of edge-sharing MnO_6 octahedra with vacancies. The resulting negative layer charge is compensated by cations such as Na^+, K^+, Ca^{2+}, Mg^{2+}, and Mn^{2+}, which may cover the vacancies. The interlayer region contains water, resulting in a d-spacing of 7 Å (Figure 2h). Three other minerals, vernadite, chalcophanite, and buserite, are structurally very similar to birnessite. Vernadite is frequently observed as an initially amorphous oxidation product of microorganisms and has been found in soils and in

Table 2 Manganese oxide minerals

Mineral name (synthetic equivalent)	*Formula*	*Structural type*	*Figure*
Birnessite[a]	(Na, Ca, Mn^{2+})$Mn_7O_4 \cdot 2.8\ H_2O$	Layer structure	2h
Vernadite[a] (δ-MnO_2)	$MnO_2 \cdot nH_2O$	Layer structure	–
Chalcophanite	$ZnMn_3O_7 \cdot 3\ H_2O$	Layer structure	2h
Buserite	$Na_4Mn_{14}O_{27} \cdot 21\ H_2O$	Layer structure	–
Lithiophorite[a]	$LiAl_2Mn_2^{4+}Mn^{3+}\ O_6(OH)_6$	Layer structure	2e
Todorokite[a]	(Na, Ca, K)$_x$(Mn^{4+} Mn^{3+})$_6O_{12} \cdot 3.5\ H_2O$ ($x = 0.3$–0.5)	3×3 tunnel	2i
Romanechite	$Ba_{0.66}$ (Mn^{4+} Mn^{3+})$_5O_{10} \cdot 1.34H_2O$	3×2 tunnel	–
Hollandite (α-MnO_2)	$Ba_x(Mn^{4+}Mn^{3+})_8O_{16}$ ($x = 1$)	2×2 tunnel	2f
Cryptomelane	$K_x(Mn^{4+}Mn^{3+})_8O_{16}$ ($x = 1.3$–1.5)	2×2 tunnel	2f
Coronadite	$Pb_x(Mn^{4+}Mn^{3+})_8O_{16}$ ($x = 1$–1.4)	2×2 tunnel	2f
Pyrolusite (β-MnO_2)	MnO_2	Rutile	2j
Manganite (γ-MnOOH)	MnOOH	Rutile	2j
Ramsdellite	MnO_2	Diaspore	2a
Groutite (α-MnOOH)	MnOOH	Diaspore	2a
Feitknechtite (β-MnOOH)	MnOOH	Boehmite	2c
Hausmannite	Mn_3O_4	Inverse spinel	2d

[a]Minerals commonly occurring in soil.

freshwater lakes. Its structure may be similar or identical to turbostratic H^+-birnessite. Chalcophanite has been observed in Zn-contaminated soils. In comparison with birnessite, it has a higher number of vacancies that are covered by Zn in octahedral or tetrahedral coordination (Figure 2h). Buserite forms in hydrothermal environments. In comparison with the former phyllosilicates, it has a larger d-spacing of 10 Å, probably due to a larger amount of interlayer water. By dehydration, buserite transforms into birnessite.

Lithiophorite has been frequently identified in soils. Its layer structure consists of sheets of edge-sharing MnO_6 octahedra which alternate with sheets of (Al, Li)$(OH)_6$ octahedra (Figure 2e). The cation sites in the Mn-O octahedral sheet are occupied with Mn^{4+} and Mn^{3+}. The net negative layer charge is counterbalanced by the positive charge of the (Al, Li)-OH octahedral sheets. Li is frequently replaced by Ni, Cu, and Zn, while Co substitutes for Mn within the octahedral Mn sheets. Asbolane forms in soils with high Co and Ni concentrations. Its structure is very similar to that of lithiophorite, with alternating sheets of Mn^{4+}-O octahedra and Co-Ni-OH octahedra (Figure 2e). The Co-Ni sheet may be discontinuous (island-like).

Tectomanganates (Tunnel Structures)

Tectomanganates are formed from double or triple chains of MnO_6 octahedra, which form tunnels (or channels). These tunnels are occupied by large foreign cations and water molecules.

Todorokite is the most frequently observed soil tectomanganate. It consists of triple chains of edge-sharing MnO_6 octahedra linked to form large 3×3 tunnels. These tunnels contain Na, Ca, K, Ba, Sr, and water molecules (Figure 2i). The occurrence of variable tunnel widths (3×2, 3×3, 3×4, and 3×5) observed in high-resolution transmission electron microscopy (HRTEM) images suggests that todorokite represents a family rather than a single mineral.

Hollandite, cryptomelane, and coronadite, sometimes grouped as α-MnO_2, have rarely been observed in soils. They consist of double chains of edge-sharing MnO_6 octahedra linked to form 2×2 tunnels (Figure 2f). The tunnels contain water molecules along with cations, which are primarily Ba^{2+} in hollandite, K^+ in cryptomelane, and Pb^{2+} in coronadite. Natural samples usually contain a variety of cations in the tunnels. The large cations are located at specific tunnel sites, and their presence is necessary to prevent the structure from collapsing. The charges of the tunnel cations are balanced by the substitution of Mn^{4+} by Mn^{3+} in the tunnel walls.

Chain Structures

Mn oxides of this group have been synthesized in the laboratory at ambient temperature and pressure, and at moderately alkaline pH. However, they have rarely been observed in soil. Pyrolusite (MnO_2) and manganite (γ-MnOOH) consist of single chains of edge-sharing MnO_6 octahedra which form 1×1 pseudotunnels (Figure 2j). Ramsdellite (MnO_2) and groutite (α-MnOOH) are isostructural with goethite (Figure 2a); feitknechtite (β-MnOOH) has a structure similar to lepidocrocite (Figure 2c). Hausmannite (Mn_3O_4) has a disordered-spinel structure analogous to magnetite (Figure 2d).

Al Oxides

In contrast to Fe and Mn, most Al released by the weathering of Al-containing silicates is taken up by

secondary, pedogenic phyllosilicates. As long as the Si concentration in soil solution is more than 0.5 mg l^{-1}, the formation of phyllosilicates out-competes the formation of Al oxides. Furthermore, the complexation of Al^{3+} by organic matter prevents Al oxide precipitation. Therefore, Al oxides are commonly observed only in organic matter-poor soils of tropical and subtropical climate, where intense weathering causes secondary phyllosilicates to dissolve and Si is lost by leaching. The (colorless) Al oxides are either evenly distributed within the soil matrix or enriched in concretions (pisolites). Horizons cemented by Al oxides are called alcretes.

Al oxides may have a high surface area (up to 600 $m^2 g^{-1}$), a high PZC (pH 8–10), and a pH-dependent surface charge. Similar to Fe and Mn oxides, they tightly sorb heavy metals (Cu, Pb, Zn, Ni, Co, Cd) and anions (phosphate, silicate, molybdate, sulfate, catechol) by forming inner-sphere sorption complexes. The adsorption is related to the reactive surface area (singly coordinated OH^- groups at crystallite edges) rather than the total surface area. Aluminum oxides, in addition to Fe oxides, are responsible for the aggregation of Oxisols and Ultisols, but the mechanism of aggregation is unclear. Aluminum oxides have a low solubility at slightly acidic and neutral pH, but at low pH soluble species form that are toxic to plants and animals. These soluble species impact food production on naturally acidic soils and contribute to forest decline in areas influenced by acid rain.

Of six different Al oxide minerals, predominantly gibbsite and, less commonly, boehmite, nordstrandite, and bayerite form under soil conditions (Table 3). Diaspore and corundum are occasionally found in bauxite deposits, but are rare in soils. Aluminum also forms highly reactive, poorly crystalline precipitates and colloids.

Gibbsite

Of the four $Al(OH)_3$ polymorphs (Table 3), only gibbsite (γ-$Al(OH)_3$) is common in soil. It consists of sheets of edge-sharing octahedra stacked along the *c*-axis (Figure 2k). Only two-thirds of the available octahedral sites are filled with Al^{3+} ions analogous to the dioctahedral sheets of phyllosilicate clay minerals. In soil, gibbsite forms thick, platy crystals. In saprolites, feldspars (plagioclase) may transform directly into gibbsite, while maintaining the original shape of the feldspar crystals (pseudomorphs).

Gibbsite is found in Oxisols on old, stable upland landscapes in association with Fe oxides (goethite, hematite) and the aluminosilicate kaolinite. Intense weathering and leaching under humid tropical and subtropical climate over long periods of time (millions of years) have depleted these soils with respect to alkali metals, earth alkali metals, and Si, thereby enhancing the formation of gibbsite. Gibbsite contents of Oxisols may be as high as 600 g kg^{-1}, but some Oxisols do not contain gibbsite at all. In Ultisols and Inceptisols of tropic to temperate climates, gibbsite is derived from the underlying saprolite, which has formed by weathering of igneous and metamorphic rocks. The gibbsite content generally decreases from the saprolite to the soil surface. Finally, gibbsite has been observed in Andisols, where it formed from rapidly weathering volcanic ash. Although these soils are often very young (hundreds of years) and Si-rich, sufficient Si was leached locally to explain the formation of gibbsite.

Gibbsite formation is conditioned by the intensity of leaching which, in turn, is affected by rainfall, temperature, parent rock, topography, ground-water table, vegetation, and time. Environments with warm temperatures, high rainfall, and free drainage favor desilication and leaching of ions, as well as mineralization of organic matter. Anions such as sulfate, carbonate, phosphate, and silicate, as well as organic ligands such as citric, malic, tannic, aspartic, and fulvic acid, which have a strong affinity for Al^{3+}, may interfere with the crystallization of $Al(OH)_3$. This explains why there is no or only small amounts of gibbsite in the A horizon of acidic soils high in organic matter and exchangeable Al, while it is found in greater quantities in deeper horizons. However, gibbsite may also decrease with soil depth. A possible explanation is an inverse gradient of $Si(OH)_4^o$ (silica) activity: rain water infiltrating at the soil surface is not yet in equilibrium with silicates and therefore has a lower silica activity as compared to the soil water at a greater depth, which had more time to equilibrate. The heterogeneous distribution of gibbsite in a landscape may be similarly explained by the silica activity of percolating waters.

Table 3 Aluminum oxide minerals

Mineral	*Formula*	*Figure*
Gibbsite[a]	γ-$Al(OH)_3$	2k
Nordstrandite	$Al(OH)_3$	2k
Bayerite	$Al(OH)_3$	2k
Doyleite	$Al(OH)_3$	2k
Diaspore	α-AlOOH	2a
Boehmite	γ-AlOOH	2c
Corundum	α-Al_2O_3	2b

[a]Minerals commonly occurring in soil.

Rare Al Oxide Minerals

The difference between gibbsite and its other polymorphs, bayerite, nordstrandite, and doyleite,

consists of a slightly different arrangement of hydroxyl groups. While gibbsite formation is favored by slow hydrolysis of Al and by a pH <6, nordstrandite and bayerite form at neutral to alkaline pH under fast hydrolysis. Correspondingly, their rare occurrences are related to limestone materials. Boehmite (γ-AlOOH), which is isostructural with lepidocrocite (Figure 2c), has been identified in lateritic materials and in bauxites. It may form from gibbsite by diagenesis or hydrothermal alteration. The oxyhydroxide diaspore (α-AlOOH) is isostructural with goethite (Figure 2a). It has been identified as a surface-weathering product formed by the desilication of kaolinitic clay. The rarely found oxide corundum (α-Al_2O_3), which is isostructural with hematite (Figure 2b), may be derived from corundum-containing parent rock or form by heating of the soil by bush fires.

Al Hydrolysis and Poorly Crystalline Al Hydroxides

Al^{3+} cations released from Al-bearing minerals are initially coordinated with six water molecules, forming the hexaquo complex $Al(H_2O)_6^{3+}$. This hexaquo complex hydrolyses in several steps, the first and most important one in soil solution being:

$$[Al(H_2O)_6]^{3+} + H_2O \Leftrightarrow [Al(OH)(H_2O)_5]^{2+} + H_3O^+$$

$$pK_s = 5.0$$

At higher Al^{3+} or OH^- concentrations, the mononuclear complexes start to polymerize to form polynuclear species. The smallest unit identified so far is $AlO_4Al_{12}(OH)_{24}(H_2O)_{12}^{7+}$. This so-called Al_{13} or Keggin polymer consists of a highly symmetrical, tetrahedrally coordinated Al center enclosed in a cage-like structure of 12 octahedrally coordinated Al centers. Further polymerization of this colloidal species leads to the precipitation of a poorly crystalline solid phase.

Si Oxides

The Si oxides differ from the oxides of Fe, Mn, and Al in several aspects. The most prevalent Si oxide, quartz (SiO_2), is a primary mineral in many rocks, contributing approx. 120 g kg^{-1} to the average composition of the Earth's upper mantle. In soils, dissolved Si associates with Al to form aluminosilicate clay minerals. However, secondary Si oxides precipitates may also form in soil, sometimes to such an extent that they cement the soil matrix to form duricrusts and silcretes. Primary quartz crystals are very pure, hard, and recalcitrant, and occur in soil as large particles (sand). Their small surface area, their small amount of reactive bonds at the surface, and the lack of isomorphic substitution make them chemically very inert. Quartz has the lowest ion exchange capacity of all soil minerals. The solubility of Si oxides is low at usual soil pH (2–8), but strongly increases at pH >8. The main soluble species, $Si(OH)_4^o$ (monosilicic acid), is taken up by higher plants, especially grasses, to increase the mechanical strength and the resistance against fungal pathogens and herbivores, and is later recycled back into the soil as biogenic opal. Radiolarians, diatoms, silica sponges, and cyanobacteria are known to accumulate large amounts of Si oxides in fresh and marine waters, and microorganisms may exert a similar function in soils, competing with the inorganic formation of aluminosilicate clay minerals.

The Si oxides are tectosilicates. The repeating unit is a SiO_4 tetrahedron in which each O is linked to Si atoms of adjacent tetrahedra, forming a 3D framework structure. This contrasts with the Fe, Al, Mn, and Ti oxides, in which the basic unit is a cation in octahedral coordination. In the following sections, only α-quartz, α-cristobalite, and opal, which commonly occur in soil, will be described (Table 4).

Table 4 Silicon oxide minerals

Mineral	*Formula*	*Structural type*	*Occurrence*	*Figure*
α-Quartz[a]	SiO_2		Ubiquitous	2m
α-Tridymite	SiO_2			–
α-Cristobalite[a]	SiO_2		Volcanic deposits	2n
Opal-C[a]	$SiO_2 \cdot x H_2O$	Well-ordered α-crystobalite	Volcanic deposits	–
Opal-CT[a]	$SiO_2 \cdot x H_2O$	Disordered α-crystobalite and α-tridymite	Cherts, porcelanites, fossil wood, silcretes	–
Opal-A[a]	$SiO_2 \cdot x H_2O$	Nearly amorphous	Biogenic ubiquitous; inorganic in strongly weathered soils, duripans, silcretes	–

[a]Minerals commonly occurring in soil.

Quartz and Cristobalite

α-Quartz consists of paired helical chains of corner sharing SiO_4 tetrahedra that spiral along the *c*-axis. The intertwined chains produce open, hexagonal channels (Figure 2m). In cristobalite, the tunnels are larger (Figure 2n), and the density is smaller ($2.32\,g\,cm^{-3}$) than quartz ($2.65\,g\,cm^{-3}$).

Quartz is ubiquitous in soil environments, while cristobalite is restricted to soils developed from volcanic materials. The quartz in soils is mainly a primary mineral, inherited from the parent material. In soil solution, quartz may form from soluble $Si(OH)_4^o$ by polymerization, precipitation of amorphous, hydrous SiO_2, which slowly transforms into crystalline quartz. Direct precipitation from solution seems to be unlikely. Quartz is generally concentrated in the sand and silt fractions of soils, with a lower frequency in the coarse-clay fraction (0.2–2 μm). Quartz may comprise more than 90% of the inorganic fraction of Quartzipsamments. By contrast, highly weathered and leached Oxisols have only very low amounts of quartz left. The solubility of quartz ranges from 1 to $4\,mg\,l^{-1}$ Si depending on particle size.

Opal

Opal is classified into three structural groups, opal-C (well-ordered α-cristobalite), opal-CT (disordered α-cristobalite, α-tridymite) and opal-A (highly disordered, nearly amorphous). Opal contains 40–90 $g\,kg^{-1}$ water, and the specific gravity ranges from 1.5 to $2.3\,kg\,dm^{-3}$.

Opal-C and α-cristobalite occur in Andepts and other soils derived from volcanic material. Opal-CT has been identified in soils derived from bentonites, siliceous shales, and indurated silicates. Opal-CT is also a primary constituent of many cherts, porcelanites, fossil wood, silcretes, and some duripans. Opal-A may form by both organic and inorganic processes in pedogenic environments. Biogenic opal-A originates from Si accumulated by plants and aquatic organisms, and thus occurs under a wide range of environmental conditions. Biogenic opal is a nearly ubiquitous constituent of soils. Opal phytoliths are mostly derived from grasses, while sponge spicules, diatoms, and radiolaria, which form in aquatic environments, may reach soils via weathering of sediment bedrocks. Amounts of opal phytoliths in soils commonly range from less than 1 to $30\,g\,kg^{-1}$, usually occurring in the 5- to 50-μm size fraction, and decreasing with soil depth. Inorganic opal-A forms from supersaturated soil solution. It occurs as nodules and as the primary cement of indurated soil horizons (duripans) and silcretes.

At ambient temperature and neutral pH, the solubility of amorphous Si is approximately 50–60 mg $Si\,l^{-1}$, and therefore much higher than that of quartz. The presence of organic molecules greatly enhances the dissolution rates of Si oxides. Coatings of chemisorbed Al and Fe, however, reduce the dissolution rates. Oxides of Fe and Al, by acting as a sink for soluble Si, greatly increase the dissolution rates of amorphous Si oxides and of reactive, uncoated quartz surfaces. Dissolution (weathering) of Si oxides, which starts below approximately 3 mg $Si\,l^{-1}$, is induced by the reduction of Si levels in soil solution by leaching and plant uptake.

Ti Oxides

Titanium oxides are common in many igneous and metamorphic rocks and sediments. Due to their resistance against weathering, they are often inherited to soils. However, Ti oxides also form in soil after weathering of less-resistant, Ti-bearing minerals (biotite, amphibole). Titanium oxide minerals are relatively heavy (density more than $2.9\,g\,cm^{-3}$). Their concentration in soils is usually low; because of their small surface area, they do not contribute significantly to soil reactivity. Only in tropical soils, where they become increasingly enriched during pedogenesis, has some influence on phosphate and arsenate retention been observed. Old sand dunes may contain enough Ti oxides to be mined as a commercial source for TiO_2, which is used to produce white paint. Due to their recalcitrant nature, Ti oxides can be used as reference minerals to study weathering and soil genesis.

Titanium occurs primarily in octahedral coordination, and like the Fe and Mn oxides, the structures of the various Ti oxide minerals can be described in terms of the different arrangement of the Ti-containing octahedra. The principal Ti minerals are listed in Table 5. Rutile (TiO_2) is isostructural with pyrolusite and manganite, consisting of single chains of edge-sharing TiO_6 octahedra (Figure 2j). Rutile is a common residual mineral occurring in the sand and silt fractions of a variety of soils.

Table 5 Titanium oxide minerals

Mineral	*Formula*	*Figure*
Rutile[a]	TiO_2	2j
Anatase[a]	TiO_2	2l
Ilmenite[a]	$Fe^{2+}\,TiO_3$	2b
Pseudorutile[a]	$Fe_2O_3 \cdot nTiO_2 \cdot mH_2O$ ($n = 3$–5, $m = 1$–2)	–
Ulvöspinel	Fe_2TiO_4	2d
Titanomagnetite	$Fe_{3-x}\,Ti_xO_4$	2d
Titanomaghemite	$Fe_{2-x}\,Ti_xO_3$	2d

[a]Minerals commonly occurring in soil.

Anatase consists of edge- and corner-sharing octahedra that outline a three-dimensional framework, rather than distinct chains (Figure 2l). The platy crystals found in soils are believed to be a weathering product of titanite ($CaTiSiO_5$) and Ti-bearing silicates. Anatase is easily synthesized at room temperature. Ilmenite ($FeTiO_3$) is almost isostructural with hematite, with one-half of the octahedral centers occupied by Fe^{2+}, the other half by Ti^{4+} (Figure 2b). By oxidation of structural Fe^{2+} to Fe^{3+}, ilmenite weathers to pseudorutile ($Fe_2O_3 \cdot n\,TiO_2 \cdot mH_2O$; $3 < n < 5$ and $1 < m < 2$), a structurally disordered and poorly characterized mineral. Pseudorutile has been widely identified in soils, primarily as the weathering product of ilmenite. Ilmenite itself is relatively unstable in soils, because it easily weathers to pseudorutile and mixtures of rutile, anatase, and Fe oxides.

Titanomagnetites ($Fe_{3-x}\,Ti_xO_4$) are solid solutions of magnetite with ulvöspinel (Fe_2TiO_4), with Ti^{4+} occupying only octahedral sites of the magnetite structure (Figure 2d). Their formation is probably related to that of magnetite. Titanomaghemites ($Fe_{2-x}\,Ti_xO_3$) form by weathering of titanomagnetites. Their composition is between those of magnetite, maghemite and ulvöspinel. Many aspects of the Ti oxides, such as whether the minerals are relict or pedogenic, and the conditions responsible for their weathering and formation in soils, are still unresolved.

Further Reading

Banfield JF and Nealson KH (eds) (1997) Geomicrobiology: interactions between microbes and minerals. In: *Reviews in Mineralogy,* vol. 35. Washington, DC: Mineralogical Society of America.

Bigham JM, Fitzpatrick RW, and Schulze DG (2002) Iron oxides. In: Dixon JB and Schulze DG (eds) *Soil Mineralogy with Environmental Application*pp. 323–366. Madison, WI: Soil Science Society of America.

Cornell RM and Schwertmann U (2003) *The Iron Oxides: Structure, Properties, Reactions, Occurrence and Uses*, 2nd edn. Weinheim, Germany: VCH Verlagsgesellschaft.

Dixon JB and Schulze DB (2002) *Soil Mineralogy with Environmental Applications*. Madison, WI: Soil Science Society of America.

Dixon JB and Weed SB (1989) *Minerals in Soil Environments*. Madison, WI: Soil Science Society of America.

Heaney PJ, Prewitt CT, and Gibbs GV (1994) Silica: physical behavior, geochemistry and materials applications. In: Haney PJ (ed.) *Reviews in Mineralogy,* vol. 29, pp. 375–443. Washington, DC: Mineralogical Society of America.

Kämpf N, Scheinost AC, and Schulze DG (2000) Oxide minerals in soils. In: Sumner ME (ed.) *Handbook of Soil Science*pp. 125–168. Boca Raton, FL: CRC Press.

Waychunas GA (1991) Crystal chemistry of oxides and oxyhydroxides. In: Lindsley DH (ed.) *Reviews in Mineralogy,* vol. 25, pp. 11–68. Blacksburg, VA: Mineralogical Society of America.

METALS AND METALLOIDS, TRANSFORMATION BY MICROORGANISMS

S M Glasauer and T J Beveridge, University of Guelph, Guelph, ON, Canada

E P Burford, F A Harper, and G M Gadd, University of Dundee, Dundee, UK

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Fungi and bacteria from all major taxonomic groups have been found in metal-rich habitats such as mineral ores, acidic soils, or polluted environments. The survival of these microorganisms may depend on a number of biochemical, structural, and physiological properties that modify metal bioavailability and toxicity, which are constitutive to a species and/or have been selected by adaptation.

All microorganisms have essential metabolic and structural requirements for specific metals in metabolism, including K, Na, Mg, Ca, Mn, Fe, Cu, Zn, Co, and Ni, while some metals, e.g., Rb, Cs, Al, Cd, Ag, Au, Hg, and Pb, and metalloids, e.g., As and Te, have no apparent essential function. Intense competition for some metals, coupled with low bioavailability, has driven the evolution of tightly regulated mechanisms for metal acquisition. These include highly specialized, metal-specific membrane transport systems and metal chelates to scavenge metals effectively from the environment. All metals and metalloids, regardless of their metabolic function, are, however, toxic in excess and can inhibit growth, affect sporulation and adhesion, induce morphological changes, and affect biochemical activities. This potential toxicity has led cells to develop diverse mechanisms of defense. Both genetically coded mechanisms of resistance and nonspecific mechanisms such as oxidation or reduction of the metals to

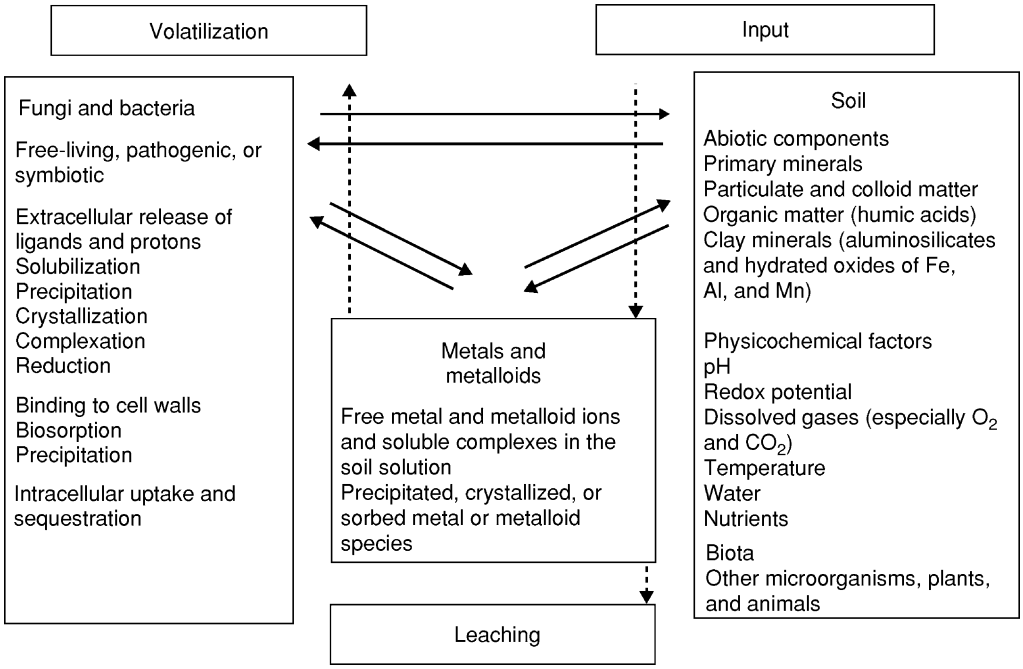


Figure 1 Fungal interactions with metals and metalloids in the soil environment. Metal(loid) species lost from the soil environment are indicated by broken arrows. Interactions mediated by microorganisms are indicated by solid arrows.

produce less toxic chemical species, or sorption and precipitation on cell surfaces, protect cells from toxic metals.

The chemical form, or species, of metals is also affected by microbial processes other than those related to nutrition and toxicity. Metal speciation is important because different metal species can have widely differing chemical and biological properties, including mobility in the environment and toxicity. Excreted cellular metabolites transform metal compounds by nonspecific reactions such as dissolution by protons and organic acids. Some bacteria use metals as electron donors and acceptors during respiration, a process that is a major pathway for the cycling of metals in soils and sediments. Such microbial transformations both mobilize and precipitate metals, leading to alteration and modification of soils and sediments. Many of these processes also have potential for the bioremediation of metal-contaminated sites.

Microbially Induced Metal Transformations

Interactions between microorganisms with metals and soil components take many forms; unravelling their complexity is integral to understanding metal speciation in the soil. The relative role of fungi and bacteria depends on associative biota (other microorganisms, plants, and animals) as well as on abiotic factors such as physicochemical components of the environmental matrix, e.g., pH, water, inorganic and organic ions, molecules, compounds, colloids, and particulates (Figure 1).

The ability of microorganisms to affect metal speciation in the soil environment stems from two opposing phenomena: mobilization and immobilization, which influence the distribution of metal species between soluble and insoluble phases. The solubilization of metals can be achieved by various mechanisms: protonation, chelation, and chemical transformation. Metal immobilization can occur by precipitation or crystallization of insoluble organic or inorganic compounds, or by sorption, uptake, and intracellular sequestration. In addition, redox reactions can mobilize or immobilize metals, depending on the metal species.

Mobilization

Fungi and bacteria can facilitate the transformation of metals into soluble forms via the dissolution of metal compounds, including oxides, phosphates, sulfides, and more complex mineral ores, or by desorption from exchange sites on clay minerals or organic matter. Metal leaching by Fe- and S-oxidizing bacteria is discussed further on. A number of other mechanisms by which microbes can leach metal ions have been proposed.

Acidification

Competition between protons and the metal (in a metal–anion complex) leads to protonation of the

anion under conditions of increased acidity, and results in the release of free metal cations. Fungi and bacteria can acidify their environment in a number of ways; for example, proton efflux via plasma membrane H^+-ATPases, maintenance of charge balance, or as a result of respiratory carbon dioxide accumulation.

Complexation by Ligands

Microorganisms produce a number of extracellular metabolites that can complex metals in solution, including polysaccharides, pigments, organic acids, and siderophores. Organic acids are important in energy production as intermediates in the tricarboxylic (TCA) cycle and in cellular pH homeostasis. They can correct internal perturbations in charge balance and pH resulting from nutrient uptake by their ability to produce and consume protons via the formation and removal of carboxylic groups, and by efflux from the cell.

Citric and oxalic acids are the most commonly reported organic acids in soils and are released into the soil by fungal hyphae, lichens (a symbiotic relationship between fungi and a photosynthetic cyanobacterium or alga), bacteria, and plant rootlets (mycorrhizae). Organic acids are believed to have an essential role in nutrient acquisition as a result of their acidification and metal-binding properties. For example, phosphate is released from common soil compounds such as $CaHPO_4$, $Ca_3(PO_4)_2$, $FePO_4.4H_2O$, and $AlPO_4$, either by direct exchange of chemically binding species, or ligands, between phosphate and the organic anion or by binding of the metal to the anion. Metal complexation depends on metal-to-anion availability, the stability of the particular metal–ligand complex, and pH. Anion concentration (which will depend on the pK_a of the acid) is important, because fully protonated organic acids chelate metals poorly. The affinity depends on the stability constant of the organic acid–metal complex. For example, the efficiency of organic acids to solubilize phosphate from soils often follows the order: citrate > oxalate > malate > acetate. The bioavailability of a solubilized metal–ligand complex will depend ultimately on its long-term fate in the environment, including sorption on to components of the soil matrix or potential utilization by biota. For example, metal citrates can be highly resistant to microbial degradation, which can result in effective leaching of metals from soil. In contrast, metal oxalates crystallize.

Many soil bacteria, fungi, and lichens are particularly efficient at solubilizing metal phosphates and other metal-bearing compounds, which has been attributed to their ability to produce protons and organic ligands. This can be tested by simple laboratory experiments: if insoluble metal compounds are incorporated into solid growth medium, clear zones may appear around microbial colonies if they are capable of mineral dissolution.

In addition, organic acid production by microbes is an important agent of mineral deterioration, playing a role in both biogenic chemical weathering and soil formation. Fungi can also physically weather minerals by etching and fracturing mineral grains during hyphal growth.

Siderophores

Siderophores are the largest class of known compounds that can bind and transport, or shuttle, Fe. They are highly specific Fe(III) ligands (formation constants often greater than 10^{30}). These low-molecular-weight coordination molecules are excreted by a wide variety of fungi and bacteria to aid Fe assimilation. Although the mechanism could be used to acquire other metals, Fe is the only known essential element for which these specific organic shuttles operate. This is most likely because Fe is needed in larger amounts by cells than other poorly soluble metals, and, given the low solubility-product constant of ferric hydroxide (less than 10^{-38}), the concentration of free Fe^{3+} is too low to support microbial growth at pH values where most life exists.

In a typical soil, many different species, including plants, compete for Fe via siderophores. Organisms have most likely evolved mechanisms to ensure that Fe demand is met through the production of species-specific siderophores, or attachment to solid mineral Fe (e.g., Fe oxides) to shorten the pathway between the Fe substrate and the site of uptake.

Biomethylation

Methylation of Hg, As, Se, Sn, Te, and Pb is mediated by a range of microorganisms, including clostridia, methanogens, and sulfate-reducing bacteria under anaerobic conditions, and principally fungi under aerobic conditions, such as *Penicillium, Cephalosporium*, and *Alternaria* spp. as well as a variety of bacteria, including pseudomonads. Methyl groups are enzymatically transferred to the metal, and a given species may transform a number of different metal (loid)s. Methylated metal compounds formed by these processes differ in their solubility, volatility, and toxicity. Volatile methylated species, e.g., $(CH_3)_2Se$ and $(CH_3)_2Se_2$, are often lost from the soil.

Redox Reactions

Microorganisms mobilize metals, metalloids, and organometallic compounds by reduction and oxidation processes, exemplified by pathways for metal

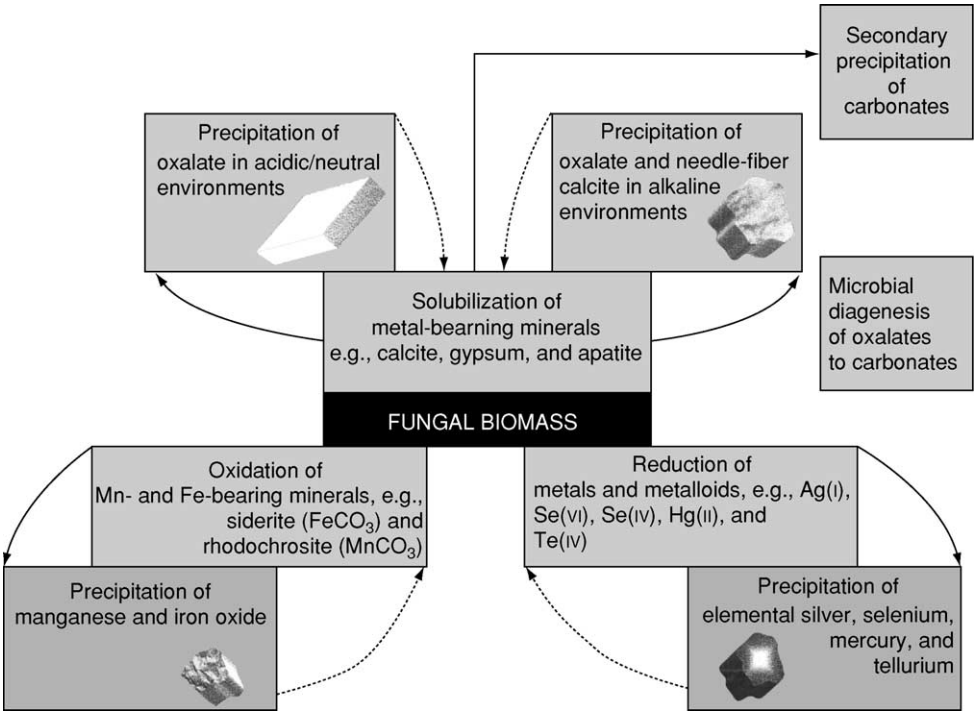


Figure 2 Metal precipitation by microorganisms. Solid arrows represent precipitation by fungi and bacteria. Broken arrows represent precipitation and association with microbial biomass (living or dead).

cycling by fungi (Figure 2). For example, oxidation of metal-complexing dimethylsulfide, dimethylsulfoxide, or thiosulfate increases metal availability if metal sulfates are formed, and the solubilities of Fe and Mn increase upon reduction of Fe(III) to Fe(II) and Mn(IV) to Mn(II). While many microorganisms reduce small amounts of Fe during metabolism, most Fe reduction is carried out by specialized anaerobic bacteria such as *Shewanella* spp. and *Geobacter* spp. that use Fe for respiration, and there is evidence that reduction of Fe(III) was one of the first forms of microbial respiration. In this process, oxidized forms of metals are used as the final electron acceptor in place of the oxygen consumed during aerobic respiration. The dissimilatory metal-reducing bacteria respire and thereby reduce oxidized metal species at or near the outer cell wall. The membrane-bound electron transport system provides energy to cells by cycling electrons through an electrochemical cascade of metalloenzymes, expelling e^- and H^+ from the cell. Energy is generated by the flux of H^+ into the cell through specific membrane-bound channels (Figure 3). Dissimilatory metal-reducing bacteria can use a variety of metal(loid)s with an appropriate redox couple, including Fe(III), Mn(IV), Se(IV), Cr(VI), and U(VI). While Fe and Mn increase their solubility upon reduction, the solubility of other metals such as U(VI) to U(IV) and Cr(VI) to Cr(III) decreases, resulting in

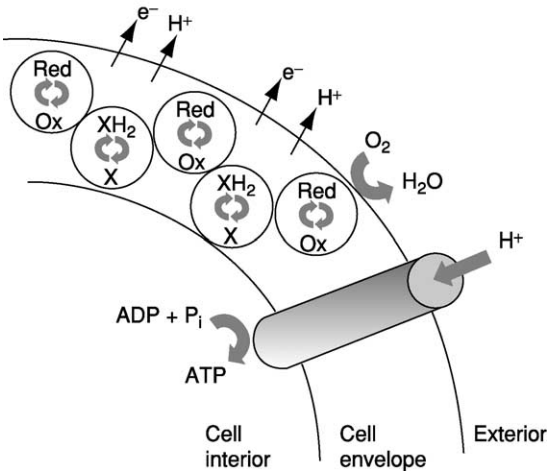


Figure 3 Simplified bacterial electron transport chain, showing linked carriers in the cell wall that cycle electrons down an electrochemical gradient to the final electron acceptor (in this case O_2). Energy is generated by the flux of protons through specific channels across the membrane. Red, reduction; Ox, oxidation; P_i, inorganic phosphorus; ADP, adenosine diphosphate; ATP, adenosine triphosphate.

immobilization (see Precipitation of Metals and Minerals, below). Reduction of Hg(II) to Hg(0) by bacteria and fungi results in diffusion of elemental Hg out of cells, and conferment of Hg resistance on organisms that possess this property since Hg(0) is considerably less toxic than Hg(II).

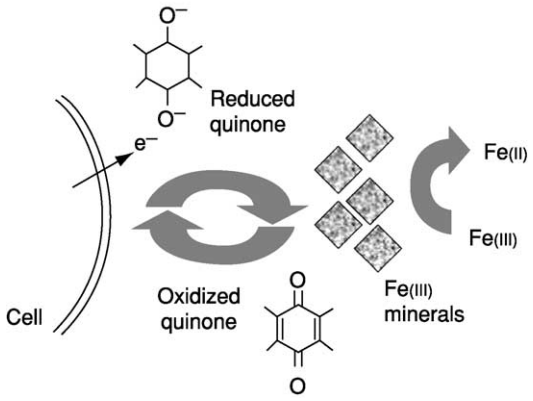


Figure 4 Proposed mechanism for electron transfer from metal-reducing cells to oxidized Fe minerals via chemical reduction of a quinone shuttle. The shuttle regenerates when Fe(III) is reduced to Fe(II).

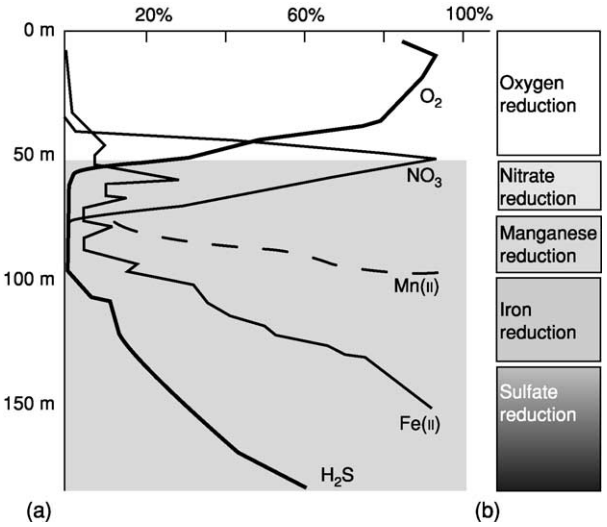


Figure 5 Reductive processes in the environment: (a) representative electron acceptor profiles (percentage of maximum) with water depth in meters. Data are from the Black Sea; (b) major processes occurring in the redox zone with depth. (Reproduced with permission from Nealson KH and Saffarini D (1994) Iron and manganese in anaerobic respiration: environmental significance, physiology and regulation. *Annual Review of Microbiology* 48: 311–43. © Annual Reviews, Inc.)

Dissimilatory metal reduction requires that bacteria access relatively large amounts of insoluble metals. Bacteria can use redox-active organic compounds to shuttle electrons between cells and the metal being respired under anaerobic conditions. Many quinoid compounds, common components of humic substances, can function as electron shuttles. These compounds move electrons from the cell membrane to the target metal, such as iron contained in ferric oxides. The shuttle molecule regenerates when the metal is reduced and the cycle continues (Figure 4). While the use of electron shuttle mechanisms appears to be widespread among bacteria, the impact and ecologic significance of the process in natural subsurface environments are not fully understood.

In the absence of processes that mix water columns, sediments, and soils, redox potentials tend to decrease with depth, reflecting a change in the dominant electron acceptors from O_2 to NO_3^-, Mn, Fe, and (finally) S. The interplay between microbial and abiotic metal reduction forms characteristic element profiles (Figure 5). *Shewanella*, *Geobacter*, and *Desulfovibrio* spp. have been intensively studied for their metal-reducing capability. Molecular studies reveal that many as yet uncultured species exist in the subsurface which can also contribute, and improved culturing methods are likely to show novel dissimilatory metal reducers amongst these organisms. The poorly crystalline iron fraction is generally considered to be the 'reactive' fraction for these bacteria; however, iron in some clay minerals and crystalline Fe oxides could be available for reduction.

Immobilization

There are a number of processes whereby microbial biomass can immobilize metals in the soil environment. Although immobilization reduces the external free metal activity, it may also shift the equilibrium to release more metal into the soil solution.

Biosorption

The strong interaction of metals with cell surfaces stems from cell wall characteristics, including component pigments and polysaccharides. Bacterial cells depend on the diffusion of nutrients for survival, and their small size, shape, and characteristic membrane structures have evolved to maximize their interaction with nutrient solutes. There is remarkable similarity in the general format of bacterial wall structure, which falls into two broad classes based on reaction with Gram's stain, which utilizes crystal violet. Gram-positive walls consist of a thick matrix of peptidoglycan and secondary polymers, which provides structural support and gives the cell its shape and form (Figure 6). The phosphate and carboxyl groups within the wall fabric and exposed at the surface confer pH-dependent negative charge to the cell in the types of environments where microbes generally exist. The point of zero charge (PZC), or pH where positive charge equals negative charge, is approximately pH 2–3 for bacterial cell walls. This charge provides the basis for metal sorption. At pH values above the PZC, the net charge on the cell wall is negative, which facilitates sorption of positively charged metal species. Gram-negative cells are also negatively charged, although structurally more

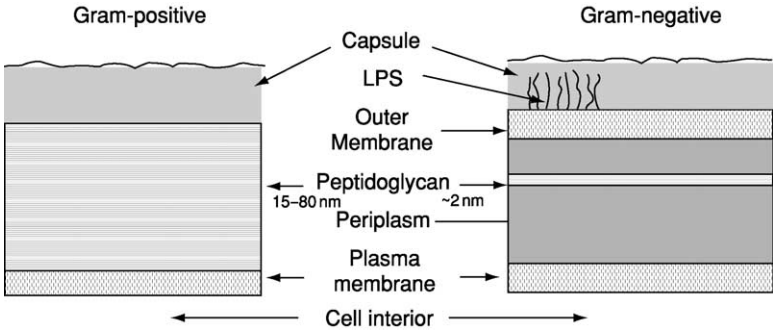


Figure 6 Cross-section view showing the cell wall structures characteristic for Gram-positive (left) and Gram-negative (right) bacteria. LPS, lipopolysaccharide.

complex. In this case, a thin layer of peptidoglycan floats in a gel-like periplasm, sandwiched between the plasma membrane and the outer membrane (Figure 6). The outer membrane has characteristic surface lipopolysaccharides (LPS) containing ionizable phosphoryl, carboxyl, hydroxyl, and amine groups; the phosphoryl groups are the most reactive to metals. In addition to these basic structures, surface capsules, sheaths, and S-layers provide additional sites for metal interaction and the heterogeneous nucleation of minerals.

Fungal cell walls are predominantly composed of polysaccharides (80–90%) with glycoproteins and some lipids. The polysaccharide component includes microcrystalline fibrils of β-linked polysaccharides (chitin, chitosan, and β-glucans) that are responsible for wall strength and rigidity. As for bacteria, the chemical properties of the functional groups associated with fungal walls, including carboxyl, amine, hydroxyl, phosphoryl, and sulfhydryl groups, provide the basis for the attraction of metals to cell walls. In addition, pigments such as melanin contribute to rigidity and have significant metal-binding properties, preventing metal entry into melanized structures.

Precipitation of Metals and Minerals

Partitioning in the cellular environment Metals are found in all parts of bacterial and fungal cells: in the cytoplasm, the cell wall, or sorbed to the cell surface. At sufficient concentrations, metal precipitates begin to develop or nucleate on or within cells, or in the immediate extracellular environment (demonstrated for a bacterial cell in Figure 7). In addition to immobilizing metals, this also reduces the external free metal activity and may shift the equilibrium to release more soluble metal to the soil. Precipitation and crystallization restrict metal leaching and bioavailability, and can potentially release nutrients such as sulfate and phosphate.

The formation of extracellular and sorbed precipitates depends on the chemical composition of the

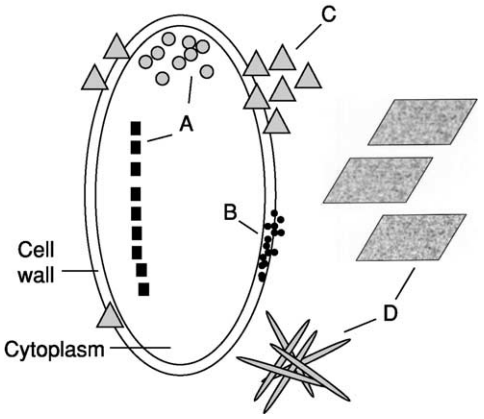


Figure 7 Possible locations of metal precipitates and minerals associated with a bacterial cell: A, biogenic precipitates within the cell cytoplasm; B, metal precipitates formed by nucleation in the cell wall, e.g., periplasmic; C, mineral nanocrystals formed independently of cell activities and subsequently sorbed to the cell wall; D, minerals formed outside cells due to the influence of cellular activities.

extracellular environment and is often mediated by cellular processes. The precipitates can be copious and may completely enmesh the cells, effectively mineralizing them (Figure 8). Environmental anions, such as OH^-, SO_4^{2-}, PO_4^{3-}, and CO_3^{2-}, provide counterions for mineral hydroxides, sulfate and sulfides, phosphates, carbonates, and oxalates. It is possible that cellular metabolism and excretion of such anions also contributes to these minerals. Metal oxide and hydroxide minerals formed by abiotic reactions also attach to cells, and it can be difficult to distinguish these phenomena (Figure 9). This is, however, another important mechanism by which cells can concentrate metals.

Inside cells, metals can be built into cellular constituents. Excess or toxic metals may be effluxed, form intracellular precipitates, bind to metal-complexing molecules, and, in eukaryotes, partition into vacuoles. The formation of internal precipitates often results from a change in metal oxidation state, which does not require an active cell; for example, gold is

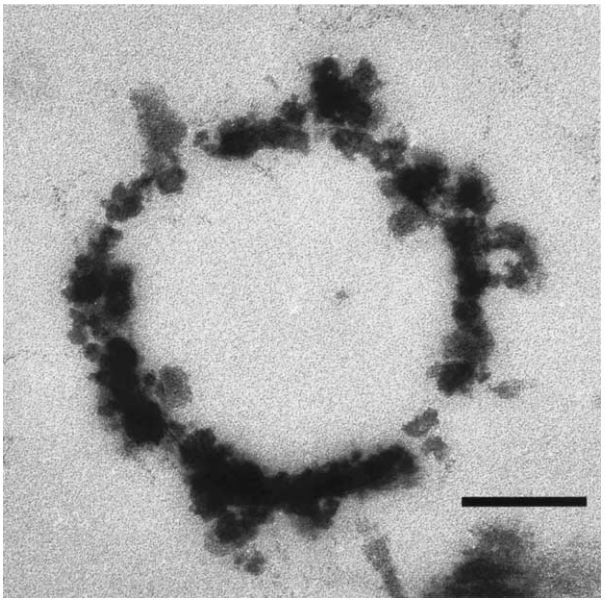


Figure 8 Transmission electron micrograph of a thin-section sample taken from a Yellowstone hot spring, showing an unknown bacterial cell embedded in metal-rich precipitates. Scale bar 150 nm.

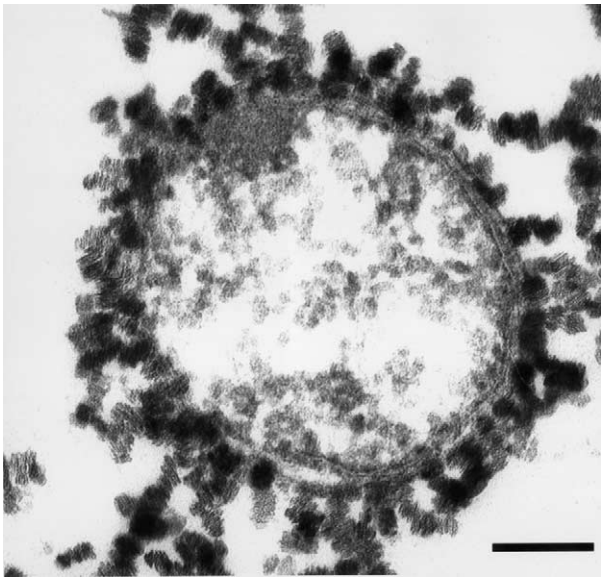


Figure 9 Transmission electron micrograph of a thin section showing abiotically precipitated goethite nanocrystals sorbed to the cell wall of *Shewanella putrefaciens*. Scale bar 150 nm.

highly toxic to cells, yet intracellular precipitates of elemental gold form in bacteria after cells are killed. The formation of oriented chains of magnetite (Fe_3O_4) crystals by magnetotactic bacteria, found in sediments and hydric soils, is a unique example of directed, intracellular mineral precipitation. In this case, the formation of mineral precipitates is a mechanism that enables the bacteria to orient in the Earth's magnetic field.

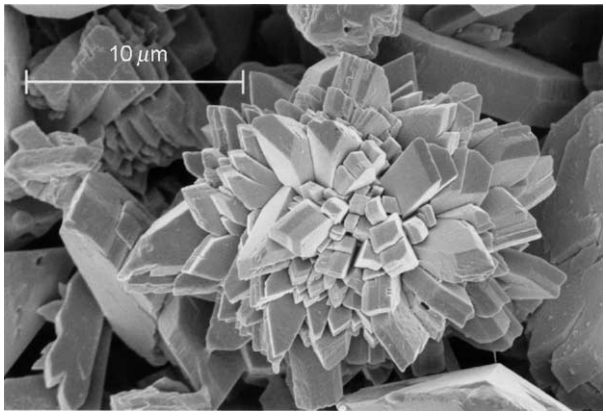


Figure 10 Calcium oxalate precipitated by *Aspergillus niger*. Scale bar 10 μm.

Oxalates Calcium oxalate is the most common form of oxalate associated with soils and leaf litter, occurring as dihydrate (weddellite) or the more stable monohydrate (whewellite). Calcium oxalate crystals are commonly associated with free-living, pathogenic, and plant-symbiotic fungi and are formed by the reprecipitation of solubilized calcium as the oxalate. Fungal-derived calcium oxalate can exhibit a variety of crystalline forms (tetragonal, bipyramidal, plate-like, rhombohedral, or needles), which often associate with the outer surfaces of hyphae (Figure 10). The formation of calcium oxalate by fungi has an important influence on biological and geochemical processes in soils, acting as a reservoir for calcium in the ecosystem, and also influencing phosphate availability.

Fungi can also produce other metal oxalates with a variety of different metals and metal-bearing minerals, e.g., Cd, Co, Cu, Mn, Sr, and Zn. The formation of metal oxalates may provide a mechanism whereby fungi can tolerate environments containing potentially high concentrations of toxic metals. A similar mechanism occurs in lichens growing on copper sulfide-bearing rocks, where precipitation of copper oxalate occurs within the thallus.

Carbonates In many arid and semiarid regions, calcareous soils and near-surface limestones (calcretes) are often secondarily cemented with calcite ($CaCO_3$). This phenomenon has been partly attributed to physicochemical processes; however, the abundance of calcified fungal filaments in weathered profiles of chalky limestone and Quaternary calcretes indicates fungal activity. Calcite precipitation on fungal hyphae has also been observed with the ectomycorrhizal fungus *Pisolithus tinctorius* in calcareous soil.

Calcite has been proposed to form indirectly through fungal excretion of oxalic acid and precipitation of calcium oxalate, which results in dissolution

of the internal pore walls of the limestone matrix. This enriches the matrix solution in free carbonate ions. During passage of the solution through the pore walls, calcium carbonate recrystallizes as a result of a decrease in CO_2, hardening the limestone. Calcite crystals act as sites of further secondary calcite precipitation when the hyphae decompose. Through these processes, fungal activity is important in lithification (hardening), alteration, and diagenesis of subsurface limestone formations.

Mineralized carbonate precipitates are also found in association with bacterial biofilms. One striking example of microbial carbonate precipitation is the towering microbialites at Lake Van (Turkey), formed by thick biofilms of cyanobacteria (*Pleurocapsa* spp.) that colonize crusts of inorganic calcite precipitated in the alkaline soda lake water. Groundwater forced up and through the carbonate matrix raises the towers of aragonite and calcite to heights up to 40 m.

Redox Reactions

Bacterial and fungal oxidation The range of bacterial species with the capability to oxidize Fe and Mn spans sedimentary and soil environments. For bacteria to use Fe(II) specifically for growth, however, they must recover the energy produced in the oxidative process. Demonstration of a respiration-linked, energy-yielding reaction is particularly difficult for iron due to its rapid chemical oxidation at neutral pH and atmospheric O_2. It has been more readily shown for bacteria that grow at low pH and/or low O_2, conditions that drastically slow chemical oxidation of Fe(II). Bacterial Fe oxidation is probably ubiquitous in environments with sufficient Fe^{2+} and conditions to support the growth of bacteria that can make a living by oxidative respiration of reduced metal species. Specific environments include drainage waters and tailing piles in mined areas, pyritic and hydric soils (bogs and sediments), drain pipes and irrigation ditches, and plant rhizosphere zones.

Iron-oxidizers found in acidic soil environments are acidophilic chemolithotrophs, organisms that use inorganic sources of chemical energy, such as *Thiobacillus ferrooxidans*, which is significant for its role in generating acid mine drainage. Facultative chemolithotrophic microaerophiles such as *Leptothrix ochracea*, *Sphaerotilus natans*, and *Gallionella ferruginea* are common in mildly acidic to neutral environments, and accumulate abundant Fe hydroxide precipitates on their filament or stalk structures that can age to more crystalline minerals.

After Fe, Mn is the second most abundant transition metal in the Earth's crust and is actively cycled in the same environments as Fe. Mn oxidation is carried out by many of the organisms that are implicated in Fe oxidation, including *Bacillus*, *Pseudomonas*, and *Leptothrix* spp. In contrast to Fe^{2+}, substantial chemical oxidation of Mn^{2+} does not occur below pH 8, and solutions containing Mn^{2+} can be stable for weeks. This fact has allowed bacterial Mn oxidation to be identified more readily than its Fe counterpart.

It is not yet clear why cells carry out metal oxidation if not for gaining energy. Oxidation may help the cell rid itself of toxic intracellular concentrations of the metal, or the precipitates that coat the bacterial surface subsequent to oxidation may protect the cells from attack by other microbes. Protection from UV radiation would have been especially crucial during the early evolution of life on Earth.

Fungi also oxidize metals in their environment. Desert varnish is an oxidized metal layer (patina) a few millimeters thick found on rocks and in soils of arid and semiarid regions, and is believed to be of fungal and bacterial origin. *Lichenothelia* (a fungus) can oxidize manganese and iron from metal-bearing minerals such as siderite ($FeCO_3$) and rhodochrosite ($MnCO_3$), and metals from rainfall or windblown dust.

Dissimilatory metal-reduction by bacteria When dissimilatory metal-reducing bacteria utilize relatively soluble metals such as Cr(VI) and U(VI) as electron acceptors, less-soluble Cr(III) and U(IV) species may result. Once reduced, soluble metals sorb to the cells or partition to the bulk extracellular solution, where new minerals can form from the metals and suitable counterions. Postreduction partitioning of metals depends on the element. For example, U(IV) accumulates as fine precipitates in the periplasm of Gram-negative metal reducers (**Figure 11**). Fe^{2+} and Mn^{2+}, generated by dissimilatory reduction, can form copious extracellular precipitates and nodules. The fate of the reduced metal depends on the environment, and it is feasible that a metal such as Fe, along with any associated redox-active metals, could be cycled hundreds of times between reducing and oxidizing zones before being incorporated into refractory minerals (**Figure 12**).

Fungi and other microorganisms also precipitate reduced forms of metal species, e.g.: Ag(I) to Ag(0); selenate (Se(VI)) and selenite (Se(IV)) to Se(0); and tellurite (Te(IV)) to Te(0), by a variety of mechanisms.

Biotechnological Applications of Microbial Metal Transformations

Ore Leaching

Metals in ores that are not worth recovering by smelting can sometimes be effectively leached by

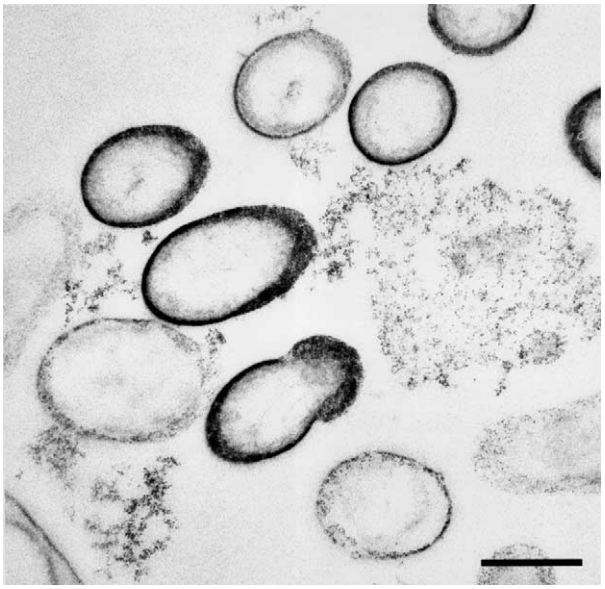


Figure 11 Transmission electron micrograph of a thin section showing uranium precipitates formed in the periplasmic space of *Geobacter metallireducens* cells during dissimilatory reduction of U(VI). No metal stains were used, so that the contrast in the image is from the uranium precipitates only. Scale bar 150 nm.

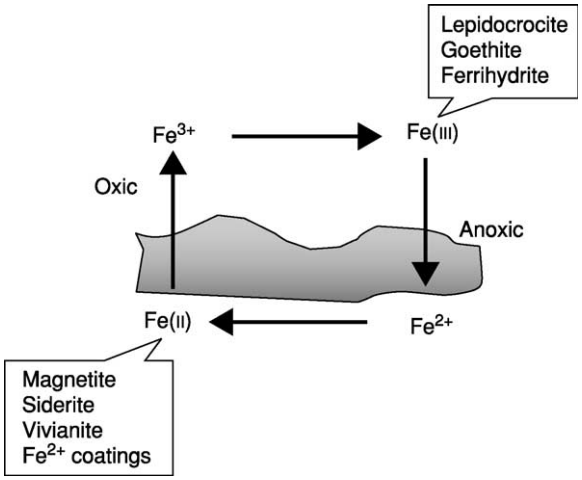


Figure 12 Cycling of Fe between soluble and insoluble mineral, oxidized and reduced forms mediated by microorganisms in soils and sediments. Although by quantity Fe is the major element to undergo redox transformations, other metals can be cycled by the same processes and associate with the precipitating minerals.

bacteria and fungi. Important species of bacteria in such processes are Fe and S oxidizers such as *Thiobacillus ferrooxidans*, *T. thiooxidans*, and *Leptospirillum ferrooxidans*. Their activities solubilize Fe and S contained in metal sulfides, with the concomitant release of associated metals such as Cu and Zn. Large amounts of Fe^{2+} and sulfuric acid are generated in the process. It has been accepted most commonly that bacteria catalyze the oxidation of Fe^{2+} to regenerate Fe^{3+}, but do not directly attack the mineral:

$$14Fe^{2+} + 3.5O_2 + 14H^+ \rightarrow 14Fe^{3+} + 7H_2O \quad \text{bacteria} \quad [1]$$

$$FeS_2 + 8H_2O + 14Fe^{3+} \rightarrow 15Fe^{2+} + 2SO_4^{2-} + 16H^+ \quad \text{abiotic} \quad [2]$$

where eqn [1] gives the reaction for bacteria, and eqn [2] is the abiotic reaction. There is, however, increasing evidence that *T. ferrooxidans* may directly enhance dissolution of metal sulfides via an enzymatic mechanism. Approximately 90 mol Fe^{2+} is oxidized to assimilate 1 mol C, which explains the tremendous amount of acidity – 'acid mine drainage' – that is generated. Industry can take advantage of the acidic leachate by recirculating it through the ore body to dissolve and harvest more metal. Here, the acidic leachate is recirculated through the ore body to solubilize more metal, and the soluble metals are then separated by sedimentation, solvents, or electrodialysis. When uncontrolled, such as at poorly managed or abandoned mining sites, the same process can have disastrous environmental consequences.

Extracellular ligands produced by fungi, including *Aspergillus* and *Penicillium* spp., have also been used to leach metals such as Zn, Cu, Ni, and Co from a variety of solid materials, including low-grade mineral ore, solid waste, and electronic scrap materials.

Bioremediation

Toxic metals in natural, industrial, and agricultural soils may be a risk to biota at all trophic levels if they exist at potentially toxic levels and are, or become, bioavailable. Some of the processes outlined above have potential for application to treat contaminated land. Solubilization processes provide a route for removal of metals from soil matrices, whereas immobilization processes enable metals to be transformed to insoluble chemically inert forms. Although some processes could be used *in situ* (e.g., leaching using S-oxidizing bacteria), many are probably most suitable for *ex situ* use in bioreactors, where the mobilized or immobilized metal can be separated from soil components. Living or dead fungal and bacterial biomass and metabolites have been used to remove metals and metalloids from solution by biosorption or chelation.

Recently there has been renewed interest in the ability of some mycorrhizal fungal symbionts to improve plant metal tolerance and increase metal translocation to the aboveground parts of plants. Such associations could be used to enhance the

phytoextraction capabilities of plants grown for removing toxic metals from soil by accumulation in their shoots. Amelioration of metal phytotoxicity by mycorrhizal fungi has been widely demonstrated, particularly when inoculated with mycorrhizae isolated from metalliferous or acid soils.

The increased attention to microbial activities in anaerobic, subsurface environments in the past few years has opened up new possibilities for the bioremediation of metal contaminants. Metal(loid)s that form insoluble precipitates when reduced may be of particular interest for *in situ* treatment, such as Se(0), Cr(III), Tc(IV), and U(IV). Natural processes in subsurface soil may immobilize contaminants otherwise predicted to leach. Thorough knowledge and monitoring of the system are essential for such treatment schemes, because reoxidation can result in mobilization.

Summary

Dissolution, precipitation, oxidation, and reduction processes are vital to metal cycling in aerobic and anaerobic surface and subsurface environments. As active agents in cycling, bacteria and fungi are fundamental biotic components of natural biogeochemical cycles for metals and metalloids. Some important processes catalyzed by microorganisms also have important applications in environmental biotechnology in the areas of ore leaching and bioremediation.

Acknowledgments

S.M.G. and T.J.B. acknowledge support from the US Department of Energy Natural and Accelerated Bioremdiation program. G.M.G. gratefully acknowledges financial support from the Biotechnology and Biological Sciences Research Council (94/MAF12243; 94/BRE13640), and British Nuclear Fuels plc. E.P.B. gratefully acknowledges receipt of a Natural Environment Research Council postgraduate studentship.

See also: **Heavy Metals**; **Metal Oxides**

Further Reading

Banfield JH and Nealson KH (eds) (1997) Geomicrobiology: interactions between microbes and minerals. *Reviews in Mineralogy* 35.

Beveridge TJ and Doyle RJ (eds) (1989) *Metal Ions and Bacteria*. New York: John Wiley.

Burford EP, Fomina M, and Gadd GM (2003) Fungal involvement in bioweathering and biotransformation of rocks and minerals. *Mineralogical Magazine* 67: 1127–1155.

Gadd GM (1993) Interactions of fungi with toxic metals. *New Phytologist* 124: 25–60.

Gadd GM (1999) Fungal production of citric and oxalic acid: importance in metal speciation, physiology and biogeochemical processes. *Advances in Microbial Physiology* 41: 47–92.

Gadd GM (2000) Bioremedial potential of microbial mechanisms of metal mobilization and immobilization. *Current Opinion in Biotechnology* 11: 271–279.

Gadd GM (ed.) (2001) *Fungi in Bioremediation*. Cambridge, UK: Cambridge University Press.

Hernandez ME and Newman DK (2001) Extracellular electron transfer. *Cellular and Molecular Life Science* 58: 1562–1571.

Korenevsky A, Glasauer S, and Beveridge TJ (2002) Biomineralization by bacteria. In: Bitton G (ed.) *Encyclopedia of Environmental Microbiology*. New York: John Wiley.

Lovley DR (ed.) (2000) *Environmental Microbe–Metal Interactions*. Washington, DC: ASM Press.

McLean JS, Lee J-U, and Beveridge TJ (2002) Interactions of bacteria and environmental metals, fine-grained mineral development, and bioremediation strategies. In: Huang PM, Bollag J-M, and Senesi N (eds) *Interactions between Soil Particles and Microorganisms*, pp. 228–261. New York: John Wiley.

Verrecchia EP (2000) Fungi and sediments. In: Riding RE and Awramik SM (eds) *Microbial Sediments*, pp. 69–75. Berlin, Germany: Springer-Verlag.

Verrecchia EP and Dumont J-L (1996) A biogeochemical model for chalk alteration by fungi in semiarid environments. *Biogeochemistry* 35: 447–470.

Metals, Heavy *See* **Heavy Metals**

MICROBIAL PROCESSES

Contents
Environmental Factors
Community Analysis
Kinetics

Environmental Factors

P G Hartel, University of Georgia, Athens, GA, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

The activity of all living organisms in soils is affected by environmental factors. These environmental factors are commonly divided into abiotic and biotic factors: the primary abiotic factors include nutrients, water, oxygen, temperature, pH, and light; the primary biotic factors include organism–organism interactions of mutualism, commensalism, amensalism, competition, predation, and parasitism. In addition to these primary factors, other factors (e.g., ionizing radiation in a uranium-rich soil) exist in a few specialized soil habitats, but these rare factors are not considered here.

Before one can predict how an environmental factor will affect the activity of microbes in a particular soil, there are a number of points to consider with respect to the soil, the soil organisms, and the environmental factors. With respect to the soil, it is important to consider how soil differs from other environments. A typical mineral soil is 50% solid material and 50% pore space, and has a bulk density (the weight of the soil divided by the total volume of the soil) ranging from 1.00 to 1.80 $g\,cm^{-3}$. Because of its density, soil can offer a much broader spectrum of varying environmental factors than other environments. For example, bacteria incapable of tolerating oxygen can exist only micrometers apart from bacteria incapable of existing without oxygen. The creation of these microsites allows soil to contain an exceptional diversity of microorganisms over a small space, and gives soil a complexity unlike all other habitats.

Further, all these organisms live in the soil pores. Under normal circumstances, chemical and physical forces limit movement of most soil organisms in these pores. In terms of chemical forces and bacteria, these organisms are negatively charged at typical soil pH values (5–8) and are held in place by negatively charged clay and organic matter through ion exchange, coordination bonding, hydrogen bonding, and van der Waals forces. In terms of physical forces, soil pores, particularly micropores (less than 30 μm), offer only a tortuous path to movement. Only when soils are saturated, or when plant roots or larger soil organisms actively or passively transport smaller soil organisms, do smaller soil organisms overcome these physical and chemical forces to move great distances (centimeters or more). Therefore, when environmental factors become adverse, most soil organisms are stuck in place and cannot escape to better conditions elsewhere.

With respect to soil organisms, many soil organisms are unidentified. Although the percentage of identification is comparatively good for soil algae and many soil protozoa, nematodes, and other soil fauna, only 0.1–10% of all soil bacteria and archaea, and 5% of all fungi are speciated. Molecular techniques, especially nucleic acid analyses, will increase these percentages, but at present, generalizations about how environmental factors affect microbial activity must be interpreted with caution when the identity of so many species is unknown.

With respect to environmental factors, these factors range from essential to unimportant. Like Goldilocks, soil organisms prefer essential factors to be "just right." For an organism to be outside the boundary of an essential environmental factor means diminished microbial activity, and, if sufficiently far outside the boundary, death (Figure 1). Understanding how environmental factors affect microbial activity is more difficult than it appears, because environmental factors are often dependent on each other.

Abiotic Factors

Nutrients

Nutrients are substances that soil organisms must obtain from the environment to survive and grow. Soils vary in their ability to contain and to hold these nutrients. On a chemical basis, the ability of a soil to hold nutrients is based on its anion exchange capacity and cation exchange capacity, and it is the

type and amount of clay and organic matter in soil that determine this.

Some nutrients provide energy, whereas others provide cell constituents. In terms of energy, all soil organisms are either chemotrophs, which derive all their energy from organic and inorganic sources (sometimes also called organotrophs and lithotrophs, respectively), or phototrophs, which convert light energy into chemical energy. Because carbon is the most common element in all soil organisms, chemotrophs and phototrophs are typically further divided into autotrophs, which convert inorganic carbon (i.e., CO_2 or carbonates) into organic compounds, or heterotrophs, most of which are saprophytes and feed on nonliving plants and animals or on humic substances in soil. Table 1 shows examples of this classification.

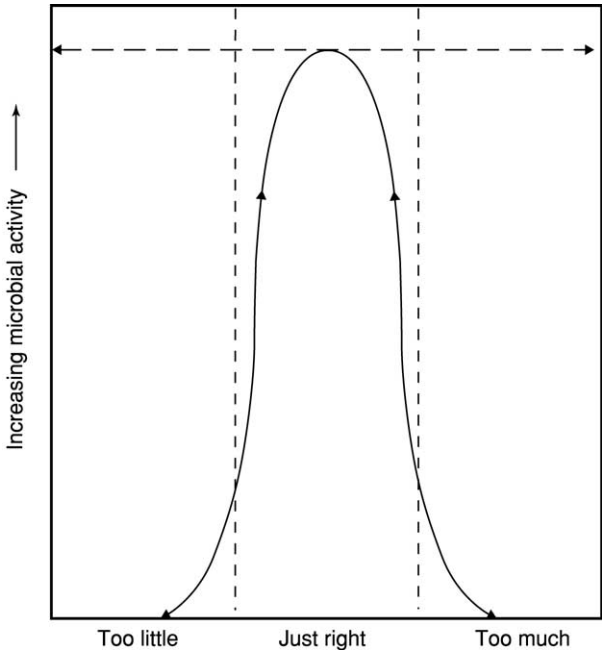


Figure 1 Idealized response of a soil organism to a critical soil environmental factor (solid line). The response of the organism does not have to be bell-shaped, and can be asymptotic to the left or to the right (horizontal dashed line).

In terms of cell constituents, nutrients are divided into macronutrients, substances required in large quantities by organisms, and micronutrients, substances required in small quantities (nanograms per gram of dry weight). Besides carbon, the major macronutrients for soil organisms are nitrogen, phosphorus, and sulfur, and, to a lesser extent, calcium, magnesium, potassium, and sodium. The major micronutrients are cobalt, copper, iron, manganese, molybdenum, and zinc. Macronutrients comprise the vast majority of the structural components such as carbohydrates, proteins, lipids, and nucleic acids, while micronutrients are typically found in growth factors and enzymes.

There are only three natural conditions during which soil organisms see nutrients in abundance. The first is governed by the natural life cycle of plants. As annual plants die and perennial deciduous plants shed their leaves, the accumulated nutrients are released into the soil.

The second condition is the relatively steady exudation of organic compounds from plant roots. Because some soil organisms can promote plant growth through the production of plant growth hormones and other mechanisms, plants have evolved to exude a complex mixture of organic compounds through their roots, primarily organic acids, amino acids, and carbohydrates. The rhizosphere, or zone of soil under the influence of the plant, has sufficient nutrients such that the colony-forming units of bacteria and fungi per gram of rhizosphere soil are typically 10- to 100-fold higher than bulk soil, and the numbers of soil animals that prey on these microbes are doubled or tripled. With such high numbers, it is unsurprising that microbial activity is often intense.

The third condition is when something disrupts the first two conditions. Natural disruptions include conditions such as fire and the episodic drying and rewetting of soil. Fire, by consuming aboveground biomass, minimizes the rhizosphere effect by reducing the energy supply to the soil organisms. Episodic drying causes soil organisms to die, unless these organisms have developed a mechanism to avoid desiccation (e.g., they form a resistant structure such as a spore or cyst). When soil is rewetted, the survivors

Table 1 Classification of all soil organisms based on their source of energy ('photo-,' light; 'chemo-,' organic or inorganic sources) and carbon ('auto-,' CO_2 or carbonates; 'hetero-,' organic C sources)

Energy source	*Carbon source*	*Classification*[a]	*Example(s)*
Light	Inorganic carbon	Photoautotroph	Algae, cyanobacteria
	Organic carbon	Photoheterotroph	Photosynthetic bacteria
Organic or inorganic sources	Inorganic carbon	Chemoautotroph	Nitrifying and sulfur-oxidizing bacteria
	Organic carbon	Chemoheterotroph	Fungi, most bacteria, and soil fauna

[a]The suffix '-troph' means to nourish.

dine on the deceased. Human disruptions include conditions such as clear-cutting and tillage. Clear-cutting has the same effect as fire, while tillage disrupts the soil structure, releasing occluded organic matter and causing a temporary glut of nutrients to the soil community.

Bioavailability, or the degree to which nutrients are accessible to soil microorganisms or extracellular enzymes, depends not only on chemical characteristics of the nutrients (e.g., whether they are polar or nonpolar) and their physical place (e.g., whether they are in inaccessible nanopores or not), but also on abiotic factors. For example, in the case of leaf fall in northern temperate regions, microbial activity reaches a peak in autumn, diminishes in winter with its cooler temperatures, and then resurges again in spring when favorable abiotic factors recur. Activity in summer is low because water is often limited.

Not surprisingly, soil organisms have developed an array of different mechanisms to deal with this periodic, episodic, or steady flow of nutrients. Some soil organisms opt for rapid growth only when nutrients are plentiful. By placing all their efforts into reproduction, these organisms attempt to occupy a niche as quickly as possible and then die to low numbers awaiting the next flush of nutrients. Organisms like this are termed 'r-selected' or 'zymogenous' strategists, and this strategy is characteristic of many bacterial species and soil fauna. Some organisms, by virtue of being left with 'crumbs from the table,' opt for a slower growth rate and tend to metabolize more complex nutrients. Organisms like this are termed 'K-selected' or 'autochthonous' strategists, and this strategy is characteristic of many fungal and actinomycete species. Finally, other organisms lie on the continuum between these two strategies.

For autotrophs, there is no problem of limited carbon, because CO_2 and bicarbonate are usually present in excess of demand. For heterotrophs, microbial activity is dependent on the quality and quantity of available carbon. Readily available natural organic compounds such as simple sugars are degraded quickly and microbial activity is high. For natural organic compounds that are recalcitrant or resistant to degradation (e.g., lignin), microbial activity is low. Toxic compounds such as antibiotics also have the capacity to diminish microbial activity. There is one special case for xenobiotics, anthropogenic compounds that have no biological parallel in the environment. These nonnatural compounds typically have either halogen substituents (e.g., Cl) or are branched polymers (e.g., polypropylene). Soil organisms, by virtue of their enzymatic virtuosity, degrade most xenobiotics as they would natural compounds. Sometimes, however, an organism accidentally metabolizes a xenobiotic in the process of metabolizing a second compound. This process, called cometabolism or fortuitous metabolism, is beneficial to humans in terms of biodegradation of xenobiotics in soils, but is detrimental or at best neutral to the activity of the organism, because the process does not yield any energy or carbon. For example, a bacterium cometabolizing polychlorinated biphenyls (PCBs) yields chlorinated aliphatic and aromatic acids, both dead-end products that the bacterium cannot assimilate.

Water

All soil organisms – without exception – require water for microbial activity. The major force governing water in soil is the total water potential, which is composed primarily of gravitational, matric, and osmotic potentials, and is measured in units of pressure, pascals (Pa). Gravitational potential is the force of gravity pulling water toward the Earth's center, matric potential is the adsorption of water to surfaces of soil particles and capillary forces, and osmotic potential is the attraction of solute ions for water molecules. Soil organisms have their own water potentials, composed primarily of osmotic and turgor potentials. Before considering these potentials, it is important to consider the relationship among these potentials, soil oxygen, soil texture (the mixture of sand, silt, and clay in soil), and soil porosity. The relationships among temperature, pH, light, and total water potential are considered in later sections.

The major mechanism by which oxygen moves into soil is diffusion. The diffusion coefficient of oxygen in air is $0.189\,cm^2\,s^{-1}$, but the diffusion of oxygen in water is $0.000\,025\,cm^2\,s^{-1}$. Thus, oxygen diffuses 10 000-fold slower in water than in air. How much water a soil holds is dependent on soil texture and soil porosity. Sandy soils have less pore space (35–50%) than clayey soils (40–60%), but the pore space in sandy soils is mostly macropores (30–100 μm), whereas the pore space in clayey soils is mostly micropores (less than 30 μm). For this reason, sandy soils drain quickly and hold less water than clayey soils, which drain more slowly and hold more water.

The theoretical limits for microbial activity in soil are between 0 and -81 MPa. At 0 MPa, soil is completely saturated; at -81 MPa, DNA is denatured. Besides the effect of each water potential on microbial activity, it is also important to consider oxygen, movement, and nutrients. When the soil is saturated, diffusion of oxygen into the soil is slowed, and if the oxygen consumption rate exceeds the diffusion rate, the soil, regardless of its texture and porosity, becomes anaerobic. Heterotrophs incapable of anaerobic respiration (e.g., most fungi) fare poorly. Heterotrophs capable of anaerobic respiration (e.g.,

many bacteria) fare better but show diminished activity because the amount of energy produced under anaerobic conditions is less than under aerobic conditions. Phototrophs on the soil surface (e.g., eukaryotic algae and cyanobacteria) fare well and even prefer saturated conditions. Movement by aquatic soil organisms is maximal because the water film thickness around soil particles is maximal. Texture plays a role because soil organisms move more easily in sandy soils than in clayey soils (again, because sandy soils have more macropores). Concentrations of water-soluble nutrients are low. Water stress for almost all soil organisms is minimal.

Once water drains from the macropores, the water potential becomes negative (approximately −0.01 MPa for sandy soils and −0.03 MPa for other soils), and oxygen diffusion increases. Sandy soils, which contain more macropores than clayey soils, are more oxygenated than clayey soils. Except for movement by aerial (i.e., living in air-filled spaces) soil animals such as microarthropods (mites and collembolans), movement by aquatic soil organisms such as nematodes, protozoa, oomycete zoospores, and bacteria becomes more difficult as the water-film thickness dwindles to only a few micrometers. However, the combination of aerobic conditions, concentrated nutrients, and good water relations means microbial activity is near maximal.

As the total soil water potential decreases from −0.01 to −0.1 MPa, oxygen diffusion improves, but water films become discontinuous, and consequently nutrient availability and movement by aquatic organisms decrease. Because matric potential is related to capillary forces and these forces are directly related to the soil pore diameter, clayey soils, by virtue of their more numerous micropores, contain more water than sandy soils and dry more slowly. This drying rate is important, because the faster the rate of drying, the greater the mortality of the organism. Because microbial activity depends on each soil organism maintaining proper turgor pressure for cell functioning, soil matric and osmotic potentials increase in their importance. Except in saline soils, matric potentials are more important than osmotic potentials.

By −1.5 MPa, water is gone from the micropores and only water bound to soil solids is available for microbial activity. The soils are relatively oxygenated; movement stops. Diffusion of nutrients is so retarded that starvation conditions are created. Furthermore, because all soil organisms contain semipermeable membranes that allow water to enter and exit in near equilibrium with the environment, osmotic potential becomes critical as soil ions become more and more concentrated. To avoid plasmolysis (contraction of the cytoplasm from the cell wall), most soil organisms divert their energy to accumulating stress solutes (e.g., K^+) and compatible solutes (e.g., glycerol; so named because they must be compatible with cell functioning) to counter this negative potential. By −4.0 MPa, most biological processes (e.g., nitrification) stop and only variously named osmophiles, halophiles, or xerophiles show appreciable microbial activity. Unless an organism can form a resistant structure (e.g., spores in bacteria and fungi, akinetes in the cyanobacteria, cysts in the protozoa), most organisms desiccate and die.

Oxygen

As already noted, oxygen in soil depends on soil water, soil texture, and soil porosity. Both soil depth and temperature can be added to this mix: generally, oxygen decreases and CO_2 increases with increasing depth, and gas solubility, particularly in water, decreases with increasing temperature. Most soil organisms are obligate aerobes and use oxygen exclusively as a terminal electron acceptor. Because aerobic respiration produces more energy per mole of substrate than anaerobic metabolism, obligate aerobes recover more energy for biomass production than anaerobic organisms. For this reason, population densities of obligate aerobes are higher than those of anaerobes. Without a source of oxygen, aerobic respiration ceases and no energy is produced. In contrast, obligate anaerobes cannot survive the presence of oxygen because aerobic respiration temporarily yields hydrogen peroxide (H_2O_2) and superoxide (O_2^-), toxic products to which obligate anaerobes lack the enzymes (catalase and superoxide dismutase, respectively) to break down. As a result, obligate anaerobes live strictly by fermentation. Some obligate aerobes that prefer oxygen as the terminal electron acceptor can respire anaerobically by substituting nitrate (NO_3^-), sulfate (SO_4^{-2}), or carbonate (CO_3^{-2}) for oxygen. This anaerobic respiration is restricted to bacteria and for this reason these bacteria dominate in anaerobic soils. Many bacteria and some fungi are facultative anaerobes, which grow in the presence or absence of oxygen. When oxygen is available, the organisms grow aerobically; when oxygen is unavailable, the organisms grow by either fermentation or anaerobic respiration. Finally, aerotolerant anaerobes lack an electron transport system and function solely by fermentation.

Redox (reduction–oxidation) potential, the tendency of a compound to accept or donate electrons, can also give some insight to oxygen tensions in soil and which organisms are likely to be active in a particular soil environment. From a standpoint of processes, soils with a positive redox potential (0–400 mV) are 'oxidizing' habitats and are

conducive to aerobic processes, whereas soils with a negative redox potential (0 to −400 mV) are 'reducing' habitats and conducive to anaerobic processes such as methane production. However, it would be a misstatement that highly aerobic soils contain few or no anaerobes. The ability of soils to contain water-filled micropores restricts oxygen diffusion and guarantees that anaerobic microsites exist. Thus, it is common to find obligate anaerobes in the upper soil horizons. Also, some anaerobic processes, such as denitrification, can occur in soils with a positive redox potential when molecular oxygen is absent but other terminal electron acceptors are present.

Temperature

For soil organisms, it is important to consider not only the temperature at which microbial activity is optimal, but also the minimum and maximum temperature that the organism can tolerate. Soil organisms are divided into psychrophiles, mesophiles, and thermophiles. Psychrophiles do not grow above 20°C, have minima of 0°C or lower, and optima around 16°C. Mesophiles have minima of above 0°C, maxima below 50°C, and optima between 15 and 40°C. Finally, thermophiles have minima at or above 20°C, maxima at or above 50°C, and optima between 40 and 50°C.

There is an important relationship between temperature, water potential, and microbial activity. Both water potential and temperature must be within critical range for microbial activity to occur. Assuming water potential is within this range, at one extreme, reduced temperatures slow cellular processes. At the other extreme, excessive temperatures denature proteins and alter cell membrane permeability. At either extreme, soil organisms typically alter the fatty acyl composition of their membrane lipids and produce either cold-shock or heat-shock proteins to withstand these stresses. Between these extremes is another relationship, often referred to as the 'Q_{10} rule for biological systems,' where Q_{10}, the temperature quotient, equals 2. This rule states that, within a limited range, microbial activity doubles for each 10°C increase in temperature. For example, mesophilic organisms at 30°C will have twice the activity as the same organisms at 20°C. Again, this rule is a range, and there are examples, such as petroleum mineralization in Alaska, where Q_{10} values much greater than 2 have been measured.

Because most soil temperatures correspond well with the temperature optima of mesophiles, it is not surprising that most soil organisms are mesophiles. However, some specialized habitats (e.g., volcanic soils) do favor thermophilic actinomycetes and fungi.

pH

As the negative log of the hydrogen ion concentration, pH is intimately connected to water in the soil and the soil organisms. With regard to soil, wherever rainfall is abundant, water leaches bases from the soil and the soils are predominantly acid; wherever rainfall is scant, there is not enough water to leach the bases from the soil and the soils are predominantly alkaline. Because soils have a pH-dependent charge originating from organic matter and the broken edges of clays, this charge decreases as the soil becomes more acid, reducing not only the cation exchange capacity of the soil, but also the ability of the soil to retain nutrients. Reducing the pH does not necessarily reduce the entire cation exchange capacity in soil; some charge may be associated with isomorphous substitution in clay minerals and this charge is pH-independent. With regard to soil organisms, when soils are saturated, the hydrogen ion concentration is diluted; when soils are dry, the hydrogen ion concentration is concentrated and this concentration affects the ability of soil organisms to control anions and cations and offset high osmotic potential.

Although some bacteria and soil fauna tolerate low pH (acidophiles; able to grow at pH 3 or lower) or high pH (alkaliphiles; able to grow at pH 9 or higher), most bacteria and soil fauna are neutrophiles and prefer a near-neutral pH range of 6–7. The same is true of soil algae, which again are found in acid and highly alkaline environments, but prefer pH ranges of nearly neutral to slightly alkaline. Many bacteria and algae prefer this pH range because soil nutrients are most available. Furthermore, a neutral pH avoids heavy metal toxicity, because almost all these metals (a density of greater than 5 $Mg\,m^{-3}$) are only active in acid soils (pH <5). In contrast, fungi are acid-tolerant (pH 4–7) and tend to be abundant in acid soils, where there is less competition from bacteria.

Light

For phototrophs such as cyanobacteria and algae, light is an essential factor because it is required as a source of energy. Soil affects light because soil constituents absorb or reflect it (the albedo effect), and soil disturbances (e.g., tillage) affect how far light penetrates into the soil. For these reasons, the photic zone in soil is variable. Nevertheless, phototrophs decrease with increasing soil depth and decreasing light intensity. Except for a few phototrophs that are facultatively capable of heterotrophy, phototrophs deep in the soil are likely to be metabolically inactive.

For chemotrophs, light is important for a different reason. Although visible light may be important for certain phototrophic responses (e.g., spore formation

in fungi), ultraviolet-B (UV-B) light (290–320 nm) is a source of nonionizing radiation. This irradiation splits water molecules into free radicals that damage DNA. This irradiation links light to water potential: the higher the water content, the more free radicals produced, and the higher the DNA damage. Although ozone screens out most UV-B light, ozone depletion in the atmosphere offsets this. Therefore, it is likely that this abiotic factor will increase its deleterious effect on microbial activity in the photic zone, and only soil organisms that produce UV-absorbing pigments or develop rapid mechanisms of DNA repair will be relatively unaffected.

Biotic Factors

Soil organisms can also interact with other organisms, and, with the exception of neutralism (no interaction), these interactions affect their microbial activity. These biotic interactions can also include plants and larger soil fauna, and at their most complex, populations of organisms or consortia. These interactions are mutualism, symbiosis, commensalism, competition, amensalism, predation, and parasitism (Table 2).

Symbiosis or Mutualism

In symbiosis or mutualism, both organisms benefit. This association can be specific or general and, in some cases, so highly specialized that one or both organisms are obligate symbionts and cannot live without the other. Relationships where there is mutual benefit but no obligation on the part of either organism are often termed 'synergistic.' Because of its importance to agriculture, the best-known symbiosis is the presence of nitrogen-fixing bacteria in plants. In return for the bacteria fixing atmospheric nitrogen, the plant supplies the bacteria with nutrients. An agricultural example is the symbiosis between *Bradyrhizobium japonicum* and the legume soybean (*Glycine max*); a forestry example is the symbiosis between the actinomycete *Frankia* and a range of angiosperms such as alder (*Alnus* sp.). A similar situation is observed with several genera of soil fungi, the mycorrhizae, which colonize plant roots, thereby enhancing uptake of water and mineral nutrients by the plant, principally phosphorus. Again, the plant supplies the mycorrhizae with nutrients. Symbiosis or mutualism is much more frequent than commonly realized. For example, 95% of all plants are mycorrhizal. These kinds of mutually cooperative ventures – not competitive ones – are the norm in nature.

Table 2 All possible positive interactions (+), no effect (0), and negative interactions (−) between two organisms. The effect of these interactions on microbial activity corresponds with the symbol

Organism No. 1	*Organism No. 2*		
	+	*0*	−
+	Mutualism, symbiosis	Commensalism	Predation, parasitism
0	Commensalism	Neutralism	Amensalism
−	Predation, parasitism	Amensalism	Competition

Commensalism and Amensalism

In commensalism, one organism benefits while the other is unaffected. For example, one organism can provide an essential growth factor, such as a vitamin, for another organism. This type of cross-feeding is common in soil organisms. The opposite of commensalism is amensalism, where one organism is harmed while the other is unaffected. A good example of this interaction is when one organism produces an antibiotic against another organism. Such an interaction is often the basis of biological control. For example, some isolates of the bacterium *Pseudomonas fluorescens* can suppress the fungal pathogen *Gaeumannomyces graminis*, responsible for 'take-all' in wheat (*Triticum aestivum*).

Competition

In competition, both organisms vie for the same resource. This resource can be for any one of the abiotic factors as well as space. In the case of nutrients, the three factors that govern competitive success are growth rate, substrate affinity, and efficiency. If nutrients are unlimited, then microbial activity is governed by the maximal growth rate. This maximal growth rate is rarely the case in soil; the norm for most organisms in soil is starvation. Under these circumstances, substrate affinity and efficiency govern the ability to compete. In the case of substrate affinity, assuming an organism can metabolize a substrate, an organism that recognizes a substrate at a low concentration outcompetes an organism that only recognizes a substrate at a high concentration. In the case of efficiency, an organism that is more efficient in its carbon and energy use outcompetes a less efficient organism.

Predation and Parasitism

Finally, in predation and parasitism, one organism feeds on another. The most common example of predation in soil is protozoa and nematodes preying on bacteria and fungi, particularly in the rhizosphere,

where numbers of bacteria and fungi are high. In Figure 2, two types of soil protozoa, a ciliate (*Colpoda* sp.) and a flagellate (*Bodo* sp.), as well as a bacteriovorous nematode, are preying on bacteria attached to the clay particles and organic matter, while, above, a mycophagous mite is preparing to eat some mycorrhizal hyphae. Predator responses follow classic predator–prey curves: increases and decreases in numbers of predators lag behind increases and decreases in numbers of prey, respectively. Numbers may not be a good reflection of microbial activity, however. The C:N ratio of soil fauna is approximately 20:1, whereas the C:N ratio ranges from 4 to 5:1 for bacteria and up to 15:1 for fungi. When protozoa prey on bacteria and fungi, nitrogen is in excess and is excreted into the environment. This nitrogen is then available to promote microbial activity.

The most common example of parasitism in soil is viruses. Even though viruses are nonliving (because they have no intrinsic metabolism or ability to replicate), they parasitize numerous soil bacteria, fungi, and soil fauna. This parasitism is normally a negative interaction, but viruses can sometimes interact positively because they are capable of transduction (adding a new gene or modifying an existing gene in the host DNA). If this addition or modification makes the host more competitive, then the change benefits microbial activity.

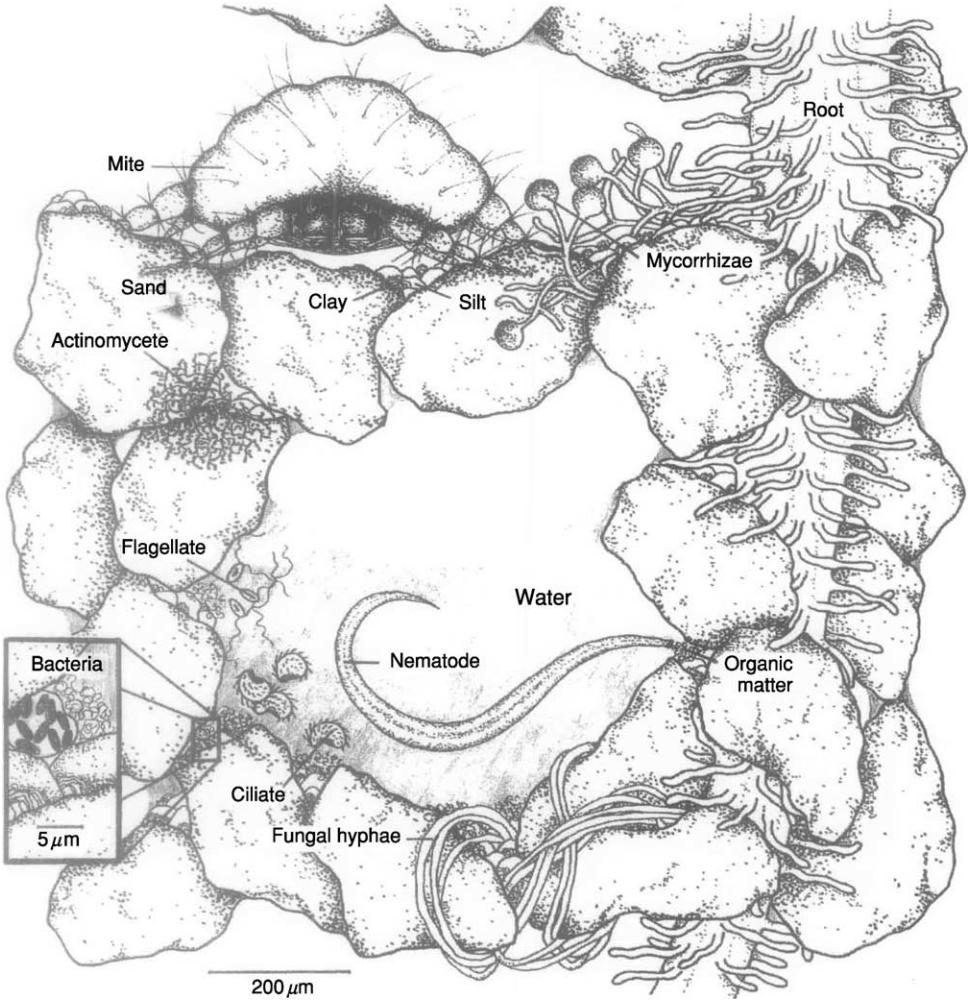


Figure 2 A sample of well-aggregated soil, slightly greater than 1 mm^2, showing a variety of biotic and abiotic factors. As gravitational water flows out of the two largest pores, water potential and oxygen vary from top to bottom, while a small root creates a rhizosphere effect, a gradient of nutrients from high (right) to low (left). Root transpiration affects the water potential depending on whether it is day (more) or night (less). The root alters the soil pH by releasing H^+, HCO_3^-, and organic compounds, depending on the type of nitrogen in the soil (NO_3^- increases, NH_4^+ decreases), symbiotic relationships (here, mycorrhizae), and microbial activity. Temperature also affects the rate of microbial activity. Biotic factors, and interactions of biotic and abiotic factors are noted in the text. In all, a wide variety of microsites are being created, many only micrometers apart. (The mite has been turned head-on to fit.) (An original drawing by Kim Luoma.)

Integrating Abiotic and Biotic Factors

In addition to interactions between and among soil abiotic factors, and interactions between soil organisms, there are also interactions between the abiotic and biotic factors. In one direction, soil abiotic factors affect biotic interactions. For example, nutrient availability, water potential, and oxygen profoundly affect microbial interactions. In Figure 2, the aerial soil mite would not be preparing to eat the mycorrhizae unless the limited soil nutrients (e.g., phosphorus) had promoted the symbiotic relationship between the plant and the fungus, and the gravitational water had drained from the pore, permitting sufficient oxygen diffusion for the mite to breathe. In the other direction, biotic interactions affect soil abiotic factors, both physically and chemically. In Figure 2, actinomycetous and fungal hyphae, and bacteria in a extracellular polysaccharide 'plug' are binding the soil particles, thus promoting soil aggregation, one of the most important physical effects of microorganisms on soil. This aggregation affects soil porosity, which, in turn, affects soil water and oxygen relations. As an example of a chemical effect, the nematode and protozoa eat the bacteria and excrete excess ammonia, which chemoautotrophic bacteria convert to nitrite, which, in turn, is converted commensalistically by nitrite-oxidizing bacteria to nitrate. These acidifying reactions lower the soil pH.

The interactions between and among abiotic and biotic factors are the equivalent of an intricate and elegant dance with many partners. Our understanding of this dance, particularly among microbial communities, is still rudimentary. The major barriers to understanding this dance are our poor understanding of the identity of most soil archaea, bacteria, fungi, and undoubtedly many of the soil fauna, and the nature of the soil matrix to provide a multitude of microsites, each with their properties, over distances that may be only a few micrometers apart. This is what gives soil a spatial and temporal dynamicism and heterogeneity unlike all other habitats.

See also: **Microbial Processes:** Community Analysis; Kinetics

Further Reading

Brady NC and Weil RR (2004) *Elements of the Nature and Properties of Soils*, 2nd edn. Upper Saddle River, NJ: Pearson/Prentice–Hall.

Coleman DC and Crossley DA Jr (1996) *Fundamentals of Soil Ecology*. New York: Academic Press.

Coyne M (1999) *Soil Microbiology: An Exploratory Approach*. Albany, NY: Delmar Publishers.

Darbyshire JF (1994) *Soil Protozoa*. Wallingford, UK: CAB International.

Havlin J, Tisdale SL, and Beaton JD (1998) *Soil Fertility and Fertilizers: An Introduction to Nutrient Management*, 6th edn. New York: Macmillan.

Horikoshi K and Grant WD (1998) *Extremophiles: Microbial Life in Extreme Environments*. New York: Wiley–Liss.

Madigan MT, Martinko JM, and Parker J (2000) *Brock's Biology of Microorganisms*, 9th edn. Upper Saddle River, NJ: Prentice-Hall.

Nester EW, Roberts CE, and Nester MT (1998) *Microbiology: A Human Perspective*. New York: WCB/McGraw-Hill.

Parr JF, Gardner WR, and Elliott LF (1981) *Water Potential Relations in Soil Microbiology*. SSSA Special Publication No. 9. Madison, WI: Soil Science Society of America.

Sylvia DM, Fuhrmann JJ, Hartel PG, and Zuberer DA (1998) *Principles and Applications of Soil Microbiology*. Upper Saddle River, NJ: Prentice-Hall.

Torsvik V, Sørheim R, and Goksøyr (1996) Total bacterial diversity in soil and sediment communities – a review. *Journal of Industrial Microbiology* 17: 170–178.

Wicklow DT and Söderström B (1997) *Environmental and Microbial Relationships*, vol. 4. In: Esser K and Lemke PA (eds) *The Mycota: A Comprehensive Treatise on Fungi as Experimental Systems for Basic and Applied Research*. Berlin, Germany: Springer-Verlag.

Community Analysis

C H Nakatsu, Purdue University, West Lafayette, IN, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

The ecological diversity of soil microbial communities is a measure of microbial community richness and abundance. Richness is an indication of the number of different species present, whereas abundance is the number or biomass of each species present. Microbial communities can also be analyzed for their functional role in soil. There is not a single method available that can provide complete information on a microbial community; the method chosen is dependent on the question being posed. In the past, soil scientists relied mainly on cultivation and measuring substrate transformation to assess and monitor microbial communities. Now a common view in soil microbiology is that cultivation-dependent methods represent only a small fraction of the microbial community present in soil. Beginning in the 1980s, Vigdis Torsvik and her advisor, Professor Jostein Goksøyr, began testing the use of deoxyribonucleic acid (DNA) reassociation kinetics to demonstrate the true extent

of soil microbial richness. They have concluded that 1 g of soil has an estimated 4000–10 000 different bacterial genomes present, of which less than 1% can be readily cultivated. This work has prompted largely the move from isolation and phenotypic characterization of microbial communities to direct analysis using genetic and biochemical tests to survey a more representative group. Using a variety of molecular techniques, investigators are studying soil microbial communities from a wide range of habitats and geographic locations. Their results generally demonstrate that soil environments are far more diverse, even under harsh environmental conditions, than was expected and that previously undescribed groups are important components of the soil community. Analytical approaches adopted are crucially dependent on the questions being addressed and availability of equipment and resources.

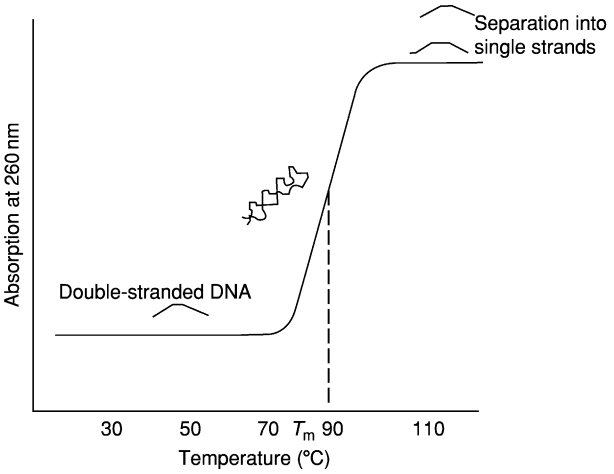


Figure 1 Melting curve of DNA showing the melting temperature, T_m, when half the DNA is denatured.

Background

Nucleic Acids

Nucleic acids are polymers of nucleotides joined by covalent (phosphodiester) bonds formed between their sugar and phosphate groups. There are two types of nucleic acids that differ chemically because of their sugar component, deoxyribose in DNA and ribose in ribonucleic acid (RNA). Physically the two differ because DNA is composed of two strands that form a helix and RNA is single-stranded. Attached to the sugar and phosphate groups are nucleotide bases of which there is a total possibility of five: cytosine (C), thymine (T, in DNA only), uracil (U, in RNA only), adenine (A), and guanine (G). The two strands comprising DNA are joined by hydrogen bonds between complementary nucleotide bases, where G always bonds with C (by three hydrogen bonds) and A with T (by two hydrogen bonds). The relative content of G + C compared with A + T in the genome of an organism can be used to differentiate taxonomic groups. The different order or sequence of nucleotide bases codes the information to determine appearances and functions in every organism. Nucleotide sequences that code for proteins are called genes. After nucleotide sequences are obtained, one can compare them with sequences in a databank (e.g., GenBank, EMBL) to determine whether any similar sequences have been found. This can provide information about potential function and phylogeny of a gene.

Denaturation and Renaturation

Denaturation of DNA occurs when the weak hydrogen bonds between the double strands are disrupted and the molecule becomes single-stranded. Thus the rate of denaturation is dependent on the proportion of G + C versus A + T bases. This process can be reversed in a process called renaturation or annealing. In the laboratory, DNA can be denatured by applying heat, increasing pH, or by adding denaturing chemicals (e.g., urea, detergents). In addition to these factors, viscosity and ionic strength can be adjusted to influence rates of denaturation and renaturation. As the DNA strands become denatured, different measurable properties of the molecule changes over time. Often DNA is quantified spectrophotometrically by measuring its ultraviolet (UV) light absorption at 260 nm. This approach can be used to track the melting characteristics of DNA, because in the single-stranded form it absorbs more light than when it is double-stranded. The data when graphed reveal the melting curve of the DNA (Figure 1). The melting temperature, T_m, is the transition temperature at which half the strand is denatured, as reflected in the halfway point in the change in absorbance at 260 nm. This temperature is important when carrying out experiments such as hybridization and polymerase chain reaction (PCR), where denaturation and/or renaturation of DNA strands is required.

Soil Nucleic Acid Extraction

Nucleic acids can be extracted directly from soil to analyze the microbial community. Alternatively cells can first be separated from the soil then the nucleic acids extracted. The latter approach decreases contamination by soil components but may also be biased toward cells that are separated more easily from soil. Most researchers isolate DNA from soil because it is more stable than RNA; however, protocols have been developed for soil RNA isolation. All

methods for nucleic acid extraction and purification follow the same principles. Cell lysis is the first step, followed by separation of the nucleic acids from the other cellular components, and finally concentrating the nucleic acids into a working volume. There are many papers that outline these methods, and commercial kits are marketed by a number of companies. An important factor when extracting nucleic acids from soil is to ensure the approach achieves equivalent lysis of all cell types so that the analysis is representative of the community. This is the main source of bias in direct-analysis methods of soil-extracted nucleic acids. Many laboratories have chosen to use mechanical methods such as bead-beating to ensure cell lysis. Lysis efficiency can be determined by microscopic examination of the sample before and after the extraction procedure. The other factor that must be considered is soil composition, because nucleic acids are charged and will bind to components, particularly clay. After the nucleic acids are extracted, a number of methods can be considered that examine nucleic acids directly or indirectly after amplification by PCR.

Ribosomal RNA

Ribosomes are comprised of proteins and ribosomal RNA (rRNA) and are found in all organisms, because they are the sites for protein synthesis. Presently, the genes coding for rRNA are used most extensively to obtain phylogenetic information. These genes were chosen because: they are present in all organisms; rRNA function is conserved; rRNA sequences have regions with high and low evolutionary conservation, making them phylogenetically informative; and, in actively growing cells, ribosomes can occur in high copy numbers (more than 10 000). The ribosome is composed of two subunits, referred to as small and large rRNA, differentiated by their sedimentation coefficient, *S*. In prokaryotes the small subunit contains 16*S* rRNA and the 70*S* large subunit contains both 23*S* and 5*S* rRNA. In eukaryotes the small subunit is slightly larger and contains 18*S* rRNA and the 80*S* large subunit contains 28*S*, 5.8*S*, and 5*S* rRNA. The 16*S* rRNA gene has been used the most in prokaryotic soil-community analysis. Many laboratories are determining 16*S* rRNA sequences from environmental sources and adding them to databanks, increasing the utility of using these genes for community analysis. In eukaryotic organisms the sequences between the rRNA genes, i.e., the intergenic region, are often included in analyses, because the gene sequence itself is often too conserved to differentiate species. The rRNA or rDNA can be isolated directly from soil and examined directly or after amplification using PCR. Many of the soil-community analytical techniques discussed herein are based on analysis of these genes or gene regions. It is important to note that the number of rRNA gene copies in an organism is variable, therefore one must not assume that determining the number of rDNA copies is equivalent to counting the number of cells in a community.

Direct Community Analysis

G + C Content

The G + C method is a coarse scale method, because different populations with similar G + C content will group together. It is used to determine compositional differences between microbial communities in different soils or after the soils have been perturbed, for example by the addition of a pesticide. A dye that preferentially binds to A + T or G + C nucleotides is added to DNA extracted from a soil community. Presently, the most commonly used dye is bisbenzimidazole (Hoechst 33258), which binds to A + T-rich regions. The buoyant density of the DNA decreases as more dye is bound to AT regions. Populations with different A + T content are then separated by cesium chloride density gradient centrifugation. Populations with similar G + C content will layer at the same gradient concentration. These layers can be collected in fractions and quantified by spectrophotometry to estimate G + C group sizes. Each of these fractions can be differentiated again by density gradient centrifugation at a finer scale, or some of the methods described below can be used to obtain more information about community composition. This method is not recommended if large quantities of DNA (more than 50 μg) and an ultracentrifuge are not available. The number of samples examined can be limited by the capacity of sample handling in the laboratory and amount of soil available for DNA extraction.

DNA Reassociation

A method to determine species richness in a soil community is reassociation of DNA strands after they have been completely denatured. It is based on the assumption that the rate of reassociation is dependent on the number of different genomes in a sample. Strand reassociation takes longer when the genetic heterogeneity of a community is greater. This method can be very time-consuming, because soils of high heterogeneity can take weeks to reassociate. The heterogeneity of the sample is proportional to the $Cot_{1/2}$ value, where Co is the initial molar concentration of DNA and $t_{1/2}$ is half the time required for complete strand reassociation. The number of different species can then be estimated by comparing the sample $Cot_{1/2}$ value to that of a reference strain with a known genome size. This method requires large quantities

(25 μg) of highly purified DNA, sheared to equivalent lengths. Due to the length of time required for reassociation, it is mandatory to have a UV spectrophotometer equipped with a thermally controlled sample cell dedicated to the analysis.

Nucleic Acid Hybridization

Nucleic acid hybridization methods can be used in soil community analysis to assess the composition, physical distribution, abundance, and/or relatedness of nucleotide sequences to DNA or RNA extracted from soil samples or left intact in cells (discussed in the next section). Nucleic acid hybridization methods rely on the ability of complementary single-stranded nucleotide sequences to anneal to form nucleic acid duplexes (discussed in the section Denaturation and Renaturation). This is commonly accomplished by hybridization (or annealing) of a previously characterized nucleic acid sequence (probe) to its counterpart (target) within a mixture of DNA or RNA (Figure 2). Probes are labeled with a radioisotope or fluorescent marker for detection. The target is

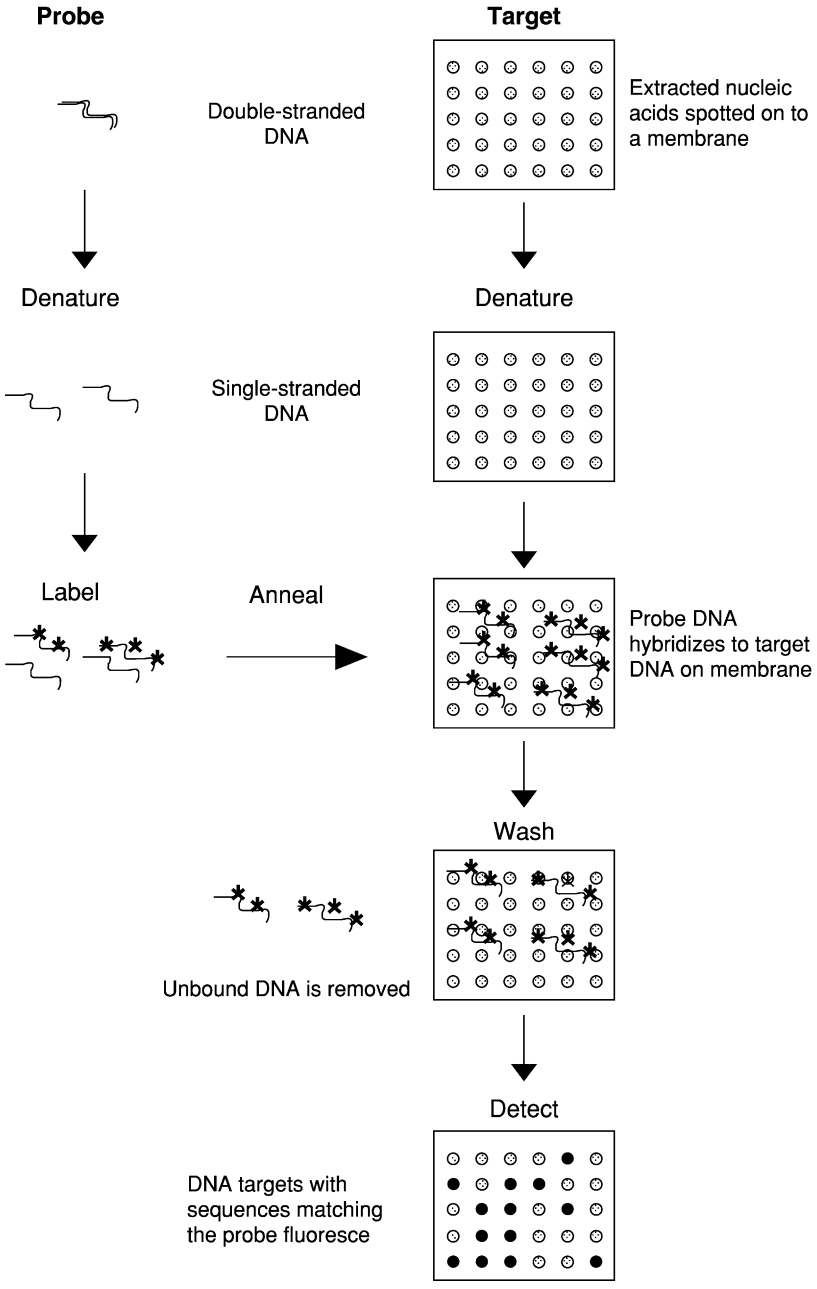


Figure 2 Hybridization and detection of a fluorescently labeled probe to target DNA that has been attached to a nylon membrane.

immobilized on to a solid support such as a nylon membrane or glass slide to facilitate hybrid detection. For these assessments, the nucleic acid sequences need not match exactly, because a limited number of base-pair mismatches can be tolerated if hybridization conditions (temperature, salt concentrations) are properly manipulated. Quantification of the active community members can be accomplished by using rRNA as a target instead of DNA, because the rRNA numbers increase with cell activity. Microarray technology is now making it possible simultaneously to hybridize to thousands of targets on a relatively small surface area. If adequate quantities of target nucleic acids are not readily obtained by extraction, potential targets can first also be amplified by PCR.

To obtain a relatively coarse measure of community composition, 16*S* rRNA gene probes are often used. Probe sequences have been determined that can differentiate community members at phylogenetic levels, ranging from broad (all organisms, kingdom, classes) to narrow (species or strain; examples given in Table 1). The broad probes can provide numbers of different groups present in a community. It is also possible to use all these sequences as targets for microarray analysis. However, due to the large number of microbial species still not described from soil, it is presently not possible to use this method to determine species richness. The species- and strain-specific probes are also useful to track the fate of a specific microorganism within a soil community. This approach has been used extensively to detect the presence of specific functions or functional groups such as ammonia-oxidizing bacteria in a soil community. If microarrays are not being used, then this method does not require specialized equipment other than those found in standard molecular biology laboratories.

Microscopic Examination and Fluorescent *In Situ* Hybridization

To examine microorganisms in the context of their soil environment, analysis can be made microscopically. An *in situ* hybridization-based procedure is employed in which fluorescently labeled probes diffuse into the cell, hybridize to the target rRNA, and can then be observed microscopically. Microscopic observations of bacterial cells are typically made using phase-contrast, epifluorescent, or confocal laser microscopes. Cells are usually killed then stained to detect all organisms or specifically to detect certain groups. General stains such as acridine orange, Hoechst 33258, and ethidium bromide bind to nucleic acids and thus can be used to detect all organisms. Specific groups are detected using the rRNA as the target, because this target has a high copy number in active cells and it is single-stranded to allow hybridization to take place. Sequences have been determined that are specific to different phylogenetic levels, ranging from all organisms to specific species. The availability of different-colored fluorescent markers makes it possible simultaneously to observe cells hybridized to different probes. In addition to determining the spatial distribution of cells in a community, this approach is beneficial because it is cultivation-independent and does not suffer from the biases of PCR. However, as in all methods, there are other shortfalls that must be considered when deciding the appropriateness of its use. Some of the limitations are: background fluorescence from the soil; interference by soil components; low cellular ribosome content making it difficult to see labeled cells; poor penetration of probe into the cell; and differential penetration of probe into different cell types.

Table 1 Examples of oligonucleotide probes targeting Bacteria at different phylogenetic levels

Probe name	*Target organism*	*Sequence*
UNIV1390	All organisms	5′ GAC GGG CGG TGT GTA CAA 3′
EUB338	Domain Bacteria	5′ GCT GCC TCC CGT AGG AGT 3
PLA46	Planctomycetes	5′ GAC TTG CAT GCC TAA TCC 3′
ALF968	α-Proteobacteria	5′ GGT AAG GTT CTG CGC GTT 3′
BET42*	β-Proteobacteria	5′ GCC TTC CCA CTT CGT TT 3′
GAM42*	γ-Proteobacteria	5′ GCC TTC CCA CAT CGT TT 3′
ACID20	Acidobacteria	5′ CAG CGT TGA TTC TGA GCC A 3′
GPOS1192	Gram-positive bacteria	5′ TAA GGG GCA TGA TGA TTT GAC G 3′
HGC69*	Actinobacteria	5′ TAT AGT TAC CAC CGC CGT 3′
LGC353	Low G + C Gram-positive	5′ AAG ATT CCC TAC TGC T 3′
BACIL1136	Bacillus	5′ AGT CAC CTT AGA GTG CCC AAC T 3′

All probes target the 16*S* rRNA with the exception of probes with * designations, which target 23*S* rRNA.
Source: El Fantroussi S, Urakawa H, Bernhard AE, Kelly JJ, Noble PA, Smidt H, Yershov GM, and Stahl DA (2003) Direct profiling of environmental microbial populations by thermal dissociation analysis of native rRNAs hybridized to oligonucleotide microarrays. *Applied and Environmental Microbiology* 69: 2377–2382.

Fatty Acid Analysis

Although many researchers are using nucleic acid-based analyses, phospholipid fatty acids (PLFA) have also been used extensively as a cultivation-independent method for soil-community analysis. With the exception of Archaea, phospholipids are found in cell membranes of all living organisms and are rapidly degraded upon their death. Thus this analysis can provide a measure of viable community biomass, structure, and at a crude level, composition. Unlike most of the other methods discussed that are genotypic, this is a phenotypic approach. There are many different PLFA that can be identified by gas chromatography after their conversion to fatty acid methyl esters (FAME). The PLFA profile of a microorganism is phylogenetically informative. The mixture of different PLFA profiles extracted from a soil can be used as a fingerprint of the community structure. Changes in community structure can be monitored by following changes in their PLFA fingerprint profiles. There are 'signature' PLFA profiles found in certain groups, which allow the identification of the presence and quantity of PLFA in a community. For example, there are fatty acids specific to fungi, Gram-positive bacteria, Gram-negative bacteria, and plants, which can then be used to calculate diversity of the community at a coarse level. However, there are overlapping fatty acids between microorganisms in a community, therefore this method cannot be used to identify or track specific species. Also, due to the ubiquity of lipids, it is essential that great care is taken during sample handling and processing in the laboratory. The method requires an evaporator to dry samples for storage, a spectrophotometer for biomass measures, and a gas chromatograph for FAME analysis.

Indirect Community Analysis

Polymerase Chain Reaction

PCR is extremely useful in soil-community analysis, because an organism does not need to be cultivated for the procedure to be used. It mechanizes the steps in nucleic acid replication in the laboratory, thus allowing the amplification of a specific nucleotide sequence by approximately 10^6-fold. DNA replication is semiconservative. Each strand of DNA serves as a template, and the newly replicated daughter strand is complementary to the parent strand because of basepair specificity. The nucleotides are added to the growing strand by an enzyme called DNA polymerase. Polymerization begins at a primer site and always occurs by joining the carbon 5 (5′) of one sugar to carbon 3 (3′) of the adjacent sugar (5′ to 3′ direction). Primers are short lengths of nucleotides, oligonucleotides, ranging from 8 to 24 basepairs in length for PCR depending on the desired specificity, which increases with primer length. PCR amplifies a region in the genome that is delimited by the oligonucleotide primers chosen for the initiation of DNA replication (Figure 3). The sequence of steps in the replication of genetic material is controlled by changes to the temperature of the solutions. The temperature of the solution is first raised (usually 94–95°C) to denature the two strands of DNA. Then it is lowered to allow primers to anneal to the template DNA (typical range is 50–65°C). It is then raised to a temperature optimal for activity of the thermal tolerant polymerase (e.g., Taq polymerase is best from 75 to 80°C) to synthesize new strands of DNA. The cycling of temperatures is repeated many times, resulting in amplification of discrete regions of DNA. This technology is used for nucleotide sequencing, synthesis of probes for hybridization experiments, detection of mutations, production of site-specific mutations, and synthesis of genes. As with all techniques, PCR-based methods also have limitations. The primers chosen for PCR can bias the populations amplified. The specificity of the primers is dependent on the sequence information available, thus different results can occur when different primer sets are used. Other limitations to the technique include: overestimates of diversity due to the formation of heteroduplexes, chimeric molecules, or sequence heterogeneities; and underestimation of diversity due to unequal amplification, the resolving power of the gel, or sequences comigrating in the gel. Nevertheless, this approach is very effective if the limitations of the primers are understood and PCR conditions are optimized to minimize errors in the PCR process. All the methods require access to a thermal cycler for PCR and gel electrophoresis units.

Community Richness

The identity of species and a crude measure of richness in a soil community can be determined by nucleotide sequence determination of rRNA genes. The method is independent of cultivation by using PCR directly to amplify rRNA genes from soil DNA extracts. The PCR products are cloned, unique clones identified by restriction fragment length polymorphism (RFLP), and then the nucleotide sequences determined. RFLP is sometimes referred to as amplified ribosomal DNA restriction analysis (ARDRA). The number of unique clones obtained can give an estimation of species richness in systems that are not too complex. The method is labor-intensive therefore an

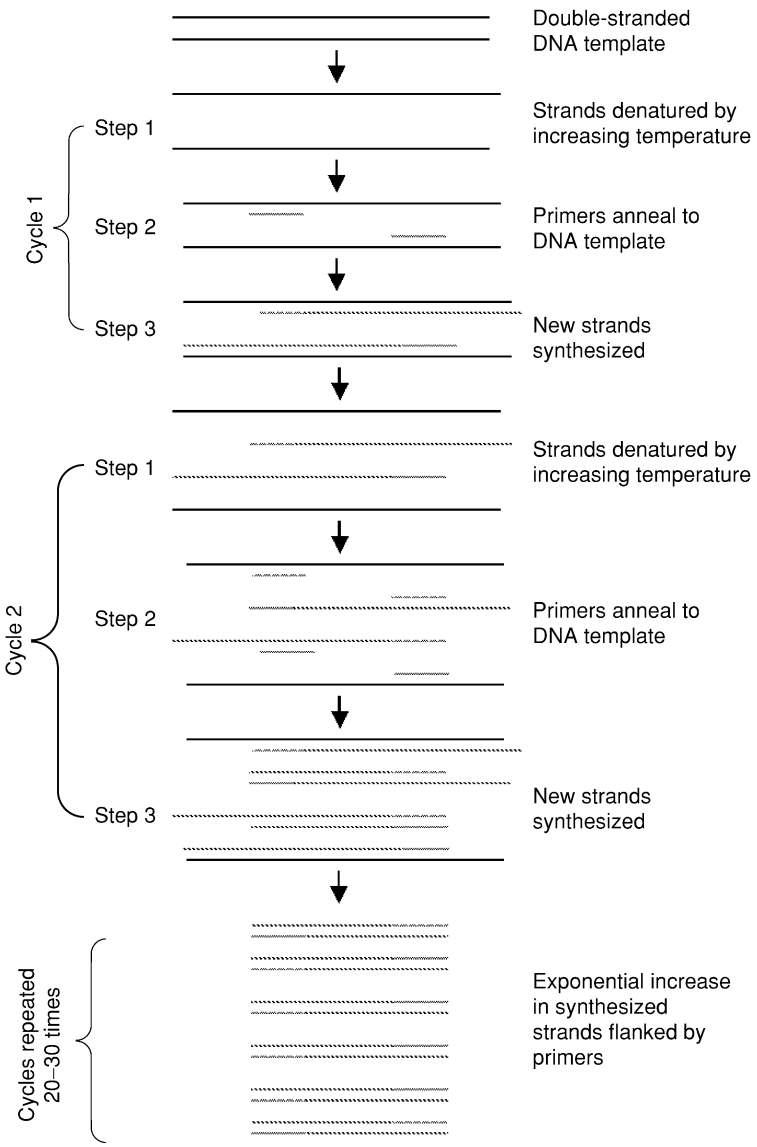


Figure 3 Steps in the polymerase chain reaction.

absolute number of species present in most soils is too time-consuming to obtain; for example, a variety of soils have been examined in which over 100 clones were screened and each clone was unique. However, the use of this approach has added new sequences to the rRNA database, revealed the presence of previously undescribed groups important in soil functions, and been used to determine unique oligonucleotide sequences to track specific strains in the community. For community analysis this method is best used in conjunction with other methods that can reduce the complexity of the sample. This method requires access to nucleotide-sequencing equipment or funds to send samples to a facility. Both are becoming more available to researchers.

PCR-Based Community Structure Analysis

To study differences in soil microbial community structure, researchers are using genetic profiles or fingerprints based on differences in rDNA sequences. The rRNA gene from the community mixture is first amplified by PCR. The primers chosen for PCR can be used to determine the level of discrimination between groups in the community. For example, primers can be used to amplify all Bacteria, or only the Actinomycetes. The products are then distinguished using gel electrophoresis, a means of separating nucleic acids by applying an electric current. The negative charge of the nucleic acids at neutral pHs allows the molecule to travel contained within a solid

matrix, agarose or polyacrylamide gels, through an electric field. The nucleic acids typically separate according to their size, but denaturants can be added that will separate the molecules by their melting nature. The fractionated nucleic acids can be viewed after staining with a dye, e.g., ethidium bromide or SYBR green, and visualized by UV illumination. There are two electrophoresis methods that are being used to distinguish between PCR products of different populations present in a community. One approach, terminal restriction fragment length polymorphism (TRFLP), separates products by size and the other approach, by sequence composition, denaturing gradient gel electrophoresis (DGGE), or thermal gradient gel electrophoresis (TGGE).

Terminal restriction fragment length polymorphism TRFLP analysis is based on size differences of the terminal fragment of the rRNA gene after restriction enzyme digestion. In this method one or both of the primers used for PCR are labeled with a fluorescent molecule or a radioisotope. If both primers are labeled, then different fluorescence colours are used to distinguish the two ends. The PCR products are digested using a frequently cutting enzyme, e.g., *Rsa*I, *Hha*I. The more enzymes that are used to digest the fragments the finer a distinction can be made between numbers of different populations present. However, as the number of enzymes are increased, the time for analysis and interpretation increases. The products are separated in a polyacrylamide gel that can distinguish 1-bp differences in fragment length. The fluorescent fragments can then be visualized and counted. To facilitate the handling of a large number of samples and accuracy of fragment length distinction, automated nucleotide sequencing apparatuses are being used for analysis. Each fragment length represents a group that may consist of one or more populations. The band intensity can be used to quantify the number of PCR products represented by each fragment length, but this does not equate to the numbers of cells, because rDNA copy numbers vary between species. As more rRNA sequence information is accumulated, populations can potentially be identified based on their terminal fragment lengths.

Gradient gel electrophoresis This method uses electrophoresis in a denaturing gradient gel to distinguish PCR products of the same size by differences in their nucleotide sequence composition. There are two techniques that are based on the same principles, but one uses chemical denaturants, DGGE, while the other uses temperature, TGGE. The hydrogen bonds holding the two DNA strands are disrupted under denaturing conditions; the extent of disruption is dependent on the denaturant concentration and nucleotide sequence of the fragment. The extent of strand separation will affect the electrophoretic mobility of the DNA molecules. The different distances that the PCR products migrate can produce a distinct fingerprint profile. The band intensity can be used to quantify the number of PCR products. However, as stated earlier, potential PCR biases and variable rDNA copies should be considered when interpreting the data. In this method it is possible to identify populations represented by each band by nucleotide sequence determination of excised bands or by hybridization with specific probes.

Genetic fingerprinting methods are a useful first step in community analysis because multiple samples can be simultaneously and quickly analyzed to get an indication of richness, then subsequent analytical methods can be chosen if needed. They are an effective technique to discriminate between communities and monitoring dynamics of communities after environmental perturbation. Many soil communities (e.g., agricultural soils) have numerous populations in relatively equivalent proportions, thus resulting in a smear of bands that makes it difficult to differentiate communities (**Figure 4**, lanes 7 and 8). They can be differentiated if dominant bands are seen because of strong selective conditions, e.g., polyaromatic hydrocarbons (PAH), that give certain populations a competitive advantage or the overall richness has been decreased (**Figure 4**, lanes 1–6). Then the number of dominant bands in profiles with low complexity can be counted roughly to estimate richness. Care must be taken, because one band may represent one or more populations. Changes in the microbial community structure can be monitored by the appearance and disappearance of bands in the profiles over time. The presence and absence of bands can be used in quantitative analysis to estimate the similarity between two communities. However, care should be taken when interpreting these data because of the biases of PCR discussed earlier.

Data Analysis

The various methods described here need statistically valid analysis of the data produced. Similarity indices, such as Dice or Jaccard, are currently being used to quantify the relationship between two profiles. These data can be represented by cluster analysis, e.g., unweighted pair group method with arithmetic means (UPGMA), which provides a graphic illustration of the relationship between communities. Principal components analysis (PCA) is often used on large datasets. This method reduces the number of variables, determines whether there is a structure to the relationship between these variables, and can identify

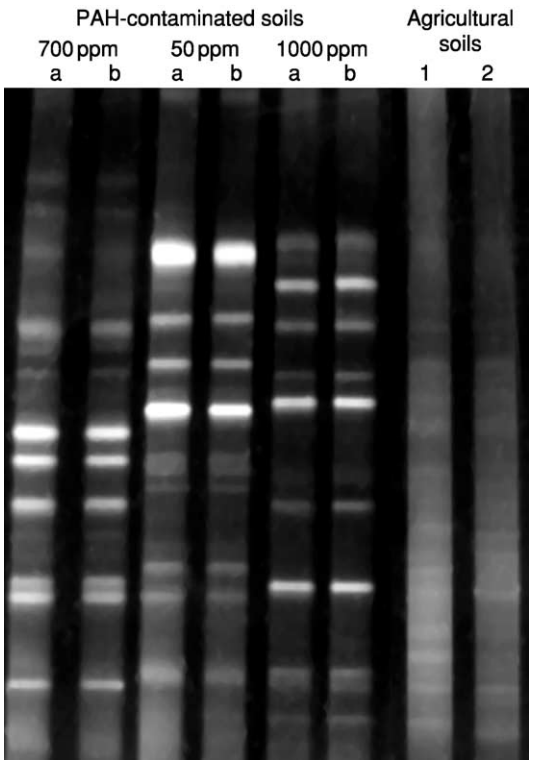


Figure 4 Soil community fingerprint profile of 16*S* rRNA gene V3 region amplified by polymerase chain reaction (PCR) and separated by denaturing gradient gel electrophoresis. Comparison of community profiles from soils contaminated with different concentrations of polyaromatic hydrocarbons (PAHs) to two different agricultural soils (labeled 1 and 2). Lanes a and b are replicate PCR reactions amplified from the same soil DNA extract. Primers used for PCR are bacterial primers BA338F (5′ AC TCC TAC GGG AGG CAG CAG 3′), with a GC clamp attached to the 5′ end (5′ CGC CCG CCG CGC GCG GCG GGC GGG GCG GGG GCA CGG GGG G 3′), and BA518R (5′ ATTACCGCGGCTGCTGG 3′).

main factors contributing to the variation. These techniques just begin to address specific research questions quantitatively, and now research groups are in the process of using these data to construct predictive statistical models of soil microbial communities. It is important that the proper experimental design and appropriate statistical analysis are employed to answer the research question being addressed.

Further Reading

Akkermans ADL, van Elsas JD, and de Bruijn FJ (eds) (1996) *Molecular Microbial Ecology Manual* (including supplement 5). Dordrecht, The Netherlands: Kluwer Academic Publishers.

El Fantroussi S, Urakawa H, Bernhard AE, Kelly JJ, Noble PA, Smidt H, Yershov GM, and Stahl DA (2003) Direct profiling of environmental microbial populations by thermal dissociation analysis of native rRNAs hybridized to oligonucleotide microarrays. *Applied and Environmental Microbiology* 69: 2377–2382.

Gerhardt P (ed.) (1994) *Methods for General and Molecular Bacteriology*. Washington, DC: ASM Press.

Hurst CJ, Crawford RL, Knudsen GR, McInerney MJ, and Stetzenbach LD (eds) (2001) *Manual of Environmental Microbiology*, 2nd edn. Washington, DC: ASM Press.

Johnson AR and Wichern DW (1998) *Applied Multivariate Statistical Analysis*, 5th edn. Upper Saddle River, NJ: Prentice Hall.

Kozdroj J and van Elsas JD (2001) Structural diversity of microorganisms in chemically perturbed soil assessed by molecular and cytochemical approaches. *Journal of Microbiological Methods* 43: 197–212.

Lederberg J (ed.) (2000) *Encyclopedia of Microbiology*, 2nd edn. London, UK: Academic Press.

Maier RM, Pepper IL, and Gerba CP (2000) *Environmental Microbiology*. London, UK: Academic Press.

Theron J and Cloete TE (2000) Molecular techniques for determining microbial diversity and community structure in natural environments. *Critical Reviews in Microbiology* 26: 37–57.

Tiedje JM, Asuming-Brempong S, Nusslein K, Marsh TL, and Flynn SJ (1999) Opening the black box of soil microbial diversity. *Applied Soil Ecology* 13: 109–122.

Torsvik V, Daae FL, Sandaa RA, and Øvreås L (1998) Novel techniques for analysing microbial diversity in natural and perturbed environments. *Journal of Biotechnology* 64: 53–62.

Kinetics

N S Panikov, Stevens Institute of Technology, Hoboken, NJ, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Kinetics (Greek κινηματική, forcing to move) is a branch of natural science that deals with the rates and mechanisms of any process – physical, chemical or biological. Kinetic studies in soil microbiology cover all the quantitative and dynamic aspects of microbial life in the soil: growth and multiplication, mutation, survival and death of soil microorganisms, microbial interactions with the soil environment, other microbes, plants, and animals. Often studies of microbiological kinetics are part of interdisciplinary environmental research on soil processes on which microorganisms play a unique or essential role: organic matter decomposition, biogeochemical cycles of methane, S, N, Fe, P. Kinetics provides a theoretical framework for a better understanding of soil function and for the optimal design of environmental biotechnologies. These technologies are based on the release of beneficial microbes into the environment or activation of the indigenous microbial community and

may benefit research into soil bioremediation, sustainable agricultural systems, and an integrated plant defense system. Recently the kinetics of microbial processes has increasingly focused on the construction of various 'soil modules' in large-scale ecosystem and biosphere models aimed at quantitative prediction (e.g., global climatic changes) and optimal resource management.

Contrary to simple monitoring of soil dynamics (time-series record of the soil state), kinetic studies require an understanding of the mechanisms underlying processes of interest. For example, a complex process such as plant litter decomposition must be viewed as the interplay of several simpler reactions: the conversion of individual litter components, growth of particular degraders, and specific enzymatic transformations. Ideally, mechanistic studies infer the coupling of experimental measurements with simulation mathematical models. These models formalize postulated mechanisms, so that the comparison of observations and the model's predictions allows incorrect hypotheses to be discarded.

Microbial kinetics is tightly linked to growth stoichiometry (Greek στοιχείο, element), which is the quantitative relationship between substrates and the products of microbial growth. In practical terms, stoichiometry mainly addresses problems of a static nature, like 'how much?' and 'in what proportion?' while kinetics considers the dynamics and deals with questions like 'how fast?' and 'by which mechanism?'

The Main State Variables and Growth Parameters

There are two types of kinetic models, deterministic and stochastic. The former describe clear, determined, and regular processes. The latter models deal with random processes. The main variables used in deterministic models are the same as in chemical kinetics – concentrations of cell mass, substrates, and products – while stochastic models consider instead probabilities, frequency distributions, and variance. For example, a deterministic model can describe the process of N_2-fixation as a function of energy and nutrient supply to the rhizosphere, while a stochastic model considers the probability of rhizobia to infect legume roots for a given frequency distribution of bacteria across the soil space. Any real soil process has both deterministic and stochastic components, so ideally the kinetic model should be a 'hybrid' and contain a deterministic core with a stochastic interface, generating probability distributions and confidence corridors rather than single values or curves.

The major state variables of kinetic models describing microbial processes in the soil are summarized in Table 1 and described in detail below.

Amount of Soil Microorganisms

Amount of soil microorganisms can be expressed as biomass (x) or cell numbers (N) of unicellular organisms (bacteria, yeasts, protozoa, fungal spores) per unit soil mass, volume, or surface area. The filamentous organisms (fungi, actinomycetes) are characterized by length (L) of mycelium and number of hyphal tips, n, initiated by hyphal branching. Note that n is equivalent to N, as soon as branching is equivalent to cell division.

The relationship between x and N or L is not straightforward because the mass of individual cells and hypha varies among different organisms and even for the same species depending on growth conditions. Generally, a plentiful supply of nutrients fosters the formation of bigger cells or wider filaments while starvation causes the formation of smaller cells and thinner filaments. To estimate soil biomass from direct microscopic counting, one should generate several arrays of data: the number $\boldsymbol{N} = \{N_1, N_2, \ldots N_n\}$ of cells with specifying biovolume $\boldsymbol{V} = \{V_1, V_2, \ldots V_n\}$, where N_i is the number of cells in the ith volume class V_i and n is the number of such classes, the volumes being calculated from lengths and widths of all detected cell shapes (spherical, ellipsoid, rod-shaped, irregular). For filamentous microorganisms one should have arrays of hyphal length

Table 1 The major dynamic variables used in deterministic kinetic models

Variable	*Symbol*	*Dimension (examples)*
Density/concentration of cell biomass	X	μg cell mass (g soil^{-1}) or g cell mass m^{-2} or μg cell mass (cm^3 soil)$^{-1}$
Cell number[a]	N	10^6 cell (g soil)$^{-1}$ or 10^6 cell (ml soil solution)$^{-1}$
Single-cell mass	m	g dry weight per single cell
Mycelium length[b]	L	m (g soil)$^{-1}$
Tips number	n	10^6 (g soil)$^{-1}$
Concentration of limiting substrate	s	mg (g soil)$^{-1}$, g m^{-2} or g l^{-1} of soil solution
Concentration of product (CO_2, CH_4)	p	mg (g soil)$^{-1}$, g m^{-2} or g l^{-1} of soil solution

[a]For unicellular organisms (bacteria, yeasts).
[b]For filamentous organisms (fungi, actinomycetes).

$\boldsymbol{L} = \{L_1, L_2, \ldots L_m\}$ and corresponding hyphal diameters $\boldsymbol{D} = \{D_1, D_2, \ldots D_m\}$. The total soil biomass, x, can be calculated as a sum

$$x = \boldsymbol{V} \times \boldsymbol{N} \times (1 - \Theta) + 1/4\pi \boldsymbol{D}^2 \times \boldsymbol{L} \times (1 - \Theta) \quad [1]$$

where Θ is the water content in wet cell mass (usually taken from lab data).

Advanced microscopic (ultraviolet, electron, and atomic force microscopy) and computer-aided image-analysis techniques make cell counting and sizing less tedious. Additionally, fluorescence *in situ* hybridization (FISH) allows targeting of specific taxonomic or functional groups based on the use of RNA-oligonucleotide probes. The total soil biomass, x, can be quantified by extracting from soil samples and quantifying one of the following integral and unique cell constituents: (1) DNA; (2) adenosine triphosphate (ATP); (3) microbial phospholipids; (4) microbial cell wall components (chitin for fungi and muramic acid for bacteria); or (5) the sum of labile/extractable C released after soil fumigation with CCl_3.

Selection of x or $N(L)$ depends on the specific research question. Biomass, x, has the obvious advantage in studies of carbon and nutrient cycling, while cell number, N, is preferred in population studies when, say, mutation or plasmid transfer is crucial for an understanding of the studied soil processes.

Substrate Concentration

Substrate concentration in the soil, s, stands for the amount of some essential soil constituents used by soil microbes for growth and maintenance. There are two groups of such constituents: (1) catabolic substrates, which are sources of energy; and (2) anabolic or conserved substrates, which are sources of biogenic elements forming cellular material. Examples of catabolic substrates are H_2 for lithotrophic hydrogen bacteria, NH_4^+ and NO_2^- for nitrifying bacteria, sulfur for sulfur-oxidizing bacteria, and oxidizable or fermentable organic substances for heterotrophic species. Their consumption is accompanied by oxidation and dissipation of chemical substances into waste products which are no longer reusable as an energy source (H_2O, NO_3^-, SO_4^{2-}, CO_2, etc.). (Fermentation products like acetate, ethanol, butyrate, H_2 etc. seem to be an exception as they do contain reusable oxidation potential, but it is not available under anaerobic conditions supervising fermentation.) After uptake, anabolic substrates are incorporated into *de novo* synthesized cell components, being conserved in biomass; thus, sometimes anabolic substrates are referred to as conserved substrates. In contrast to catabolic substrates, the conserved substrates can be reabsorbed after excretion or cell lysis. They include nearly all noncarbon sources of biogenic elements (N, P, K, Mg, Fe, and trace elements), CO_2 for autotrophs, as well as amino acids and growth factors.

Two variables, cell mass x and substrate concentration s, are linked via stoichiometric parameter growth yield:

$$Y = -\frac{dx}{ds} \approx -\frac{\Delta x}{\Delta s} \quad [2]$$

where Δx is the increase in microbial biomass consequent on utilization of the amount Δs of substrate. Dividing both parts of eqn [2] by $x\delta t$, gives the relationship between growth rate and substrate production:

$$Y = -\frac{\delta x}{\delta s} = -\frac{\delta x}{x\delta \mathrm{t}} : \frac{\delta s}{x\delta \mathrm{t}} = -\mu/q \quad [2a]$$

where μ is specific growth rate and q is specific rate of substrate consumption.

The reason for Y variation is different for catabolic and anabolic substrates. In the case of energy sources, some fraction of the total substrate flux is diverted from growth *per se* to meet so-called maintenance functions, such as resynthesis of self-degrading cell proteins, nucleic acids and other macromolecules, osmotic work to keep the concentration gradient between cell interior and environment, and cell motility. The energy balance is described by the following equation:

$$\begin{array}{lcll} \text{Total energy source consumed} & & \text{Consumption for growth} & \text{Consumption for maintenance} \\ q & = & \mu/Y^{\max} & + \quad m \end{array} \quad [3]$$

where m is the maintenance coefficient or the specific rate of catabolic substrate consumption by nongrowing cells (i.e., $m = q$ when $\mu = 0$). (With some rare exceptions (fungal exospores and bacterial cysts), the microbial cells are not stable at $\mu = 0$ and either grow ($\mu > 0$) or decay ($\mu < 0$). Therefore the maintenance coefficient is found by linear extrapolation of a series of $q(\mu)$-measurements to the point where μ is zero.) Under chronic starvation the maintenance coefficient, m, becomes progressively smaller in contrast to higher values associated with intensive growth; as a result, when $\mu \to 0$, the growth yield Y approaches a low, nonzero value $Y^{\min}$ (**Figure 1**).

There is also wasteful oxidation of substrate under at least three specific circumstances: (1) when growth is nutrient-limited and energy-sufficient; (2) when starving cells are brought to rich nutrient medium (famine-to-feast transition); and (3) when cells are exposed to uncoupling inhibitors. In all these cases,

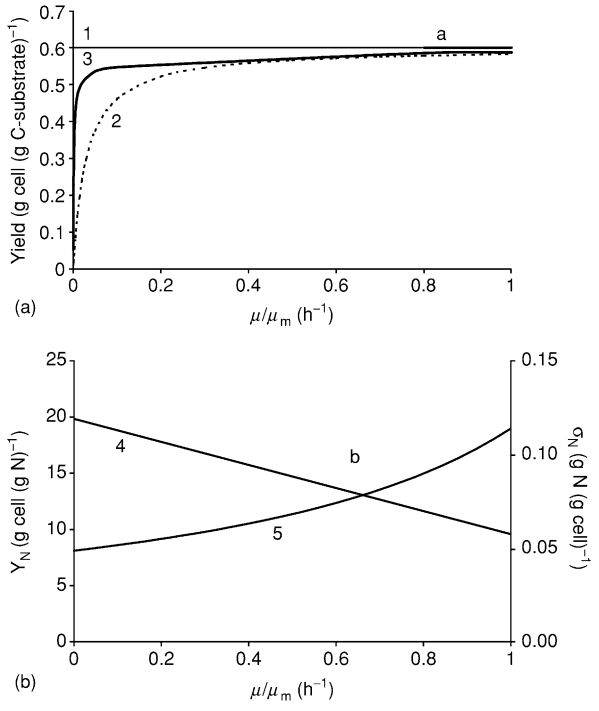


Figure 1 Variation of the cell mass yield on catabolic (a) and anabolic (b) substrates related to specific growth rate of soil microorganisms. Curves calculated assuming no maintenance (1), constant maintenance coefficient (2), and adaptive decrease in maintenance in slowly growing cells (3), Curve 4 is cell yield per g N, and curve 5 is cell quota, g N/g cell. Curves 3–5 were calculated using SCM.

the cell's catabolic machinery produces more energy that can be used for ATP generation. Such wasteful catabolism should be differentiated from maintenance *per se*.

Cell yield on anabolic substrates varies mainly as a result of alterations in biomass chemical composition as expressed by the parameter σ_s, the intracellular content of deficient element or cell quota. The variation in N content in bacteria from 5 to 15% gives the σ_s diapason 0.05–0.15 g N (g cell mass)$^{-1}$. For most known cases, the quota σ_s increases in parallel to growth acceleration, because the higher growth rate requires higher intracellular content of proteins and RNA (containing N, P, S) as well as K^+, Mg^{2+}, and vitamins participating in all primary metabolic reactions. The yield and cell quota are inversely related to each other, e.g., low N-content ($\sigma_N = 0.05\,g\,N\,g^{-1}$) corresponds to high cell yield ($Y_N = 1/\sigma_N = 20$ g cell per g of N utilized), high N-content in rapidly growing cells can be attained only with low cell yield ($Y_N = 1/0.15 = 6.67$ g cell per g N) (Figure 1).

Products

There may be intermediate and end metabolic products accompanying microbial growth. The typical intermediates are organic acids forming in glycolysis/Etner–Doudoroff pathways or various fermentations:

1. glucose → 2 pyruvate$^-$ + 2H
2. glucose → gluconic acid → pyruvate
3. pyruvate → acetate

The only end product of aerobic decomposition is CO_2, whereas products formed under anaerobic conditions include organic acids, alcohols and ketones, CO_2, CH_4, and various reduced inorganic compounds (S^{2-}, Fe^{2+}). The CO_2 evolution rate is probably the most popular parameter to characterize soil biological activity due to the wide availability of gas analyzers (gas chromatography, infrared, and mass spectrometry). In particular cases of aerobic decomposition, the measured CO_2 evolution is a significant fraction of the C budget, as described by the following equation:

$$\begin{array}{cccccccc} \text{Total carbon consumed} & & \text{C incorporated into cell} & & \text{C oxidized to } CO_2 \text{ to provide energy} & & \text{C in extracellular products} \\ \Delta C - s & = & \Delta C - x & + & \Delta C - CO_2 & + & \Delta C - p \end{array} \quad [4]$$

Dividing by time increment δ/t gives:

$$\begin{array}{cccccccc} \text{Uptake of C-substrate} & & \text{Microbial growth} & & \text{Respiration} & & \text{Products formation} \\ \dfrac{\delta s}{\delta t} & = & \dfrac{\delta x}{\delta t} & + & \dfrac{\delta [CO_2]}{\delta t} & + & \dfrac{\delta p}{\delta t} \end{array} \quad [4a]$$

Thus, the microbial consumption of C-substrate is the sum of growth, respiration, and product formation. Based on such C-balance, we can reconstruct the missed variables. For example, cell mass x in a laboratory soil incubation study can be calculated as the difference between the amount of initially added substrate and current s, p, and CO_2 values. (The priming effect of added organic matter on decomposition of soil humus must be evaluated independently – it is usually small enough. The priming is fully excluded when using labeled amendment.) However, it would be incorrect to identify the CO_2 evolution rate alone with total soil decomposition, as some degraded C can be (and really is) incorporated into soil organisms or their extracellular products.

The Models of Microbial Growth

Exponential Growth

Under favorable growth conditions (temporary excess of nutrient substrates, absence of inhibition), the bacterial growth rate should be proportional to

the instant cell mass, x, the quotient μ remaining constant:

$$\frac{\delta x}{\delta t} = \mu x \qquad [5]$$

The integration of eqn [5] at initial condition $x = x_0$ at time $t = 0$, gives the exponential equation:

$$x = x_0 e^{\mu t} \quad \text{or} \quad \ln x = \ln x_0 + \mu t \qquad [5a]$$

The rate of CO_2 or CH_4 production (gas production rate: GPR) is proportional to the growth rate of the respective group of soil microorganisms:

$$\underset{\frac{\delta p}{\delta t}}{\text{GPR}} = \underset{\frac{1}{Y}}{\text{Yield factor}} \times \underset{\frac{\delta x}{\delta t}}{\text{Cell growth}} = \frac{\mu x}{Y} = \frac{\mu x_0}{Y} e^{\mu t} = A e^{\mu t} \qquad [6]$$

where p is gaseous product concentration, expressed for simplicity in the same units as biomass, say mg C per g soil. Integration of eqn [3] under initial condition $p = 0$ at $t = 0$ gives:

$$p = \frac{x_0}{Y}(e^{\mu t} - 1) = B \times (e^{\mu t} - 1) \qquad [6a]$$

Comparing eqns [5a] and [6a], we see that x and p dynamics are not identical, although both imply exponential cell growth; the $p(t)$ curve deviates from the exponential law, especially at small t. In particular the '$\ln x - t$' plot is linear, but '$\ln p - t$' is not. As t increases, the difference becomes smaller and smaller.

In contrast to the exponential growth equation described above, the logistic equation sets an upper limit for microbial growth by empirically assuming that the specific growth rate μ is dependent on microbial concentration:

$$\frac{\delta x}{\delta t} = \mu(x)x = \mu x\left(1 - \frac{x}{x_m}\right) \qquad [7]$$

where x_m is the upper limit of cell mass. In ecological and demographic literature, x_m is called the carrying capacity of ecosystem, K. This equation produces the S-shaped growth curve frequently observed in nature. However, this equation is no more than empiric, and parameters μ and x_m are not 'fundamental,' i.e., they cannot be measured and tabulated for particular soils or ecosystems or predicted before actual observation.

Monod Model

This model, developed by the famous French scientist in 1941, is still very popular due to its elegant simplicity. There are two basic premises: (1) the specific growth rate, μ, is dependent on the concentration of one nutrient component which happens to be the most deficient, the so-called limiting substrate: and (2) biomass formation is linked to substrate uptake by mass-conservation conditions (eqn [2a]). Therefore, according to these premises:

$$\frac{\delta x}{\delta t} = \mu(s)x = \mu_m \frac{s}{K_s + s}x$$
$$\frac{\delta s}{\delta t} = -\frac{1}{Y}\frac{\delta x}{\delta t} \qquad [8]$$

Eqn [8] contains three parameters: yield Y, maximal specific growth rate μ_m, and saturation constant K_s (substrate concentration at which $\mu = 0.5\ \mu_m$), which can be thought of as the 'ID' for particular organisms and used to predict their growth dynamics.

Remarkably, this model was used to develop a chemostat theory before actual experiments with continuous culture had been undertaken, one of the very rare events in the history of mathematical biology! The model predicted a number of counterintuitive features of the chemostat, e.g., that specific growth rate μ can be set up by an experimentalist by changing the medium flow at any values between 0 and μ_m (previously, exponential growth was believed to occur only at $\mu = \mu_m$) and that μ-values do not depend on the feed-substrate concentration and are governed solely by the residual substrate concentration in the culture.

However, the Monod model fails to explain a number of essential growth phenomena observed experimentally, namely: lag-phase, death of starving cells, product formation and any kind of adaptive changes in microbial population, such as induction-repression of enzymes, yield variation, and changes in the cell RNA content. These gaps were filled in by so-called structured models.

Structured Models

Structured models explicitly describe variations in cell composition. They usually include mass balance equations not only for external substrate(s) but also several intracellular components, C_1, C_2, ... C_n. For each variable C_i a differential equation is written which takes into account all sources, r_+, and sinks, r_-, as well as its dilution due to cell mass expansion (growth):

$$\frac{\delta C_i}{\delta t} = r_+(s, C_1 \ldots C_n) - r_-(s, C_1 \ldots C_n) - \mu C_i \qquad [9]$$

The earliest structured models accounted for no more than 3–5 cell constituents, e.g., total cell proteins, RNA and DNA, reserved polysaccharides, and ATP-pool. Modern models contain up to hundreds

and even thousands of internal variables taken directly from available genomic/metabolic databases. Currently the challenge is to create a virtual cell (so-called cell *in silico*) that is complex enough to function as a real cell. However the predictive capability of these intricate models is rather modest: they are still a 'caricature parody' of the real cell, but too complex to be efficiently studied mathematically (stability analysis, parameters identification) or to improve our understanding of the biosystem. The best choice of a mathematical model lies, apparently, midway between the unstructured and highly structured models outlined here. One of the known examples is the synthetic chemostat model (SCM).

Synthetic Chemostat Model

According to the SCM, microbial growth occurs by converting exosubstrate, S, into a number of cell macromolecules, X', via a pool of intermediates, L, part of which are respired to CO_2 (Figure 2).

Macromolecular cell components are susceptible to degradation (turnover), and intermediates L can leak out. The array X', the cell composition, is not fixed and varies in response to the changing environment. The heart of the SCM is the solution of the problem: how to characterize these variations without going too deeply into details?

For this purpose, all the macromolecular cell constituents are divided into two groups: Primary cell constituents necessary for intensive growth (P-components), and components needed for cell survival under any kinds of growth restriction (U-components). The content of P-components (ribosomes and all enzymes of the primary metabolic pathways) increases in parallel with growth acceleration. The contribution of U-components under good growth conditions decreases (to comply with conservation conditions $P + U = \text{constant}$), and attains the maximum under chronic environmental stress to improve cell resistance. Typical U-components are enzymes of secondary metabolism, protective pigments, reserved substances, and transport systems of high affinity. An interdependent variation of individual P- and U-components is approximated by linear function of some master variable r:

$$\mathrm{P}+\mathrm{U}=\begin{vmatrix}P_1\\P_2\\\cdots\\P_n\end{vmatrix}+\begin{vmatrix}U_1\\U_2\\\cdots\\U_m\end{vmatrix}=\begin{vmatrix}P_1^{\min}\\P_2^{\min}\\\cdots\\P_n^{\min}\end{vmatrix}+\begin{vmatrix}U_1^{\min}\\U_2^{\min}\\\cdots\\U_m^{\min}\end{vmatrix}+r\begin{vmatrix}P_1^{\max}-P_1^{\min}\\P_2^{\max}-P_2^{\min}\\\cdots\\P_n^{\max}-P_n^{\min}\end{vmatrix}+(1-r)\begin{vmatrix}U_1^{\max}-U_1^{\min}\\U_2^{\max}-U_2^{\min}\\\cdots\\U_m^{\max}-U_m^{\min}\end{vmatrix} \quad [10]$$

where $P^{\max}$, $U^{\max}$, and $P^{\min}$, $U^{\min}$ are respectively upper and lower boundaries for P and U adaptive variations, and r is the scalar function, not array! The r-value depends directly on environmental factors, e.g., on the limiting substrate concentration and is included in all kinetic expressions (q_s, m, μ) to simulate the combined effects of current environmental factors and the cellular physiological state, as determined by past growth conditions.

The predictive and explanatory capabilities of SCM are higher than that of more complex structured models. It usually gives good agreement with most available experimental steady-state and transient growth data, such as batch culture from lag- to decline phase or continuous culture after abrupt changes in dilution rate, pH, or temperature. For ecological applications, SCM realistically describes and predicts not only growth *per se* but also many other dynamic phenomena: survival dynamics under starvation, formation of dwarf cells under growth restriction, the adaptive adjustment in cell maintenance requirements, variation of growth potential and affinity to substrate, and utilization of substrate mixture. Below we will use SCM to explain microbial growth kinetics both in culture and in the soil.

Growth Kinetics *In Situ* and *Ex Situ*

Soil as a microbial habitat is a system of relatively independent microsites, each of which can be drastically dissimilar in the amount and nature of growth substrates, aeration level, texture, and moisture. However, there are several general features which

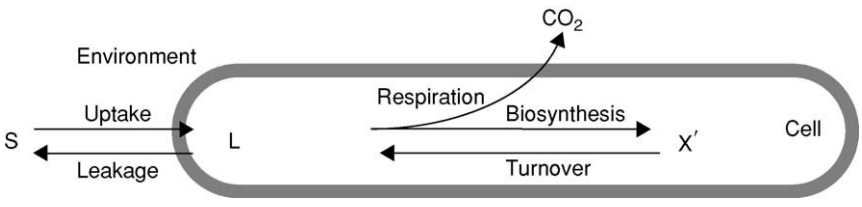


Figure 2 Flow diagram of major cell processes used to stimulate microbial growth by basic SCM.

Table 2 Matrix of growth patterns *in situ* and *ex situ*

	Substrate input			
	Continuous (1)		*Single-term (2)*	
Spatial organization	*Cell removed (α)*	*Cell retained (β)*	*Cell removed (α)*	*Cell retained (β)*
Homogeneous (a)	1aα	1aβ	2aα	2aβ
Heterogeneous (b)	1bα	1bβ	Forbidden combination	Forbidden combination

are most essential for determining the population dynamics of microorganisms occupying a respective microsite:

1. Substrate input may be continuous (1) or discontinuous (2).
2. Growing cells can be retained within a microhabitat (α) or eliminated (β). Examples of elimination include: cell washout, predation (consumption of microbial cells by protozoa, nematodes, microarthropods) or lysis by parasites, active migration due to taxis (motile bacteria) and tropism (vectorized apical growth of hyphal organisms).
3. Habitats vary in spatial organization, being (a) homogeneous and (b) heterogeneously organized. In the first type of system, there is an even or at least random distribution of cells, substrate, and metabolic products in space, whereas in the second type of habitat there are gradients in x, s, and p. The homogenization in soil occurs due to tillage and fossorial animals. Homogeneous microenvironments are also formed when the growth substrates are mobile: gases and volatiles.

These three independent characteristics of ecotopic conditions produce eight possible combinations (Table 2): two of them are logically forbidden because any kind of spatial heterogeneity implies continuous rather than momentary single-term delivery of substrate from some parts of habitat across the *s* concentration gradient. Figure 3 depicts several characteristic microbial habitats occurring in the soil. The most spacious and poorest/oligotrophic habitat is called the dispersion zone (1aβ); the subsoil and patches of bare topsoil are devoid of plant roots or fresh litter. Numerous microbial populations inhabiting these habitats grow very slowly on volatile or readily soluble compounds, which continuously diffuse from other soil loci where monomeric concentration is high due to intensive decomposition (plant litter) or excretory activity of plants (rhizosphere). The lack of elimination (no motility in the majority of soil bacteria and no predation due to low prey density) combined with slow but uninterrupted continuous growth eventually results in a significant buildup of a half-dormant cell mass.

The highest and most continuous microbial activity is localized in the soil litter layer, rhizosphere, and digestive tract of soil animals (habitats of the types 1aβ–1bβ). The C-substrates for microbial populations are the monomeric labile compounds (sugars and organic acids) derived from plants either as root exudation in the rhizosphere or released by extracellular hydrolytic enzymes from the lignocellulose and other polymeric material in plant litter. The litter layer on the soil surface is formed from deposition of aboveground plant remnants, while belowground plant senescence (root litter and root sloughing) provide microbial C-substrates in the rhizosphere. Root exudation is closely related to plant photosynthesis and changes diurnally, whereas hydrolytic release is monotonous. Spatially, all these habitats are rather heterogeneous; however, the random distribution of microloci combined with macroscopic sampling size allows us to use homogeneous kinetic models. Clear vertical gradients are formed in the litter layer from uncolonized fresh plant debris on the top to highly decomposed sublayer at the interface with mineral soil. At the microscale, the spatial heterogeneity of microbial colonies (mainly fungi) is manifested in its differentiation into growing extension zone and the nongrowing reproductive compartments (Figure 4). Within the rhizosphere, there are also several spatial gradients of different scales: (1) the vertical gradient of the root phytomass, which reflects the spatial pattern of belowground allocation of photosynthate; (2) the horizontal gradient between distant trees or tussocks; and (3) the micro-scale gradient around root hair with maximal concentration of microbial cells and substrates on the root surface (rhizoplane) and exponential decline outwards. (Some researchers found a statistically significant decline in cell density of rhizosphere bacteria on the plant surface, probably due to excretion of antibiotic compounds by plants.)

Soil millipedes, isopods, some earthworms and other primary decomposers inhabit the litter layer and feed on plant debris. The ingested lignocellulose material is mechanically disrupted by the mandible,

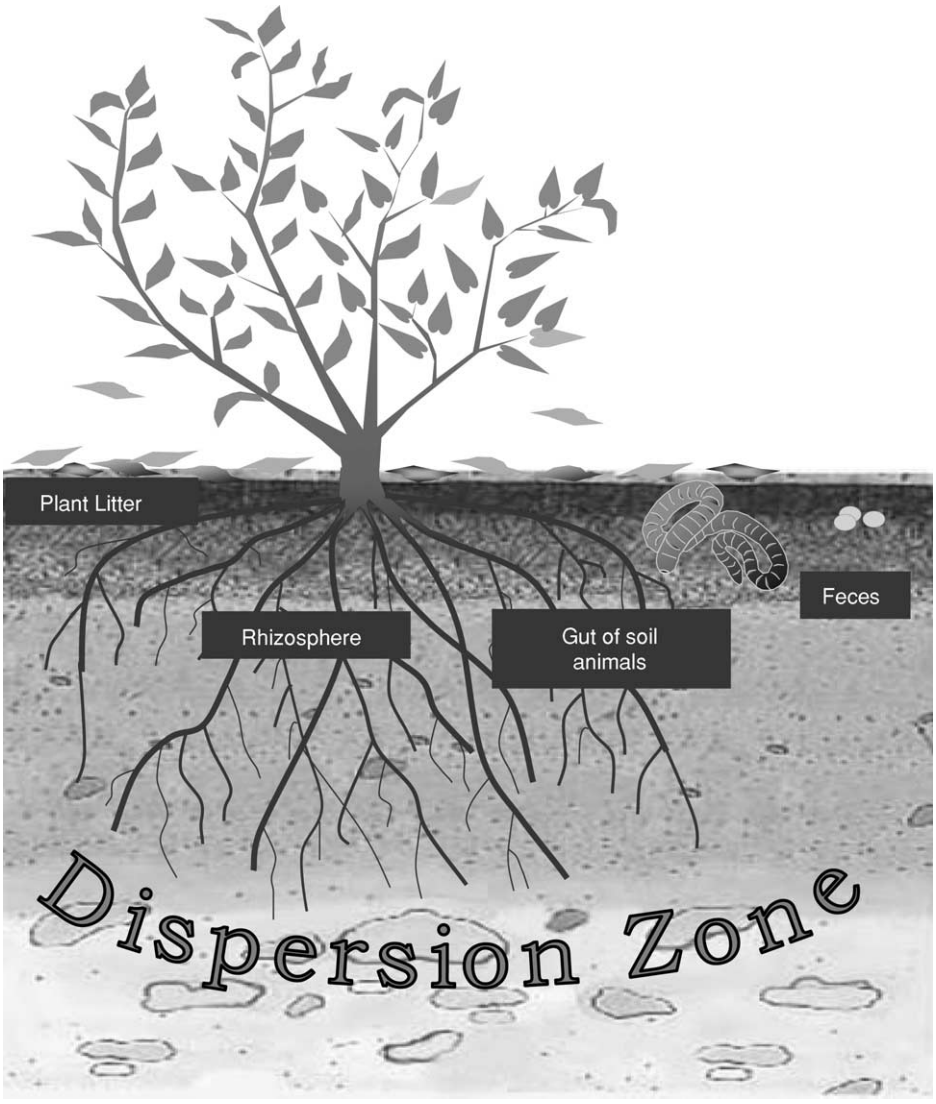


Figure 3 The principal microbial habitats in the soil (see text for explanation).

moistened with saliva, and then passed to the mid- and hindgut. The digestive tracts of various soil invertebrates harbor not only specific symbionts, but also the normal free-living microorganisms occurring in soil or plant litter, e.g., *Pseudomonas*, *Flavobacterium*, *Vibrio*, *Enterobacter*, and *Streptomyces*. The acceleration of their growth in the hindgut is due to favorable conditions: neutral pH, optimal moisture, elevated concentrations of nutrient and growth factors (amino acids, peptides, vitamins), as well as the continuous input of fresh substrate and concomitant removal of digestion products (glucose) to prevent the negative feedback (so-called catabolic repression) on cellulase synthesis. Due to peristaltic motion the content of the gut is mixed and homogeneous. Secondary decomposers (i.e., the earthworms *Allolobophora chlorotica*) feed on amorphous humus containing bacteria and fungal mycelia; they eliminate some microbial species and greatly stimulate growth of others.

Discontinuous, explosive microbial growth occurs within hot spots initiated by sudden increases in the availability of organic substrates/nutrients in the soil (2aα–2aβ), such as feces and carcasses of animals, rain washing of organic matter from plant foliage, dry–rewet or freeze–thaw cycles, manure application, and soil fumigation. Growth is usually accompanied by elimination in the form of grazing, myco- and bacteriolytic activity, as well as the active migration of microbial cells.

The dynamics of microbial growth in any type of habitat is described by the mass balance equations:

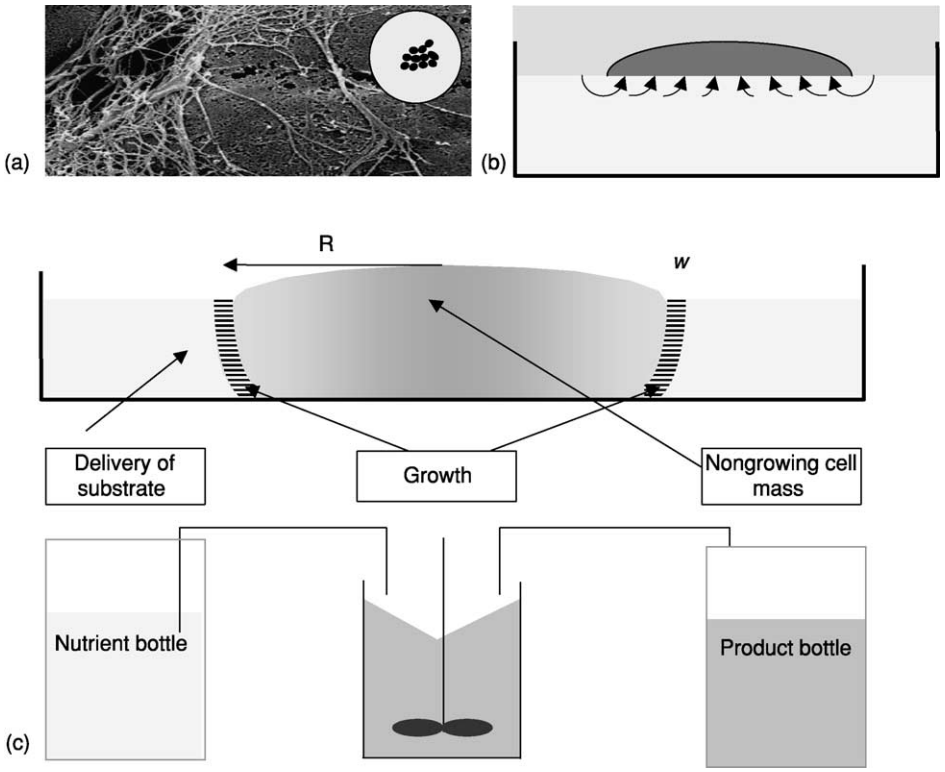


Figure 4 The colony growth: (a) in the soil; (b) bacterial growth on agar plate; (c) fungal growth in nutrient agar. Note that bacteria grow only on the agar surface and the colony expansion is controlled by diffusion of substrates from outside the colony. The fungi and actinomycetes are able to penetrate into the depth of the agar layer, so their mycelium expansion is not dependent on nutrient diffusion within agar layer. The fungal colony follows a chemostat-type growth pattern, being: (1) continuous; (2) steady-state; and (3) limited by substrate availability. The role of the fermentation vessel of the conventional chemostat is played by the peripheral zone of the colony; the product bottle is analogous to the central part of the colony, while the pump is substituted by the chemotropic movement of hyphae tips along the substrate concentration gradient.

$$\begin{array}{lcccccccc} \text{Net change} & & \text{Substrate input} & & \text{Substrate loss} & & \text{Uptake for growth} & & \text{Uptake for maintenance} \\ \dfrac{\delta s}{\delta t} & = & F(t) & - & G(s) & - & \dfrac{\mu(s)x}{Y} & - & mx \\ \text{Net change} & & \text{Immigration} & & \text{Elimination} & & \text{Growth} & & \\ \dfrac{\delta x}{\delta t} & = & V(t) & - & H(x) & + & \mu(s)x & & \end{array} \quad [11]$$

To specify kinetic expressions, we can rely on kinetic models developed for the laboratory cultivation system which is closest to the respective soil habitat (Table 3). For instance, growth of soil bacteria in the intestine of earthworms is remarkably similar to a tubular (plug-flow) reactor, the rhizosphere community is related to a chemostat or phased-batch culture with injection of the growth-limiting substrate once per day (simulation of photosynthesis-dependent root exudation), the slow microbial growth in dispersion zone is adequately simulated by dialysis culture, and growth of fungi degrading plant litter contains kinetic patterns similar to colonies developing on agar.

Growth models based on SCM for various homogeneous cultivation systems are shown in Figure 5, and colonial growth is illustrated in Figure 4. One of the best criteria of adequacy of the selected cultivation system is specific growth rate, μ, of the studied microbial population. If the values of μ for *ex situ* and *in situ* populations coincide, then the cultivation technique may be regarded as appropriate.

Diversity of Growth Characteristics: Life Strategy Concept

In soils, there are always multiple microbial populations performing the same function. They have different growth parameters and regulators, but they require and compete for the same nutrient substrates. Their coexistence, which contradicts the well-known principle of competitive exclusion, can be explained by at least two factors:

1. spatial segregation of soil microsites, which leads to a lack of direct interactions between potentially competing species in the soil

Table 3 Examples of typical soil habitats and their counterparts in laboratory cultivation systems

		Soil habitats		
Kinetic type	*Lab cultivation system*	*Locus*	*Substrate flux*	*Elimination factor*
1aα	Chemostat, turbidostat, pH-auxostat	Rhizosphere	Rhizodeposition: exudation and sloughing	Root extension, predation by protozoa and nematodes
1aβ	Fed batch culture, dialysis culture	Dispersion zone	Diffusion of gases, volatile and soluble organic matter	No or very weak
1bα	Plug-flow, colonies on plate	Digestion tract of invertebrates	Predigested plant litter	Excretion
1bα	Colonies, solid-phase fermentation	Plant litter	Litter fall, root sloughing	Predation by microarthropods, apical hyphal extension
1bβ	Column packed with microbe attached, biofilms	Vertical network of the soil water channels	Trickling solutions	Cell motility
2aα	Phased culture	Temporary hot spots of microbial activity	Excrements and corpses soil animals, manure	Predation/grazing and exposure to harmful factors
2aβ	Simple batch	Hot spots without intensive microbial grazing	Washout of labile organic matter with rainwater	No or weak predation[a]

[a]Low probability of intense predation if no buildup of microbial biomass (prey population density) occurs.

2. perpetual environmental changes which cause variation in the direction/intensity of natural selection, so the fittest competitor today can be different from the day before, none of them being excluded completely

Studying microbial communities from a kinetic perspective can describe quantitatively why the pressure of natural selection can differ. One can distinguish the following three types of natural selection which correspond to three types of life strategy:

1. *K*-selection for high competitive power ('lions'); operates in climax ecosystems
2. *r*-selection for reactivity and ability to recover after perturbations ('jackal'); operates on the pioneer stages of succession
3. *L*-selection for resistance to stress-factors ('camel'); operates under adverse environmental conditions (**Figure 6**)

Figure 7 summarizes results of kinetic studies of the typical microbial *r*-, *K*-, and *L*-strategists. The *Enterobacteria*, *Pseudomonas*, Baker and Fodder yeast are mainly *r*-selected organisms. Their dynamic behavior is erroneously considered to be typical for all microbes: rapid and balanced growth, short lag-periods, and smooth transitory processes. They dominate in those soil habitats that are frequently 'rejuvenated' to the pioneer succession stage: hot spots of substrate amendment, animal guts and feces, rhizosphere with diurnal fluctuations in exudation rates and perpetual changes in 'address' of exudation loci due to apical extension of root hair. Plating is a good simulation of such hot spots; that is why *r*-selected species are easily isolated from soil.

K-selected bacterial species are much less amenable to isolation and cultivation. Probably most unculturable microbial species belong to this type of life strategy. When cultivated under artificial laboratory conditions, they are fastidious and unpredictable (a 'lion' in captivity). The best option for their cultivation is continuous culture with cell retention: fed-batch, dialysis culture, or batch culture with C-substrate delivered via gas phase (volatile C-substrates, such as ethanol or volatile fatty acid). Under these cultivation conditions the *K*-selected species display high yield and almost 100% viability, even at extremely slow cell division (generation time up to months). In other cultivation systems, their growth patterns are characterized by: (1) biphasic growth on the single substrate in batch culture (false diaxia); (2) long survival under starvation; and (3) oscillations and bistability (two or more steady states at one dilution rate) in chemostat. To simulate mathematically the described abnormal behavior of *Arthrobacter* and other oligotrophic species, the SCM was elaborated to include intermediates with autoinhibition functions (e.g., peroxides as respiratory by-products) and the possibility of direct incorporation of deficient C-substrate to a pool of reserved compounds (**Figure 7**). The main feature of growth control in oligotrophic species is the relative autonomy/independence of several metabolic domains: transport, respiration, and biosynthesis.

Many *L*-selected microorganisms (bacilli, actinomycetes, some fungi) share the following common

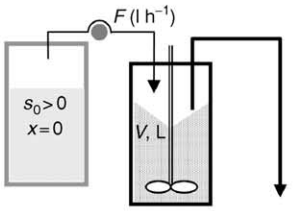
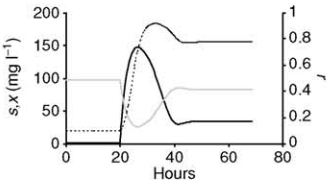
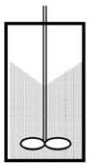
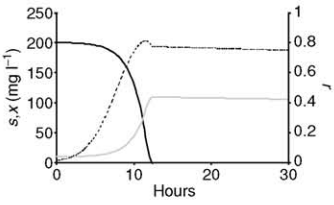
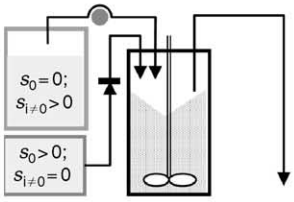
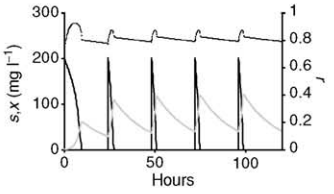
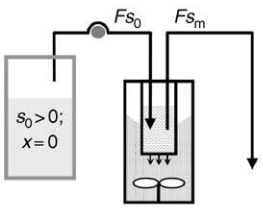
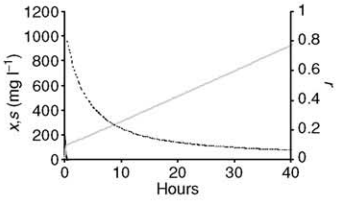
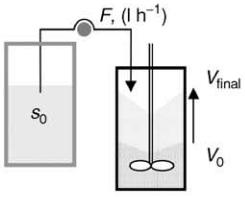
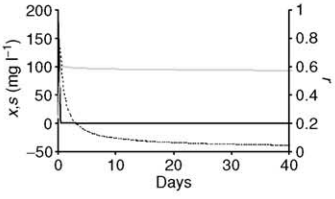
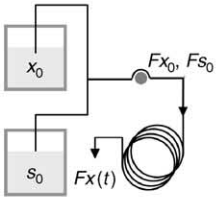
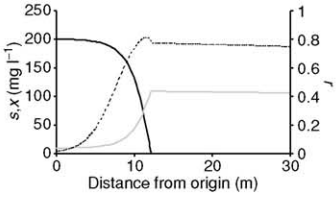
Cultivation method	*Design*	*Equations of mass balance*	*Growth dynamics*
Chemostat		$\frac{\delta s}{\delta t} = D(s_0 - s) - q_s x$ $\frac{\delta x}{\delta t} = \mu x - Dx$ $\frac{\delta r}{\delta t} = \mu\left(\frac{s}{K_r + s} - r\right)$ [12] $q_s = rQ\frac{s}{K_s + s}$; $D = F/V$ $\mu = Y[q_s - mr]$	
Batch		Dilution rate $D = 0$ and $s = s_0$ at $t = 0$; the rest is according to eqn [12]	
Phased culture		D = constant $s_0 > 0$ at $t = i\Delta t$, where $i = 0, 1 \ldots n$, and $s_0 = 0$ at other t; the rest is according to eqn [1]	
Continuous dialysis		$\frac{\delta s}{\delta t} = \frac{F}{V_m}(s_0 - s_m) - q_S x$ $\frac{\delta x}{\delta t} = \mu x$ [13] The rest is according to eqn [12]	
Fed-batch culture (FBC)		$D = F/(V_0 + Ft)$ The rest is according to eqn [12]	
Continuous plug-flow		$\frac{\partial s}{\partial t} + f\frac{\partial s}{\partial z} = D\frac{\partial^2 s}{\partial z^2} - q_s x$ $\frac{\partial x}{\partial t} + f\frac{\partial x}{\partial z} = D\frac{\partial^2 x}{\partial z^2} - \mu(s)x$ The rest is according to eqn [12]	

Figure 5 The major cultivation systems relevant to the purposes of microbial ecology. *s*, *x*, and *r*, limiting substrate concentration, biomass, and variable of physiological state; q_s, specific rate of substrate uptake; *mr*, maintenance requirement; *m*, specific growth rate; *V*, reactor volume; *z*, distance along tubing reactor in the plug-flow system.

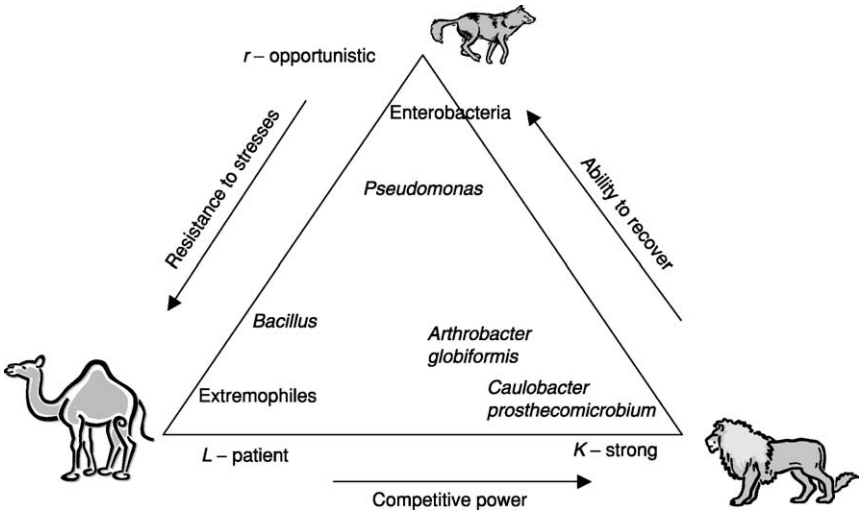


Figure 6 The concept of life strategies. All soil microorganisms are located along three axes characterizing the survival triad: the ability to compete for resources (*K*-axis), recover after stresses (*r*-axis), and resist unfavorable environment (*L*-axis). One can distinguish the following three types of natural selection which correspond to three types of life strategy:

1. *K*-selection operates in climax ecosystems under stable and predictable conditions with no or rare perturbations and stresses. The products of *K*-selection are highly competitive species with 'lion' type of life strategy. The bacterial 'lions' are oligotrophic species that have evolved powerful transporters to absorb the nutrients from diluted solutions (K_s in the range 10^{-6}–10^{-9} mol l^{-1}).
2. *r*-selection operates on the pioneer stages of succession initiated by some perturbation of a climax ecosystem (fire, nutrients flash, cataclysmic elimination of competitors) that has resulted in a temporary relief from competition for nutrient resources. *r*-selected species survive in ephemeral habitats due to mobility and high reproduction rates (opportunistic or 'jackal' type of strategy).
3. *L*-selection operates under adverse environmental conditions caused by abiotic stresses (high/low salt concentration, temperature, pH, water content) or biotic negative interactions (antagonism, starvation caused by depletion of substrate by more successful competitors). The products of *L*-selection are the patient species resistant to a particular stress factor ('camel' type of strategy).

features: endospore formation, production of antibiotics, and synthesis of hydrolytic enzymes. All these features help them to survive even if they lose in direct combat with competitors for deficient nutrients. Kinetic studies allow us to understand why they are weak competitors. The most striking feature was observed in *Bacillus* dialysis culture: the bacteria stopped growing after 2–3 weeks when the residual glucose level dropped below the threshold value of 20–50 $\mu g/l^{-1}$. In rich environments, even with an intermittent supply of nutrients (feast-to-famine transitions simulated in batch culture), these bacteria perform well. They rapidly deplete the available substrates, which triggers sporulation and transition to a dormant state, preserving the bacteria from extinction. However, the chronic starvation typical of most oligotrophic environments is the 'trap' for bacilli; they are provoked to sporulate but are not able to finish it in a normal way. The slow feed via the dialysis membrane provides glucose levels too high to allow termination of normal sporulation and too low for growth due to the uncoupling action of metabolite H and acceleration of turnover rate (Figure 7).

Decomposition Kinetics

Decomposition of organic matter is the most essential soil process. There are several ways to quantify kinetics of decomposition.

Decomposition as Self-Decay

The simplest way is to use a first-order kinetic equation:

$$\frac{\delta s}{\delta t} = -ks \qquad [14]$$

where k is a constant dependent on environmental factors. To account for the heterogeneity of detritus, the same equation is written for n (usually $n = 2$ or 3) arbitrarily selected organic matter fractions and then summed as follows:

$$\frac{\delta s}{\delta t} = -(k_1 s_1 + \ldots + k_n s_n) = -\sum_{i=1}^{n} k_i s_i \qquad [14a]$$

Literally, eqn [14a] stands for spontaneous (without involvement of soil organisms or enzymes)

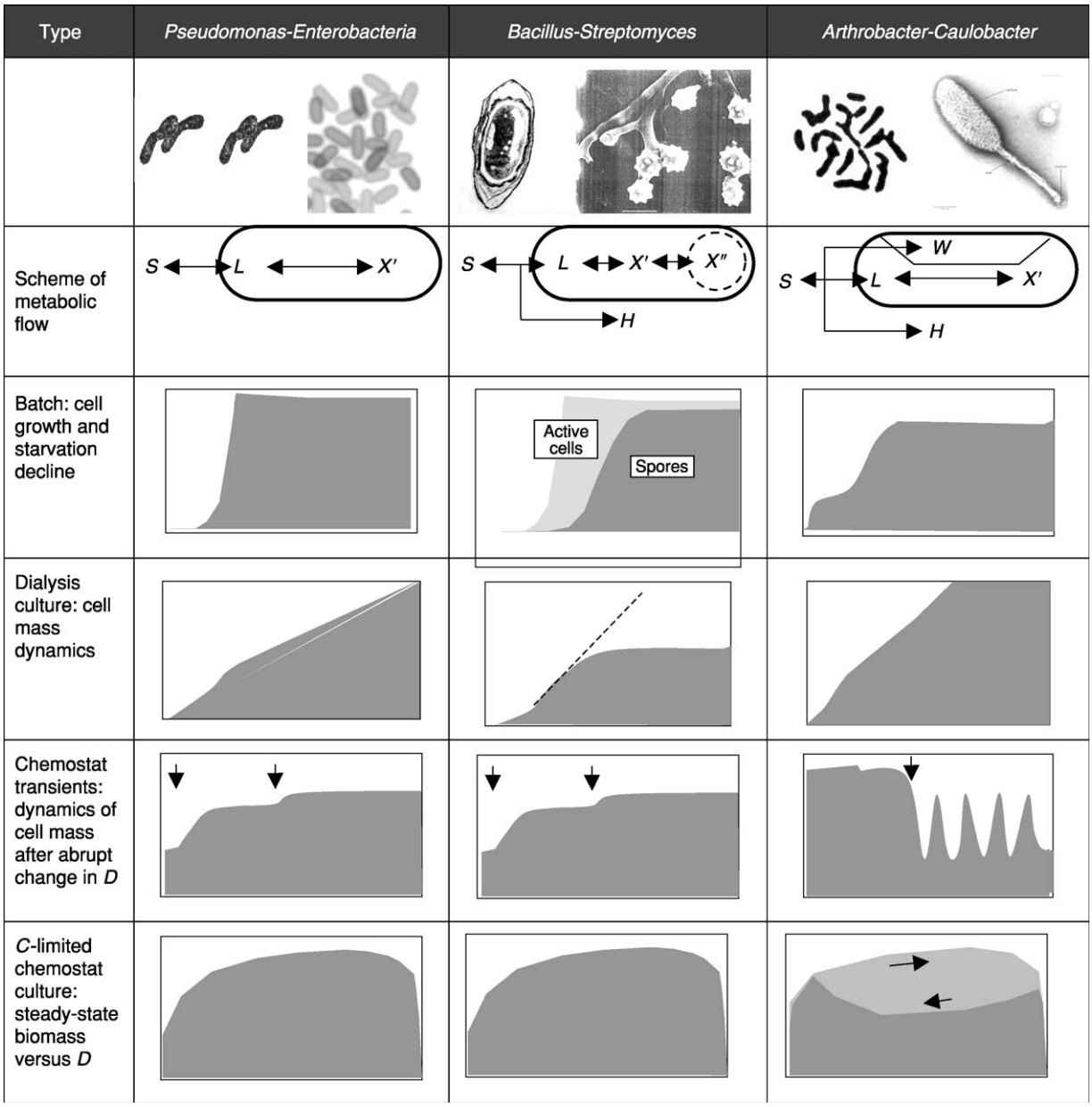


Figure 7 Diversity of growth patterns displayed by soil bacteria in various life strategies in several contrasting types of laboratory cultivation (1) and dialysis (2) cultures as well as chemostat during transient (3) and under steady-state conditions (4). Upper row shows typical morphology of selected microbial groups; next row demonstrates cell growth chart and the main state variables of the respective modification of SCM used for simulation (X″ is prospore compartment, H is autoinhibitor, and W is poly-b-oxybutyrate – reserved compounds): for S, L, and X′. see **Figure 2**. Arrows indicate the time of switching from one dilution rate, D, to another in chemostat culture (3) or derection of sequential D-changes (4, from low D to high or reverse) in chemostat culture displaying multistability.

decomposition of organic matter as well as independent and parallel decay of different individual compounds; no transformation reactions are assumed. All the listed assumptions are unfair; but eqn [14a] is the most popular in modeling of the soil decomposition due to its simplicity.

To account for the transformations accompanying decomposition of organic matter in the soil, e.g., the synthesis of microbial cells and exometabolites from original plant litter, condensation reactions, and formation of organomineral complexes, eqn [14] is modified as follows:

$$\frac{\partial f(z,t)}{\partial t} + \frac{\partial V(z,t)}{\partial z} = -U(z,t)f(z,t) \qquad [15]$$

Here $f(z,t)$ is the distribution of detritus components with respect to their susceptibility to microbial

degradation, z, where z ranges between 0 and the maximum value z_m; $V(z,t)$ is the amount of detritus converted into more or less available compounds; and $U(z,t)$ is the degradation rate.

Decomposition of Simple Individual Compounds Added to Soil

More realistic models explicitly account for the soil organisms responsible for biodegradation. Fortunately, such a task does not require a mandatory microbiological analysis, because microbial growth can be evaluated indirectly, from the decomposition dynamic patterns. Such an approach is especially useful in soil incubation studies, when decomposition of added individual organic compounds (pollutants, ameliorators, model compounds) is recorded as residual concentration of chemicals and/or as the CO_2 formation rate. As shown in Figure 8, nonmicrobial degradation should follow zero- or first-order kinetics with a monotonous decline of soil respiration. If enzymes or nongrowing cells are involved, then decomposition follows the Michaelis–Menten equation with its mixed order. The growth of degraders is manifested as a temporary increase in the degradation rate. Overall, more biomass (x_0) before amendment and a faster growth (higher μ) leads to a steeper response and shorter lag-phase. The approximation of substrate-induced soil respiration dynamics with SCM allows us to find the exact value of x_0 and r_0, characterizing the physiological state of studied microorganisms *in situ*.

As another example, decomposition of glucose in bare soil kept for a long time without plants or fresh litter always shows a characteristic wavy dynamic pattern (Figure 8, line 5): rapid glucose uptake the first 1–3 h, followed by deceleration and then acceleration of glucose uptake until all the added compound is consumed. In this case, the respiration dynamics remains monotonic. A remarkably similar pattern was found in a batch culture of *Arthrobacter globiformis* and other *K*-selected oligotrophic species; the first phase of intensive uptake is attributed to filling of storage cell compartment, while the second one is due to normal exponential growth.

Decomposition of Cellulose

Significantly more complicated is the decomposition kinetics of polymeric plant material, e.g., cellulose. In contrast to monomeric substrates, plant polymers are degraded by extracellular enzymes, and the depolymerization products are shared by the whole community. In addition, senescent plant tissues are poor in biogenic elements (N, P, Mg, Fe), while production of extracellular enzymes requires extra amounts of N, Mg, Ca, and P (direct incorporation into protein molecules and support of ribosomal machinery). The deficiency of biogenic elements coupled with C-sufficiency provides a higher risk of metabolic imbalance: catabolic repression of enzyme synthesis, accumulation of inhibitory products, and 'substrate-accelerated death.' In the sphagnum bog, these obstacles are expressed in the highest degree. In addition to metabolic imbalance, these systems are partially anoxic and decomposition rates are so low that peat accumulates. A microbial decomposition under extremely unfavorable conditions is made possible only by the 'cooperative efforts' of the entire community.

The minimal structure of the cellulolytic community consists of three microbial groups: (1) *L*-selected hydrolytic populations producing extracellular cellulases; (2) *r*-selected copiotrophic organisms capable of explosive growth and some way to cope with deficiency of mineral nutrients (e.g., N_2 fixation); and (3) the *K*-selected oligotrophic populations capable of scavenging monomeric products diffusing from decomposition hot spots. In addition, free cellulases immobilized on soil particles provide some background level of depolymerization activity.

The chain of events after cellulose is added to soil is as follows:

1. induction of microbial cellulase synthesis, splitting polymers into mono- and oligosaccharides which are accumulated within the hot spot of microbial activity
2. sugar uptake and ephemeral microbial growth until one of the nutrients, for example N, becomes depleted
3. N-starvation stops hydrolytic activity by driving these organisms into a dormant stage (most hydrolytic microbes form spores), or possibly switching to N_2 fixation
4. growth of this group is stopped when the available C-source is depleted and double N and C starvation occurs, then the third group of *K*-selected oligotrophs continue to function as a sink for depolymerization products (generated by free enzymes).

The oligotrophic bacteria are not affected by N-starvation due to location outside the decomposition hot spot. This verbal description is translated into a mathematical model which gives fair agreement with observations (Figure 9). Fertilization of soil with NPK considerably stimulates cellulose decomposition; however, the highest activity is achieved if soils are air-dried and rewetted before incubation (compare the left and right parts of Figure 9). The

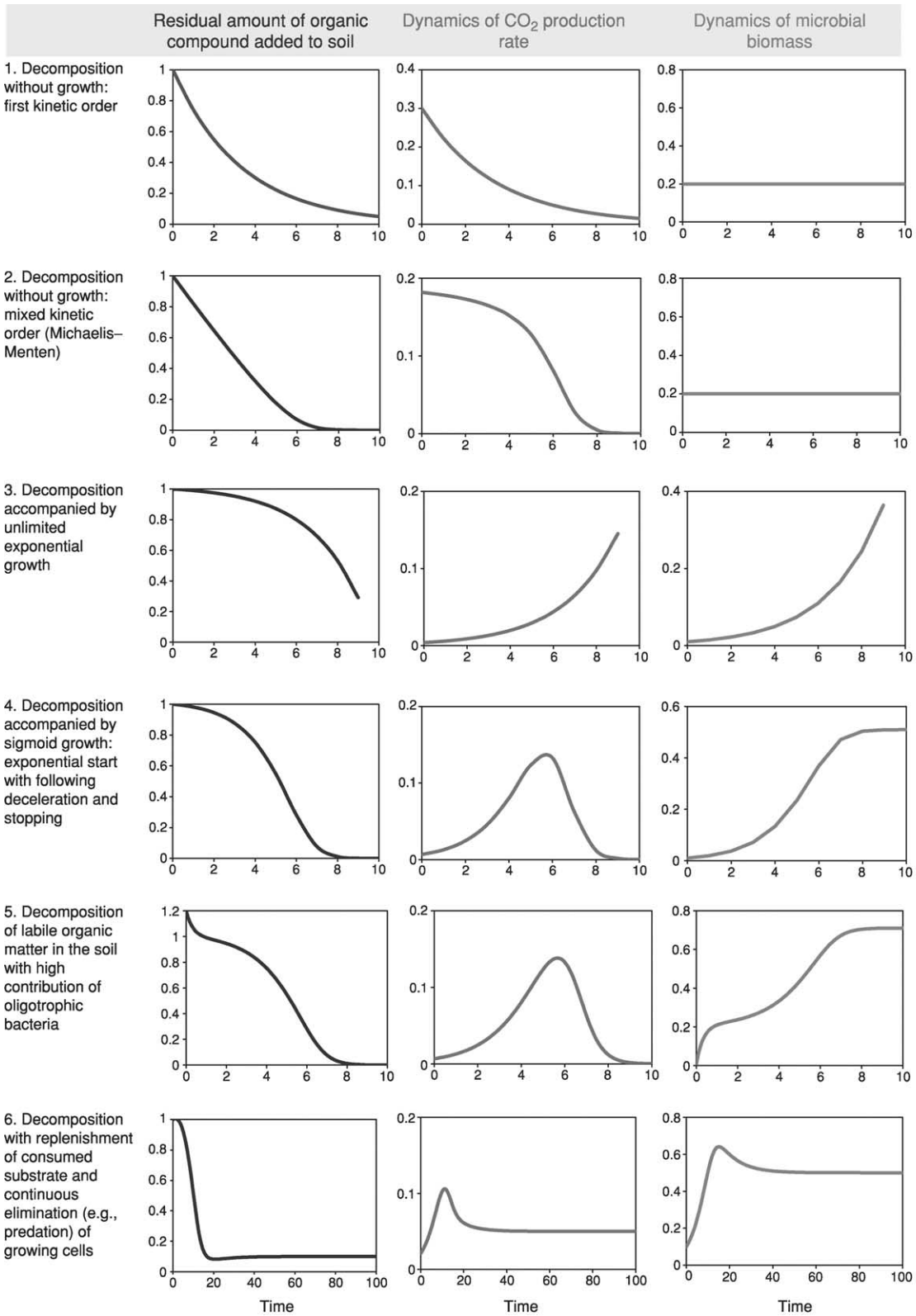


Figure 8 Biodegradation dynamics recorded in soil incubated with added organic substrates. In typical soil incubation studies degradation of labile organics (glucose, amino acids, degradable pesticides, etc.) is followed by recording of residual amount of added compound (first column), CO_2 evolution rate (second column) and microbial biomass (third column). The shape of dynamic curves is a sensitive indicator of the biodegradation mechanisms as indicated by six shown examples.

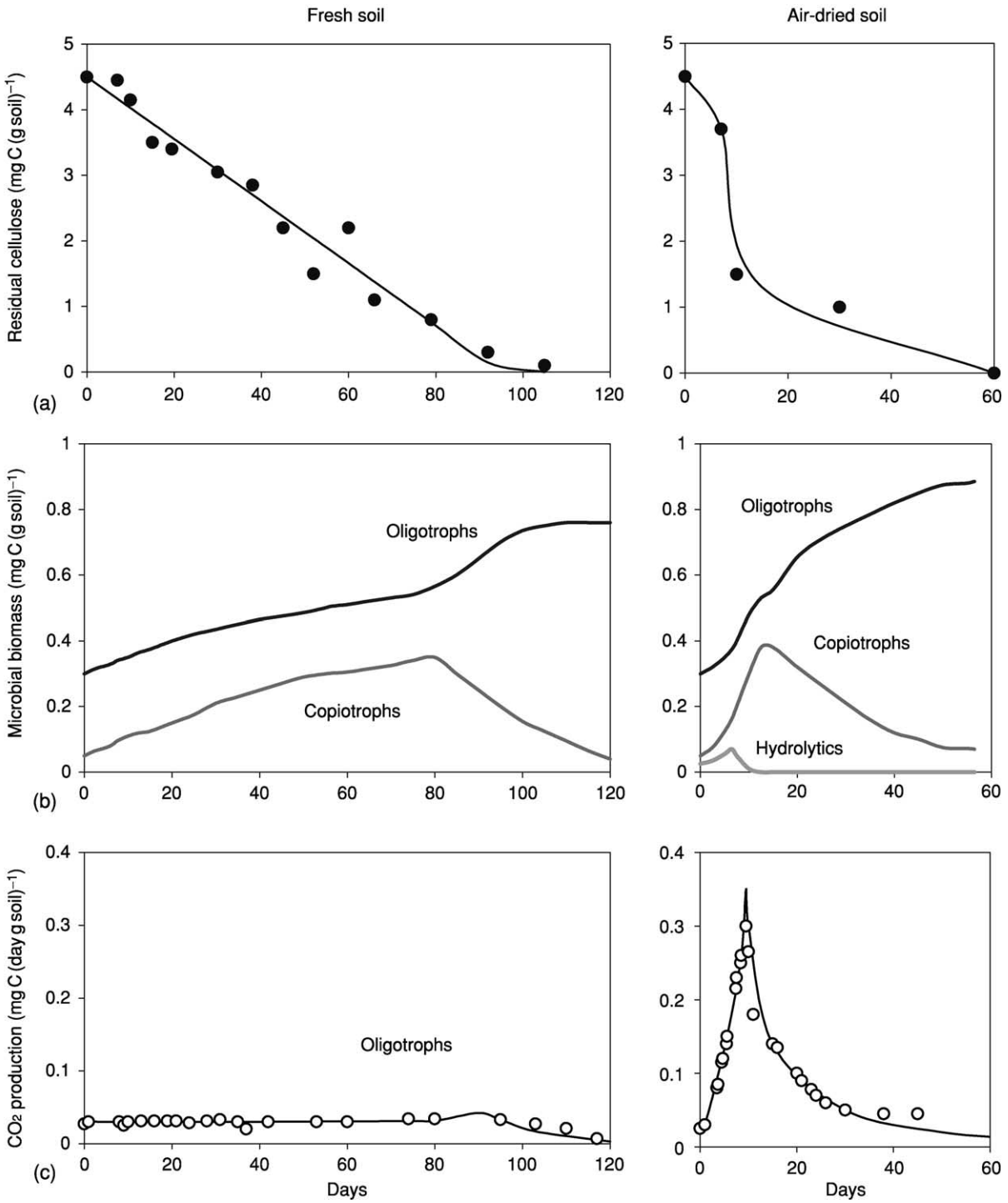


Figure 9 Cellulose decomposition in the soil as related to population dynamics of cellulolytic microbial community. Left: fresh forest soil; right: the same soil after air-drying and rehydration 5 days before cellulose amendment. (a) Residual cellulose; (b) simulated dynamics of: (1) hydrolytics producing extracellular cellulases; (2) copiotrophic N_2-fixing organisms; and (3) the oligotrophic populations capable of scavenging monomeric products diffusing from decomposition hot spots; (c) respiration dynamics. All curves}were simulated by a synthetic chemostat model-based model. Source: Panikov NS (1995) *Microbial Growth Kinetics*. London: Chapman & Hall.

drying–rehydration causes partial damage to microbial cells (similar to fumigation) and release of cell constituents into soil solution. In addition, this treatment removes inhibitors and signaling metabolites, which arrest the biosynthetic activity of hydrolytic microorganisms. It is interesting that healthy soils have restricted hydrolytic activity (slow zero-order decomposition due to action of 'free' soil enzymes). An excessive activation of soil hydrolases takes place more often in pathological cases, e.g., in arable soil after continuous use of NPK and development of soil toxicity.

Decomposition of Plant Litter

The final step from cellulose to lignocellulose and entire plant litter adds the following complications: (1) in lignocellulose the polysaccharide component is further preserved from microbial enzyme by lignin; (2) the heterogeneity of plant litter as a substrate is essentially higher due to the presence of the complex mixture of heteropolysaccharides, waxes, and monomeric phenolic compounds (often considered as inhibitors of decomposition process); and (3) a more complex soil community. However, degradation of cellulose and lignin is carried out jointly, the carbohydrate component being used as available C-source and to generate peroxides driving the oxidative destruction of lignin. Although lignin slows the decay rate, lignocellulose is decomposed as a whole, and there is no significant enrichment of lignin in the old residual substrate. Therefore the kinetic mechanisms of decomposition revealed for cellulose alone are also applicable for lignocellulose.

See also: **Microbial Processes:** Community Analysis; **Mycorrhizal Fungi**

Further Reading

Brock TD (1971) Microbial growth rates in nature. *Bacteriological Review* 35: 39–58.

Gause GF (1934) *The Struggle for Existence*. Baltimore, MD: Williams & Wilkins. Available online at www.ggause.com/Contgau.htm.

Gerson U and Chet I (1981) Are allochtonous and autochthonous soil microorganisms *r*- and *K*-selected? *Revue d'Ecologie et de Biologie du Sol* 18: 285–289.

Hattori T (1973) *Microbial Life in the Soil; An Introduction*. New York: Marcel Dekker.

Lynch JM (1983) *Soil Biotechnology: Microbial Factors in Crop Productivity*. Oxford: Blackwell.

Lynch JM (1990) *The Rhizosphere*. Chichester: Wiley.

Panikov NS (1995) *Microbial Growth Kinetics*. London: Chapman & Hall.

Paul EA and Clark FE (1996) *Soil Microbiology and Biochemistry*. San Diego, CA: Academic Press.

Pirt SJ (1975) *Principles of Microbe and Cell Cultivation*. Oxford: Blackwell Science.

Poindexter JS (1981) Oligotrophy. Fast and famine existence. *Advances in Microbial Ecology* 5: 63–89.

Winogradsky SN (1949) *Microbiologie du Sol*. Paris: Masson.

MICRONUTRIENTS

L M Shuman, University of Georgia, Griffin, GA, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Micronutrients in soils and their plant availability are critical to plant nutrition. The supply of nutrients to plants is included in the discipline of soil fertility. Micronutrients are elements found in plants at concentrations less than $100\,mg\,kg^{-1}$ and are generally found at low concentrations in soils, with the notable exception of Fe. Four are present as cations (Zn, Cu, Mn, and Fe), one is an uncharged molecule (B), and one is an anion (Mo) in soil solution and in plants. Several other elements have been shown to be essential to plant nutrition, such as Ni and Cl, but they are not controlled through soil amendments, and deficiencies have only been found in research settings. Not only are the element concentrations important to soil fertility, but more so are the soil properties and processes that influence availability to plants.

Zinc

Zn is taken up by plants in the divalent form, which is the form most often found in soil solution at normal soil pH values. Concentrations in the soil solution that are usually adequate for plant growth are $0.05–0.25\,mol\,l^{-1}$, but will become toxic

at concentrations greater than 3–6 mol l^{-1}. Soil-solution concentrations of Zn^{2+} are usually low at 0.01–1.0 mol l^{-1} in calcareous soils, but higher in acid soils. There is still uncertainty about whether Zn uptake by plants is an active or passive process. The main interaction with other micronutrient ions is with Cu^{2+}, which can inhibit Zn uptake as well as compete for transport sites within the plant. Zinc interacts with the macronutrient P, in that high P availability can cause Zn deficiency. It is still not known for certain whether the P–Zn interaction takes place mostly in the soil or the plant. Divalent cations such as Mg^{2+} and Ca^{2+} can also interfere with Zn uptake by competition. Zinc deficiencies increase when the soil is waterlogged as for rice-growing. These deficiencies could come from the precipitation of Zn or by increases in Zn–organic complexes.

Zinc is found at higher concentrations in basic rocks than in acidic or sedimentary rocks such as limestone and occurs naturally as sulfides, carbonates, and silicates. The mean concentration in the Earth's crust is 80 mg kg^{-1}. As for most elements, Zn is found at higher concentrations in the topsoil than the subsoil, because plant roots bring Zn to the surface. Of all the soil properties affecting Zn availability to plants, pH is certainly the most important. There are several causes for the increase in availability with decreasing pH: first, the adsorption of Zn by colloids increases with increasing pH; second, Zn can precipitate at high pH values as in calcareous soils. The availability is very low at pH > 7.0; finally, diffusion to the plant root is decreased by increasing pH. Zinc usually does not become toxic unless the pH is low (less than 4.0) or the plants are susceptible, as in the case of peanuts. Soil organic matter also has a profound effect on Zn availability. The two forms of organic matter, solid and soluble, have opposing effects on Zn, as they do for most metals. Soluble organic matter–Zn complexes in solution assist in bringing Zn to plant roots, thus enhancing availability. Solid organic matter provides negative adsorption sites, as well as specific adsorption sites that bind Zn and decrease plant availability. As a result, Zn availability is low in organic soils; however, the action of soil microbes in breaking down solid soil organic matter releases Zn and increases plant uptake. Soil secondary clay minerals as well as Fe and Al oxides also have high bonding energies for Zn and high adsorption capacity. These can bind Zn and hinder plant uptake. Zinc is not susceptible to oxidation and reduction in soils, so redox status per se does not affect availability. In soils that undergo oxidation–reduction cycles, the freshly precipitated amorphous (or microcrystalline) Fe oxides are in abundance, and these have approximately 10 times more adsorptive capacity than aged, crystalline Fe oxides.

Copper

Cu is adsorbed by plants as the divalent ion Cu^{2+}. Of the micronutrients considered here, it is most strongly bound with organic chelating materials (Cu > Zn > Mn > Fe) and thus is strongly associated both with solid soil organic matter and with soluble organics in soil solution. Up to 98% of Cu found in soil solution is bound by low-weight organic compounds. For adequate growth and development, plants need at least 0.02 mmol l^{-1} in solution. Since the free Cu^{2+} ion is the form taken up by plants, Cu bound by organics is not readily available. Soil moisture has little effect on uptake except that flooding decreases both water-soluble Cu and Cu available to plants.

More Cu is found in soils developed from basic rocks and shales than from sands, sandstones, and acid igneous rocks; basalts are higher in Cu, while granites are low in Cu. Cu deficiencies are prevalent in organic soils such as peats and mucks, and the mean total Cu found in soils is 70 mg kg^{-1}. Cu is strongly associated with oxides of Fe and Al according to adsorption and fractionation studies. As with most metals, it is more strongly adsorbed by soil colloids at higher pH values, but plant availability is not as sensitive to pH as for some other metals because of Cu–organic complexing. Soil-solution Cu concentration in terms of chemical activity is a linear function of soil pH (Figure 1). Plants can develop Cu toxicity symptoms when grown in acid soils of pH less than 6.5. High organic matter content can lead to Cu deficiencies due to the formation of insoluble organic complexes. On the other hand, in calcareous soils, where high pH could decrease Cu availability, soluble organic complexes can help to keep availability above deficient levels. The soluble organics are evidently able to shift the equilibrium from the solid organic complexes to the soluble, so that even at higher pH values the soluble organics can enhance Cu concentrations in solution. Also, over time, the Cu associated with solid organic matter is released through microbial breakdown of the organic matter. Cu does not change oxidation values readily in soils, so the decrease in availability in flooded soils could come from the chemical reduction of crystalline Fe and Mn oxides to amorphous forms, which provide surfaces for Cu adsorption.

Manganese

Mn and Fe are considerably different in their soil chemical reactions from Zn and Cu, because they

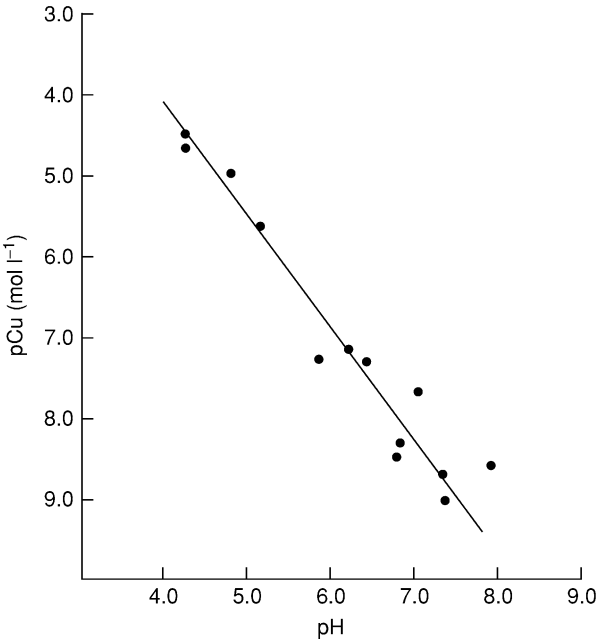


Figure 1 Activity of Cu^{2+} versus soil-suspension pH. pCu, negative $\log_{10}$ of Cu concentration in solution. (Adapted from McBride MB and Blasiak JJ (1979) Zinc and copper solubility as a function of pH in an acid soil. *Soil Science Society of America Journal* 43: 866–870, with permission.)

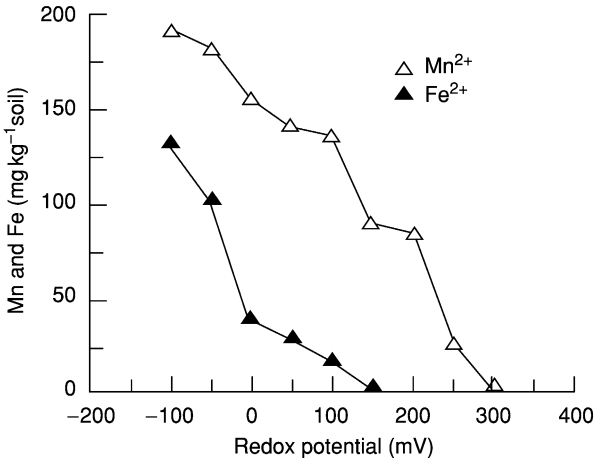


Figure 2 Soluble Mn^{2+} and Fe^{2+} in soil suspensions versus redox potential during an oxidized-to-reduced cycle. (Adapted from Patrick WH Jr and Jugsujinda A (1992) Sequential reduction and oxidation of inorganic nitrogen, manganese, and iron in flooded soil. *Soil Science Society of America Journal* 56: 1071–1073, with permission.)

are easily oxidized and reduced in soils. Mn is taken up by plants as Mn^{2+}, although most of the Mn in soils is in the Mn^{4+} oxidation state. Mn concentrations in soil solution can vary from less than 0.1 mmol l^{-1} in well-aerated soils to as high as 400 mmol l^{-1} in submerged soils. Plants need a range of 0.01 to approximately 50 mmol l^{-1}. Chelation of Mn occurs in soils as for the other metals, but the uptake of chelated Mn by plants is poorly understood. High soil-moisture content tends to reduce the oxidized Mn^{4+} form to the reduced Mn^{2+} form, making it more available. Thus, deficiency is not a problem on wet soils, and toxicity may result, especially for susceptible plants.

Mn is the 11th most common element in the Earth's crust and averages 0.09% or 900 mg kg^{-1}. The oxides or sulfides are often associated with Fe. Mn can be found in soil minerals such as MnO_2, MnOOH, and Mn_3O_4. Mn is low in soils that are derived from crystalline shales and acid igneous rocks, but is higher in soils derived from basalt, limestone, and shales, especially limestones. Deficiencies occur in organic, calcareous, and poorly drained acid sandy soils. Usually the higher the clay content of a soil, the greater will be the Mn content. For example, the soils of the US southeastern piedmont have significant amounts of clay and some are derived from limestone. These soils are so high in Mn that toxicities can result. However, on the southeastern lower coastal plain, the sandy soils are alternately wet and dry and can cause Mn deficiencies. Manganese availability is tied closely to the oxidation–reduction status of soils. Soils that are waterlogged are reduced and release Mn^{2+} from the oxidized forms. The more negative the redox potential, the higher the Mn concentration in soil solution (Figure 2).

Besides oxidation–reduction, the other major influence on Mn availability is soil pH. At pH values greater than 6 and under high-oxidation conditions, Mn is mostly in the insoluble Mn^{4+} form. At lower pH and/or under wet conditions, the Mn is in the soluble Mn^{2+} form and is readily available to plants. At pH values less than 5, oxidation–reduction is not as important as pH in determining Mn availability, but at pH values of 6–8, oxidation–reduction becomes very important. Plant roots themselves exude protons and chelating agents which can assist in mobilizing Mn. Solid soil organic matter can complex Mn with stability increasing at increasing pH.

Iron

Approximately 5.1% of soil on the average is Fe, but availability can be low in well-aerated soils. Soil Fe is derived from ferromagnesium silicates, which includes a great number of rock types. Mineral forms are oxides, hydroxides, and silicates, as well as amorphous oxides. These minerals take on forms such as goethite and hematite depending on the hydration status and crystalline form. The Fe released by weathering is redistributed into secondary

minerals. The solubility of soil Fe depends very much on pH. Iron is solubilized from solid Fe oxides and hydroxides by the following reaction with protons:

$$\text{Soil Fe} + 3H^+ = Fe^{3+} + 3H_2O \qquad [1]$$

In solution, Fe hydrolyzes to Fe^{3+}, $Fe(OH)_2^+$, and $Fe(OH)_4^-$, which is again dependent on the pH. As for Mn the oxidation–reduction status of the soil also affects Fe solubility. Figure 2 shows that, as for Mn, the more positive the redox potential, the less the Fe concentration in soil solution. The divalent form is more soluble than the trivalent form and is the form that plants absorb. But in well-aerated soils at pH values of approximately 6–8, the Fe is in the Fe^{3+} state and is relatively insoluble, with concentrations around 10^{-8}–10^{-3} mol l^{-1}. Submergence of soils can cause chemical reduction of Fe^{3+} to Fe^{2+}, resulting in Fe toxicity. However, damp soils can cause a decrease in the Fe absorption capacity of plant roots and affect bicarbonate in calcareous soils, decreasing uptake. Fe is strongly complexed by soil organic matter, and the soluble organic matter can increase the soil-solution Fe concentrations to 10^{-5}–10^{-3}. This soluble organic matter comes from solid organic matter degradation and from release of siderophores (nonprotein complexes) by plant roots and soil bacteria and fungi. Plants need Fe concentrations of the order of 10^{-4}–10^{-2}, so plants must have mechanisms to increase soil Fe availability. Plants have two strategies to increase Fe availability. What has been named strategy I is release of reductants to reduce Fe^{3+} to Fe^{2+}, and strategy II is release of Fe-binding organic ligands and protons to help solubilize soil Fe. Strategy II is used by grasses and strategy I by dicotyledons and nongraminaceous monocotyledons.

Boron

B and Mo are found in combination with oxygen in nature. A common mineral for B is borax, with the formula $Na_2B_4O_7 \cdot 10H_2O$. B is present in soil solution and in plants as boric acid ($B(OH)_3$), a compound that is considered an acid because it hydrolyzes according to the following reaction, producing hydronium ions:

$$B(OH)_3 + 2H_2O = B(OH)_4^- + H_3O^+ \qquad [2]$$

This ionic form is a very weak acid, which accepts OH^- ions; at usual soil and plant pH values, the acid is undissociated. The mechanism of uptake is not well understood. B has a narrow range of sufficiency to toxicity for plants. Plants need 0.01 mmol l^{-1} in soil solution, but 0.03 mmol l^{-1} is toxic to B-sensitive plants, which include citruses and peaches. Drought conditions in soil increase B deficiency in many crops, which was first noticed in sugar beet; B deficiency restricts root growth, having more severe effects than any of the other micronutrients. Drought also restricts the mineralization and availability of organically bound B.

One mineral that is known for its B content is tourmaline; however, it is so insoluble that it cannot be considered a B source for plants. Boron is found as a substitute for Si in other minerals, and the rate of weathering of these minerals is slow. Soils derived from granite and other igneous rocks are usually low in B content, whereas those derived from sedimentary rocks of marine origin are usually higher in B content. The mean concentration of B in the crust is 50 mg kg^{-1}. B is water-soluble and is easily leached from soils, so sandy, porous soils can often be B-deficient. Sandy soils have more B available to the plant at a given hot water-extractable level than finer-textured soils, because the B is bound more strongly by the clays and oxides in the finer-textured soils (Figure 3). B is sometimes toxic in arid and semiarid soils; the B content in irrigation water can cause B buildup in semiarid soils, leading to toxicity. As for most other micronutrients, high soil pH (6–9) causes low B availability. Higher pH levels mean higher adsorption on soil clays and oxides, thus liming can have a large effect on B availability. At lower pH values, the uncharged boric acid is not readily adsorbed and is held by ligand exchange; at higher pH values, the $B(OH)_4^-$ ion forms and is strongly adsorbed by electrostatic anion exchange. At low pH, soil B can become toxic to plants. Most B in the soil is contained in the organic matter and becomes more available as the conditions for microbial degradation become favorable, such as warmer temperatures, higher moisture, and better aeration. In organic soils such as peats and mucks, the availability can become too high as the B in the organic matter is

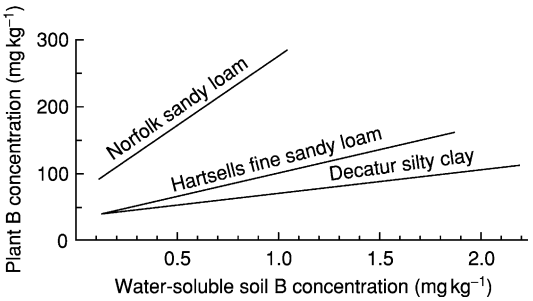


Figure 3 Plant boron (B) concentration versus water-soluble soil B concentration for three soils of different textures. (Adapted from Wear JJ and Patterson RM (1962) Effect of soil pH and texture on the availability of water-soluble boron in the soil. *Soil Science Society of America Proceedings* 26: 344–346, with permission.)

mineralized from the microbial biomass. B does not undergo oxidation–reduction reactions in soils.

Molybdenum

Mo concentrations are the lowest in soils of any of the micronutrients, being only 2.3 mg kg^{-1} on average in the crust. Soil Mo is derived more from shale and granite rocks, which are higher in content than sandstones or limestones. Soils that are very high in organic matter content, are alkaline, or are poorly drained are generally higher in Mo availability than other soils. Mo is different in its response to soil pH from the other micronutrients. As soil pH increases, the Mo plant availability also increases (Figure 4). Mo is adsorbed by Fe and Al oxides as the molybdate ion, MoO_4^{2-} at pH > 4.2, and this adsorption decreases at higher pH values. Thus, liming soils increases the uptake of Mo. Mo toxicity is rare in agricultural crops, but has been observed in controlled experiments and on peats and mucks that are poorly drained. Mo availability is higher in soils as the organic matter content or the Fe and Al oxide content decreases. Since Mo is so strongly associated with Fe and Al oxides through adsorption and precipitation, Mo availability is negatively correlated with the contents of Fe and Al oxides, whereas total Mo content may be positively correlated with Fe and Al oxide content. Oxidation and reduction processes do not affect Mo, so any oxidation–reduction effects on availability are indirect. Mo is needed by plants for nitrate reduction and symbiotic N-fixation by legumes. Deficiencies are usually on acid soils and are usually in legumes. Increases in soil moisture increase Mo availability due to decreased soil adsorption of MoO_4^{2-} under the higher soil pH conditions resulting from waterlogging. The reducing conditions caused by waterlogging reduce Fe^{3+} to Fe^{2+}, and Mo is less associated with the reduced form of Fe so therefore more available. Also, under wet conditions, Mo moves to the plant root more by mass flow than it does under drier conditions.

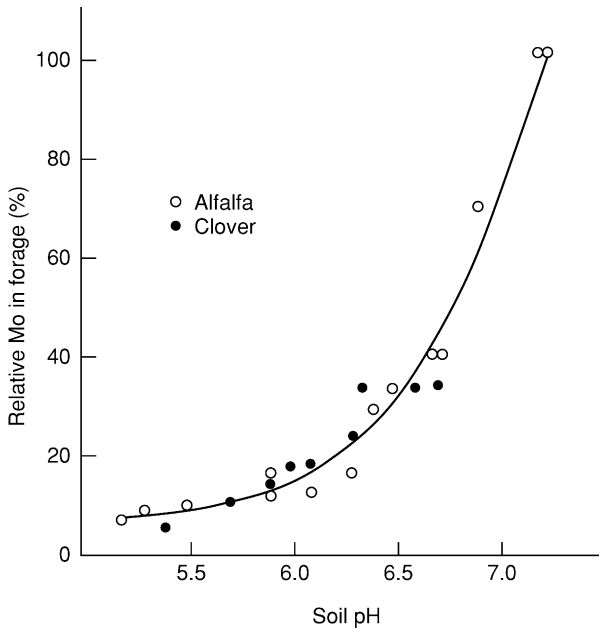


Figure 4 Molybdenum in forage as a percentage of maximum versus soil pH. (Adapted from Mortvedt JJ and Anderson OE (1982) Forage legumes: diagnosis and correction of molybdenum and manganese problems. Southern Cooperative Series Bulletin 278. Athens, GA: University of Georgia, with permission.)

Micronutrient Soil Tests

This discussion will be limited to soil chemical testing, which is the major diagnostic tool used for both macronutrients and micronutrients, designed to predict micronutrient availability to the crop. If the soil is found to be nutrient-deficient, the test should give an indication of the severity of the deficiency and possible yield losses. Field calibration studies are carried out to estimate the amounts of micronutrient fertilizers to be added in response to a certain soil-test result. In interpreting soil-test data, a 'critical level' is used below which are expected yield responses to micronutrient fertilizers. One approach is the Cate–Nelson procedure, where a scatter diagram is formed from soil-test results on the *x*-axis and yields on the *y*-axis. Crossed lines are drawn such that a minimum number of points are in the upper left and lower right quadrants (Figure 5). The value found on the *x*-axis

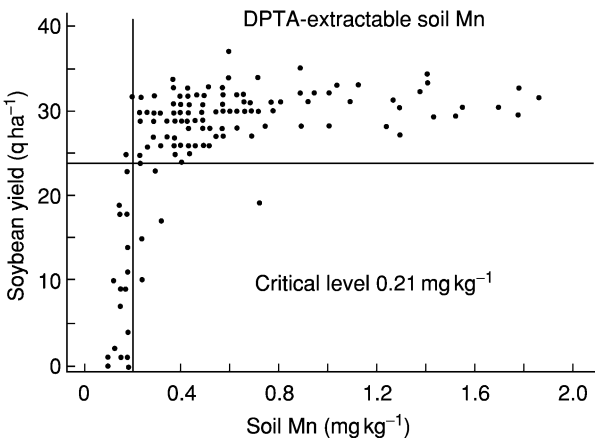


Figure 5 Cate–Nelson plot for soybean yield versus diethylenetriaminepentaacetic acid (DTPA)-extractable soil Mn. (Adapted from Shuman LM, Boswell FC, Ohki K, Parker MB, and Wilson DO (1982) Critical soil manganese deficiency levels for four extractants for soybeans grown in a sandy soil. *Soil Science Society of America Journal* 44: 1021–1025, with permission.)

is the critical level below which fertilizer is applied to keep yields at maximum. Another approach is correlation analysis. Since some micronutrients are very sensitive to pH, an approach often used is to have different critical levels for various pH levels. At times other soil properties are combined with soil-test results, but pH is usually the only value used.

Micronutrient soil tests should be able to extract, from the soil samples of the micronutrients in question in proportion to those taken up by plants; in other words, the samples should contain micronutrients in the same forms as those available to the plants. Soil is usually sampled in the root zone at approximately 0 to 15–20 cm. Soil tests should be accurate, repeatable, rapid, and correlated well with crop response from severe deficiency to sufficiency. Ideally a test will show differences in availability due to changes in soil properties such as pH or oxidation–reduction conditions. Soil tests are usually designed to remove the soluble and exchangeable forms, but in some instances they dissolve mineral forms that contain precipitated or occluded forms, as for Mo. These chemical extractants are usually fairly simple reagents such as water, dilute acids or bases, and salt solutions, and more recently include chelating agents such as diethylenetriaminepentaacetic acid (DTPA) or ethylenediaminetetraacetic acid (EDTA). Dilute acids commonly in use are 0.1 mol l^{-1} HCl, and an acid mixture of 0.025 mol l^{-1} H_2SO_4 and 0.05 mol l^{-1} HCl (Mehlich 1). These acids are used on acid soils to dissociate, displace, or dissolve the cations from soil clays, oxides, or organic matter. The chelating agent DTPA is found in Mehlich 3, DTPA, and AB-DTPA; and EDTA is in the acid ammonium acetate–EDTA extractant. All these tests extract multielements. Chelates promote extraction from surfaces of clays, oxides, and organic matter. See Table 1 for commonly used micronutrient soil tests, with the deficiency critical ranges for each.

Zn is usually extracted using a multielement extractant, e.g., Mehlich 1 and 3, AB-DTPA, and 0.1 mol l^{-1} HCl. Some of these are used in conjunction with soil pH tests. Cu deficiencies are rare, but Cu is routinely tested using the same extractants as for Zn. Since Cu binds readily with organics, the better extractants are those with chelating agents. For Mn, dilute acids and chelating agents are both used, such as in Mehlich 1 and 3, 0.1 mol l^{-1} HCl, 0.03 mol l^{-1} H_3PO_4, and DTPA tests. Soil pH is often used with these extractants, using a 'sliding scale' for critical levels based on soil pH. Some chelating reagents such as DTPA and AB-DTPA were designed for use with Fe. The dilute acids mentioned are also used for Fe extraction as a soil test. Hot water was found to be a good soil test extractant for B in 1939 and it is still in use today. Mo is associated with oxides and is extracted by acid ammonium oxalate, which also extracts Fe and Al oxides. Soil pH is a big factor in Mo deficiency, so is often used in conjunction with the soil test.

Table 1 Major soil diagnostic tests and critical ranges for deficiency

Micronutrient	*Soil-test method*	*Critical range*
Zinc	AB-DTPA	0.5–1.0 mg kg^{-1}
	DTPA	0.2–2.0 mg kg^{-1}
	Mehlich 1	0.5–3.0 mg kg^{-1}
	Mehlich 3	1.0–2.0 mg kg^{-1}
	0.1 mol l^{-1} HCl	1.0–5.0 mg kg^{-1}
Copper	AB-DTPA	0.5–2.5 mg kg^{-1}
	DTPA	0.1–2.5 mg kg^{-1}
	Mehlich 1	0.1–10 mg kg^{-1}
	Mehlich 3	0.3–15 mg kg^{-1}
	0.1 mol l^{-1} HCl	0.1–2.0 mg kg^{-1}
Manganese	AB-DTPA	0.5–5.0 mg kg^{-1}
	DTPA	1.0–5.0 mg kg^{-1}
	Mehlich 1	5.0 mg kg^{-1} at pH 6 10 mg kg^{-1} at pH 7
	Mehlich 3	4.0 mg kg^{-1} at pH 6 8.0 mg kg^{-1} at pH 7
	0.1 mol l^{-1} HCl	1.0–4.0 mg kg^{-1}
Iron	AB-DTPA	4.0–5.0 mg kg^{-1}
	DTPA	2.5–5.0 mg kg^{-1}
Boron	Hot water	0.1–2.0 mg kg^{-1}
	Mehlich 3	0.7–3.0 mg kg^{-1}
Molybdenum	Ammonium oxalate–pH 3.3	0.1–0.3 mg kg^{-1}

Reproduced with permission from Sims JT (2000) Soil fertility evaluation. In: Sumner ME (ed.) *Handbook of Soil Science*, pp. 113–153. Boca Raton, FL: CRC Press.

Micronutrient Fertilizers

Micronutrient fertilizers include four major types: inorganic salts, synthetic chelates, natural organic complexes, and frits. Certainly the most commonly used are the inorganic sources, since they are the least expensive (Table 2). These include oxides, carbonates, and metallic salts such as the sulfates, chlorides, and nitrates. Manganese oxide and zinc oxide are used to some extent, but are relatively insoluble and become available over a long period of time. The sulfates of Zn, Cu, Mn, and Fe are the most common fertilizers in use, being quite soluble and immediately available to the crop. For B, borax is the fertilizer of choice, and for Mo it is sodium molybdate.

The synthetic chelates include *N*-hydroxylethylenediaminetriacetic acid (HEDTA) and EDTA, with EDTA salts being the most commonly used (Table 3). The rates of chelate application are about two to five times less than inorganic sources, but the cost is 5–100 times greater. The chelates are usually

Table 2 Inorganic fertilizer sources for micronutrient fertilizers

Micronutrient source	*Solubility in water*	*Element concentration (%)*
Zinc		
$ZnSO_4 \cdot H_2O$	Soluble	36
$ZnSO_4 \cdot 7H_2O$	Soluble	22
$ZnCl_2$	Soluble	47
$ZnSO_4 \cdot 4Zn(OH)_2$	Slightly soluble	55
$ZnCO_3$	Insoluble	52
ZnO	Insoluble	60–78
Zinc oxysulfate	Variable	25–60
Copper		
$CuSO_4 \cdot H_2O$	Soluble	35
$CuSO_4 \cdot 5H_2O$	Soluble	20
$CuSO_4 \cdot 3Cu(OH)_2 \cdot H_2O$	Insoluble	37
CuO	Insoluble	75
Manganese		
$MnSO_4 \cdot XH_2O$	Soluble	24–30
$MnCl_2 \cdot 4H_2O$	Soluble	28
$MnCO_3$	Insoluble	31
MnO	Insoluble	41–68
Manganese oxysulfate	Variable	30–50
Iron		
$FeSO_4 \cdot H_2O$	Soluble	33
$FeSO_4 \cdot 7H_2O$	Soluble	20
$Fe(SO_4)_3 \cdot 9H_2O$	Soluble	20
$FeSO_4 \cdot (NH_4)_2SO_4$	Soluble	22
Boron		
$Na_2B_4O_7$ (anhydrous borax)	Soluble	20
$Na_2B_4O_7 \cdot 5H_2O$ (fertilizer borate)	Soluble	14
$Na_2B_4O_7 \cdot 10H_2O$ (borax)	Soluble	11
H_3BO_3	Soluble	17
$Ca_2B_6O_{11} \cdot 5H_2O$ (colemanite)	Slightly soluble	10
Molybdenum		
Na_2MoO_4 (anhydrous)	Soluble	47
$Na_2MoO_4 \cdot 2H_2O$	Soluble	39
$(NH_4)_2MoO_4$	Soluble	49
MoO_3	Slightly soluble	66
$CaMoO_4$	Insoluble	48

Reproduced with permission from Mortvedt JJ, Cox FR, Shuman LM, and Welch RM (eds) (1991) *Micronutrients in Agriculture.* Madison, WI: Soil Science Society of America.

Table 3 Organic fertilizer sources for micronutrient fertilizers

Micronutrient source	*Formula*	*Element concentration (%)*
Zinc		
Zinc chelate	$Na_2ZnEDTA$	14
Zinc chelate	NaZnHEDTA	9
Zinc chelate	NaZnNTA	9
Zinc lignosulfonate	—	5–8
Zinc polyflavonoid	—	5–10
Copper		
Copper chelate	$Na_2CuEDTA$	13
Copper chelate	NaCuHEDTA	9
Copper lignosulfonate	—	5–8
Copper polyflavonoid	—	5–7
Manganese		
Manganese chelate	$Na_2MnEDTA$	5–12
Manganese lignosulfonate	—	5
Manganese polyflavonoid	—	5–7
Iron		
Iron chelate	$Na_2FeEDTA$	5–14
Iron chelate	NaFeHEDTA	5–9
Iron chelate	NaFeEDDHA	6
Iron chelate	NaFeDTPA	10
Iron lignosulfonate	—	5–8
Iron polyflavonoid	—	9–10
Iron methoxyphenylpropane	FeMPP	5

Reproduced with permission from Mortvedt JJ, Cox FR, Shuman LM, and Welch RM (eds) (1991) *Micronutrients in Agriculture.* Madison, WI: Soil Science Society of America.

sold as liquid preparations so are sprayed either on the soil or directly on the crop. Elements applied as chelates are Zn and Fe. Manganese chelate can be sprayed on a growing crop, but, when applied to the soil, the Mn is quickly replaced by Fe and the Mn is ineffective. The high concentrations of Fe in the soil solution cause Fe to replace the Mn on the organic chelate molecule.

Natural organic sources most often in use are the lignosulfonate complexes made from a by-product of the wood-pulping process in paper mills. These complexes are cheaper than synthetic chelates, but are also not as effective because they are readily decomposed in the soil. Inorganic salts are reacted with these organic by-products to form complexes. The elements applied with natural chelates are Zn and Fe.

The last major fertilizer type is frits. These are glass beads with micronutrients occluded in them by melting sand with the inorganic salts added. Like the oxides, these slowly release the micronutrients and are used for maintenance. They are most often applied in sandy soils with a high water table, where leaching of soluble salts is likely to move elements below the root zone before they can be taken up by the plant.

Finally, another source is industrial waste products that contain micronutrients, such as wastes from the steel industry. Even farm manures can be micronutrient sources; for example, pig manure is high in Cu and poultry manure high in Zn.

List of Technical Nomenclature

AB-DTPA A soil-testing method using a solution of $1.0\,mol\,l^{-1}$ ammonium bicarbonate (NH_4HCO_3) and the DTPA soil test solution buffered at a pH of 7.6

DTPA Diethylenetriaminepentaacetic acid

DTPA	A soil-testing method using a solution of $0.005\,mol\,l^{-1}$ DTPA and $0.01\,mol\,l^{-1}$ $CaCl_2$ buffered at a pH of 7.3 by $0.1\,mol\,l^{-1}$ triethanolamine (TEA)
EDDHA	Ethylenediaminedi(*o*-hydroxyphenylacetic) acid
EDTA	Ethylenediaminetetraacetic acid
HEDTA	*N*-Hydroxyethyl-ethylenediaminetriacetic acid
Mehlich 1	A soil-testing method using a solution of $0.0125\,mol\,l^{-1}\,H_2SO_4 + 0.05\,mol\,l^{-1}$ HCl
Mehlich 3	A soil-testing method using a solution of $0.2\,mol\,l^{-1}\,CH_3COOH + 0.25\,mol\,l^{-1}$ $NH_4F + 0.013\,mol\,l^{-1}\,HNO_3 +$ $0.001\,mol\,l^{-1}$ EDTA $+ 0.25\,mol\,l^{-1}$ NH_4NO_3
NTA	Nitrilotriacetic acid
pCu	Negative $\log_{10}$ of the Cu concentration in solution in moles per liter
pH	Negative $\log_{10}$ of the H^+ concentration in solution in moles per liter

See also: **Fertilizers and Fertilization**; **Iron Nutrition**; **Nutrient Availability**

Further Reading

Adriano DC (1986) *Trace Elements in the Terrestrial Environment*. New York: Springer-Verlag.

Barber SA (1984) *Soil Nutrient Bioavailability*. New York: John Wiley.

Mengel D and Rehm G (2000) Fundamentals of fertilizer application. In: Sumner ME (ed.) *Handbook of Soil Science*, pp. 155–175. Boca Raton, FL: CRC Press.

Mortvedt JJ (2000) Bioavailability of micronutrients. In: Sumner ME (ed.) *Handbook of Soil Science*, pp. 71–88. Boca Raton, FL: CRC Press.

Mortvedt JJ, Cox FR, Shuman LM, and Welch RM (eds) (1991) *Micronutrients in Agriculture*. Madison, WI: Soil Science Society of America.

Rendig VV and Taylor HH (1989) *Principles of Soil–Plant Interrelationships*. New York: McGraw-Hill.

Shuman LM (1998) Micronutrient fertilizers. In: Rengel Z (ed.) *Nutrient Use in Crop Production*, pp. 165–195. New York: Food Products Press.

Sims JT (2000) Soil fertility evaluation. In: Sumner ME (ed.) *Handbook of Soil Science*, pp. 113–153. Boca Raton, FL: CRC Press.

MINERAL–ORGANIC–MICROBIAL INTERACTIONS

P M Huang, University of Saskatchewan, Saskatoon, SK, Canada
M C Wang, National Chung Hsing University, Taichung, Taiwan
M K Wang, National Taiwan University, Taipei, Taiwan

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Minerals, organic components, and microorganisms are three major solid components of the soil. They profoundly affect the physical, chemical, and biological properties, and processes of terrestrial ecosystems. They should not be considered as separate entities but rather as a united system constantly in close association and interactions with each other in soil environments.

Interactions of these three solid components mediated by soil solution and atmosphere govern mineral weathering transformation, the formation of organo–mineral complexes, microbial and enzymatic activities, soil structural stability and dynamics of aggregate turnover, biogeochemical cycling of C, N, P, and S, and transformation and dynamics of organic pollutants and metals. Therefore, interactions of these components have enormous impacts on terrestrial processes critical to environmental quality and ecosystem health.

Mechanisms of Binding of Organic Matter by Soil Mineral Colloids

Nature of Mineral Colloid Surfaces

The surface properties of soils are dependent on the weathering processes of the parent materials that are initiated by geological, hydrological, and biological agents. Therefore, soil formation processes have a tremendous impact on the formation and modification of surface properties of soils. The surface of 2:1 expansible layer silicates is negatively charged. The ditrigonal cavities of the surface can thus adsorb positively charged ions and polymers to form hydroxy-interlayered minerals in soil and related environments

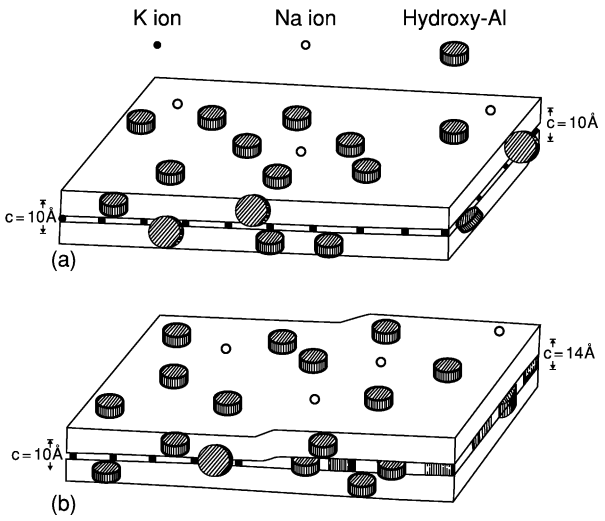


Figure 1 Structural models depicting the hydroxy-Al coatings on external surfaces (a) edge and planar surfaces of micas; and (b) of mica-vermiculite and on interlamellar surfaces of K-depleted layers of mica-vermiculite. Reproduced with permission from Huang PM (1980) Adsorption processes in soil. In: Hutzinger O (ed.) *The Handbook of Environmental Chemistry*, vol. 2/part A. *Reactions and Processes*, pp. 47–59. Berlin: Springer-Verlag.

(Figure 1). These hydroxy-interlayered minerals occur in soils throughout the world and are present in soils of several orders of soil taxonomy systems. In addition, positively charged hydroxy polymers can be adsorbed on the edges and external surfaces of phyllosilicates (Figure 1).

The complexity of surface properties of soil mineral colloids is further manifested by the formation of iron oxyhydroxide or calcium carbonate coatings on the external surfaces of layer silicates, the growth of two-dimensional solid solutions of trace metal oxides on the periphery of Mn oxide nodules, and the coating of calcium carbonate on Mn oxides. Similar effects can occur on the surfaces of other minerals in soils. The surfaces of soil mineral colloids can also complex with organic matter through the formation of inner-sphere and/or outer-sphere complexes. The surface complexation reactions may result in the formation of thick envelopes of colloidal organics on metal oxides and aluminosilicates. Therefore, the surface chemistry of soil minerals is mainly that of the functional groups exposed on the coatings rather than that of the groups in the underlying matrix. The surface chemistry of soil mineral colloids has a very significant effect on the transformation and dynamics of organic matter in soils.

Binding of Organic Matter

Organic molecules interact with the surfaces of mineral colloids by a variety of mechanisms, as summarized below. The reaction mechanisms mainly depend on the intrinsic nature and properties of the organic species, the kind of exchangeable cation on the surface of the mineral, the water content of the system, and the properties of mineral colloids.

Ion exchange Organic cations such as protonated amines and quarternary ammonium cations will enter into exchange reactions with metal cations on the surfaces of mineral colloids. Organic cations are often preferred over metal cations because of size considerations.

Protonation Some organic molecules may become cationic after adsorption at the surfaces of mineral colloids by protonation. The relative basicity of the organic species and the Brønsted acidity of the mineral surface determine whether or not protonation happens. The degree of protonation of the organic molecule is determined by its intrinsic basicity and the ability of the surfaces of the mineral colloids to donate protons. The nature of exchangeable cation, the water content, and the nature of mineral colloids play an important role in influencing the degree of protonation of organic molecules.

Coordination and ion dipole Coordination or ion–dipole interaction is another kind of mineral colloid–organic interaction. Where exchangeable cations are of the transition metal type, unfilled d orbitals will allow the formation of coordinate covalent bonds wherein an electron donor such as N or O on functional groups of organic molecules is available as a ligand. Where exchangeable cations are not able to form classical coordination complexes, they will interact with polar molecules through ion–dipole interaction processes.

Hydrogen bonding Hydrogen bonding is another type of interaction which is possible between organic molecules and mineral surfaces. Where a polar organic molecule cannot displace a water molecule solvating an exchangeable metal cation, it settles for a secondary role by H bonding with the directly coordinated water. This type of arrangement is referred to as 'water bridge,' where water acts as the linking entity between an exchangeable metal cation and a polar organic molecule.

Hydrophobic bonding on clay–organic complexes Natural clay–organic complexes may have adsorption properties quite different from pure clays. Many organic molecules, including aromatics and particularly the halogenated types such as chlorinated and brominated phenyls and biphenyls, are adsorbed a little, if at all, on clay surfaces. The surfaces of clays

can be made hydrophobic after reaction with organic molecules which have some hydrophobic properties. When this happens, the surfaces of clay–organic complexes become hydrophobic and in turn organophilic. Therefore, hydrophobic portions of the organic adsorbate can react with the hydrophobic organic portions of clay–organic complexes.

Anion adsorption Although pure clay minerals are negatively charged, hydroxy-interlayered clay minerals have positively charged sites (Figure 1). Organic anions can interact with polyvalent cations such as Al and Fe on the surface coatings of clay minerals.

Adsorption on external mineral surfaces The extent of adsorption of organic matter on external mineral surfaces depends on the physical and chemical characteristics of the surface, the pH of the system, and its water contents. Complexation reactions of humic acid (HA) and fulvic acid (FA) with structural cations of edges of mineral colloids are important binding mechanisms. Hydrogen bonding, which apparently involves H and O of COOH and OH groups of HA and FA and O and H of external planar surfaces and edges, is also of considerable importance.

Adsorption in clay interlayers Interlayer adsorption of FA by expansible silicates is pH-dependent, being greatest at low pH, and no longer occurring at pH 5.0 (Figure 2). An inflection occurs in the adsorption–pH curve near the pH corresponding to the pK of FA. Since the degree of dissociation of the FA is very low at low pH, undissociated FA can penetrate the clay interlayer and displace water coordinated to the cation in the interlayer. Therefore, interlayer adsorption of FA is greatest at low pH. One of the well-defined precursors for the formation of humic substances, hydroquinone, can be transformed in aqueous solution even at near-neutral pH (6.5) to humic macromolecules and adsorbed in the interlayers of nontronite (Fe(III)-bearing smectite) saturated with Ca (Figure 3); this is the most common and most abundant exchangeable cation in soils and sediments. The decrease in the organic carbon contents of the complex on heating between 150 and 500°C is accompanied by a decrease in basal *d* values. Most of the interlayer humic macromolecules are highly resistant to alkali extraction and are, thus, humin-type materials. Therefore, the formation of humic substance interlayers in 2:1 expansible-layer silicates, through polymerization of phenol monomers and associated reactions in soils and sediments, deserves close attention.

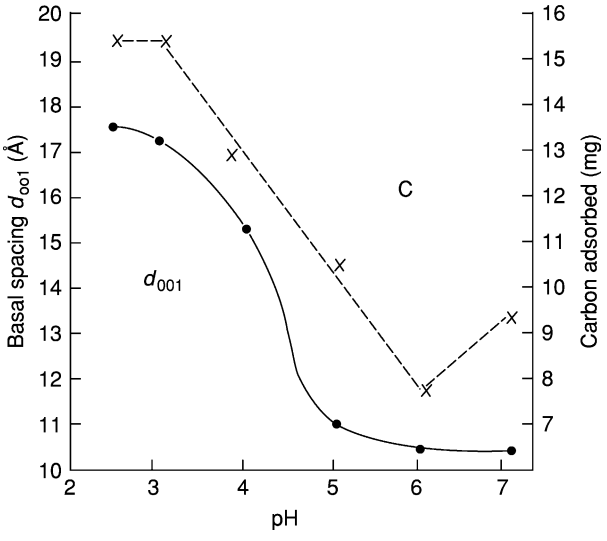


Figure 2 Effect of pH on basal spacing (d_{001}) and adsorption of fulvic acid. Reproduced with permission from Schnitzer M and Kodama H (1966) Montmorillonite: effect of pH on its adsorption of a soil humic compound. *Science* 153: 70–71.

Influence of Organic Matter on the Formation and Surface Chemistry of Metal Oxides

Metal oxides play a significant role in influencing soil behavior. They may exist as crystalline minerals, short-range ordered minerals, or noncrystalline precipitates which form surface coatings on clay minerals and humic substances. Organic components exert an important influence on the formation and surface properties of metal oxides.

Aluminum Oxides

The influence of a particular organic acid is generally related to the stability constant of the complex that the acid forms with Al (Table 1). Therefore, *p*-hydroxybenzoic acid, which forms the least stable complex with Al, does not inhibit the crystallization of Al hydroxides, whereas aspartic, tannic, malic, and citric acids increasingly retard crystallization. Besides the stability constant of the complex, the concentration of the organic acid is important. At certain low concentrations, the presence of some organic acids actually promotes the crystallization of particular $Al(OH)_3$ polymorphs by catalysis, but above a critical concentration, it disrupts crystallization through complexation with Al.

Humic substances also influence the transformation of Al by hampering the formation of more crystalline phases and promoting the formation of microcrystalline boehmite. FA and HA resemble aliphatic acids such as citric and malic acids, in that they contain COOH and aliphatic OH groups. They also resemble quercetin and tannic acid, because they contain phenolic

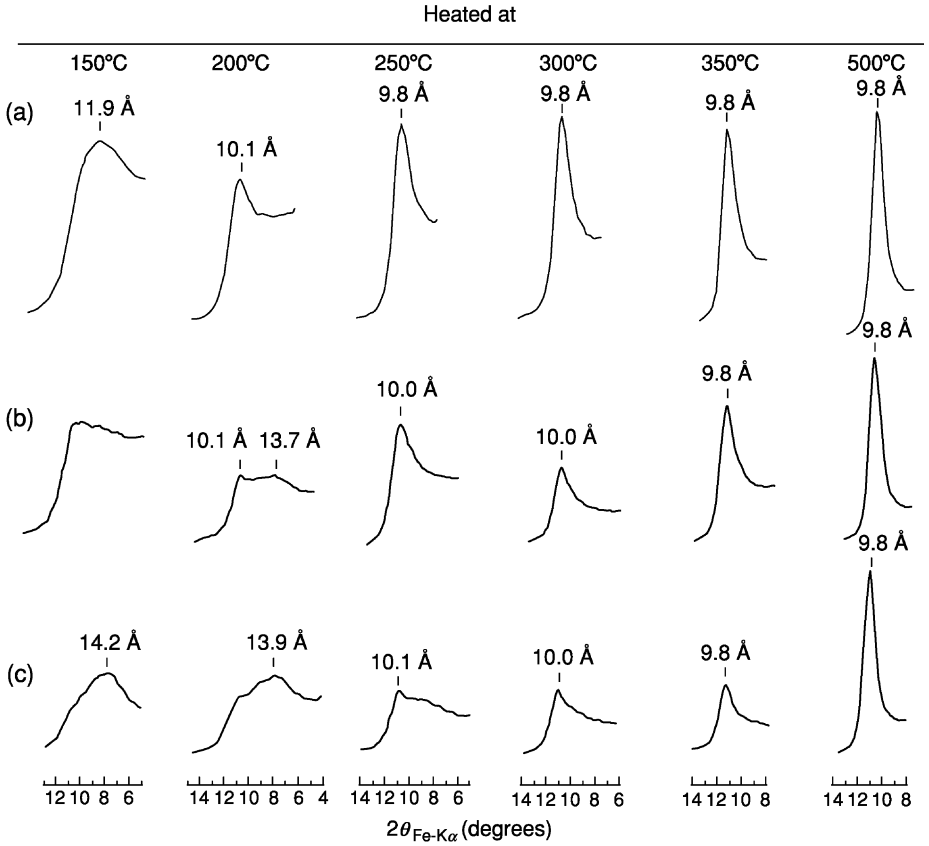


Figure 3 X-ray diffractograms of Ca-nontronite, influenced by deposition of humic macromolecules in the interlayers through polymerization of hydroquinone and the associated reactions. Diffractograms were recorded on a Philips X-ray diffractometer using Fe-Kα radiation with an Mn filter. (a) Ca-nontronite; humic macromolecule–Ca–nontronite complex before (b) and after (c) NaOH extraction. Reproduced with permission from Wang MC and Huang PM (1986) Humic macromolecular interlayering in nontronite through interaction with phenol monomers. *Nature* (London) 323: 529–531.

Table 1 Stability constants of the complexes formed between Al and five organic acids at 25°C

	Stability constants of the complexes	
Organic acids	$logK_1$	$logK_2$
p-hydroxybenzoic acid	1.66	–
Aspartic acid	2.60	–
Tannic acid	3.78	–
Malic acid	5.14	8.52
Citric acid	7.37	13.90

Reproduced with permission from Kwong KF, Ng Kee, and Huang PM (1979) The relative influence of low molecular weight, complexing organic acids on the hydrolysis and precipitation of aluminum. *Soil Science* 128: 337–342.

hydroxyl and ketonic C=O groups. Through these functional groups, FA and HA form stable complexes and inhibit the crystallization of Al hydroxides.

Organic components have an enormous impact on the surface properties of Al transformation products. For example, the presence of organic acids during aging of Al hydroxide gel for 40 days increases the specific surface of the products up to 30-fold over that of the control and higher organic acid concentrations result in higher specific surface. The surface charge characteristics of the products also dramatically change. The intermediate transformation products of Al, which include soluble mononuclear and polynuclear Al species and colloidal short-range ordered Al hydroxides, are the most reactive Al species in influencing physical, chemical, and biological processes of soil and related environments.

Iron Oxides

Soil organic matter plays a vital role in the formation of pedogenic Fe oxides (Figure 4). In soil environments where the amount of organic matter is low, for example in subsoils, the Fe supplied will form goethite and hematite depending on environmental factors. As organic matter content increases, more of the Fe will be complexed with organics, leading to decreases of Fe activity. The activity of Fe(III) ions

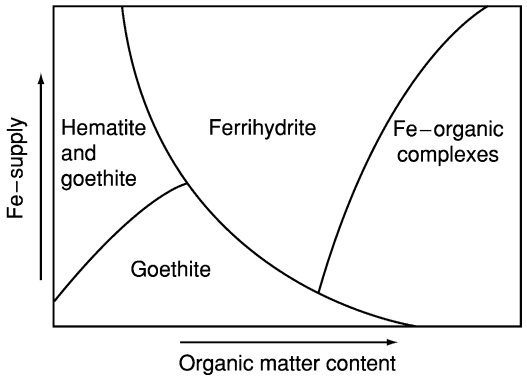


Figure 4 Schematic representation of the effect of organic matter content and rate of soluble Fe supply on the formation of various Fe forms in soils. Reproduced with permission from Schwertmann U, Kodama H, and Fisher WR (1986) Mutual interactions between organics and iron oxides. In: Huang PM and Schnitzer M (eds) *Interactions of Soil Minerals with Natural Organics and Microbes*, pp. 223–250. SSSA Special Publication 17. Madison, WI: Soil Science Society of America.

is so low that only the solubility product of goethite (10^{-41} to 10^{-42}), but not the solubility product of ferrihydrite (10^{-37} to 10^{-39}) is exceeded. Consequently, goethite but not ferrihydrite may form. Therefore, no hematite will form in an environment where the organic matter is high, since ferrihydrite is deemed to be a necessary precursor for hematite. This trend is generally observed in soils in temperate and cool regions as well as in wet depressions and surface soils of subtropical and tropical regions. At a higher content of organic matter, the rate of Fe supply is high; ferrihydrite will form and may survive for pedogenic times. If the content of organic matter is even higher, such as in O horizons or in peaty environments, all of the Fe may be in the form of Fe–organic complexes.

In an aqueous weathering environment, Fe oxides generally form via solution transformation. The oxidation products of Fe(II) solutions are thus important because it is in this valence state that Fe is commonly mobilized during weathering under the *E*h-pH region of natural soil environments. The kinetics of Fe(II) oxidation and the nature of Fe oxides are greatly influenced by a series of organic ligands. The rate constant of Fe(II) oxidation at the same ligand/Fe molar ratio generally decreases with the increase in the stability constant of the Fe(II)–ligand complex. The rate constant of Fe(II) oxidation also decreases with the increase in the perturbing ligand/Fe(II) molar ratio due to the increasing formation of the Fe(II)–ligand complex. The retardation of Fe(II) oxidation inhibits the nucleation of goethite and promotes the crystal growth of lepidocrocite. A further increase in the amount of the perturbing ligand distorts the structural order, resulting in the increasing inhibition of crystal growth of lepidocrocite and the formation of noncrystalline Fe oxides.

Surface properties of Fe oxides have a profound impact on microaggregate formation, water flux, nutrient and pollutant flux, and the ability of soils to promote plant growth, to maintain reasonable soil biotic habitat, and to respond to management, and resist degradation. The specific surface, charge characteristics, mean surface roughness, fractal dimension, and microporosity of Fe oxides depend on the nature and concentration of organic substances in the solution in which the Fe oxides are formed.

Role of Soil Minerals in Abiotic Catalysis of the Formation of Humic Substances

Polyphenol Pathway

Soil minerals play an important role in catalyzing the abiotic polymerization of phenolic compounds and the polycondensation of phenolic compounds and amino acids and the subsequent formation of humic substances. Layer silicates have the ability to catalyze the abiotic formation of humic substances. The promoting effect of 2:1 layer silicates is higher than 1:1 layer silicates. This is attributed to the higher specific surface and higher lattice imperfection of the former than the latter. The edges of kaolinite are virtually the only catalytic sites for the formation of phenol-derived humic substances. The edges of nontronite have a very important role as catalytic sites in the formation of hydroquinone-derived humic macromolecules. The Fe(III) in the octahedral sheet of nontronite also serves as a Lewis acid site to catalyze the oxidative polymerization of hydroquinone.

Aluminum, Fe, and Mn oxides, hydroxides, and oxyhydroxides also catalyze the oxidative polymerization of phenolic compounds. Among these minerals, Mn oxides are the most powerful catalysts in the transformation of phenolic compounds. Manganese oxides (birnessite, cryptomelane, and pyrolusite), which are common in soils, act as Lewis acids that accept electrons from phenolic compounds, leading to their formation of semiquinone and their oxidative polymerization and formation of humic substances. The catalytic power of metal oxides in affecting the C turnover and humic substances formation via abiotic processes in soil and related environments merits attention.

Poorly crystalline aluminosilicates are common in soils. Allophane has the ability to catalyze the polymerization of polyphenols. Proto-imogolite sol, which is present in soil solution, also plays a role in the abiotic formation of humic substances.

Primary minerals differ in their ability to catalyze the abiotic polymerization of hydroquinone. The sequence of the catalytic power of primary minerals is tephroite > actinolite > hornblende > fayalite > augite > biotite > muscovite ≃ albite ≃ orthoclase ≃ microcline ≃ quartz.

Maillard Reaction

The Maillard reaction is considered to be a major pathway in humification because of significant similarities between humic substances and melanoidins formed through this pathway, involving sugar–amino acid condensation. Birnessite (δ-MnO_2) significantly increases the extent of humification of the glucose–glycine system over the pH range of 6–8 (Figure 5). The chemical shifts of FA formed in the Maillard reaction system resemble those of natural humic substances. Further, light intensity of 168 μE $s^{-1}m^{-2}$ exerts an increase in the browning of the glucose–glycine–birnessite system compared with the same system kept in the dark. Equally important is that, even in complete darkness, birnessite catalyzes the Maillard reaction between glucose and glycine. Therefore, birnessite catalysis can occur in soil and sediment environments at any depth, but the presence of sunlight should accelerate the reaction.

Interactions of Soil Mineral Colloids with Microorganisms and Enzymes

Surface Interactions of Soil Minerals with Microorganisms

Soil is a habitat for myriads of microorganisms. The microbial biomass constitutes only a very small proportion (3%) of the total organic carbon in soil, but it is the most active and dynamic fraction of the living organic pool. Among the inorganic components of soil, mineral colloids are the most reactive because of their large specific surface and reactive charge characteristics. Being enriched in ions, water, and organic matter relative to the bulk soil, the surface of mineral colloids serves as a preferred habitat for soil microorganisms. Complex formation between mineral colloids and microorganisms can and does occur.

The surfaces of bacterial cells and crystalline clay minerals are both negatively charged. However, bacteria have a propensity for producing extracellular polysaccharides (EPS) which bind simultaneously

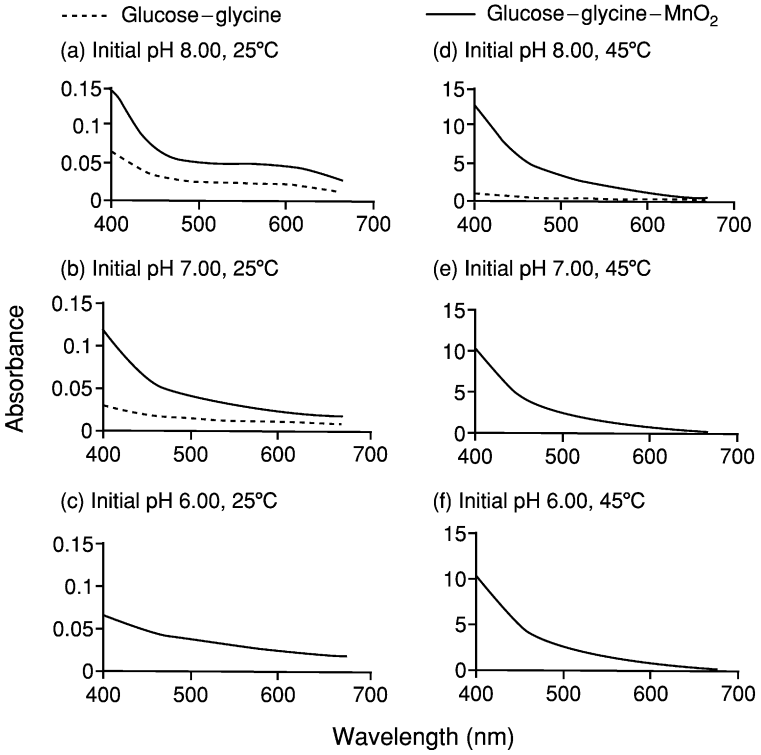


Figure 5 Absorbance versus wavelength plots in the Maillard reaction between glucose and glycine as influenced by birnessite catalysis. (a), (b), and (c): 30-day reaction period. (d), (e) and (f): 15-day reaction period. Reproduced with permission from Jokic A, Frenkel AI, Vairavamurthy MA, and Huang PM (2001) Birnessite catalysis of the Maillard reaction: its significance in natural humification. *Geophysical Research Letter* 28: 3899–3902.

to cell and clay surfaces through 'cation bridging,' involving polyvalent cations (Figure 6). Further, mineral colloids in soils are generally coated with hydroxy Al (or Fe) polymers (Figure 1). These coated minerals thus behave as a positively charged species or display amphoteric characteristics. Therefore, mineral colloids can strongly interact with negatively charged microbial cells in soil environments. This type of bonding, which is much stronger than cation bridging, is also expected to occur with Al and Fe oxides over the pH range of soils. The attachment of microorganisms to short-range minerals and the crystal edges of layer silicates through electrostatic interactions would also be predicted to occur when the soil pH falls below 6 because all of these surfaces would then be largely positively charged.

In the majority of cases, minerals in topsoils are partially covered with organic materials, especially humic substances (HS), which are microbially resistant and the most prevalent. The most common mode of mineral colloid–organic material–microorganism interactions may be represented by the following scheme:

$$[\text{Mineral colloid} - \text{HS}]^{-} \ldots . \text{M}^{n+} \ldots .^{-} [\text{EPS} - \text{B}]$$

In humic-rich calcareous Mollisols, Ca would be the predominant bridging cation. In Andisols, Oxisols, Ultisols, and the B horizon of spodosols, HS largely occur as complexes with Al and Fe, or their respective poorly crystalline oxides. In soils with little organic matter and in subsoils, mineral colloid–microorganism interactions are largely influenced by the mineralogical composition and pH of the system.

Mineral colloids can directly or indirectly influence the activity of microorganisms in their immediate vicinity. The effect of mineral colloids may be positive, negative, or sometimes so small as to escape detection. Mineral colloids have a stimulating effect on the activity of adhering bacteria by keeping the pH of microhabitats within the optimum physiological

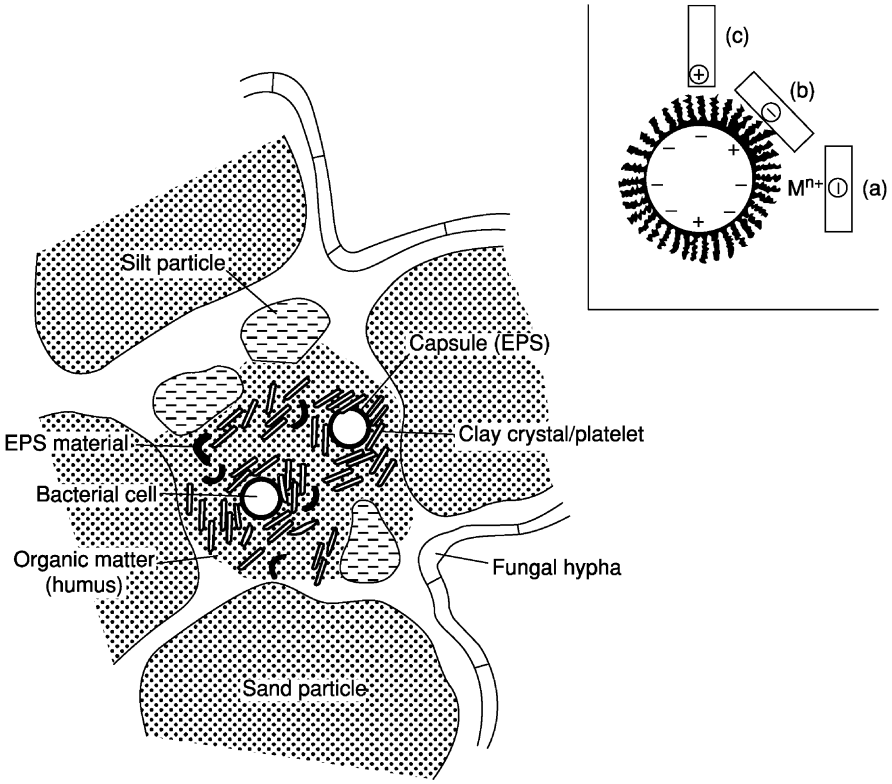


Figure 6 Diagram illustrating the interaction of bacteria and fungi with mineral particles in a soil aggregate. Bacterial cells with a coat of extracellular polysaccharides (EPS) are enveloped by clay particles. The pore space where clays and bacteria interact, bounded by silt- and sand-size particles, is relatively enriched in organic matter, including EPS residues. Fungal hyphae are attached to the outside surface of an aggregate. Inset: an enlarged view of a bacterial cell with its complement of EPS. At normal soil pH conditions, the cell has a net negative surface charge. Most clay particles adhere to the cell surface by bridging through polyvalent cations, represented by M^{n+} (a), although some may be attached directly by electrostatic interactions, either in a face-to-face (b), or edge-to-face (c) association. Reproduced with permission from Theng BKG and Orchard VA (1995) Interactions of clays with microorganisms and bacterial survival in soil: a physicochemical perspectives. In: Huang PM, Berthelin J, Bollag J-M, McGill WB, and Page AL (eds) *Environmental Impact of Soil Component Interactions*, vol. II. *Metals, Other Inorganics and Microbial Activities*, pp. 123–143. Boca Raton, FL: CRC/Lewis.

range for growth. Mineral colloids can also stimulate microbial activity by sorbing metabolites that would otherwise have an adverse effect on microbial growth. Further, the surfaces of mineral colloids can adsorb other toxic substances to microorganisms such as antibiotics and pesticides.

By forming an envelope around bacterial cells, mineral colloids may provide protection from extreme fluctuations in physicochemical environments, and thus enhance bacterial survival. The protective effect of mineral colloids, especially montmorillonite, is manifested in the ability of soil microorganisms to withstand exposure to hypertonic osmotic pressure, desiccation, and ultraviolet radiation. Mineral colloids can also act as a cementing agent of soil particles. Therefore, the addition of clay to soils, especially those of light texture, would modify the spatial arrangement of particles and especially the pore size distribution within soil aggregates. Such a modification of aggregate structure often benefits the bacterial population by increasing the proportion of pores of a certain size range ($<6\ \mu$m) from which bacteria could freely enter to colonize pores. However, bacterial predators – notably protozoa – are effectively barred from entering such pores due to steric hindrance.

The content and type of mineral colloids are influential in determining the balance between different microbial populations in soil. A well-known example is the failure of some fungi to thrive and spread in certain soils. This is largely attributed to the presence of montmorillonite in the 'suppressive' soils. Montmorillonite can serve as a proton sink and is thus able to promote the growth of acid-sensitive bacteria in these soils. This gives bacteria a selective advantage over fungi in competing for available nutrients, and fungal growth and proliferation are, thus, effectively suppressed.

Formation of Mineral Colloid– and Humic–Enzyme Complexes

Enzymes are rapidly sorbed at mineral colloids and humic substances in soil environments. Mineral colloids have a high affinity for enzymes, although this is not always synonymous with the retention of their enzymatic activity. Enzymes associated with soil humic substances retain their catalytic activity for long periods.

The mechanisms whereby enzymes are retained by mineral colloids have usually been assumed to include cation exchange reactions but not all observations have supported this view. Therefore, besides cation exchange reactions, adsorption of enzymes by mineral colloids may proceed through ionic, covalent, hydrophobic, and hydrogen bonding, and van der Waals forces. Although individual van der Waals forces are regarded as weak, retention may be cumulative if the enzyme is in close contact with the adsorbing surface. Recent studies show that enzymes such as tyrosinase have a very high affinity to the surfaces of hydrous manganese oxide birnessite. Analysis by atomic force microscopy (AFM) clearly shows the unfolding in an ellipsoidal shape of the tyrosinase protein after adsorption and immobilization on the mineral surfaces, especially the frayed edges of the phyllomanganate layer. AFM pictures along with X-ray diffraction and infrared absorption data of the tyrosinase protein complex show the coating of the mineral surfaces by the protein molecules as well as the changes in protein conformation.

Enzymes adsorbed or immobilized on mineral colloids operate in a structurally restricted system. The behavior of such enzymes toward their respective substrate is, thus, different from that occurring in a homogeneous, aqueous solution. When enzymes are adsorbed on mineral colloids, changes in the tertiary structure of the enzymes and their active sites decrease the activity or eliminate it altogether. However, there are notable exceptions to the 'adsorption-decline in activity' rule. Various supports show different enzyme immobilization capabilities (Table 2). There is considerable variation in the retained activities of the enzymes immobilized. The residual specific activities of laccase and peroxidase immobilized on all supports are high. Further, laccase immobilized on montmorillonite shows specific activity higher than that of the free enzymes. This is attributed to steric modification of the immobilized enzymes.

Many mechanisms have been proposed to account for the stability of enzymes which are complexed with soil humic substances. These include ion exchange, H-bonding, entrapment within three-dimensional micelles, lipophilic reactions, and covalent bonding during organic matter genesis. Functional groups of enzymes implicated in covalent bonding to humic polymers include terminal and basic amino, carboxyl, sulfhydryl, phenolic, and imidazole groups. All these may be involved in the stabilization process provided they do not form part of the active sites of the enzymes and not crucial to the retention of its tertiary structure. Enzyme–humic complexes can also be attached to mineral colloids and in some instances this further enhances enzyme stability.

Soil Mineral Weathering Transformations as Influenced by Microbial Activity

Mineral Weathering

Mineral weathering reactions can be enhanced by microbial activity by a factor as high as 10^6.

Table 2 Immobilization of a laccase (from *Trametes versicolor*) and a peroxidase (from horseradish) on different supports

		Enzymatic activity		
Enzyme and support	*Protein adsorbed*[a] *(mg/%)*	*Units adsorbed*[b]	*Specific activity*[c]	*Residual specific activity*[d] *(%)*
Laccase				
Glass beads	0.452/56	28.8	63.7	236
Montmorillonite	0.622/71	19.8	31.8	118
Kaolinite	0.566/64	13.1	23.1	85.5
Soil	0.644/73	15.7	24.4	90.4
Peroxidase				
Glass beads	0.092/17	8.4	91.6	93.8
Montmorillonite	0.224/43	23	102.8	105.2
Kaolinite	0.120/23	9.5	78.9	80.7
Soil	0.162/31	15	92.6	94.8

[a]Difference between proteins initially added to 200 mg of support (0.88 mg laccase and 0.52 mg of peroxidase) and those recovered in the supernatant and washings.
[b]Expressed as μmol O_2 consumed min^{-1} for laccase and μmol guaiacol transformed min^{-1} for peroxidase.
[c]Units adsorbed/protein adsorbed.
[d]Calculated as a percentage of the specific activity (sa) of the free enzyme (laccase, sa = 27 $\mu mol^{-1}\, min^{-1}\, mg^{-1}$; peroxidase, sa = 97.7 μmol $min^{-1}\, mg^{-1}$).
Reproduced with permission from Glantreda L and Bollag J-M (1994) Effects of soils on the behaviour of immobilized enzymes. *Soil Science Society of America Journal* 58: 1672–1681.

Microorganisms can dissolve minerals by direct and indirect action under aerobic and anaerobic conditions. The modes of attack of minerals by microorganisms include: (1) direct enzymatic oxidation or reduction of a reduced or oxidized mineral components; (2) indirect attack with a metabolically produced redox agent or inorganic and organic acids; (3) indirect attack by metabolically produced alkali, usually in the form of ammonia; (4) indirect attack with a metabolically produced ligand that forms a highly soluble product with a mineral component; and (5) indirect attack by biopolymers. The mode of microbial attack of a mineral may involve a combination of some of these mechanisms.

Uptake and concentration of elements by absorption and/or adsorption were observed with bacteria and fungi for P, Ca, Fe, and K. Fungi can adsorb K from solution, shift K equilibrium, and transfer micas to vermiculites. Such process can also occur for major and trace elements. When oxidized metal compounds such as Fe(III), Mn(IV), and As(V) act as electron acceptors, anaerobic respiration becomes an example of direct dissolving action under anaerobic conditions. Oxidation of sulfur entities of metal sulfides to obtain energy is an example of direct dissolving action under aerobic conditions. Volatilization of metals and metalloids or biomethylated metal and metalloid compounds from the soil into the atmosphere can be a mechanism of detoxification for toxic elements such as Hg, As, and Se.

Fine-Grained Mineral Formation

The physical and chemical characteristics of bacteria, such as their large surface area-to-volume ratio, serve to increase the metal-binding capacity of their charged surfaces, leading to precipitation and formation of mineral phases on their cell walls or surface polymers. The mechanisms by which bacteria initiate the formation of minerals in bulk solution vary widely between species. There may be a combination of biochemical and surface-mediated reactions during the process. Bacteria surface layers may passively adsorb and indirectly serve as a nucleation template. Bacteria can also more directly initiate mineral precipitation by producing reactive compounds, such as enzymes, metallothioneins, and siderophores, which bind metals and metalloids and catalyze their transformations. In addition, bacteria can instigate the spontaneous precipitation of metals and metalloids by altering the chemistry of their microenvironments.

An example of microbially mediated fine-grained mineral development is the formation of Mn oxides. Microbial oxidation of Mn(II) is a major process that can produce Mn oxide coatings on soil particles 10^5 times faster than abiotic oxidation. Manganese oxides are highly reactive minerals and can restrict the mobility of metals in soil and related environments. Biogenic Mn oxides have significantly larger specific surface and higher Pb adsorption capacity than abiotically precipitated Mn oxides. Bioformation of minerals thus merits attention in the remediation of metal contamination in the environment.

Impacts of Mineral–Organic Matter–Microbe Interactions on Soils in the Environment

Soil Structural Stability

The processes by which soil organic matter strengthens the bonds between soil mineral particles

are complicated. Root exudation and microbial action produce organic compounds with a range of composition and molecular weights. These compounds interact with the mineral particles which also vary in size, shape, crystallinity, and electric charge. Interactions between soil mineral particles, organic substances, and microorganisms can occur at many different size scales, since the materials have a large size range in soils (Figure 7). The mechanisms for stabilization of soil structure vary with aggregate size.

The adsorption of organic molecules such as microbially derived polysaccharides and other unaltered

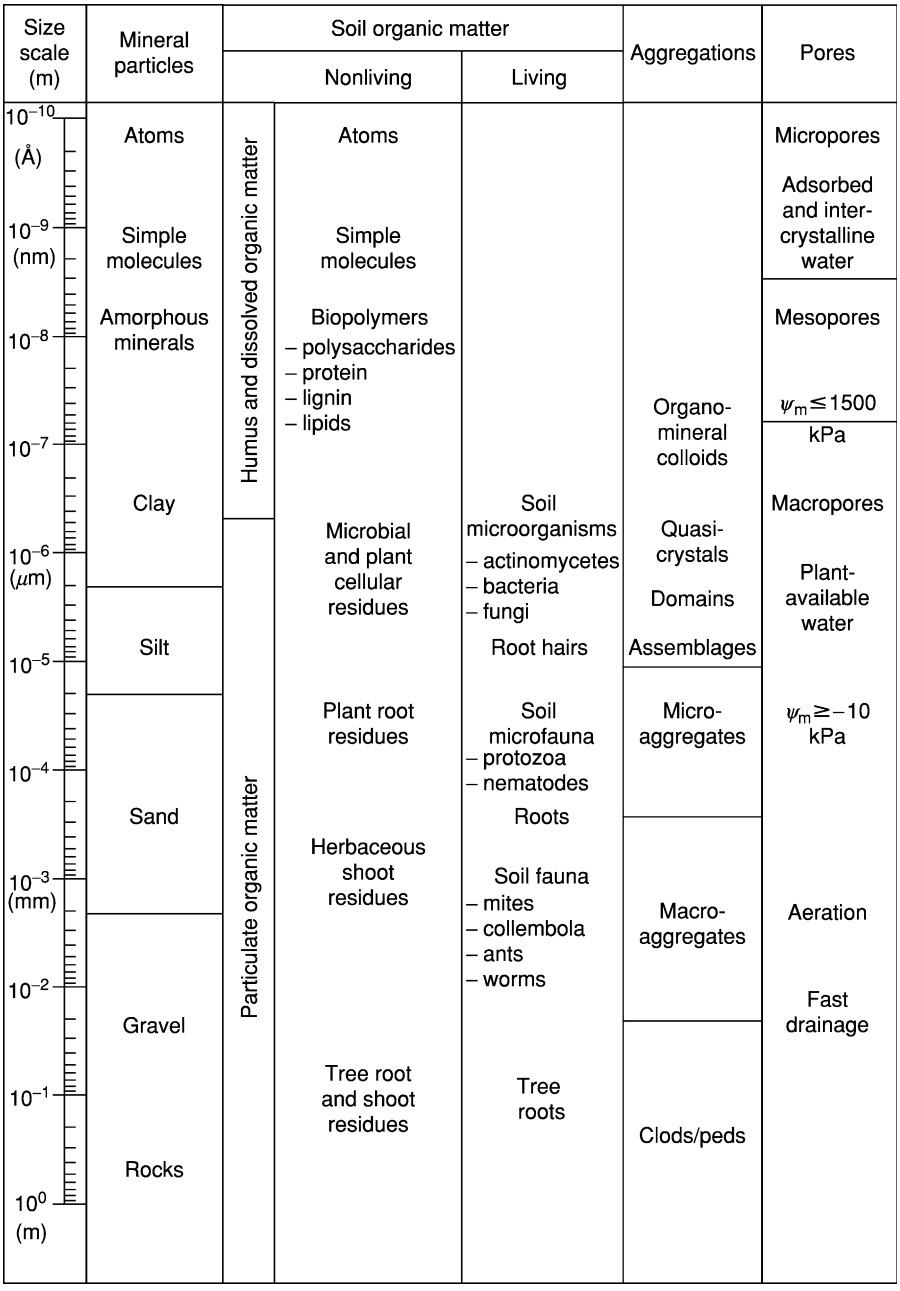


Figure 7 Size scales associated with soil mineral particles, organic components, pores, and aggregations of mineral and organic components. The definitions of pore size have used those developed by International Union of Pure and Applied Chemistry (IUPAC) (micropores <2 nm, mesopores 2–50 nm and macropores >50 nm). Alternatively the pore sizes corresponding to the lower ($\psi_m = -1500$ kPa) and upper ($\psi_m = -100$ kPa) limits of water availability to plants may be used to define the boundaries between the different classes of pore size. ψ_m is soil water metric potential. Reproduced with permission from Baldock JA (2002) Interactions of organic materials and microorganisms with minerals in the stabilization of structure. In: Huang PM, Bollag J-M, and Senesi N (eds) *Interactions Between Soil Particles and Microorganisms. Impact on the Terrestrial Ecosystem*, pp. 85–131. IUPAC Series on Analytical and Physical Chemistry of Environmental Systems. Chichester, UK: John Wiley.

and altered biomolecules on mineral surfaces can enhance the stability of individual clay microstructures. Adsorption is also important in the binding-together of clay microstructures and silt particles into small microaggregates with 2–50 μm diameters and densities $>20\,\mathrm{Mg\,m^{-3}}$. The high stability of these small microaggregates is demonstrated by their resistance to ultrasonification.

Mechanisms of stabilization of soil structure can operate over larger distances to bind microaggregates together to form macroaggregates (Figure 7). In view of the distances involved, the stabilization of macroaggregates is related to the presence of nonliving particulate organic matter (POM) capable of spanning distances $>100\,\mu$m or the existence of a network of fungal hyphae and plant roots that physically enmeshes microaggregates. The death of roots and hyphae growing within and through microaggregates results in the formation of biochemical binding agents capable of stabilizing the structure of macroaggregates.

Biochemical Cycling of C, N, P, and S in Soil

Chemically stabilized organic C in soils is slowly degraded to inorganic forms on time scales from centuries to millennia. Soil minerals are known to play a stabilizing role in soil organic matter. The interaction of Al and Fe with organic matter is of primary importance in the determination of the content of organic matter in tropical and temperate soils. The Al and Fe that complex and stabilize humus against microbial decomposition are released from soil minerals during soil formation and present as polymeric hydroxy cations. Their supply rates apparently control the content of soil organic matter to a great extent, as indicated by the relationship between pyrophosphate-extractable C and pyrophosphate-extractable Al and Fe.

The data from the radiocarbon analyses to explore interactions between soil mineralogy and soil organic C in volcanic soil environments show that the abundance of noncrystalline minerals accounts for >40% of the variation in organic C content across all the mineral horizons, substrate age, and soil orders (excluding the O and A horizons dominated by fast-cycling plant litter). Noncrystalline minerals also strongly influence turnover of organic matter.

Generally, more than 95% of the N and S and between 20 and 90% of the P in surface soils are present in soil organic matter. The close relationship between the organic forms of C, N, P, and S is well-established. The turnover of organic C is closely associated with the dynamics of N, P, and S in soils. The release of the stored organic nutrients depends on the mean residence time of soil organic matter associated with the mineral phase. The mean residence time of soil organic matter varies widely with the type of organomineral associations and the spatial location within the aggregate structure of soil (Table 3). Pools I and II are generally related to plant fragments, divided into easily available cell constituents and lignocelluloses at various degrees of degradation, respectively. Pool III consists of soil biomass and readily available organic matter within aggregates. On average, pools I–III account for about 20–30% of the total C in soil organic matter. These pools must be renewed continuously by fresh plant residues to maintain a relatively constant nutrient level and release by mineralization. The balance between decay and renewal processes, which is sensitive to management, controls the availability of N, P, and S. Pool IV is physically protected soil organic matter and is also affected by cultivation, because physical disturbance such as ploughing destroys macroaggregates and large microaggregates. Pool V is chemically stabilized soil organic matter which has the longest mean residence time. This pool represents 50–70% of the total C in soil organic matter and is seldom affected by management practices. Noncrystalline minerals have the ability to stabilize soil organic matter and thus reduce the turnover of C, N, P, and S.

Transformations of Organic Pollutants

Organic pollutants can be degraded through biotic and abiotic catalysis. Enzymes are biotic catalysts and mineral colloids are abiotic catalysts. It is frequently difficult to determine whether an organic pollutant is transformed abiotically or biotically. In many cases, significant abiotic and biotic catalytic reactions take place simultaneously.

The degradation of organic pollutants by extracellular enzymes is well-documented. Malathion, for instance, is hydrolyzed to a monocarboxylic acid (via loss of an ethyl group) by a carboxyl esterase isolated from soil. Enzymes commonly occurring in soil, such as esterases, amidases, phosphatases, and proteases, catalyze the hydrolysis of the respective chemical bonds in xenobiotic molecules.

Most organic chemicals, including xenobiotics, exhibit a strong affinity to humic substances. However, transformation of xenobiotics in terrestrial systems is greatly influenced by mineral components of soil. Mineral colloids, which are abundant in soil and have large specific surface and relatively high-charge density, contribute to the overall xenobiotic transformation at least as much as does the organic matter.

Soil organic matter and especially humic substances exert catalytic or inhibiting effects on abiotic hydrolysis of organic pollutants. Humic and fulvic acids have the ability to enhance the acid hydrolysis

Table 3 Comparison of estimated mean residence times of soil organic matter (SOM) in soil physical fractions

	Mean residence time (years)			
Pool	*Jenkinson and Rayner (1977)*[a]	*Parton et al. (1987)*[b]	*Buyanovsky et al. (1994)*[c]	Carter (1996)[d]
I	Decomposable plant material, 0.24	Metabolic plant residues, 0.1–1	Vegetative fragments 2–0.2 mm, 0.5–1	Litter, 1–3
II	Resistant plant material, 3.33	Structural plant residues, 1–5	Vegetative fragments 0.05–0.025 mm, 1–3	Free particulate organic matter (light fraction), 1–15
III	Soil biomass, 2.44	Active SOM pool, 1–5	OM in aggregates 2–1 mm, 1–4	Microbial biomass, 0.1–0.4
IV	Physically stabilized OM, 72	Slow SOM pool, 25–50	OM in aggregates 1–0.1 mm, 2–10	Intermicroaggregate OM[e], 5–50
V	Chemically stabilized OM, 2857	Passive SOM pool, 1000–1500	OM in fine silt, ~400	Intramicroaggregate OM[f] physically sequestered, 50–1000
			OM in fine clay, ~1000	Chemically sequestered, 1000–3000

[a]Jenkinson DS and Rayner JH (1977) The turnover of soil organic matter in some of the Rothamsted classical experiments. *Soil Science* 123: 298–305.
[b]Parton WJ, Schimel DS, Cole CV, and Ojima DS (1987) Analysis of factors controlling soil organic matter levels in Great Plains grasslands. *Soil Science Society of America Journal* 51: 1173–1179.
[c]Buyanovsky GA, Aslam M, and Wagner GH (1994) Carbon turnover in soil physical fractions. *Soil Science Society of America Journal* 58: 1167–1173.
[d]Carter MR (1996) Analysis of soil organic matter storage in agroecosystems. In: Carter MR and Stewart BA (eds) *Structure and Organic Matter Storage in Agricultural Soils*, pp. 3–11. Advances in Soil Science. Boca Raton, FL: CRC Press.
[e]Organic matter stored within macroaggregates but external to microaggregates; includes coarse occluded particulate organic matter and microbial organic matter.
[f]Organic matter stored within microaggregates; includes fine occluded particulate organic matter and microbial-derived organic matter.
Reproduced with permission from Guggenberger G and Haider KM (2002) Effect of mineral colloids on biogeochemical cycling of C, N, P and S in soil. In: Huang PM, Bollag J-M, and Senesi N (eds) *Interactions between Soil Particles and Microorganisms. Impact on the Terrestrial Ecosystem*, pp. 267–322. IUPAC Series on Analytical and Physical Chemistry of Environmental Systems, vol. 8. Chichester, UK: John Wiley.

of phenoxy acetic acids and esters and chloro-s-triazines and to retard the alkaline hydrolysis of the *n*-octyl ester of 2,4-D. The mechanisms proposed to explain the catalytic effects of humic and fulvic acids on dechlorohydroxylation of the chloro-s-triazines, simazine, atrazine, and propazine, consist of the interaction through H-bonding between the surface carboxyl groups of humic and fulvic acids and the heterocyclic nitrogen atoms of the triazine ring.

Soil minerals play an important role in catalyzing the abiotic transformation of organic compounds. Their ability to catalyze the transformation substantially varies with their structural configuration and surface chemistry and the structure and functionality of the organic matter involved. Manganese oxides are most powerful soil minerals in catalyzing abiotic oxidation of phenolic compounds. Organic compounds such as 2,4-D and ethyl ether can be degraded by the catalysis of birnessite (δ-MnO_2). These organic compounds can be adsorbed on birnessite and rapidly oxidized, both producing CO_2 as a major product (**Figure 8**), but by somewhat different mechanisms. Therefore, the solid degradation of organochlorine herbicides can occur by catalysis of birnessite, which is common in the environment.

Mineral surfaces may also detoxify adsorbed organic pollutants by catalysis through their ability to behave as Brønsted acids and donate protons or to act as Lewis acids and accept electron pairs. Brønsted acidity derives essentially from the dissociation of water molecules coordinated to surface-bound cations. Therefore, this acidity is strongly influenced by the hydration status and polarizing power of surface-bound and structural cations on mineral colloids. Certain organic pollutants such as organophosphate and s-triazine pesticides can be degraded by catalysis of mineral colloids through their surface acidity. This type of reaction deserves attention in a wide range of other organic compounds in soil environments. Besides the Brønsted acidity, the Lewis acidity of metals such as Al and Fe exposed at the edges of minerals is important in mineral-catalyzed hydrolysis reactions, at least for those that are hydroxyl ion-catalyzed. The ability of a metal ion to catalyze the hydrolysis varies with its ability to complex with reactant molecules and shift electron density and conformation in ways favorable to the reaction.

The degradation of organic pollutants may be considerably reduced when they are retained by soil mineral and humic colloids. The major reason for reduced biodegradation rates is the diminished bioavailability of chemicals involved in binding processes. The availability of sorbed xenobiotics to microorganisms varies with the chemical properties of the pollutant, the

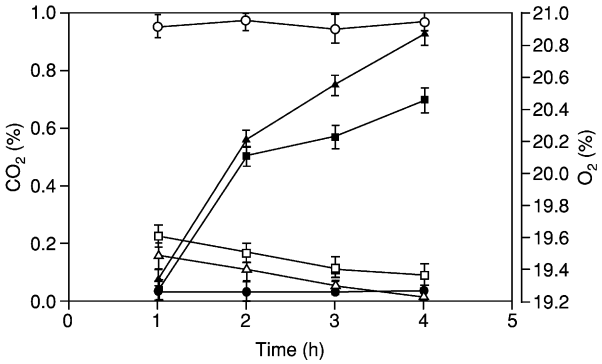


Figure 8 Gas chromatographic measurements of CO_2 and oxygen in the headspace of reaction vials and 2,4-D at 1, 2, 3, and 4 h, respectively. The values indicate the mole percentage of CO_2 and oxygen in the headspace gas. Filled circles, CO_2 from air control; filled squares, CO_2 from ethyl ether control; filled triangles, CO_2 from 0.3 μmol 2,4-D plus ethyl ether sample; open circles, oxygen from air control; open squares, oxygen from the ethyl ether control; open triangles; oxygen from 0.3 μmol 2,4-D plus ethyl ether. The error bars indicate the standard error for two measurements. Reproduced with permission from Cheney MA, Sposito G, McGrath AE, and Criddle RS (1996) Abiotic degradation of 2,4-D (dichlorophenoxyacetic acid) on synthetic birnessite: calorespirometer method. *Colloids and Surfaces. A. Physicochemical and Engineering Aspects* 107: 131–140.

nature of the sorbent, the mechanism of sorption, and the properties of the organisms. The impact of structural configuration and surface properties of soil mineral and organic components on the activity of microorganisms and their ability to degrade organic pollutants with different structure and functionality merit increasing attention for years to come.

Transformations of Metals

The transformation of metals in the environment is influenced by soil mineral–organic matter–microbe interactions. The impacts of these interactions on metal transformation are especially important in the soil rhizosphere because of root exudation. This leads to colonization by different populations of bacteria, fungi, protozoa, and nematodes. Plant–microbe interactions result in intense biological processes in the rhizosphere. These interactions, in turn, affect physicochemical reactions in the rhizosphere. Physicochemical properties that can be different in the rhizosphere include acidity, concentration of complexing biomolecules, redox potential, ionic strength, moisture, and nutrient status. The total rhizosphere environment is governed by an interacting trinity of the soil, the plant, and the organisms associated with the root. Therefore, reactions in the soil rhizosphere, which is the bottleneck of metal contamination of the terrestrial food chain, can only be satisfactorily integrated based on physical, chemical, and biological interactive processes.

A series of complexation reactions in soils affect metal transformation in the rhizosphere. Complexation reactions of metals with organic ligands are significant in determining the chemical behavior, availability, and toxicity of metals in the rhizosphere. The plant and prolific microbial activity results in increased amounts of organic ligands at the soil–root interface. Therefore, a larger fraction of the metal contaminant is in a complexed and usually soluble form in the rhizosphere soil than in the bulk soil.

The transformation of metals is significantly influenced by adsorption–desorption reactions in soil environments. These reactions are affected by biochemical and biological processes, especially in the rhizosphere. There is increasing evidence that organic ligands modify the adsorption–desorption reactions of metals in soils. For most of the trace metals, direct precipitation from solution through homogeneous nucleation appears to be less likely than adsorption–desorption by virtue of the low concentrations of these metals in soil solutions in well-aerated dryland soils. In aerobic soils, although precipitation of trace metals through homogeneous nucleation is not likely, heterogeneous nucleation may play a significant role in metal transformation because of the presence of mineral, organic, and microbial surfaces that catalyze the nucleation set of crystallization. The energy barrier to nucleation is reduced or removed by surfaces. The catalytic processes reduce the extent of supersaturation necessary for precipitation to occur.

Besides physicochemical reactions, metals have easy access to microbial surfaces through diffusion. All microorganisms contain biopolymers, such as proteins, nucleic acids, and polysaccharides, which provide sites where metal ions can bind. These binding sites include negatively charged groups such as carboxylate, thiolate, phosphate, and groups such as amines which coordinate to the metal center through lone pairs of electrons. Because of the ability of these biopolymers to bind metals, large concentrations of metals are frequently associated, not only with living microbial biomass, but also with dead cells. Many metals bind with various degrees of tenacity to the largely anionic outer-surface layers of microbial cells. Some metals are bound by cell walls to a greater extent than by clay minerals (Table 4), indicating that bacterial cell walls and membranes may act as foci for accumulation of metals in soils. Surface metal concentrations frequently exceed the stoichiometry expected per reactive chemical sites within cell walls. The sorption of metals can be so great that precipitates can be formed and distinct metallic minerals are eventually formed by biomineralization (the

Table 4 Metals bound by native *Bacillus subtilis* walls, *Escherichia coli* envelopes, kaolinite, and smectite

	Amount of metal bound (μmol g^{-1}) (oven-day weight)[a]			
Metal	*Walls*	*Envelopes*	*Kaolinite*	*Smectite*
Ag	423 ± 15	176 ± 3	0.46 ± 0.02	43 ± 0.3
Cu	530 ± 13	172 ± 9	5 ± 0.03	197 ± 4
Ni	654 ± 25	190 ± 3	4 ± 0.2	173 ± 10
Cd	683 ± 19	221 ± 6	6 ± 0.2	1 ± 0.02
Pb	543 ± 11	254 ± 5	3 ± 0.2	118 ± 6
Zn	973 ± 13	529 ± 32	37 ± 1	65 ± 2
Cr	435 ± 37	102 ± 2	8 ± 0.5	39 ± 5

[a] The data represent the average of three to five determinations for each sample from duplicate experiments and the standard error.
Reproduced with permission from Walker SG, Flemming CA, Ferris FG, Beveridge TJ, and Bailey GW (1989) Physicochemical interaction of *Escherichia coli* cell envelopes and *Bacillus subtilis* cell walls with two days and ability of the composite to immobilize heavy metals from solution. *Applied and Environmental Microbiology* 55: 2976–2984.

formation of minerals by microorganisms). Biomineralization has global consequences in dynamics, toxicity, and fate of metal pollutants.

See also: **Microbial Processes:** Environmental Factors; Community Analysis; Kinetics; **Organic Matter:** Principles and Processes; Genesis and Formation

Further Reading

Cheney MA, Sposito G, McGrath AE, and Criddle RS (1996) Abiotic degradation of 2,4-D (dichlorophenoxyacetic acid) on synthetic birnessite: calorespirometer method. *Colloids and Surfaces. A. Physicochemical and Engineering Aspects* 107: 131–140.

Gianfreda L and Bollag J-M (1994) Effect of soils on the behavior of immobilized enzymes. *Soil Science Society of America Journal* 58: 1672–1681.

Huang PM, Bollag J-M, and Senesi N (eds) (2002) *Interactions between Soil Particles and Microorganisms. Impact on the Terrestrial Ecosystem*. IUPAC Series on Analytical and Physical Chemistry of Environmental Systems, vol. 8. Chichester, UK: John Wiley.

Huang PM, Wang MK, Kämpf N, and Schulze DG (2002) Aluminum hydroxides. In: Dixon JB and Schulze DG (eds) *Soil Mineralogy with Environmental Applications*, pp. 261–289. SSSA Book Series 7. Madison, WI: Soil Science Society of America.

Jastrow JD and Miller RM (1997) Soil aggregate stabilization and carbon sequestration: feedbacks through organomineral associations. In: Lal R, Kimble JM, Follett RF, and Stewart BA (eds) *Soil Processes and the Carbon Cycle*, pp. 207–223. Boca Raton, FL: CRC Press.

Jokic A, Frenkel AI, Vairavamurthy MA, and Huang PM (2001) Birnessite catalysis of the Maillard reaction: its significance in natural humification. *Geophysical Research Letter* 28: 3899–3902.

Naidja A, Liu C, and Huang PM (2002) Formation of protein–birnessite complex: XRD, FTIR, and AFM analysis. *Journal of Colloid and Interface Science* 251: 46–56.

Schnitzer M and Kodama H (1966) Montmorillonite: effect of pH on its adsorption of a soil humic compound. *Science* 153: 70–71.

Schwertmann U, Kodama H, and Fisher WR (1986) Iron oxides. In: Huang PM and Schnitzer M (eds) *Interactions of Soil Minerals with Natural Organics and Microbes*, pp. 223–250. SSSA Special Publication 17. Madison, WI: Soil Science Society of America.

Stotzky G (1986) Influence of soil mineral colloids on metabolic processes, growth, adhesion, and ecology of microbes and viruses. In: Huang PM and Schnitzer M (eds) *Interactions of Soil Minerals with Natural Organics and Microbes*, pp. 305–428. SSSA Special Publication 17. Madison, WI: Soil Science Society of America.

Tebo BM, Ghiorse WC, van Waasbergen LG, Siering PL, and Caspi RC (1997) Bacterially mediated mineral formation: insights into manganese oxidation from molecular genetic and biochemical studies. In: Banfield JF and Nealson KH (eds) *Geomicrobiology: Interactions between Microbes and Minerals, Reviews in Mineralogy,* vol. 35, pp. 225–266. Washington, DC: Mineralogical Society of America.

Theng BKG and Orchard VA (1995) Interactions of clays with microorganisms and bacterial survival in soil: a physicochemical perspective. In: Huang PM, Berthelin J, Bollag J-M, McGill WB, and Page AL (eds) *Environmental Impact of Soil Component Interactions,* vol. II, *Metals, Other Inorganics and Microbial Activities*, pp. 122–143. Boca Raton, FL: CRC/Lewis.

Torn MS, Trumbore SE, Chadwiçk OA, Vitousek PM, and Hendricks DM (1997) Mineral control of soil organic carbon storage and turnover. *Nature (London)* 389: 170–173.

Walker SG, Flemming CA, Ferris FG, Beveridge TJ, and Bailey GW (1989) Physicochemical interaction of *Escherichia coli* cell envelopes and *Bacillus subtilis* cell walls with two clays and ability of the composite to immobilize heavy metals from solution. *Applied and Environmental Microbiology* 55: 2976–2984.

Wang MC and Huang PM (1986) Humic macromolecular interlayering in nontronite through interaction with phenol monomers. *Nature (London)* 323: 529–531.

MINERALS, PRIMARY

P M Huang, University of Saskatchewan, Saskatoon, SK, Canada
M K Wang, National Taiwan University, Taipei, Taiwan

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Primary minerals are formed at elevated temperatures and inherited from igneous and metamorphic rocks, sometimes through a sedimentary cycle. In most soils, the sand and silt fractions consist largely of primary minerals. In the clay fraction of weakly weathered soils, primary minerals are present but are minor constituents of the clay fraction of most agricultural soils.

The most abundant primary minerals in soils are quartz and feldspars, just as they are the dominant rock-forming minerals in the Earth's crust. Quartz consists of a continuous framework of silica tetrahedra and is the main form of free silica occurring in soils. Feldspars are anhydrous, three-dimensional aluminosilicates containing varying amounts of Na, K, and Ca, and occasionally of other large cations such as Ba. Feldspars are found in virtually all sediments and soils in quantities that vary with the nature of the parent material and the stage of weathering. Weathering of feldspars is an important Earth-surface geochemical process. Feldspars are stores of the macronutrients K and Ca; they play a substantial role in overall K dynamics of soils.

Micas are 2:1 phyllosilicates with tightly held, nonhydrated, interlayer cations balancing a high layer charge. They are largely of primary origin, occur extensively, and serve as precursors for expansible 2:1 phyllosilicates, vermiculites, and smectites in soil environments. Micas are often present in soils as components of particles that have been only partially transformed to expansible 2:1 minerals through interstratification with the other minerals or the formation of mica cores surrounded by expanded zones. Micas are important reserves of soil K for growing plants. The K-supplying power of a soil depends to a considerable extent on the nature and amount of micas present and their dynamics of K release.

Olivines, pyroxenes, and amphiboles are important accessory primary minerals that occur in small but significant amounts. Olivines are olive-green nesosilicates in which Mg and Fe^{2+} are octahedrally coordinated by O atoms. Pyroxenes and amphiboles, which are inosilicates, are ferromagnesian minerals with single- and double-chain structures, respectively, of linked silica tetrahedra. The variety of isomorphous substitution possible in olivines, pyroxenes, and amphiboles and their relative ease of weathering make them excellent source minerals for Ca, Mg, and trace elements in soils. Furthermore, the ability of nesosilicates and inosilicates to catalyze the abiotic formation of humic substances in soils merits attention.

Quartz

Silica occurs in nature as seven distinct polymorphs: quartz, cristobalite, tridymite, coesite, stishovite, lechatelerite (silica glass), and opal. Disordered cristobalite commonly occurs in soils; tridymite is rare in soils and usually associated with siliceous volcanic rocks; and coesite, stishovite, and lechatelerite are rare polymorphic forms. Opal is a hydrated, amorphous silica and is not uncommon, but, of these minerals, quartz is the most abundant in soil environments.

Quartz is one of the last minerals to crystallize from magma. Therefore, it is formed under conditions closer to present Earth-surface conditions than minerals crystallized earlier. This contributes to its high stability. Compared with other silica polymorphs, dense packing of the crystal structure and high activation energy required to alter the Si-O-Si bond are major factors contributing to the high stability of quartz. Quartz is present in essentially all soils and often constitutes the major portion of all sand and silt fractions. It is a major component of the coarse clay fraction of many soils.

The silica tetrahedron of quartz is almost symmetrical and has an Si-O distance of 0.161nm. The Si-O bond is approximately equally ionic and covalent in nature. The structure of quartz can be visualized as a spiral network of silica tetrahedra about the *z*-axis (**Figure 1**). Each tetrahedron is repeated in the network by a rotation of 120° and a translation of one-third about the *z*-axis. The tetrahedra are linked to form a hexagonal structure in quartz.

The solubility of silica minerals decreases as a function of increasing packing density of the silica tetrahedra and long-range crystal order. Solubility of quartz at ambient temperatures and neutral pH is commonly 3–7 mg Si l^{-1}, which is much lower than that of amorphous silica (50–60 mg Si l^{-1}). Freshly ground quartz commonly shows abnormally high solubility levels (37 mg Si l^{-1}). This has been attributed to the formation of a disrupted surface layer, which is believed

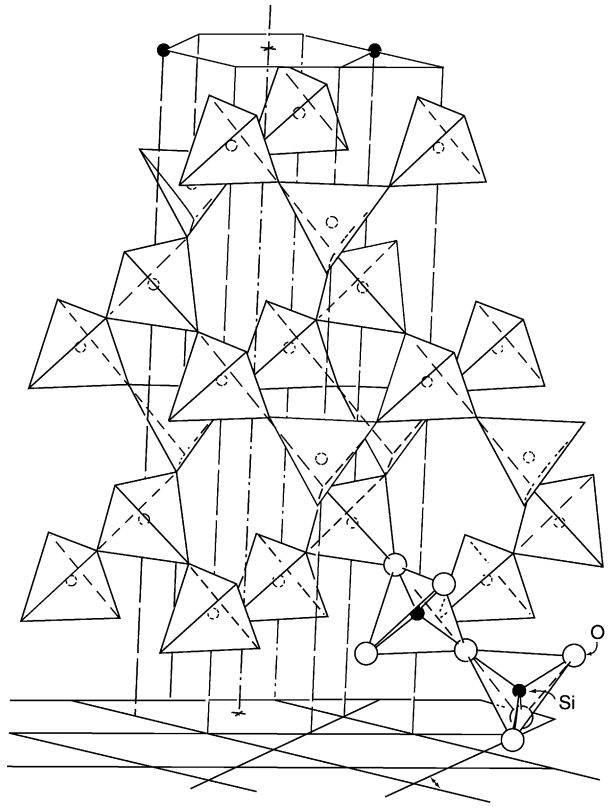


Figure 1 The structure of low quartz, SiO_2, shown as SiO_4 tetrahedra with a small silicon atom at the center of a group of four oxygens in tetrahedral coordination. Each tetrahedron shares a corner (oxygen) with an adjoining tetrahedron. For simplicity most of the tetrahedrons are shown without the circles representing oxygen. The screw axis symmetry (threefold) of the *c*-axis is shown by the dashed line, repeating each tetrahedron with a rotation of 120° and a translation of *c*/3. Two units of structure along *c* are shown. (Reproduced with permission from Berry LG and Mason B (1959) *Mineralogy*. San Francisco, CA: WH Freeman.)

to be amorphous or microcrystalline. Further, the solubility of quartz is a function of temperature, pH, and particle size. It increases with increasing temperature and is essentially constant between the pH values of 2 and 8.5, but increases rapidly above 9, due to the ionization of monosilicic acid:

$$Si(OH)_4 + OH^- = Si(OH)_3O^- + H_2O$$

The dissolution rate of quartz increases with the decrease in particle size. This is mainly attributed to an increase in specific surface of quartz as particle size decreases.

The presence of organic molecules such as amino acids greatly enhances the dissolution of silica, including quartz. The high rate of quartz dissolution in soils is attributable to leachates rich in organic molecules which react with quartz to form Si–organic molecular complexes in soil solution. The dissolution of quartz is greatest in the root zone, where production of biomolecules increases the dissolution of quartz by complexing with monosilicic acid.

Sesquioxides are a significant component of most soils and play a role in determining the concentration of dissolved silica in soils. They are responsible for much of the capacity of soils to sorb soluble silica, with the maximum adsorption occurring between pH values of 8 and 10. They act as a sink of dissolved silica and greatly increase the dissolution rate of silica. Therefore, finely divided quartz may not be stable at these pH values in the presence of excess sesquioxides.

Oxygen-isotope composition of minerals is a useful tool for determining the genesis and history of soil parent materials. The oxygen-isotope composition of quartz is dependent on the temperature of formation; delta values ($\delta^{18}O$) decrease as the temperature of formation increases. Delta values ($\delta^{18}O$) are defined as follows:

$$\delta^{18}O = \left[\frac{^{18}O/^{16}O \text{ sample}}{^{18}O/^{16}O \text{ SMOW}} - 1\right] \times 1000$$

where SMOW stands for standard mean ocean water. Quartz in igneous rocks has delta values in the order of 8–13‰; whereas quartz from metamorphic rocks has somewhat higher delta values, with a wider range that results from varying degrees of metamorphism. Oxygen-isotope ratios of quartz in shales range from 15 to 24‰; the ratios of quartz in sandstones range from 10 to 16‰. Silica minerals formed at ambient temperatures (chert, diatom, plant opal) have considerably higher delta values. The intermediate values for sedimentary deposits are attributed to the mixed origins of the sediments. Delta values of quartz from soils range from 9 to 30‰, which are generally similar to those of sedimentary rocks. Quartz is exceedingly resistant to oxygen-isotope exchange; therefore, the isotopic composition of quartz in soils, dusts, and sediments has been used as an indicator of provenance. The main use of isotopic composition of quartz in pedologic studies has been for identification of eolian additions to soils.

Feldspars

Feldspars make up an average of $595\,g\,kg^{-1}$ of igneous rock, $300\,g\,kg^{-1}$ of shale, and $115\,g\,kg^{-1}$ of sandstone. Many metamorphic rocks also contain large amounts of feldspars. Therefore, feldspars are present in virtually all sediments and soils in quantities that vary with the nature of parent material of soils and their degree of weathering. They play a substantial

role in overall dynamics of macronutrients, K and Ca, in soils.

The structure of feldspars is a three-dimensional framework of linked SiO_4 and AlO_4 tetrahedra, with sufficient opening in the framework to accommodate K, Na, Ca, or Ba to maintain electroneutrality. In the building up of the framework, four-membered rings of tetrahedra are the basic units. Chains of four-membered rings of tetrahedra are parallel to the *a*-axis and cross-linked to adjacent parallel chains by shared O atoms (Figure 2). One zigzag chain is formed by superimposing four-membered rings of tetrahedra that share some of their vertices to form new, four-membered rings. The remaining vertices form cross-links between chains. The repeat periodicity along the chain is 0.84 nm in the ideal case. The range of the *a* axis of different feldspars approximately varies from 0.81 to 0.86 nm. These four-membered rings are linked together to form a honeycomb-type arrangement.

Feldspars represent limited solid solution between the three end members (K-feldspar, albite, and anorthite) with pure K, Na, and Ca, respectively. Alkali feldspars are the K- and Na-rich members of the group with a small amount of Ca. Plagioclases are members of this group of minerals, rich in Ca and/or Na, but with small amounts of K.

Alkali feldspars have a range of chemical composition and the end members may be written as $KAlSi_3O_8$ and $NaAlSi_3O_8$. The K-feldspar polymorphs, sanidine, orthoclase, microcline, and adularia have identical chemical compositions. Sanidine is a monoclinic alkali feldspar with small optic axial angle (2V) and commonly occurs in volcanic rocks. Orthoclase is a monoclinic alkali feldspar that has the larger optical axial angle and looks homogeneous; it does not show cross-hatched twinning. Microcline is triclinic, has the larger optic axial angle (2V), and exhibits the typical cross-hatched twinning (Figure 3a). The alkali feldspars, which can be either monoclinic or triclinic but have a special crystal habit and occur in low-temperature hydrothermal veins, are referred to as 'adularia.' Anorthoclase is a triclinic, Na-rich feldspar that shows a very fine cross-hatched twinning.

Plagioclase feldspars have a chemical composition between pure albite and pure anorthite. The classification of the plagioclase series is based on the mole fractions of the albite (A_b) and anorthite (A_n) components (Table 1).

The high-temperature series of mixed crystals is quite continuous between the end-member orthoclase

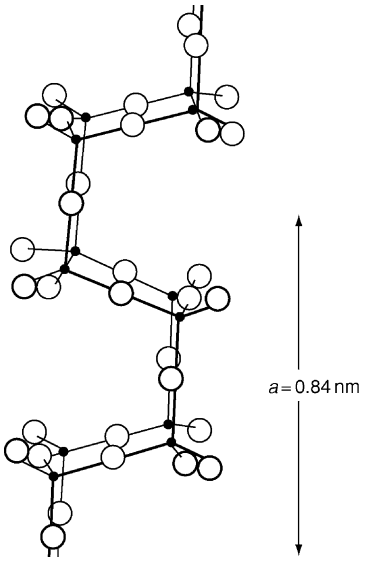


Figure 2 The essential structural feature of all feldspars projected on the (001) plane. The small black circles are Si and Al, the large circles, O. The O atoms projecting to right and left form the means of linking this chain to its neighbors. (Reproduced from Taylor WH (1965) Framework silicates: the feldspars. In: Bragg L and Claringbull GF (eds) *Crystal Structures of Minerals: The Crystalline State*, vol. IV. London, UK: G Bell.)

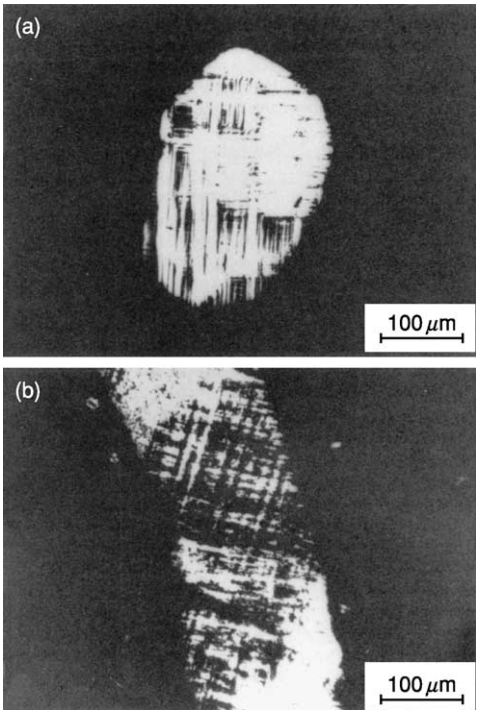


Figure 3 Photomicrographs of K-feldspar crystals from the Ap horizon of an Aridic Haploboroll (Orthic Brown Haverhill soil) in Saskatchewan, Canada: (a) a K-feldspar crystal with the cross-hatched pattern of maximum microcline of fine sand (50–250 μm) fraction; (b) a K-feldspar crystal, showing two phases of a coarse to very coarse sand (500–2000 μm) fraction. (Reproduced with permission from Somasiri S and Huang PM (1973) The nature of K-feldspars of selected soils in the Canadian prairies. *Soil Science Society of America Proceedings* 37: 461–464.)

Table 1 The classification of plagioclase feldspars

Plagioclase feldspars	*Mole fractions of albite (Ab) and anorthite (An)*
Albite	Ab_{100}–$Ab_{90}An_{10}$
Oligoclase	$Ab_{90}An_{10}$–$Ab_{70}An_{30}$
Andesine	$Ab_{70}An_{30}$–$Ab_{50}An_{50}$
Labradorite	$Ab_{50}An_{50}$–$Ab_{30}An_{70}$
Bytownite	$Ab_{30}An_{70}$–$Ab_{10}An_{90}$
Anorthite	$Ab_{10}An_{90}$–Ab_{100}

and albite and also between albite and anorthite. However, under low-temperature conditions, these phases undergo structural changes and tend to unmix. The unmixing results in the formation of a lamellar aggregate whose composition is similar to that of the end members. The lamellar aggregate, which is composed of a large amount of alkali feldspar and a subordinate amount of albite, is referred to as 'perthite.' Except for authigenic K-feldspar, which is formed through reconstitution and precipitation, most alkali feldspars usually contain varying amounts of Na in their structure and are thus more or less perthitic. The morphology of soil perthitic crystals that appear to contain a second phase, most likely albite, in a host crystal of K-feldspar is shown in Figure 3b.

Feldspars are present in nearly all soils, but their quantities vary with the intensity and capacity factors of weathering reactions. They are completely absent or present only in small quantities in strongly weathered soils, though the parent material contains considerable quantities of those minerals. However, some feldspars may be present in humid tropical soils that contain relatively fresh rock materials due to erosional and depositional processes. Feldspars are commonly present in the silt and sand fractions of young to moderately developed soils representing various soil parent materials and soil-forming conditions. Alkali feldspars are even present in the clay fraction of moderately weathered soils.

Persistence of feldspars in soils is related to the nature of the minerals, climate, topography, degree of leaching, complexation, redox potential, and certain solution ionic activities. According to the Goldich stability series, the increasing order of stability of feldspars is: anorthite < bytownite < labradorite < andesine < oligoclase < albite < K-feldspars. This order is related to the decrease in Al and Ca contents. Within the K-feldspar group, microcline is more stable than orthoclase. This is attributed to the difference in the volume occupied by the O atoms in the minerals. Compared with the monoclinic symmetry of orthoclase, the smaller volume is occupied by the O atoms of microcline by virtue of its triclinic symmetry.

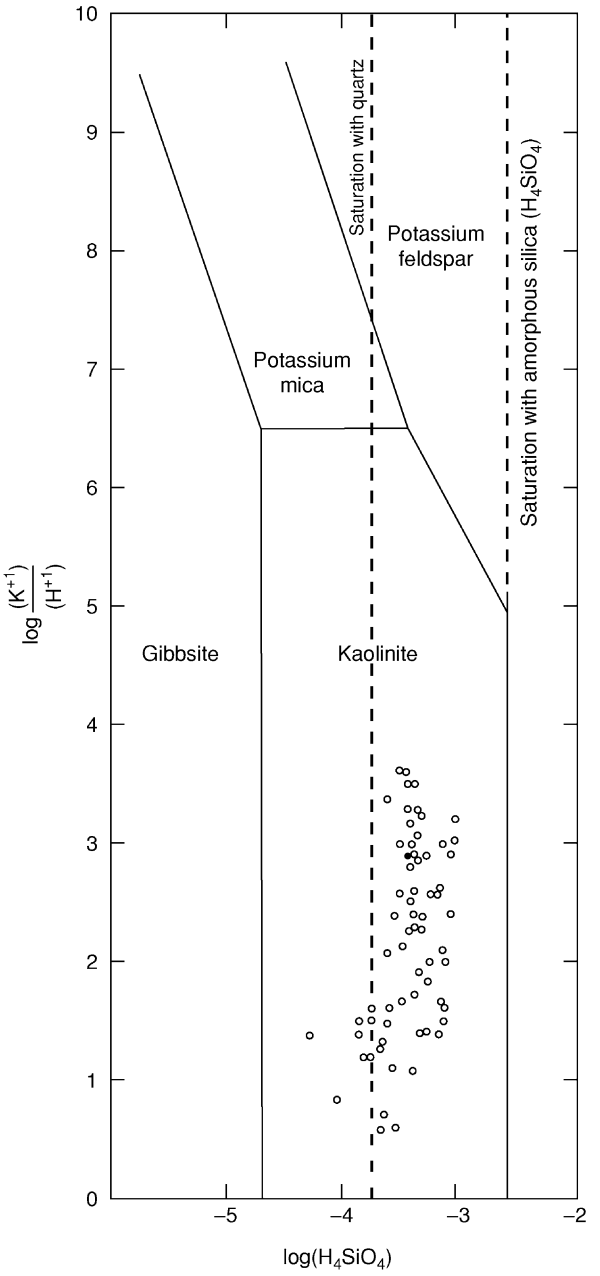


Figure 4 Stability relations of some phases in the system K_2O-Al_2O_3-SiO_2-H_2O at 25°C and 0.1 MPa pressure as functions of $(K^+)/(H^+)$ and (H_4SiO_4). Circles represent analyses of groundwaters. (Reproduced with permission from Feth JH, Robertson CE, and Polzer WL (1964) *Sources of Mineral Constituents in Water from Granitic Rocks, Sierra Nevada, California and Nevada*. US Geological Survey Water Supply Paper, 1535-I.)

Since establishment of the empirical stability series of Goldich, relative stabilities of feldspars at 25°C and 0.1 MPa have been thermodynamically shown in stability diagrams. The stability diagrams of K-feldspar and albite are shown in Figures 4 and 5, respectively. These stability diagrams are plotted as a function of the activity of H_4SiO_4 and the ratio of the

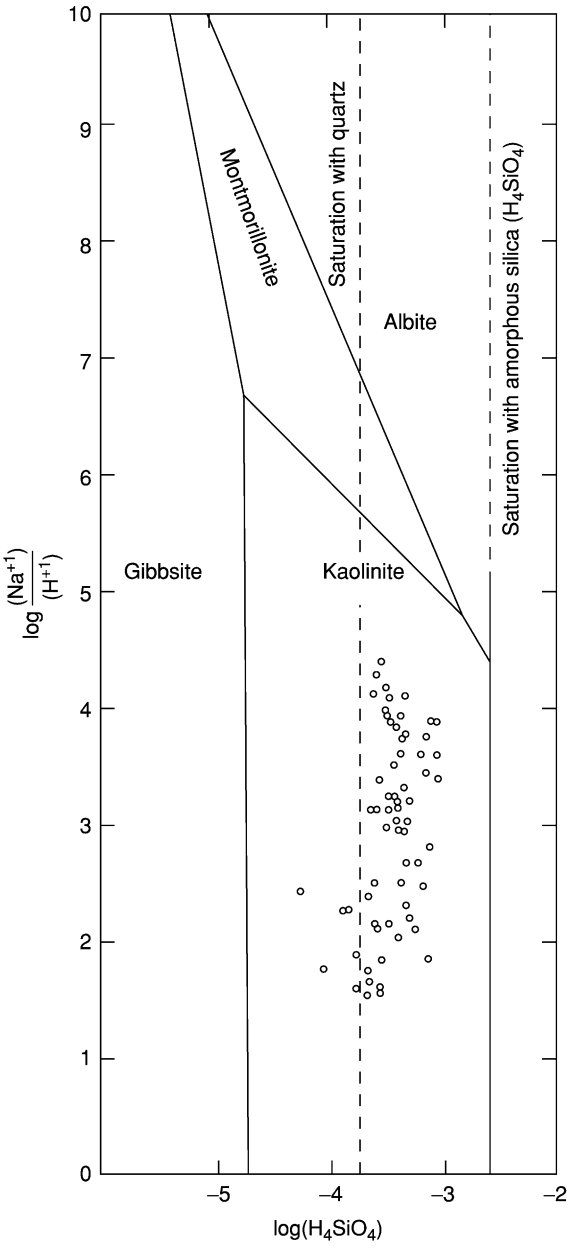


Figure 5 Stability relations of some phases in the system Na_2O-Al_2O_3-SiO_2-H_2O at 25°C and 0.1 MPa pressure as functions of $(Na^+)/(H^+)$ and (H_4SiO_4). Circles represent analyses of groundwaters. (Reproduced with permission from Feth JH, Robertson CE, and Polzer WL (1964) *Sources of Mineral Constituents in Water from Granitic Rocks, Sierra Nevada, California and Nevada*. US Geological Survey Water Supply Paper, 1535-I.)

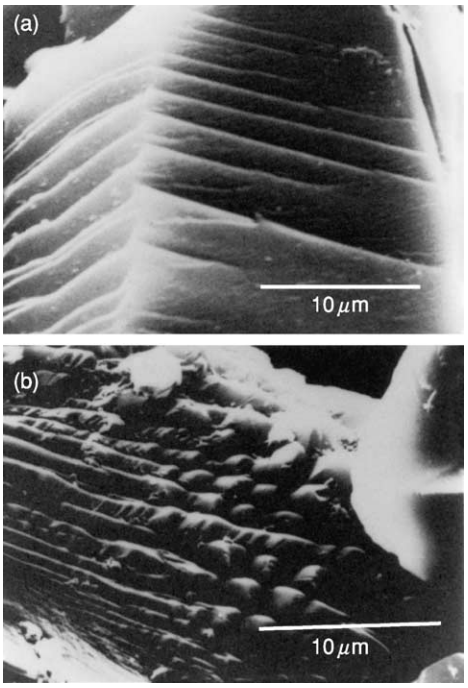


Figure 6 Scanning electron micrographs of an orthoclase particle treated with (a) water and (b) 0.01 mol l^{-1} oxalic acid. The mineral was treated for 20 days and the renewal treatments were then repeated twice for 5-day and 10-day periods. (Reproduced with permission from Song SK and Huang PM (1988) Dynamics of potassium release from potassium-bearing minerals as influenced by oxalic and citric acids. *Soil Science Society of America Journal* 52: 383–390.)

activity of each alkali ion to that of H ion. The diagrams show that mica and kaolinite are intermediate phases in the weathering of K-feldspar and that montmorillonite and kaolinite are intermediate phases in the weathering of albite. The ionic compositions of groundwater fall within the stability fields of kaolinite. This indicates that kaolinite is the stable mineral phase and mica and montmorillonite are unstable phases in groundwaters and eventually will be transformed to kaolinite. This is in agreement with the fact that groundwaters tend to transform rock minerals to kaolinite.

Feldspars are stores of the macronutrients K and Ca. Ligands of complexing organic acids may play a significant role in the release of these cations from the feldspar structure. Organic acids such as acetic, aspartic, salicylic, and citric acids have the ability to enhance the dissolution of feldspars. Their effect on dissolution increases in the order: acetic < aspartic < salicylic < citric. Citric acid is more effective than the other acids in releasing Ca, particularly from Ca-rich plagioclases, apparently because of the formation of Ca–organic complexes. In the case of the weathering of orthoclase by oxalic acid, the ridges on the cleavage surface (Figure 6) are evidently preferentially attacked by the acid and thus modified. The surfaces of orthoclase weathered by oxalic acid look much like those of naturally weathered K-feldspars. The sequence of K release from K-bearing minerals by oxalic and citric acids is biotite > microcline ≅ orthoclase > muscovite (Table 2). This differs from the sequence that occurs through cation exchange reaction

Table 2 K released from K-bearing minerals by various solutions at the end of a 10-day reaction period

Mineral	Organic acid[a]		
	Oxalic ($g\,kg^{-1}$ structural K)	Citric ($g\,kg^{-1}$ structural K)	NaTPB[b] ($g\,kg^{-1}$ structural K)
Biotite	44	11	1000
Muscovite	1.1	0.9	83
Microcline	2.3	1.4	1.6
Orthoclase	2.2	1.2	1.3

[a]The concentration of organic acid solutions was 0.01 mol l^{-1}.
[b]1 mol l^{-1} NaCl, 0.2 mol l^{-1} NaTPB, and 0.01 mol l^{-1} EDTA.
NaTPB, sodium tetraphenyl boron; EDTA, ethylenediaminetetraacetic acid.
Reproduced with permission from Song SK and Huang PM (1988) Dynamics of potassium release from potassium-bearing minerals as influenced by oxalic and citric acids. *Soil Science Society of America Journal* 52: 383–390.

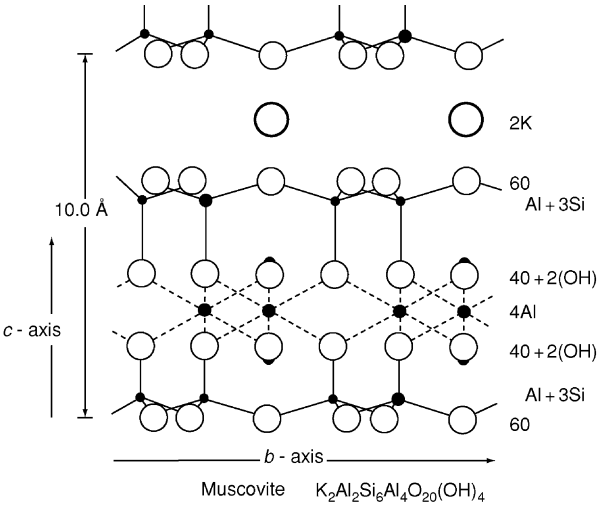


Figure 7 Schematic structure of mica. (Reproduced from Jackson ML (1964) Chemical composition of soils. In: Bear FE (ed.) *Chemistry of the Soil*. ACS Monograph Series 160. New York: Reinhold.)

using sodium tetraphenyl boron (NaTPB). The reaction of feldspars with complexing organic acids and its relation to nutrient release from soils merit attention.

Micas

Micas are 2:1 phyllosilicates with tightly held, nonhydrated, interlayer cations balancing a high layer charge. Primary attention is given to those micas with K^+ as the interlayer cation, because they are the abundant and important micas in most soils. Micas are abundant in many rocks, e.g., shales, slates, phyllites, schists, gneisses, and granites and in sediments derived from these and other rocks. The micas in soils are mainly inherited by soils from soil parent materials. Micas serve as precursors for expansible 2:1 phyllosilicates, i.e., vermiculites and smectites, to which micas may be transformed by replacement of the interlayer cations (usually K^+) by hydrated cations. Through K release, micas in soils are an important mineral source for plant growth.

The 2:1 layer of micas is composed of an octahedral sheet between two sheets of tetrahedra, as depicted in Figure 7. Micas can be classified as dioctahedral or trioctahedral, depending on the types and locations of cations in the octahedral sheets (Table 3). In trioctahedral micas such as biotite, all three octahedral positions are filled, whereas in dioctahedral micas such as muscovite, only two out of three octahedral cations are filled. Isomorphic substitution in the mica structure creates negative charge, which results in a strong coulombic attraction for charge-compensating interlayer cations such as K, which is not exchanged in the standard cation exchange-capacity determination. The negative layer charge of micas arises by some combination of three mechanisms: (1) substitution of R^{3+} (primarily Al or Fe) for Si^{4+} in tetrahedral positions, (2) substitution of R^{2+} for R^{3+} in octahedral positions, and (3) vacancies in octahedral positions. The resultant layer charge may originate entirely within the tetrahedral sheet, or may originate entirely within the octahedral sheet in some species, or may come partly from both sheets. The layer charge of micas is expressed on a formula unit basis (Table 3).

Muscovite is the most abundant dioctahedral primary mica. In muscovite, excess negative layer charge results from the substitution of Al for Si in one out of each of four Si^{4+} in the tetrahedral sheet positions. Biotite is the most common trioctahedral mica. Approximately three-quarters of the cation locations in the tetrahedral sheets of biotite are filled with tetravalent Si and about one-quarter is filled with trivalent Al. The excess negative charge generated by substitution of Al for Si results in the negative charge that is localized in the basal O atoms of the tetrahedral sheets where Al substitution occurs. In soils and sedimentary rocks, mica in the clay fraction is usually identified as illite. Illitic minerals have been called 'hydrous mica,' 'clay mica,' and 'sericite.' Illite is commonly found to be dioctahedral, but its chemical composition and layer charge differ from that of muscovite (Table 3). Glauconite structure has many similarities to that of illite, with relatively little minor substitution of Al for Si in the tetrahedral sheet. Compared with illite, the octahedral sheet of glauconite contains considerably more Fe.

After feldspars and quartz, micas are the third most extensive group of minerals in granitic and sialic

Table 3 General chemical formulas and layer charge for some dioctahedral and trioctahedral micas

Mineral	Layer charge (mol per formula unit)	Cations			Anions
		Tetrahedral sheet composition	Octahedral sheet composition	Interlayer cation	
Dioctahedral					
Paragonite	1	Si_3Al	Al_2	Na	$O_{10}(OH)_2$
Muscovite	1	Si_3Al	Al_2	K	$O_{10}(OH)_2$
Phengite	1	$Si_{3.5}Al_{0.5}$	$Al_{1.5}[MgFe^{(II)}]_{0.5}$	K	$O_{10}(OH)_2$
Illite	0.75	$Si_{3.5}Al_{0.5}$	$Al_{1.75}[MgFe^{(II)}]_{0.25}$	$K_{0.75}$	$O_{10}(OH)_2$
Glauconite	0.8	$Si_{3.67}Al_{0.33}$	$[AlFe^{(III)}]_{1.33}[MgFe^{(II)}]_{0.67}$	K	$O_{10}(OH)_2$
Margarite	2	Si_2Al_2	Al_2	Ca	$O_{10}(OH)_2$
Trioctahedral					
Phlogopite	1	Si_3Al	Mg_3	K	$O_{10}(OH)_2$
Biotite	1	Si_3Al	$Mg_{0.6–1.8}Fe^{(II)}{}_{1.2–2.4}$	K	$O_{10}(OH)_2$
Annite	1	Si_3Al	$Fe^{(II)}{}_3$	K	$O_{10}(OH)_2$
Lepidolite	1	Si_3Al	$[Li,Al]_3$	K	$O_{10}(OH,F)_2$
Clintonite	2	$SiAl_3$	Mg_2Al	Ca	$O_{10}(OH)_2$

Adapted from Bailey SW (1984) Classification and structures of the micas. In: Bailey SW (ed.) *Reviews in Mineraology, Micas*, vol. 13. Washington, DC: Mineralogical Society of America, with permission.

(acid) rocks in general but less extensive in most mafic rocks. On land areas of the world, the sialic rocks are more extensive than mafic rocks. Muscovite and biotite are the most extensive micas in igneous and metamorphic rocks. Phlogopite occurs as a product of metamorphism of magnesian limestones or dolomitic limestones and also in serpentine rocks. Micas are generally more extensive in fine-grained sediments and sedimentary rocks (e.g., shales) than in coarse-textured sedimentary rocks (e.g., sandstones). Shales and slates are usually rich in the fine-grained, illitic-type micas. Illitic-type micas are also important minerals in limestones. Much illite in sedimentary rocks is apparently detrital. However, some illite in these rocks may also have been developed or affected by authigenic or diagenetic processes. The contents and species of micas in more recent, unconsolidated parent materials such as glacial tills, loess, and alluvium vary with their origin.

Micas represent a major source of the macronutrient K and thus play an important role in plant nutrition. If the soil solution is low in K^+ activity, K near the edges of micaceous minerals may be released into solution by exchange with H^+, Na^+, Mg^{2+}, or Ca^{2+}. When the readily exchangeable K is exhausted, fixed K from the interlayer position can also be released. Estimating the power of micaceous minerals to supply K to the soil solution is thus critical for predictions of plant nutrition and growth.

Micas exhibit a strong preference for Cs^+. This strong affinity of micas for large, weakly hydrated cations such as Cs^+ has an important environmental application in determining the fate of some radioactive waste. Radioactive isotopes of Cs have been added to soils by disposal of radioactive wastes and by deposition of fallout from both nuclear weapon testing and accidents at nuclear energy facilities. The mobility and bioavailability of radiocesium in mineral soils is limited by nearly instantaneous sorption of $^{137}Cs^+$ at the frayed edges of mica particles followed by slower fixation of $^{137}Cs^+$ ions diffused into interlayer sites within mica crystals. Therefore, micas play a role in restricting the contamination of $^{137}Cs^+$ to the terrestrial food chain and the impact on human health.

Structural Fe contents of micas may be substantial. For example, Fe accounts for approximately 23% by mass of ideal biotite. Structural Fe can participate in electron-transfer reactions in soils. Three examples of mica involvement in environmentally significant electron-transfer reactions are the reduction of NO_3^- to NH_4^+, the reduction of Cr(VI) to Cr(III), and the reductive degradation of chlorinated hydrocarbons in soils and groundwater.

Common Accessory Primary Minerals

Besides quartz, feldspars, and micas, other common primary minerals are olivines, pyroxenes, and amphiboles. These minerals are accessory primary minerals, present in the sand- and silt-sized fractions of soils.

Olivines occur as the principal constituents of some igneous rocks. In the upper lithosphere, however, their concentration is certainly smaller than that of pyroxenes and amphiboles. These three groups of minerals belong to the accessory minerals of soils and are present in heavy specific-gravity fractions.

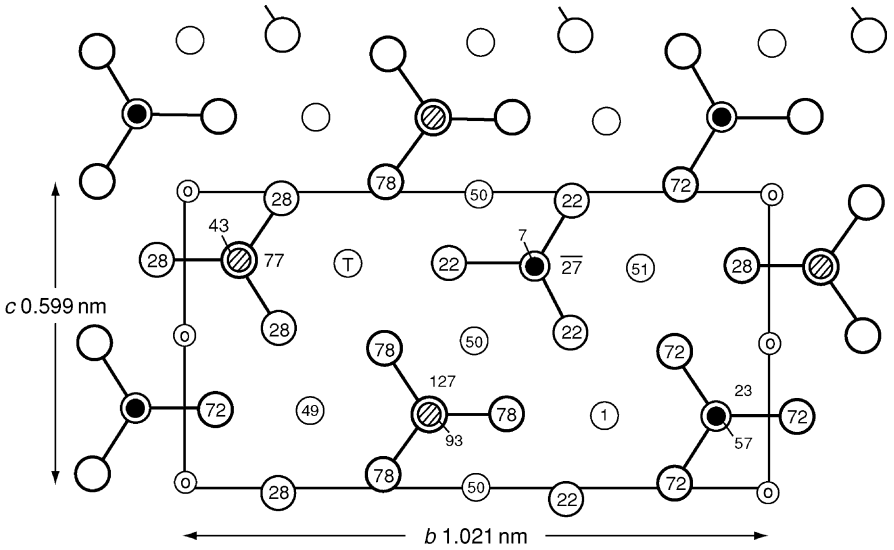


Figure 8 The structure of forsterite, Mg_2SiO_4, projected on the (100) plane. The largest circles are O, the smallest circles Mg, and the black or shaded circles Si; Si-O bonds mark out the SiO_4 tetrahedra. The kinds of positions of atoms are indicated by heights as shown by the numbers inside or beside the circles. (Reproduced from Bragg L and Claringbull GF (eds) (1965) *Crystal Structures of Minerals: The Crystalline State*, vol. IV. London, UK: G Bell.

Table 4 Certain members of olivine, pyroxene, and amphibole groups

Mineral	*Specific chemical formula*[a]
Olivines	
Forsterite	$(Mg_{0.96}Fe_{0.04})_2SiO_4$
Fayalite	Fe_2SiO_4
Tephroite	Mn_2SiO_4
Monticellite	$CaMgSiO_4$
Pyroxenes	
Enstatite	$MgSiO_3$
Hypersthene	$(Mg_{0.88}Fe_{0.12})Si_2O_6$
Orthoferrosilite	$FeSiO_3$
Clinoenstatite	$MgSiO_3$
Clinoferrosilite	$FeSiO_3$
Diopside	$CaMgSi_2O_6$
Pigeonite	$(Ca_{0.04}Mg_{0.45}Fe_{0.48})SiO_3$
Hedenbergite	$CaFeSi_2O_6$
Johannsenite	$(Ca_{0.45}Mg_{0.03}Fe_{0.05}Mn_{0.47})SiO_3$
Acmite-augite	$(Ca, Na)(Fe, Mn, Zn, Mg)(Si, Al)_2O_6$
Aegirine (acmite)	$NaFeSi_2O_6$
Jadeite	$NaAlSi_2O_6$
Spodumene	$LiAlSi_2O_6$
Amphilboles	
Anthophyllite	$(Mg, Fe)_7Si_8O_{22}(OH)_2$
Cummingtonite	$(Fe_{0.6}Mg_{0.4})_7Si_8O_{22}(OH)_2$
Tremolite[b]	$Ca_2Mg_5Si_8O_{22}(OH)_2$
Arfvedsonite	$(Ca, Na, K)_{2.64}Fe_{1.42}(Fe, Mn, Mg, Ti)_{3.54}(Si, Al)_8O_{22}(OH)_{2.15}$
Glaucophane	$Na_2(Mg, Fe, Al)_5Si_8O_{22}(OH)_2$
Riebeckite	$(Na, Ca, K)_{2.2}(Fe^{2+}, Mg, Mn, Zn, Cu, Li)_{3.0}(Fe^{3+}, Al, Ti)_{2.0}(Si, Al)_{8.0}O_{22.1}(OH, F)_{1.9}$
Hornblende	$(Ca, Na)_{2.26}(Mg, Fe, Al)_{5.15}(Si, Al)_8O_{22}(OH)_2$

[a]The specific chemical formulas are compiled from the *Data Book* (after Joint Committee on Powder Diffraction Standards (1974) *Search Manual. Selected Powder Diffraction Data for Minerals, Data Book*, 1st edn. Swarthmore, PA: International Center for Diffraction Data).
[b]The division between tremolite and actinolite is arbitrary: tremolite is the low-iron end of the series and actinolite comprises the high-iron members. Reproduced with permission from Huang PM (1989) Feldspars, olivines, pyroxenes, and amphiboles. In: Dixon JB and Weed SB (eds) *Minerals in Soil Environments*. Madison, WI: Soil Science Society of America.

Weathering of these minerals is one of the major processes in the geochemical cycles of Mg and Fe. The variety of isomorphous substitution possible in these minerals and their relative ease of weathering make them excellent source minerals for Ca, Mg, and trace elements in soils. Furthermore, the catalytic power of Mn-bearing nesosilicates and inosilicates in the abiotic formation of humic substances in soils merits attention.

Olivines

Members of forsterite-fayalite series, $(Mg, Fe)_2SiO_4$, comprise the most abundant naturally occurring olivines. The structure of forsterite is shown in Figure 8. The O atoms lie in approximately hexagonally close-packed sheets parallel to (100). The SiO_4 tetrahedra point alternately in opposite directions along the *a*- and *b*-axis. Every Mg atom is octahedrally coordinated by six O atoms. In the structures described above, Fe^{2+} may replace Mg^{2+}, because the two have identical charge and similar ionic radii. The specific chemical formulas of selected minerals in the olivine group are given in Table 4.

There is complete solid solution between forsterite and fayalite in nature. The Mg-rich olivines are predominant in basic rocks and the Fe-rich varieties in intermediate and acidic igneous rocks. Tephroite occurs mainly in metasomatic rocks and in Mn deposits.

The weatherability of olivines in soils is very high. Due to their relatively easy weatherability, olivines contribute to nutrient levels of Mg and Fe. The extent of weathering of these minerals is an indication of the intensity of chemical weathering of soils in cool humid regions. Where leaching is only moderate, olivines may alter to serpentine or trioctahedral smectite (presumably of the saponite type), nontronite, and various ferric hydrates and gels. In an environment where leaching is intense, the degradation products are poorly crystallized smectite, kaolinite,

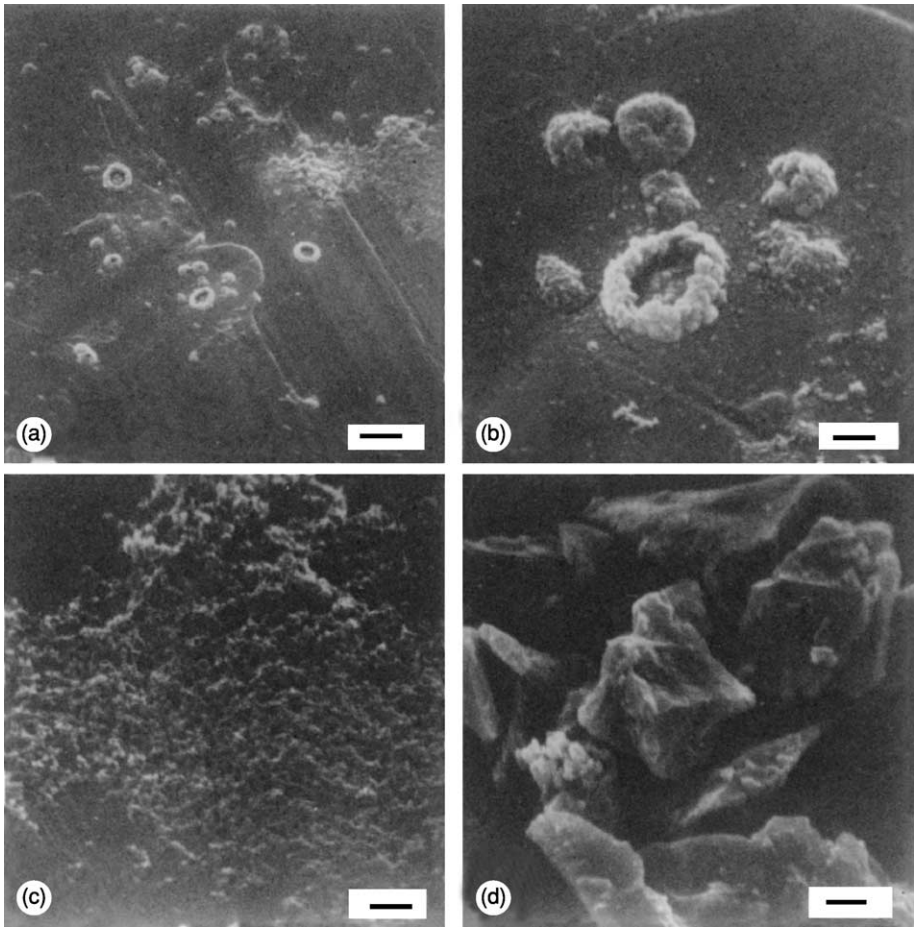


Figure 9 Scanning electron microscope micrographs of hydroquinone polymers in the supernatant and mineral particles settled in the tephroite system at the ratio of mineral to hydroquinone solution of 0.01 at the initial pH of 6.0 at the end of 7 d. (a–c) Hydroquinone polymers; (d) tephroite particles after reaction with hydroquinone. (a) Bar scale 10 μm; (b–d) 2 μm. (Reproduced with permission from Shindo H and Huang PM (1985) Catalytic polymerization of hydroquinone by primary minerals. *Soil Science* 139: 505–511.)

halloysite, and Fe oxides such as goethite, hematite, or the noncrystalline precursors of these minerals.

Olivines have the ability to accelerate the abiotic polymerization of hydroquinone to humic substances. The catalytic power of tephroite, which is an Mn-bearing olivine, in the polymerization of polyphenols such as hydroquinone to humic acids with a high degree of humification is especially noteworthy. The scanning electron micrographs of the hydroquinone-derived humic polymers formed in the presence of tephroite reveal various morphological configurations (Figure 9a–c). The surface features of these polymers are similar to those of soil humic and fulvic acids. Furthermore, the infrared (IR) spectra of these humic polymers with molecular weights of approximately 3500 and greater are similar to those of humic substances. The association of hydroquinone-derived polymers with the surfaces of tephroite is not evident (Figure 9d), indicating that most humic polymers formed in the tephroite system are present as discrete particles.

Pyroxenes and Amphiboles

Pyroxenes and amphiboles are closely related structurally. Pyroxenes consist of single chains of linked SiO_4 tetrahedra, each of which shares two O atoms with its neighbor; in amphiboles, double chains of the tetrahedra are linked together by O atoms and the tetrahedra share alternately two and three O atoms (Figure 10). The chains are indefinite in length. When viewed at right angles to the chain and parallel to the base of the tetrahedra, pyroxenes and amphibole chains look alike, as depicted in Figure 10c. The pyroxene and amphibole chains are parallel to the *c*-axis of the crystal and their planes of basal O atoms are perpendicular to (010).

The chains of pyroxenes and the double chains of amphiboles are linked together by various cations. The general chemical formula for pyroxenes is $R_2[Si_2O_6]$ and that for amphiboles is $R_{14}[(OH)_4 Si_{16}O_{44}]$. In these formulas, R is Mg, Fe^{2+}, or Ca and, in many cases, A1, Fe^{3+}, Ti^{3+}, Mn^{3+}, Na, K, or Li. The bond between the O atoms and the cations linking the chains is relatively weaker than that between the O and the Si. Therefore, cleavage takes place diagonally through the crystal and does not rupture the Si-O chains. The principal members of pyroxene and amphibole groups are listed in Table 4.

Three series of pyroxenes – the enstatite-hypersthene, diopside-jadeite, and augite series – occur in

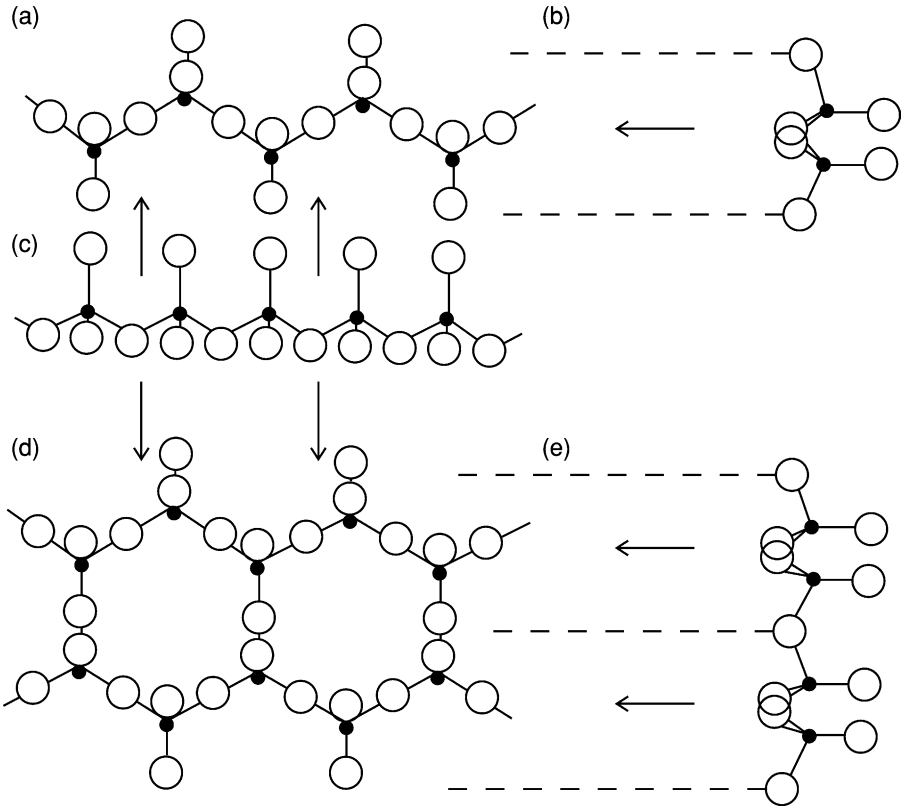


Figure 10 Si-O chains in structures of pyroxenes and amphiboles: (a–c) the pyroxene chain as seen in plane, in elevation, and end-on view, respectively; (c–e) the amphibole chain from the same three aspects. (Reproduced from Bragg L and Claringbul GF (eds) (1965) *Crystal Structures of Minerals: The Crystalline State*, vol. IV. London, UK: G Bells and Sons, Ltd.)

igneous rocks. In nature, pyroxenes are often mixtures. Among the three series of pyroxenes, the augite series is geochemically more important than the others. Its members are among the most important constituents of igneous rocks. Many amphiboles are absent in primary igneous rocks or may occur as minerals of secondary origin. The main primary amphiboles of igneous rocks are the Mg-poor riebeckites and the Si-poor basaltic and common hornblende. Hornblendes are widely distributed in igneous rocks from syenite and granite to gabbro, and in metamorphic rocks such as gneiss, hornblende schist, and amphibolites. Pyroxenes and amphiboles are largely confined to the sand and silt fractions of soils. Some of them may occur in the clay fraction of soils that are developed from glacial rock flour and have not been subjected to intensive weathering. The variety of isomorphous substitution in these minerals and their easy weatherability make them excellent sources of trace elements and Ca and Mg in soils.

Pyroxenes tend to weather to chlorite or smectite or both, through partial dissolution of Mg, Ca, and Fe^{2+}. Calcite may develop in the environment where the rate of dissolution of Ca is higher than the rate of complete breakdown of the pyroxene. As the chemical weathering proceeds further, all the Ca and Mg, and part of the Si are lost, resulting in progressive enrichment of the residue in kaolinite, ferric oxides and oxyhydroxides, and anatase. Amphiboles appear to have a weathering sequence similar to pyroxenes.

Pyroxenes (e.g., augite) and amphiboles (e.g., hornblende) have the ability to catalyze abiotic polymerization of polyphenols such as hydroquinone to humic substances. Therefore, these minerals along with olivines merit attention in the abiotic formation of humic substances in soil environments. These accessory primary minerals thus play a role in catalyzing humification and protecting soil C against mineralization and loss to the atmosphere as CO_2.

See also: **Factors of Soil Formation:** Parent Material; **Mineral–Organic–Microbial Interactions**

Further Reading

Bailey SW (1984) Classification and structures of the micas. In: Bailey SW (ed.) *Micas. Reviews in Mineralogy,* vol. 13, pp. 1–2. Washington, DC: Mineralogical Society of America.

Berry LG and Mason B (1959) *Mineralogy.* San Francisco, CA: WH Freeman.

Bloom PR and Nater EA (1991) Kinetics of dissolution of oxides and primary silicate minerals. In: Sparks DL and Suarez DL (eds) *Rates of Soil Chemical Processes*, pp. 151–189. SSSA Special Publication 27. Madison, WI: Soil Science Society of America.

Bragg L and Claringbull GF (1965) *Crystal Structure of Minerals: The Crystalline State,* vol. IV. London, UK: G Bell.

Churchman GJ (2000) The alteration and formation of soil minerals by weathering. In: Sumner ME (ed.) *Handbook of Soil Science*, pp. F3–F76. Boca Raton, FL: CRC Press.

Drees LR, Wilding LP, Smeck NE, and Senkayi AL (1989) Silica in soils; quartz and disordered silica polymorphs. In: Dixon JB and Weed SB (eds) *Minerals in Soil Environments*, pp. 914–974. SSSA Book Series 1. Madison, WI: Soil Science Society of America.

Fanning DS, Keramidas VZ, and El-Desoky MA (1989) Micas. In: Dixon JB and Weed SB (eds) *Minerals in Soil Environments*, pp. 551–634. SSSA Book Series 1. Madison, WI: Soil Science Society of America.

Feth JH, Robertson CE, and Polzer WL (1964) *Sources of Mineral Constituents in Water from Granitic Rocks, Sierra Nevada, California and Nevada*. US Geological Survey Water Supply Paper, 1535-I. Washington, DC: US Department of the Interior, US Geological Survey.

Huang PM (1989) Feldspars, olivines, pyroxenes, and amphiboles. In: Dixon JB and Weed SB (eds) *Minerals in Soil Environments*, pp. 975–1050. SSSA Book Series 1. Madison. WI: Soil Science Society of America.

Jackson ML (1964) Chemical composition of soils. In: Bear FE (ed.) *Chemistry of the Soil*, pp. 71–141. ACS Monograph Series No. 160. New York: Reinhold.

Laves FJ (1960) A1/Si-Verteilungen, Phasen-Transformationen und Namen der Alkali-feldspäte. *Zeitschrift fur Kristallographie, Kristallgeometrie, Kristallphysik, Kristallchemie* 113: 265–296.

Shindo H and Huang PM (1985) Catalytic polymerization of hydroquinone by primary minerals. *Soil Science* 139: 505–511.

Somasiri S and Huang PM (1973) The nature of K-feldspars of selected soils in the Canadian prairies. *Soil Science Society of America Proceedings* 37: 461–464.

Song SK and Huang PM (1988) Dynamics of potassium release from potassium-bearing minerals as influenced by oxalic and citric acids. *Soil Science Society of America Journal* 52: 383–390.

Taylor WH (1965) Framework silicates: the feldspars. In: Bragg L and Claringbull GF (eds) *Crystal Structures of Minerals: The Crystalline State,* vol. IV, pp. 293–333. London, UK: G Bell.

Minerals, Secondary *See* Clay Minerals

Minimum Tillage *See* **Conservation Tillage**

Miscible Displacement *See* **Solute Transport**

MORPHOLOGY

P R Owens, USDA Agricultural Research Service, Mississippi State, MS, USA
E M Rutledge, University of Arkansas, Fayetteville, AR, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Soil morphology is defined as the branch of soil science that deals with the description, using standard terminology, of *in situ* spatial organization and physical properties of soil regardless of potential use. Precise descriptions, using conventional terms, are necessary to all of the areas of science included in the field of soil science. The basic descriptors used today have been developed over the past 50 years and are continually added to. Prior to utilization of standard terminology, soils were described as clayey, sandy, stony, sedimentary, saline, marshy, dry or moist, heavy or light, soft or compact, fatty, friable, or lean. Soil horizons were described by terms such as 'gray watery sand' or 'rusty brown clay.' While these terms illustrate a fundamental understanding of soil properties, they do not impart any specific knowledge about the soils in question and cannot be compared with other soils described by other scientists. An objective, complete description of the soil is essential, because it serves as a basis for soil identification, classification, correlation, mapping, and interpretation.

The basic tools used by field soil morphologists are very simple and usually include a knife, tape measure, water bottle, Munsell color book, $\times 10$ hand lens, 10% HCl, and a clinometer or Abney level. A description of a soil profile, a vertical cross-section of the soil, represents a three-dimensional pedon. A pedon is the smallest volume that is recognized as a soil. Field morphology of a soil profile is best determined from a recently opened pit; however, a fresh roadside cut may be used. Field descriptions begin by organizing and subdividing a vertical exposure (soil profile) into reasonably distinct layers or horizons. At a minimum, a soil description includes a site characterization and the horizon depth intervals, horizon boundary characteristics, color, texture, structure, consistence, roots and pores, pH, and effervescence; and special descriptions of special features for each horizon.

Site Description

The soil pedons that are described at a site will have been affected by five soil-forming factors: parent material, biota, climate, topography, and time. A brief description of the parent material is desirable, because of inherited soil properties. It is necessary to record the observable features in the landscape that impart information about a soil. The geomorphic information needs to be described as precisely as possible, including the physiographic location, geomorphic description, and surface morphometry. The physiographic information includes: (1) physiographic division; (2) physiographic province; (3) physiographic section; (4) state physiographic area; and (5) local physiographic or geographic name. The geomorphic information includes: (1) landscape, (2) landform, (3) microfeatures, and (4) anthropogenic features. The surface morphometry includes: (1) elevation; (2) slope aspect; (3) slope gradient; (4) slope complexity; (5) slope shape; (6) geomorphic components; and (7) microrelief. The vegetation type is described, listing the common names of the plants and their scientific names when possible. The latitude and longitude, in degrees, seconds, and minutes, are most commonly used to record the geographical location of the soil as precisely as possible.

Horizons

Soil horizons may be the most distinctive visual features of a soil profile. According to the *Soil Survey Manual*, a soil horizon is a layer approximately

parallel to the surface of the soil, distinguishable from adjacent layers by a distinctive set of properties produced by the soil-forming processes. The term 'layer,' rather than 'horizon,' is used if all the properties are believed to be inherited from the parent material or no judgment is made as to whether the layer is genetic.

The depth to and thickness of the horizon should be recorded as well as the horizon designation. Horizons are labeled according to diagnostic features and interpretations. A master horizon label is given which represents characteristic properties. As an example, the system developed by the US Department of Agriculture (USDA) uses O, A, E, B, C, and R as master horizon labels. These labels correspond to:

1. O horizons and layers: Layers dominated by organic material. Some are saturated with water for long periods or were once saturated but are now artificially drained; others have never been saturated.
2. A horizons: Mineral horizons that formed at the surface or below an O horizon which exhibit obliteration of all or much of the original rock structure and that show one or more of the following: (1) an accumulation of humified organic matter intimately mixed with the mineral fraction and not dominated by properties characteristic of E or B horizons; or (2) properties resulting from cultivation, pasturing, or similar kinds of disturbance.
3. E horizons: Mineral horizons in which the main feature is loss of silicate clay, iron, aluminum, or some combination of these, leaving a concentration of sand and silt particles.
4. B horizons: Horizons that formed below an A, E, or O horizon, contain soil structure, and show one or more of the following: (1) illuvial concentration of silicate clay, iron, aluminum, humus, carbonates, gypsum, or silica, alone or in combination; (2) evidence of the removal of carbonates; (3) residual concentration of sesquioxides; (4) coatings of sesquioxides that make the horizon conspicuously lower in value, higher in chroma, or redder in hue than overlying and underlying horizons without apparent illuviation of iron; (5) alteration that forms granular, blocky, or prismatic structure if volume changes accompany changes in moisture content; or (6) brittleness.
5. C horizons or layers: Horizons or layers, excluding hard bedrock, that are little affected by pedogenic processes and lack properties of O, A, E, or B horizons. The material of C layers may be either like or unlike that from which the solum presumably formed.
6. R layers: Hard bedrock.

Soil horizons can be dominated by properties of one master horizon but have subordinate properties of another; in which case, the master horizons may be combined to denote this interpretation (i.e., AB, EB, CB, BC, etc.). Where soil material from two horizons is within the same horizon, a virgule (/) is used to indicate two materials within that same horizon that can be described separately with regard to color, texture, and structure (i.e., Bw/A, Bt/A, C/Bt, etc.).

Table 1 Suffix symbols used by the USDA for master horizons

Suffix	*Horizon*
a	Highly decomposed organic material
b	Buried genetic horizon
c	Concretions or nodules
d	Physical root restriction
e	Organic material of intermediate decomposition
f	Frozen soil or water
ff	Dry permafrost
g	Strong gleying
h	Illuvial accumulation of organic matter
i	Slightly decomposed organic material
j	Accumulation of Jarosite
jj	Evidence of cryoturbation
k	Accumulation of carbonates
m	Cementation or induration
n	Accumulation of sodium
o	Residual accumulation of sesquioxides
p	Tillage or other disturbance
q	Accumulation of silica
r	Weathered or soft bedrock
s	Illuvial accumulation of sesquioxides and organic matter
ss	Presence of slickensides
t	Accumulation of silicate clay
v	Plinthite
w	Development of color or structure
x	Fragipan character
y	Accumulation of gypsum
z	Accumulation of salts more soluble than gypsum

Reproduced from Soil Survey Staff (1998) *Keys to Soil Taxonomy*, 8th edn. USDA, Natural Resources Conservation Service. Washington, DC: US Government Printing Office.

Designations of soil horizons are supplemented with suffixes or subscripts to convey more information about a horizon. Description systems vary slightly; however, as an example the suffixes used by the USDA are listed in Table 1.

Soil horizons and layers have been developed based on central concepts observed from field investigations. The properties between soils vary greatly and are a continuum of properties. There are no definitions that are rigorously followed to label each horizon. The labeling is an interpretation based on observations of the scientist; however, the classification of the soil, which is based on the description, does follow criteria definitions.

Soil Texture

Soil texture is one of the most common features used by scientists and laymen to describe soils. Texture

refers to the relative abundance, by weight, of the size fractions – sand, silt, and clay. In all textural classification systems, the sum of all particles (by weight) less than or equal to 2 mm is equal to 100% (Figure 1), and is referred to as the 'fine-earth fraction.' There are several textural classification systems used, with varying ranges of particles sizes (Figure 2).

Textural descriptions are commonly used, because it is cumbersome to give the percentage of sand, percentage of silt, and percentage of clay within a description. Commonly, field estimates of sand, silt, and clay are used, and the soils are placed into textural classes. These textures are verified by laboratory analysis. The mineralogy of the clay may affect estimates of clay in field textures. The kaolinitc clays expand less and tend to be less expressed than smectitic clays which may result in an underestimation of the percentage of clay.

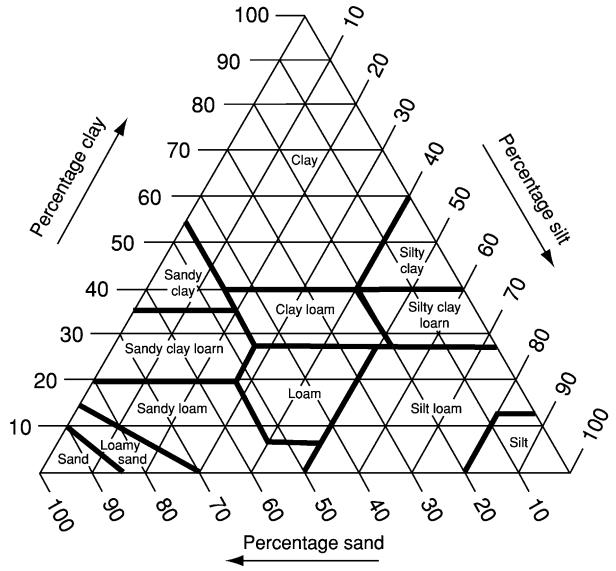


Figure 1 USDA soil textural triangle.

Soil texture relates to many of the ways a soil performs. If a soil is coarse (sandy), water tends to move through it quite well, but it may not retain enough water for plant growth. If a soil is clayey, water will probably move slowly but it should retain water for plant growth. The shrink–swell potential of clayey soils varies quite widely depending on the type of clay minerals present. If the soil is high in smectitic clay minerals, it will shrink and swell as the moisture content of the soil changes. If the shrinking and swelling are not considered in the design of infrastructure, it will cause problems for roads and foundations. However, if the clay mineral is primarily kaolinite and other low-activity minerals, construction on such soils should not be a problem. Also, if the particles larger than 2 mm are of sufficient quantity to be important, the textural classes are prefixed with modifiers. These separates are classed as gravel, cobbles, or stones (Figure 2). (*See* **Texture**.)

Soil Color

Soil colors are used to infer pedogenic processes in soils. The main pigmenting (coloring) agents in soils are organic matter, iron, and, to a lesser extent, manganese. When these agents are not covering the mineral grains, the natural color of the grains is visible. Most mineral grains are naturally gray. The contrast of color is shown in Figures 3 and 4, where two soils

USDA	Clay	Silt			Sand							Gravel				Cobbles		Stones				
		Fi.		Co.	V.Fi.		Fi.		Med.		Co.	V.Co.	Fi.		Med.		Co.					
	0.002			0.05							2						76		250 mm			
International	Clay	Silt			Sand								Gravel					Stones				
					Fi.						Co.											
	0.002			0.02							2						20 mm					
Unified	Silt or clay					Sand								Gravel					Cobbles			
						Fi.			Med.			Co.		Fi.			Co.					
	0.074										4.76						76 mm					
AASHTO	Clay			Silt				Sand						Gravel or stones						Boulders		
								Fi.				Co.		Fi.			Med.		Co.			
	0.005					0.074						2							76 mm			
Phi scale																						
	0.00195		0.0078		0.031		0.125			0.5			2			8			32		128	512 mm

Figure 2 Particle-size classes using five particle-size class systems. Fi., fine; Co., coarse; V.Fi., very fine; V.Co., very coarse; Med., medium.

Figure 3 Highly developed soil in a well-drained convex position.

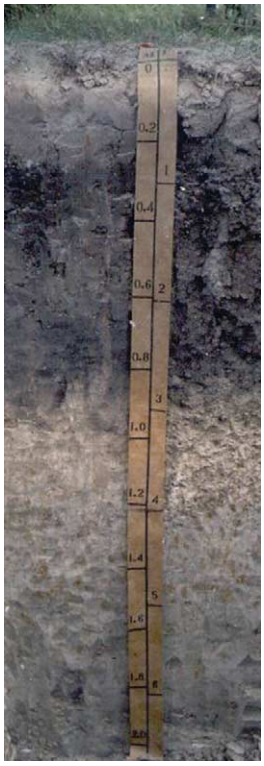


Figure 4 Highly developed soil in a poorly-drained concave position.

have similar texture and structure yet differ in color scheme. When a soil horizon has more than one color, the dominant color by volume is the matrix color.

Soil colors are most conveniently measured by comparison with a soil color chart. The one generally used is a modification of the Munsell color chart that includes only the portion needed for soil colors, about one-fifth of the entire range of color. The arrangement is by hue, value, and chroma – three simple variables combine to give all colors. Hue is the dominant spectral (rainbow) color; it is related to the dominant wavelength of the light. Value refers to the relative lightness of the color and is a function (approximately the square root) of the total amount of light. Chroma is the relative purity or strength of the spectral color and increases with decreasing grayness.

In the soil color chart, all colors on a given card are of a constant hue, designated by a symbol in the upper right-hand corner of the card. Vertically, the colors become successively lighter by visually equal steps; their value increases. Horizontally, they increase in chroma to the right and become grayer to the left. The value and chroma of each color in the chart are printed immediately beneath the color. The first number is the value and the second is the chroma. As arranged in the chart, the colors form three scales: (1) radial, or from one card to the next, in hue; (2) vertical in value; and (3) horizontal in chroma.

The matrix color is the first color described in a soil horizon or layer, since it is by definition the most dominant color. The moisture status of the soil color must be taken into account when describing soil color, i.e., moist or dry. To determine soil color, a fresh ped (a unit of soil structure) interior is used to eliminate the possibility of describing the color of a ped surface coating.

Redoximorphic Features and Mottles

The lack of uniformity of soil color may give insight regarding the pedogeneis of the soil. In the USDA description terminology, the generic term 'mottle' was changed to 'redoximorphic feature' when describing features in the soil resulting from reduction–oxidation reactions. The term 'redoximorphic feature' is a more descriptive term used when the pedogenic processes of reduction and oxidation are interpreted. 'Mottles' is a term used for spots in the soil and infers no pedogenic process.

Redoximorphic Features

Soil colors are greatly affected by the oxygen supply in the soil. As long as there is sufficient oxygen for microbial activity, the soils are usually brown or red and frequently have a single dominant color.

However, when soils become wet (saturated with water), the oxygen supply cannot flow as freely and the supply becomes insufficient for many microbes. Insufficient oxygen in the soil results in reduction – a process by which iron and manganese gain electrons.

Figure 5 Iron depletions illustrated in the subsoil.

Reducing conditions in soils are precursors to redoximorphic features, which are multicolored. Reduction of iron and manganese greatly increases their solubility, thus allowing them to dissolve into solution and move with the water, i.e., become mobile. Features caused by reduction or reduction and subsequent reoxidation are also called 'redoximorphic features.'

When the iron in the soil goes into solution, it uncoats the mineral grains, resulting in zones called 'depletions,' where the color is that of the mineral grains – usually gray. When air reenters the soil, the iron and manganese are oxidized and precipitate into brownish to reddish to blackish forms called 'concentrations.' Thus depletions are caused by removal of iron coatings, and concentrations are caused by oxidation and precipitation of previously reduced forms of iron and manganese. Both depletions and concentrations require reduction, but only concentrations require subsequent reoxidation.

Concentrations and depletions are described by kind, size, abundance, and color. Depletions are usually lighter colored than the matrix (**Figure 5**), and concentrations are usually browner, redder, or blacker than the matrix (**Figure 6**). There are several kinds of concentrations, which are described as finely disseminated materials, masses (noncemented), nodules, or concretions (cemented). Masses are noncemented concentrations of iron and manganese. Concretions and nodules are both cemented accumulations of iron and manganese, but differ in their internal structure: concretions have concentric rings, whereas nodules do not (**Figure 7**). There are two types of depletions: (1) iron depletions, which are primarily a lighter color

Figure 6 Iron concentrations.

Figure 7 A manganese nodule in a loess soil.

than the matrix; and (2) clay depletions, which have the same color change but also have lost clay and thus are grainier.

Mottles

Mottles (areas of color that differ from the matrix color) are commonly lithochromic or lithomorphic colors. Commonly these features are inherited from the geologic parent material rather than pedogenesis. Also included in mottles are relic redoximorphic features, which are features that formed in a pluvial environment or in landscape that has been naturally or anthropogenically altered and permanently affects the drainage of the soil. Description of mottles may include abundance, size, contrast (to matrix), color, and name of object boundary.

Soil Structure

An undisturbed soil forms several levels of structure that can be divided into primary, secondary (compound), and tertiary structure. The primary structural units are the basic structural units used in morphological descriptions. Secondary and tertiary structure are defined as: (1) the size, shape, and arrangement of the primary peds, their interpedal voids, and associated interpedal pedological features in a soil material; (2) the size, shape, and arrangement of the secondary peds of a sol material (compound peds resulting from the packing of primary peds), their interpedal voids and associated interpedal pedological features. Soil structure refers to units composed of primary particles. The cohesion within these units is greater than the adhesion among units and is a product of processes that aggregate, cement, compact, or unconsolidate soil particles. Under stress, the soil mass tends to rupture along predetermined planes or zones. The peds form repetitive, quasi-uniform shapes.

Clods and fragments stand in contrast to peds, for which soil-forming processes exert weak or no control on the boundaries. A clod is caused by disturbance, such as plowing or digging, which molds the soil to a transient mass that slakes with repeated wetting and drying. Fragments include: (1) units of undisturbed soil with bounding planes of weakness that are formed on drying without application of external force and which do not appear to have predetermined bounding planes; (2) units of soil disturbed by mechanical means but without significant rearrangement to a denser configuration; and (3) pieces of soil bounded by planes of weakness caused by pressure exerted during examination, with size and shape highly dependent on the manner of manipulation.

Some soils lack structure and are referred to as structureless. When structureless soils are ruptured, soil fragments, single grains, or both may result. Structureless soil may be termed either ‘massive’ or ‘single grain.’ Massive soils tend to form fragments when disturbed, and single grain soils are usually sandy and are loose in 50% or more of the mass.

Shape

Several basic shapes of structural units are recognized in soils. Supplemental statements about variations in shape of individual peds are needed in detailed descriptions of some soils. Table 2 and Figure 8 describe the possible shapes of soil structure.

Table 2 Types of soil structure

Structure	*Description*
Natural soil structural units (pedogenic structure)	
Granular	Small peds, with rounded or very irregular faces
Angular blocky	Peds with faces that intersect at sharp angles
Subangular blocky	Peds with sub-rounded shapes and lacking sharp angles
Platy	Flat plate-shaped units
Wedge	Elliptical, interlocking lenses that terminate in acute angles, bounded by slickensides; not limited to vertic materials
Prismatic	Vertically elongated peds with flat tops
Columnar	Vertically elongated peds with rounded tops
Structureless	
Single grain	No structural units; entirely noncoherent; e.g., loose sand
Massive	No structural units; material is a coherent mass

Reproduced from Soil Staff Survey (1993) *Soil Survey Manual*. USDA Soil Conservation Service, Agricultural Handbook No. 18. Washington, DC: US Government Printing Office.

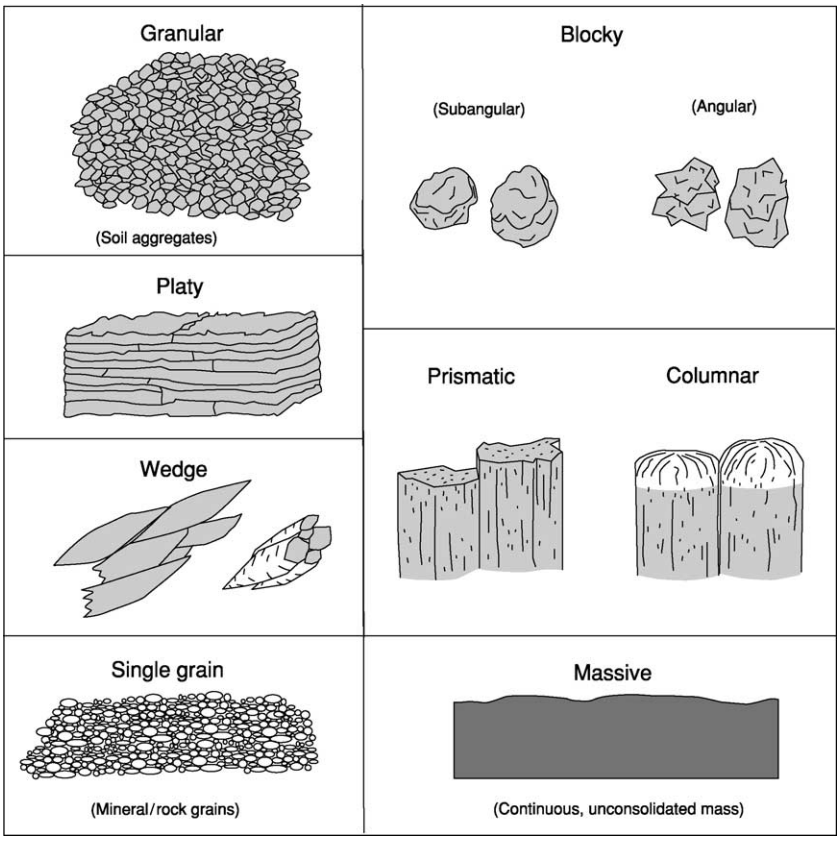


Figure 8 Shapes of soil structure. (Reproduced from Schoenberger PJ, Wysocki DA, Benham EC, and Broderson WD (eds) (2002) *Field Book for Describing and Sampling Soils*, version 2.0. Natural Resources Conservation Service. Lincoln, NE: National Soil Survey Center.)

Size

There are five classes that are used to describe the size of structural units: very fine, fine, medium, coarse, and very coarse. The size limits vary according to the shape of the unit. The size limits refer to the smallest dimension of plates, prisms, and columns.

Grade

The distinctness of units is described by the grade. Criteria are based on the ease of separation into discrete units and the proportion of units that hold together when the soil is handled.

Weak Units that are barely observable in situ are termed 'weak.' When gently disturbed, the soil material parts into a mixture of whole and broken units and much material that exhibits no planes of weakness. Faces that indicate persistence through wet–dry–wet cycles are evident.

Moderate Units graded as moderate are well formed and evident in undisturbed soil. When disturbed, the soil material parts into a mixture of mostly whole units, some broken units, and material that is not units. Peds part from adjoining peds to reveal nearly entire faces that have properties distinct from those of fractured surfaces.

Strong Units that are distinct in undisturbed soil are graded 'strong.' They separate cleanly when the soil is disturbed. When removed, the soil material separates mainly into whole units.

The distinctness of individual structural units and the relationship of cohesion within units to adhesion between units determine grade of structure. The degree of disturbance required to determine structure grade depends largely on moisture content and percentage and mineralogy of clay. Only slight disturbance may be necessary to separate the units of moist sandy loam having strong granular structure, while considerable disturbance may be required to separate units of a moist clay loam having strong blocky structure.

The designation of structure grade, size, and shape can be modified with other appropriate terms when necessary to describe other characteristics. Surface characteristics of units are described separately. Special structural units, such as wedge-shaped units of high-potential shrink–swell soils, are described in appropriate terms.

Compound Structure

'Secondary structure' is another term for 'compound structure,' where smaller structural units may be held together to form larger units. Grade, size, and shape are given for both, and the relationship of one set to the other is indicated. 'Strong medium blocks within moderate coarse prisms' and 'moderate coarse prismatic structure parting to strong medium blocky' are examples of compound structure.

Soil Consistence

Consistence refers to the cohesion and adhesion within a soil or its resistance to deformation or rupture. The consistence of soils changes greatly as the soil moisture changes. There are many expressions of consistence, such as rupture resistance, surface crust, manner of failure, stickiness, plasticity, penetration resistance, and others.

The most commonly used parameter of consistence is field evaluation (i.e., rupture resistance), which is conducted by placing a specimen 25–30 mm on each side between the thumb and first finger and applying pressure. Again, the moisture content of the soil sample is very important.

Soils in humid and subhumid climates may form brittle subsurface horizons. The degree and percentage of brittleness of the soil are described. Soils may also be hardened due to cementation. In this case, the materials accumulated causing the cementation, such as carbonates, gypsum, humus, iron, and/or silica are listed. In addition, the percentage, cementation of soil and the strength of the cementation are noted.

Roots and Pores

The roots and pores that can be observed in the field are among the larger pores in the soil. These pores are mostly noncapillary pores, i.e., pores that cannot hold water against the force of gravity. These pores are of major importance in moving water and air within the soil. After a major wetting of the soil (rainfall, snowmelt, irrigation, etc.), these pores must transmit the water throughout the soil to the smaller capillary pores. The noncapillary pores should redistribute the water in a few days, because they must drain in order also to distribute oxygen (air) into the soil. If soil oxygen is depleted for more than a few days, many plants will die or at least be harmed, although some plants thrive in oxygen-deprived soils.

Normally soils with a mixture of noncapillary and capillary pores are preferred, because they both drain well and retain water for plant growth. When describing roots and pores, at least the abundance of each size class is assigned; for example, common fine roots or many medium pores.

Chemical Response

pH

The acidity or alkalinity can infer much information about a soil. The pH of a soil can be estimated with a simple field test kit. There are several soil pH class systems used to categorize soil pH. Each horizon or layer within a pedon is tested as pH may vary dramatically due to pedogenic or lithogenic properties.

Effervescence

A simple field technique may be used to check for the presence of free carbonates. A 10% solution of HCl added to the soil produces a gaseous response when

$CaCO_3$ is present. The gaseous response of the HCl will vary depending on the amount of carbonates in the soil. The degree of effervescence is noted in the description.

Reaction to Alpha,Alpha′-Dipyridyl

In soils that are actively reducing at an Eh (redox potential) or pE (relative electron activity) low enough to produce ferrous iron, a 0.2% concentration of alpha,alpha′-dipyridyl (also listed as 2,2′-bipyridine) may be used. This solution is applied to the soil within minutes of exposure in the absence of sunlight. If Fe^{2+} is present, the alpha,alpha′-dipyridyl will change to a red or pink color; if only Fe^{3+} is present, the solution will remain clear. This solution will react with a few parts per million of Fe^{2+}, so care should be used when extracting soil. The Fe^{2+} from a knife blade or spade contact with the soil will produce a false-positive response. This solution should only be used when reducing conditions are suspected.

Special Features

Pedon Features

Pedon features occur throughout the pedon and may cross adjacent horizons. These features are described using standard technical terms; however, some added language is sometimes used to supplement the description (Table 3).

Horizon or Layer Features

Concentrations are identifiable bodies within the soil that have been formed due to the pedogenic processes of chemical dissolution and/or precipitation, transport, accrual, and physical and/or biological removal. Concentrations may be finely disseminated materials, masses, nodules, concretions, crystals, biologic concentrations, and/or inherited minerals. The kind, size, abundance, and color are used when describing concentrations. In some soils, supplemental information is needed to describe concentrations fully.

Ped Surface Features

Surface features occur from a variety of pedogenic processes and result in differences from the matrix soil in composition, orientation, or packing. Surface features include: (1) coats of a variety of substances unlike the adjacent soil material and covering part or all of surfaces; (2) material concentrated on surfaces by the removal of other material; and (3) stress formations in which thin layers at the surfaces have undergone reorientation or packing by stress or shear. Descriptions of surface features may include kind, location, amount, continuity, distinctness, and thickness of the features. Where the features are thick and there are contrasting colors, this is noted. Table 4 lists some common ped surface features found in soils.

Location

Surface features may vary in the coverage of peds or may coat other soil constituents such as gravel or rocks.

Amount

The number should be noted as a percentage or a percentage range. Some horizons have more than one surface feature. In those cases, each surface feature is noted with the percentages of each feature.

Table 3 Common pedon features

Pedon feature	*Description*
Cracks	Fissures other than those attributed to soil structure
Desert pavement	Natural concentration of polished gravel and stones at the soil surface primarily in arid and semiarid climates
Ice crystals, lenses, layers (veins), wedges	Features occurring in soils of Arctic, Antarctic, and Subarctic regions
Jarosite mottles	Yellow jarosite mineral accumulations due to oxidation of sulfidic accumulations
Krotovinas	Irregular tubular streaks of soil material within a horizon or layer transported to another horizon or layer due to biologic activity
Lamellae	Thin banded illuviated clay
Microbiotic crust	Thin surficial zone exhibiting more mechanical continuity of the soil fabric than the subjacent soil
Stone line	Natural accumulation of coarse fragments below the surface extending through the pedon
Tongues of albic material	Lobes of eluviated soil that extend into the soil horizon from overlying horizons

Reproduced with permission from Schoenberger PJ, Wysocki DA, Benham EC, and Broderson WD (eds) (2002) *Field Book for Describing and Sampling Soils*, version 2.0. Natural Resources Conservation Service. Lincoln, NE: National Soil Survey Center.

Table 4 Common ped surface features

Ped feature	*Description*
Clay films	Layers of oriented, translocated clay ('clay skins')
Clay bridges	Links between mineral grains found primarily in sandy soils
Stress surfaces, pressure faces	Smooth, shiny or smeared surfaces formed through rearrangement of particles resulting from shear forces
Slickensides	Stress surfaces that are polished and striated; direction of movement can be determined from striations
Other features	Coats of iron, aluminum, or manganese oxides, organic matter, salts, or carbonates

Reproduced from Schoenberger PJ, Wysocki DA, Benham EC, and Broderson WD (eds) (2002) *Field Book for Describing and Sampling Soils*, version 2.0. Natural Resources Conservation Service. Lincoln, NE: National Soil Survey Center.

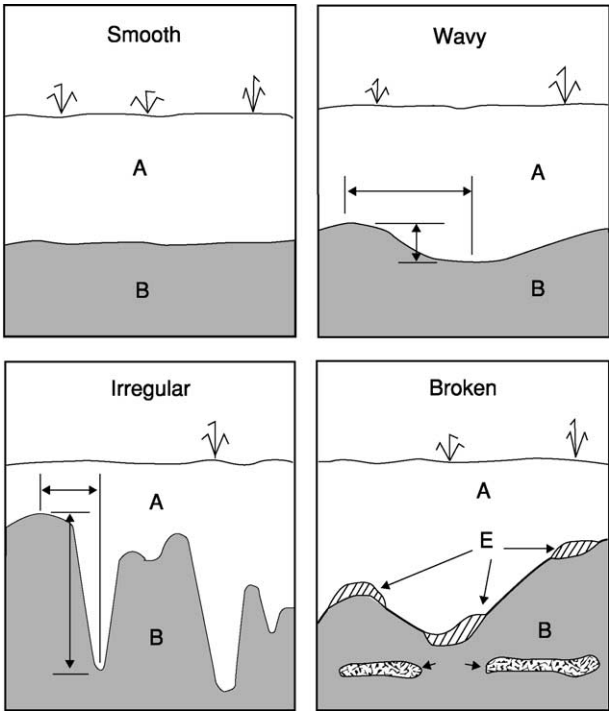


Figure 9 Boundary nomenclature of soil horizons. A, A horizon; B, B horizon; E, E horizon. (Reproduced from Schoenberger PJ, Wysocki DA, Benham EC, and Broderson WD (eds) (2002) *Field Book for Describing and Sampling Soils*, version 2.0. Natural Resources Conservation Service. Lincoln, NE: National Soil Survey Center.)

Boundary

The horizons or layers may be parallel to the surface; however, in some soils the bottom boundary may vary depending on the pedogenesis. In most description systems, the bottom boundary of the soil horizon or layer is described as smooth, wavy, irregular, or broken (Figure 9). The distance in which the boundary grades into or changes is normally described also. The boundary classes are abrupt (0–2 cm), clear (2–5 cm), gradual (5–15 cm), and diffuse (more than 15 cm).

Soil Micromorphological Description

The discussion in this chapter has focused on the macromorphological description of soils in the field. There is a discipline devoted to micromorphological descriptions in the laboratory using soil thin sections. These descriptions are valuable for classification and elucidating pedogenesis in soils. Where possible a micromorphologic investigation is used to supplement and verify features in the macromorphological description.

See also: **Texture**

Further Reading

Brewer R (1976) *Fabric and Mineral Analysis of Soils*. Huntington, NY: Kreiger.

Bullock P, Federoff N, Jongerius A, Stoops G, and Tursina T (1985) *Handbook for Thin Section Description*. Wolverhampton, UK: Waine Research.

Buol SW, Hole FD, McCraken RJ, and Southard RJ (1997) *Soil Genesis and Classification*, 3rd edn. Ames, IA: Iowa State University Press.

Childs CW (1981) Field tests for ferrous iron and ferric-organic complexes (on exchange sites or in water-soluble forms) in soils. *Australian Journal of Soil Research* 19: 175–180.

Schoenberger PJ, Wysocki DA, Benham EC, and Broderson WD (eds) (2002) *Field Book for Describing and Sampling Soils*. version 2.0. Natural Resources Conservation Service. Lincoln, NE: National Soil Survey Center.

Soil Survey Staff (1993) *Soil Survey Manual*. USDA, Soil Conservation Service, Agricultural Handbook No. 18. Washington, DC: US Government Printing Office.

Soil Survey Staff (1998) *Keys to Soil Taxonomy*, 8th edn. USDA, Natural Resources Conservation Service. Washington, DC: US Government Printing Office.

Vepraskas MJ (1992) *Redoximorphic Features for Identifying Aquic Conditions*. North Carolina Agriculture Research Service Technical Bulletin 301. Raleigh, NC: North Carolina State University.

Wilding LP, Smeck NE, and Hall GF (eds) *Pedogenesis and Soil Taxonomy. Concepts and Interactions*, part A. New York: Elsevier.

MULCHES

C L Acharya, K M Hati, and K K Bandyopadhyay,
Indian Institute of Soil Science, Bhopal,
Madhya Pradesh, India

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Mulch means a layer of dissimilar material separating the soil surface from the atmosphere, and mulching is the artificial application of mulch, practiced to obtain beneficial changes in the soil environment. Mulch may be organic (crop residues, stubble mulch etc.) or inorganic (plastic sheet, gravels etc.) in nature. It may be grown *in situ* like previous crop residue mulch left on the soil surface, produced by a cover crop or it may be living mulch like perennial legume mulch. It may also be grown or produced *ex situ* and brought in for field application, e.g., straw, sawdust, plastic products etc.

The practice of mulching has been widely used as a management tool for centuries in many ancient civilizations. The Great Plains region of the USA, which was once considered to be a dust bowl, has been turned into the granary of the world through the adoption of conservation tillage, the practice of leaving crop residues as the surface stubble or flattened straw mulch. The idea of using plastics as mulch originated in Hawaii, where it was used in pineapple cultivation with considerable success.

Mulch has a buffering effect and it dampens the influence of environmental factors on soil. The magnitude of the buffering effect of mulches depends upon the quality, quantity and durability of mulch material, soil type and the climatic conditions. Mulching influences the soil hydrothermal regime by influencing the radiation balance, rate of heat and water vapor transfer, and heat capacity of soil. Mulching improves the physical condition of soil by enhancing soil aggregation and helps in conservation of water by checking evaporation, increasing infiltration and retarding runoff loss. It favorably modifies the soil thermal regime, retards soil erosion, and improves soil chemical environment and biological activity in soil. Modification of the soil microclimate by mulching favors seedling emergence and root proliferation and suppresses weed population. Organic mulches add nutrients to soil when decomposed by microbes and help in carbon sequestration. Favorable soil edaphic environment under mulch improves crop productivity, enhances input-use efficiency and checks environmental pollution. From the conservation viewpoint, mulch simulates the effect of a plant cover. It is most useful as an alternative to cover crops in dry areas where insufficient rain prevents the establishment of a ground cover before the onset of heavy rain or strong winds, or where a cover crop competes for moisture with the main crop.

However, improper application of mulches leads to creation of an anaerobic environment under a heavy rainfall situation, leading to loss of nitrogen through denitrification. It may cause disease and pest infestation and some allelopathic materials produced in some crop residues retard the crop growth. So proper site-specific methodology needs to be followed under different soil, crop and climatic conditions for efficient mulch management to achieve sustainable crop production.

How Mulching Influences the Soil Environment

Mulching improves the physical conditions, chemical environment and biological activities of soil. Favorable modification of the soil hydrothermal regime, improvement of soil aggregation and retardation of erosion and soil loss enhance the physical condition of soil under mulch. Improvement in the soil chemical environment and biological activities by mulching is mainly attributed to the enrichment of soil organic matter through organic mulching and improvement of the physical condition of soil under both organic and inorganic mulches.

Soil Moisture Regime

Mulching favorably influences the soil moisture regime by controlling evaporation from the soil surface, improving infiltration and soil-moisture retention and facilitating condensation of water at night due to temperature reversals.

Controling evaporation from the soil surface Direct evaporation of water from the soil surface is an important process, particularly in the case of bare soils or in areas where summer fallow is practised. Mulching reduces evaporation from the soil surface by retarding the intensity of the radiation and wind velocity on the mulched surface. It is effective in retarding the evaporation mainly during the initial energy-limited stage of drying. However, mulching is not very effective in controling evaporation during the falling-rate or supply-controled stage of evaporation. During this stage, the soil surface dries under the mulch and the vapor transfer of water to

the atmosphere is inhibited more by the dry soil than by the overlying vegetative mulch. Higher mulch rate retards the energy reaching the soil surface and hence limits the evaporation at the constant-rate stage (Figure 1). The retardation of initial evaporation can also enhance the process of internal drainage and thus allows more water to migrate downward into the deeper parts of the profile, where it is conserved longer and is less likely to be lost by evaporation.

Mulching also affects the pattern of evaporation by offering resistance to water vapor flow from the soil surface to atmosphere and by increasing the thickness of the relatively non-turbulent air zone above the soil, thus changing the boundary layer conditions of the soil–air interface. The relative magnitude of evaporation reduction for different mulch rates is much greater under high evaporative demand. The cumulative evaporation reduction due to mulching follows a quadratic relationship with time. The magnitude of maximum evaporation reduction and the time when it is reached depends on the rate of mulching, soil type, magnitude of evaporative demand and the method of residue application (Figure 2). The orientation of the residue, i.e., flat or matted versus standing, affects the porosity and thickness of the layer and thus the rate of evaporation from the soil surface. As the amount of standing residue increases, greater wind speed is needed to initiate water loss through evaporation; in addition, the water loss rate at a given wind speed decreases with increased amounts of standing residue. Residue position also affects soil temperature (Table 1), which in turn affects the evaporation rate through its influence on the vapor pressure of the soil water.

Water-saving under mulching is prominent if rains are frequent. However, under extended dry spell periods, mulch may keep the soil surface moist for longer periods and hence prolong the first stage of evaporation and thus produce no net saving of water.

Improving infiltration rate Mulching with organic materials improves the infiltration rate because it serves as a barrier for runoff, which provides more opportunity for water to infiltrate into the profile. Secondly, mulch intercepts the rainwater, protects the soil from its biting action and checks splashing of the soil. It prevents crust formation due to clogging of soil pores, which increases infiltration rate. Furthermore, organic mulches improve the macroporosity and stability of the structural aggregates of soil

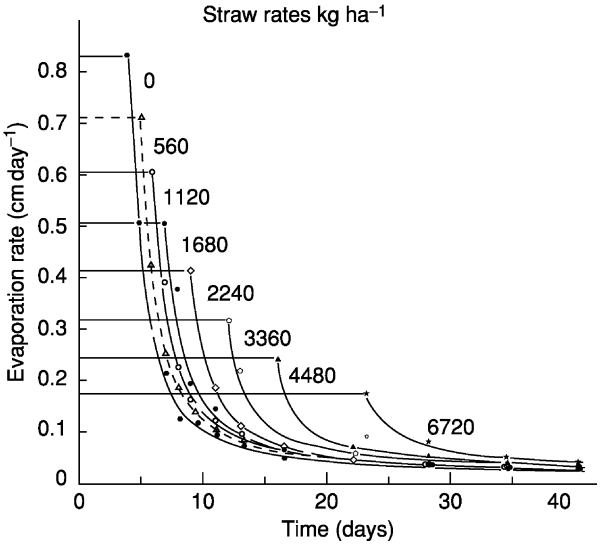


Figure 1 Temporal variation of soil water evaporation rate as influenced by rates of straw mulch application. Straw rates range from 0–6720 kg ha^{-1}. Reproduced with permission from Bond JJ and Willis WD (1969) *Soil Science of America Proceedings* 33: 446.

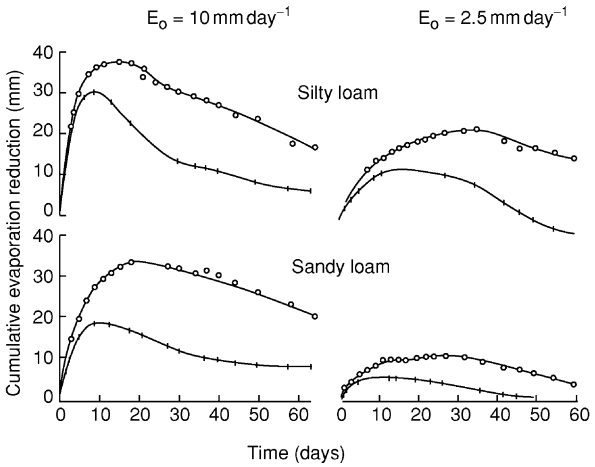


Figure 2 Time-course of cumulative evaporation reduction (CER) in two different soils under two rates of wheat straw applied as mulch at two evaporative demands (E_o). Wheat straw. ⟼ 3 t ha^{-1} as mulch; ⟜ 6 t ha^{-1} as mulch. Reproduced with permission from Prihar SS, Jalota SK, and Steiner JL (1996) *Soil Use and Management* 12: 152.

Table 1 Average daily soil surface temperature as affected by wheat straw position during August to September at Colorado, USA

Straw position	*Soil surface temperature (°C)*[a]
Bare soil	47.8 c
Flat straw	41.7 b
3/4 flat and 1/4 standing	39.6 b
1/2 flat and 1/2 standing	32.2 a

[a]Column values accompanied by different letters are significantly different at $P = 0.01$ (Duncan's multiple range test).
Reproduced from Smika DE (1983) In: Unger PW (1990) *Advances in Soil Science* 13: 27–68.

and thereby improve the water transmission properties (Table 2), which facilitate better infiltration and recharge of the soil profile. Straw mulch application increases soil water storage and storage efficiency. The amount of water storage in the soil profile, the storage efficiency, total water use and water-use efficiencies of dryland crops increase with increase in the mulch rate (Table 3). The increase in infiltration that results from mulch is found to be more important in some situations than its effect on reduction in evaporation for conservation of water in the profile.

Soil moisture retention Mulching improves the moisture retention properties of soil through its effect on pore-size distribution and soil structure. Higher mulch rates increase soil-water retention more at lower suctions (Table 4) due to an increase in macropores and interaggregate pores caused by enhanced soil organic matter content and higher activity of soil fauna, e.g., earthworms and termites in mulched plots.

Table 2 Effect of mulch rate on water transmission properties of an Alfisol

Mulch rate (Mg ha^{-1})	*Soil water sorptivity*	*Soil water transmissivity*	*Saturated hydraulic conductivity (cm h^{-1})*[a]
0	0.32	5.56	30a
2	0.57	7.81	45a
4	0.67	7.50	70b
6	0.84	10.21	132c
12	1.05	15.36	129c

[a]Column values followed by the same letter are not significantly different at the 5% level (Duncan's multiple range test).
Reproduced from Lal R, De Vleeschauwer D, and Malafa Nangje R (1980) Changes in the newly cleared Tropical Alfisol as affected by mulching. *Soil Science Society of America* 44: 827–833.

Table 3 Effect of straw mulch rate on soil water storage during fallow,[a] water storage efficiency and dryland grain sorghum yield in Bushland, Texas

Mulch rate (Mg ha^{-1})	*Water storage (mm)*	*Storage efficiency (%)*[b]	*Grain yield (Mg ha^{-1})*	*Total water use (mm)*	*Water use efficiency (kg m^{-3})*
0	72 c	22.6 c	1.78 c	320	0.56
1	99 b	31.1 b	2.41 b	330	0.73
2	100 b	31.4 b	2.60 b	353	0.74
4	116 b	36.5 b	2.98 b	357	0.84
8	139 a	43.7 a	3.68 a	365	1.01
12	147 a	46.2 a	3.99 a	347	1.15

[a]Fallow duration 10 to 11 months.
[b]Precipitation averaged 318 mm.
Column values followed by the same letter are not significantly different at 5% level (Duncan's multiple range test).
Reproduced from Unger PW (1990) *Advances in Soil Science*, 13: 27–68.

Water condensation at night Stone and gravel mulches induce lateral movement of heat and vapor, which could in turn collect water under the stones, due to condensation of vapor at night, in amounts sufficient enough to serve as the source of water for some species of desert plants.

Controlling Runoff and Soil Erosion

Mulching invariably decreases soil erosion and often reduces the rate and amount of runoff. Mulch cover protects the soil from raindrop impact and surface sealing, increases the infiltration rate and decreases runoff velocity through physical resistance to water flow. In general, loss of water through runoff decreases exponentially with increase in mulch rate (Table 5). For some soils, mulch does not substantially decrease runoff and drastically reduces soil erosion. The runoff water gets filtered through the mulch and is often clear with little sediment. Mulching decreases sediment concentration in runoff water through the protective effect of crop residues against raindrop.

The rate of soil loss decreases exponentially with the increase in the percentage area covered by mulch. The mulch factor (MF), defined as the ratio of soil loss with mulch to that without mulch, is related to the percentage residue cover (RC) or mulch cover by the following expression:

Table 4 Effect of mulch rate on moisture retention properties of 0–5 cm layer of an Alfisol

	Moisture retention (% w/w) at different suction (bar)			
Mulch rate (Mg ha^{-1})	0.03	0.06	1	15
0	27.3	26.1	5.5	3.2
2	29.1	26.1	6.5	5.1
4	29.6	28.8	5.2	3.7
6	29.6	28.5	5.4	4.2
8	31.2	30.5	7.3	5.8

Reproduced from Lal R (1987) *Tropical Ecology and Physical Edaphology*, pp. 618–685. Wiley-Interscience Publication.

Table 5 Effects of rice straw mulch rates on runoff and soil loss in Nigeria

Mulch rate (Mg ha^{-1})	*Runoff (%)*	*Soil loss (Mg ha^{-1})*	*Marginal conservation effect (Mg soil) (Mg mulch)$^{-1}$*
0	75.4	9.6	–
2	43.4	2.3	3.7
3	15.2	0.5	1.8
6	5.4	0.1	0.13
12	0	0	0.017

Reproduced from De Vleeschauwer *et al.* (1980) In: Erenstein O (2002) *Soil and Tillage Research* 67: 115–133.

$$MF = e^{-aRC}$$

where, 'a' ranges in value from 0.03 to 0.07 and 0.05 is generally used. The exponential relationship is applied only for rill erosion whereas the interrill erosion decreases linearly with the mulch cover. Since for most soils the contribution of the interrill erosion to the total soil loss is quite small for slope length in excess of 25 m and for all but very low percentage residue covers, the above equation may be used to compute the mulch factor. This mulch factor is multiplied with the 'C' factor values for calculating the effect of mulching in the universal soil loss equation. It has been observed that a minimum of 50% soil cover is required to achieve reduction in soil loss. Ideally a mulch should cover 70–75% of the soil surface. The optimum mulch rate can be computed for a pre-selected set of conditions using the Manning's equation for flow velocity. The relationship between Manning's n and mulch rate is given below:

$$\text{Manning's equation:} \quad n = (r^{0.67} \times S^{0.5})/v$$

where, r = hydraulic radius, S = slope and v = desired maximum flow velocity.

The value of n due to mulch has been described as $n_m = (n^{3/2} - n_s^{3/2})^{2/3}$ where, n_s = Manning's n for a particular soil.

Mulch rate (M, kg m^{-2}) is related to the Manning's n for mulch (n_m) by the following relationship.

$$n_m = 0.071\, M^{1.12} \quad \text{for interrill erosion}$$
$$n_m = 0.105\, M^{0.84} \quad \text{for rill erosion.}$$

Experimental evidence shows that an adequate quantity of crop residue mulch can mitigate the effect of the degree of slope in reducing soil loss.

Mulching is also effective in reducing soil erosion by wind. It reduces the wind velocity on the soil surface when the mulch is fixed to the soil, so that soil particles do not get blown up by wind. Standing crop residues are more effective than flattened surface mulches in reducing wind erosion. The effectiveness of standing crop residues depends upon the number of stalks per unit area and their size.

Modifying Soil Thermal Regime

Mulch has a moderating influence on the soil thermal regime and the effect varies among soils, climate, kind of mulch materials used and rates of application. It increases soil temperature during cooler weather and decreases it during hot spells. In general, mulch has a damping effect on the amplitude of the diurnal fluctuations in soil temperature. Organic mulching enhances the soil temperature at night and early morning hours but it decreases the daytime temperature as compared to unmulched plots (Figure 3).

Mulching is commonly practiced in some fruit crops, like strawberry, to keep the fruit clean by avoiding direct contact with soil and to moderate the soil hydrothermal regime for better productivity and quality of fruit. Transparent polythene mulch raises the maximum soil temperature whereas organic mulches like pine needle and grass mulches lower it. Black polythene mulch, however, does not alter the maximum soil temperature appreciably (Figure 4). The application of straw mulch can lower maximum soil temperature due to the interception of incoming solar radiation, high reflectivity and low heat conductivity, the magnitude of which depends upon soil wetness, incidental radiation, rate of mulch application as well as period of the year (Table 6). The lowering of maximum soil temperature by straw mulching during the early stage of growth of sugarcane can significantly improve the yield.

Soil Aeration

Crop residue mulch improves soil aeration by promoting free exchange of gases between the soil and the atmosphere. This is facilitated by improvement of structural stability, total porosity and macroporosity, decrease of surface crusting and by improving the overall soil drainage. Oxygen diffusion rate is higher under mulch than under unmulched conditions. The gaseous composition of soil air under mulch depends on the nature of the mulch material (C:N ratio), its rate of decomposition, the soil moisture regime and the climatic condition. Plastic mulch is practically impervious to carbon dioxide (CO_2), a gas that is of prime importance for photosynthesis. Very high levels of CO_2 build up under the plastic, as the film does not allow it to escape. It has to come through the holes made in the plastic for the plants and a 'chimney effect' is created, resulting in localized concentrations of abundant CO_2 for the actively growing leaves that accelerates the growth of the crop.

Improving Soil Structure

Mulching improves soil structural properties directly by preventing raindrop impact and indirectly by promoting biological activity. Organic mulching improves the total porosity, macroporosity and mean weight diameter of water-stable aggregates, because of addition of organic matter upon decomposition by soil microorganisms. The mean-weight diameter of water-stable aggregates increases with increase in the mulch rate (Table 7).

In general, under mulched conditions, the bulk density of the soil is lower than that under unmulched

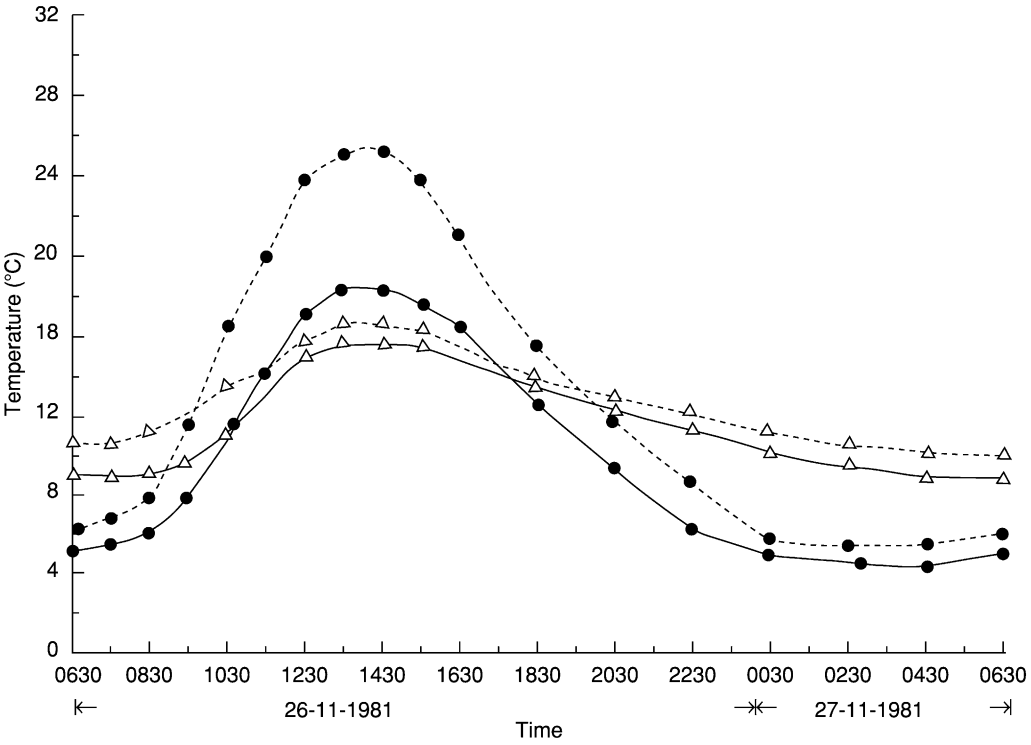


Figure 3 Diurnal soil temperature fluctuations as influenced by conventional cultivation (CC); and pine needle mulching (M) at 10 Mg ha^{-1} in wheat on a silty-clay loam soil. ● CC; △ CC+M; — 10 cm depth; - - - 5 cm depth. Reproduced from Bhagat RM and Acharya CL (1987) *Soil and Tillage Research* 9:65–77, with permission from Elsevier Science.

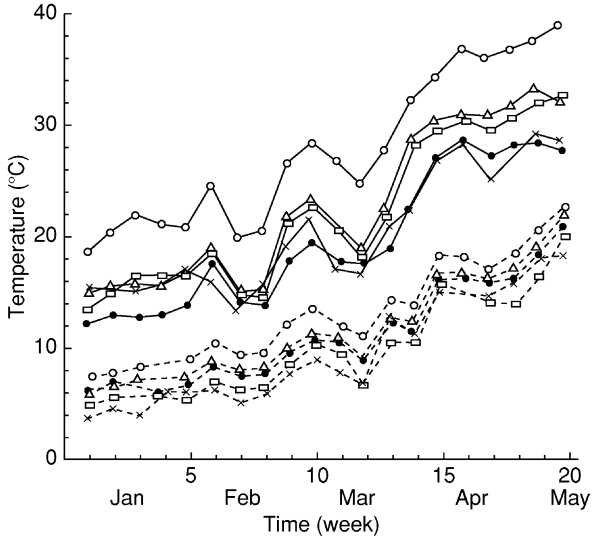


Figure 4 Mean weekly soil temperature at 5 cm depth under different mulch treatments and unmulched control in strawberry on a Typic Hapludalf. ○, transparent polythene; △, black polythene; □, unmulched control; ●, pine needle mulch; — temperature at 14.30 hours;---temperature at 06.30 hours.

Table 6 Effect of straw mulching (@ 4–6 Mg ha^{-1}) on soil temperature at 10 cm depth during the year under different crops

Month	*Crop*	*Maximum temperature of unmulched soil (°C)*	*Temperature reduction with straw mulch (°C)*
January	Autumn potato	12.5–15.5	0.5–2.0
March	Spring potato	24.5–29.0	2.5–3.8
April	Sugar cane	37.6–38.2	8.1–9.5
	Bare soil	38.3[a]	11.5
May	Fodder maize	38.0–42.5	7.5–9.8
	Bare soil	40.3[a]	6.7
June	Sugar cane	34.2–43.5	4.4
July	Grain maize	35.5–41.7	3.7–6.8
August	Grain maize	31.7–38.9	1.4–4.8
September	Grain maize	31.5–33.2	0.9–2.0
November	Autumn potato	17.0–18.5	2.0
December	Autumn potato	15.2–17.0	1.7–2.2

[a]Recorded at 5 cm soil depth.
Reproduced from Prihar SS (1986) In: *Proceedings of Seminar Organized by Fertilizer Association of India* (Northern Region) at Dehradun, India, May 10–11, 1986, pp. 31–44.

conditions. Bulk density decreases with increase in residue mulch rate. One of the reasons for the decrease in bulk density with increase in mulch rates is high earthworm activity in mulched plots.

Improving the Soil Chemical Environment

Organic mulches add organic matter and plant nutrients to soil upon decomposition. Thus, they improve carbon sequestration. The volatilization and leaching

Table 7 Effect of mulch rate on pore-size distribution of 0–5 cm layer of an Alfisol after 18 months of experiment

Mulch rate (Mg ha^{-1})	Porosity (%)			Mean weight diameter (mm)
	Macro	Meso	Micro	
0	17.9	7.1	23.1	1.2
2	20.6	7.2	22.4	1.4
4	27.7	8.4	18.7	2.1
6	28.9	8.3	17.9	2.0
12	37.0	8.2	13.2	2.4
LSD (0.05)	5.3	NS	4.2	–

Reproduced from Lal R (1987) *Tropical Ecology and Physical Edaphology*, pp. 618–685. Wiley-Interscience Publication.

Table 8 Earthworm population (number m^{-2}) in the surface (0–15 cm) layer at sowing and flowering stages of wheat

Treatments	Number of earthworms	
	Sowing	Flowering
Mulch tillage (T)		
T_1: Mulching previous standing maize with wild sage followed by its incorporation with conventional tillage immediately prior to sowing of wheat	20	8
T_2: As T_1 but wheat sowing following conservation tillage and retaining the residues of wild sage as mulch	21	10
T_3: As T_1 but mulching by eupatorium instead of wild sage	21	8
T_4: As T_2 but mulching with eupatorium instead of wild sage	20	10
T_5: Farmer's practice of conventional tillage after maize harvesting	4	1
LSD ($P = 0.05$)	1.7	1.4

Adapted from Acharya CL, Kapur OC, and Dixit SP (1998) *Soil and Tillage Research* 46: 153–163.

loss of nitrogen is reduced under mulched conditions. During decomposition of organic mulches, soil mineral nitrogen is immobilized by microbes and thus its loss is minimized. Cation exchange capacity is substantially influenced by organic matter content in soils containing predominantly low-activity clays. Improvement in the cation exchange capacity of soil improves the fertility status of these soils. Furthermore, decomposition of organic mulches adds organic acids to the soil resulting in low soil pH, which influences the bioavailability of many plant nutrients, e.g., Fe, Mn, Zn, Cu, etc. Mulches also influence the availability of nutrients through their influence on physical conditions, hydrothermal regime and biological activity of the soil.

Improving Biological Activity in Soil

Soil biological activity is either directly influenced by the supply of food substrates by organic mulches or indirectly influenced by both organic and inorganic mulches through alteration of the soil hydrothermal regime. The activity and diversity of soil organisms is substantially influenced by the quality of mulch material and the rate of its application. The soil microorganisms (bacteria, fungi and actinomycetes) and soil fauna (earthworm, millipede, centipede etc.) help in the decomposition of soil organic matter and hence the availability of nutrients. However, mulching with materials of high C:N ratio results in immobilization of nitrogen by microorganisms. Mulching with organic residues of weeds like wild sage (*Lantana camara*) or eupatorium (*Eupatorium adenophorum*) under conservation tillage practices enhances the earthworm population due to supply of food substrate and a better hydrothermal regime under mulch (Table 8).

Types of Mulches

There are different types of materials ranging from natural plant residues, to gravel to various industrial products such as paper or plastic sheets, which are used as mulch. These can broadly be categorized into two groups i.e., organic mulch and inorganic mulch.

Organic Mulches

Organic mulches include living mulch, cover crops and mulches whose materials are derived from natural sources, such as field crops, trees, or natural waste products. These include hay, straw, wood chips, bark, sawdust, wood shavings, peat moss, compost, grass clippings, paper, and materials such as cocoa shells and buckwheat hulls. Organic materials for mulching should be a stable material that will remain in place at least through the growing season of an annual crop and longer for perennial crops. More durable organic material provides longer-lasting benefits as mulches than readily decomposable plant residues. Durability of organic mulch material depends upon its composition. Factors influencing the rapidity of residue decomposition include the nature of the material like the C:N ratio, cellulose and lignin content, climatic conditions, and types and population of soil microorganisms. The decay rate of surface-applied plant residues indicates that the high lignin and cellulose-containing cereal straws produce more weather-resistant mulch than do legumes.

The effectiveness of organic or vegetative mulches also depends upon the thickness of the mulch. A lesser covering does not adequately protect the soil while a greater covering suppresses plant growth. Vegetative mulches, due to their low thermal conductivities,

decrease the conduction of heat into and out of the soil, causing the soil underneath to be cooler during the day and warmer during the night than bare soil. Thus it acts as an insulating agent on the soil surface. Vegetative mulches significantly affect soil temperature through their effect on the radiant energy balance of the system. The radiation balance is influenced by reflection of incoming radiation by surface residues, differential heating of soil and air and alteration of evaporation of soil water from the surface. Soil temperature generally decreases as reflectance increases. The insulating effect of mulches improves with increase in thickness of the surface residue layer. Age of the residue, color, geometry (standing, flattened or matted), distribution and amount are the main characteristics that affect radiation reflection. Reflection is normally greatest from bright residues like wheat straw and decreases with residue aging (discoloration) and decomposition. As a result, temperature differences between residue covered and bare soils are greatest when residues are bright and decrease as they age and decompose.

Organic mulches have the advantage of being biodegradable. In time, they are broken down and consumed by microorganisms, adding to the soil's organic matter content and improving soil structure or tilth. The vegetative mulches reduce the rate of evaporation from the soil surface, but the effectiveness of vegetative mulching in the reduction of evaporation depends upon the thickness, type, orientation and color of the residue. The rate of evaporation decreases with an increase in mulch thickness. As the material density of mulch affects the thickness and surface coverage obtained with a given weight of material, a low-density material such as wheat straw reduces evaporation more effectively than higher-density materials such as sorghum stubble or cotton stalks. To obtain a reduction in evaporation similar to that obtained with wheat straw on a weight basis, about twice as much sorghum stubble and more than four times as much cotton stalks are needed.

Living or cover crop mulch is also an important organic mulch used for some specific crops. Living mulch is a covering of living plant material, which is planted or allowed to remain living on the land after another crop has been planted. The live mulch system is based on the principles of mixed cropping. Here a perennial legume is established with the objective of smothering weeds and a seasonal grain crop is grown in small strips through the established live cover crop. The system works if the live mulch is a low-growing, nonclimber and does not compete for light, moisture or nutrients with the main crops. The live mulch system is more successful in humid and sub-humid regions with little or no water deficit than in semiarid or arid regions. This system imparts some of the benefits similar to surface-applied organic mulches such as increased soil carbon content, increased rainwater infiltration and reduced soil erosion. Cover crop mulches suppress weed growth through modification of the microclimate and release of allelochemicals. By land-surface coverage, mulches block the light stimulus required for germination of many small-seeded weed species. However, drastic yield suppression of the food crops has been reported in some studies due to allelopathic effects of the cover crop on the main crop and competition for moisture during periods of drought stress. Some important crops like cereal rye (*Secale cereale* L.) and hairy vetch (*Vicia villosa* Roth) can be used as cover crop mulch. Cereal rye is one of the best winter-hardy cover crops grown in the mid-western USA. Perennial ryegrass (*Lolium perenne* L.), a potential cover crop, is commonly used as a living mulch for some vegetables like cabbage.

Paper mulch is an important organic mulch produced from recycled paper. It has been evaluated as an available biodegradable alternative to plastic mulch. Paper does not require disposal at the end of the season and the use of recycled paper can reduce solid waste as well as energy costs. But paper may decompose too rapidly to maintain season long ground coverage. This can be checked by treating the paper with waste cooking oil, which delays the breakdown and enhances soil warming by increasing the paper's translucency to sunlight.

Inorganic Mulches

Different inorganic materials such as plastic film, gravels, crushed rocks etc. are used as mulch. Inorganic materials used for mulches do not add nutrients or humus to soil and do not decompose except after long exposure to weathering. Otherwise these materials are effective mulches and several of them are permanent and quite attractive.

Plastic mulch Plastics are the most widely used inorganic mulch materials. Mulching with the help of plastic film has played a major role in plant cultivation by creating some measure of mechanical protection at the soil surface and a microclimate favorable from the aspects of temperature distribution, retention of humidity and the supply of carbon dioxide to the stomata of lower leaves of small plants. Plastic mulch has been used commercially for vegetable production since the early 1960s. It can be used very effectively with trickle or drip irrigation, particularly for the cultivation of vines and strawberries where it encourages the lateral development of the root system.

In many of the developed countries plastic film, such as low-density polyethylene (LDPE) film is rapidly replacing conventional vegetative mulching because of its effectiveness and commercial viability, particularly in the case of high-value crops. In this method around 30–60% of the cultivated area is covered with a thick polythene film (30–150 μm). The plants are allowed to grow through the holes made in the LDPE film. In ornamental plants it is often used under gravel or stone mulches. The effectiveness of the plastic mulch depends upon the color of the mulch. Color determines the energy-radiating behavior of the mulch and affects the surface temperature of the mulch and the underlying soil temperature.

Recent new innovations in plastic technology manufacturing have resulted in the production of new plastic products for the farmers, e.g., photodegradable mulch, co-extruded mulch (black/white), and different colored mulches that serve specific purposes like maintenance of cool soil temperature during summer in tropical countries, increased warming of soil in spring in temperate regions and repelling aphids with silver-colored mulch. By raising the soil temperature in the planting bed, plant growth is accelerated, producing earlier yields in the temperate region. Black plastic mulch can result in harvest 7–14 days earlier while clear plastic accelerates harvest by 21 days in some conditions. For enhancing the earliness of vegetable crops in temperate regions, the mulch should maximize the transmission of the soil-warming portion of sunlight (the near infra-red radiation) and minimize the amount of visible light transmitted, as this visible light increases the weed growth under the mulch.

Due to the high cost of the cover and the difficulty of keeping it in place, plastic mulches are mainly used for some high-value vegetable and cash crops where these are found to be cost-effective. Muskmelons, tomatoes, peppers, cucumbers, squash, eggplants, watermelons and okra are some of the vegetable crops that have shown significant increases in earliness, yield, and fruit quality when grown on plastic mulch. Some less-valuable crops like sweetcorn, snap beans, southern peas and pumpkins have also shown similar responses in temperate climates. Based on the ability to transmit, absorb or reflect specific wavelengths, plastic mulches are divided into the following categories.

Black plastic mulch: Black is the most commonly used color of polyethylene mulch. This is chosen more often than clear polythene mulch because of its ability to control weeds by excluding photosynthetically active radiation. Black polythene placed in direct contact with the soil can warm the soil down to a depth of about 8–10 cm. Under black plastic film the temperature of the soil during the day is not significantly higher and in certain cases it is even lower than in the bare ground but during the night the temperature is always higher by 0.5–4°C than uncovered ground. This is desirable for warm-season vegetables such as peppers, tomatoes, eggplants etc. It is also widely used for the cultivation of pineapples in Hawaii, USA for reaping a higher yield.

Transparent/clear plastic mulch: Clear plastics have the greatest warming potential. They are transparent to incoming radiation and trap the longer wavelengths re-radiating from the soil, thereby producing a greenhouse effect that warms the soil underneath. Under transparent plastic film during daytime, soil temperature remains 2–10°C more than the bare ground while at night a difference of 2–4°C is observed, depending on the season, soil type and its moisture content. For warm-season crops, such as vine crops, earlier and higher yields can be achieved with clear mulch. However, as clear mulch is transparent to visible light, weed infestation under this mulch poses some serious problems, which may negate the benefits of soil warming.

White plastic mulch: Under white plastic mulch the temperature always remains lower than the uncovered ground due to its high reflectance and low transmittance of solar radiation. Since energy exchange is very poor, this film is used either in regions with a high level of sunshine where it is required to reduce the energy and lower the soil temperature or in regions of low luminosity where the amount of reflected light is to be increased on the lower and middle leaves of crop plants.

Colored plastic mulch: These are opaque, like black plastics, and are made with many different pigments. They do not allow light to pass through, but rather, reflect specific wavelengths of light up into the crop canopy. These mulches may be useful for enhancing crop production. For example, red mulches are effective in promoting growth of tomatoes. The other response observed with different mulch colors is the increase or decrease in insect population depending on reflected light and possibly temperature. Yellow appears to attract insects, especially cucumber beetle, and silver appears to repel aphids. The possibility of utilizing yellow mulch to monitor insect populations in a field of non-yellow plastic mulch shows promise. In addition, reduction of pesticide applications could be possible if only rows of yellow mulch (one row of yellow for every five or six rows of other-colored mulch laid) were sprayed with insecticide to control

high populations of pests. Thus, yellow plastic mulches can be used to attract insects, much like a trap crop.

Infrared transmitting mulches (IRT): These have been developed recently and are wavelength-selective mulches. IRT mulches warm the soil intermediate to clear and black polythene mulches. These mulches are pigmented to reduce the amount of visible light transmitted and thereby reduce weed growth underneath. Thus, they blend the soil-warming characteristics of clear mulch with the weed-control ability of black mulch. IRTs contain very specific pigments that give them the unique ability to transmit a maximum of near-infrared radiation and a minimum (14–16%) of visible light. Some visible light is required to be transmitted in order to maximize soil warming. The benefits of IRTs are greatest for early-season plantings in temperate regions. Though these mulches are more expensive than black or clear plastics, some growers find them useful for high-value, heat-loving crops like melons.

Gravel mulch Gravel mulching is an age-old practice and can be very effective in water conservation (both in enhancing infiltration and in suppressing evaporation, even in layers as thin as 5–10 mm). Pea gravel is an attractive permanent mulch that is usually applied 5–10 cm deep and can be reused indefinitely. These are particularly good for rock garden plantings. The problem of this type of mulches is that once applied, gravels cannot be removed, do not decompose and might hinder future farm operations or land uses. Another issue is its high cost of mining, transportation and application in the field.

Crushed rock Crushed volcanic rocks or stones are available in many colors or sizes and make a permanent cover. These materials are especially useful around plants subject to crown rot. The color of the rocks influences the temperature of the soil underneath. Black rock absorbs heat and keeps the soil temperature warmer than normal, while white rock reflects sunlight and thus maintains a cooler temperature underneath. However, white rock can generate too much heat for the above-ground portion of some plants to survive.

Other materials used as mulch Sprayable preparations of latex, asphalt, oil, fatty alcohols etc. are also used as mulch for certain high-income crops such as strawberries. They inhibit evaporation by forming monomolecular films on the soil surface. Some water-proofing chemicals like silicone on application to clods on the soil surface form a layer of water-repellent aggregates. These aggregates then constitute a dry mulch that serves as an evaporation barrier. This layer also enhances infiltration by allowing the treated surface zone to act as a detention reservoir to prevent runoff.

Methods of Mulch Application

Proper methodology and timing of mulch application should be followed to realize the full benefit of mulching. The methods of mulch application vary with soil type, climate, crop and type of mulch material used.

Paper and plastic mulches are generally spread and fastened down either between the rows or over the rows of the crop. The plant seedlings in the latter case are allowed to grow through suitable openings made in the mulch material. Plastic mulches in commercial crop cultivation are generally applied by mulch applicators, which are available commercially or can be manufactured. They generally consist of a plastic-roll holder, two furrow-openers, and two land wheels followed by two furrow-closers mounted on a tool bar. The furrow-openers create a trench on either side of the mulch. The two land wheels follow the furrow-openers and hold the plastic in the furrow. The two furrow-closers refill the trench, firmly anchoring the plastic. A 15–23 cm border of plastic is usually buried leaving 76–91 cm of mulch exposed. Rototilling prior to installation creates a smooth lump-free surface, which helps to ensure a tight union between the plastic and the soil for more efficient heat transfer.

Organic mulches are normally spread a few centimeters thick on the soil surface. The depth of mulch depends upon the mulch material and types of crop on which they are applied. Thicker layers of mulch are normally placed around trees and shrubs, while finer and smaller materials are generally used for small flowers and vegetables. Ten centimeters of loose fibrous materials work well around trees and shrubs. The finer and smaller the particle size, the thinner the layer needs to be. Thick layers of very fine material block air to the roots of plants. For trees, mulch cover should reach at least to the tree's drip line or the outer perimeter of the branches. It should preferably be kept at least 15 cm away from the tree trunk and a couple of centimeters away from plant stems.

Mulches are normally applied and spread uniformly on the soil surface except in the case of vertical mulching. In vertical mulching crop residues, stubbles of the previous crop or other agricultural waste materials are dumped in vertical trenches dug at few-meter intervals in the field with the main objective to recharge the profile and conserve moisture in low

rainfall areas of the semiarid tropics. The few centimeters-protruded portions of the stubble from the trenches also reduce the runoff and sediment loss from the field.

Impact of Mulching on Crop Growth and Input-Use Efficiency

Modification of Crop Growth

The effect of mulches on crop growth response depends on the combination of climate, type of crop and soil. Mulching influences three principal factors affecting crop growth, i.e., weed competition, seedling emergence and root growth, which in turn influence crop productivity.

Weed competition Straw mulching suppresses weed growth through delayed emergence, reduced weed count and smothering effects. Effectiveness of weed control is dependent on the type of mulch material used. Black plastic mulches are more effective for weed control than organic mulches in thin layers as light does not penetrate through black plastic mulches.

Seedling emergence Lack of moisture or unfavorable soil temperature in the seed zone often limits seedling emergence even though there may be sufficient water in the lower part of the profile. Mulch keeps the surface layers wetter for longer periods and thus enhances seed germination and seedling emergence. The times required for initiation of emergence and maximum emergence are shorter under mulching than under no mulch. Straw mulch significantly increases percentage of seedling emergence both under dry spell and rainfall situations (Table 9). Under dry-spell conditions, seedling emergence increases with mulching due to lowering of maximum temperature and reduction in evaporation loss of water. Under rainfall conditions the seedling emergence increases under mulching due to retardation of crust formation, which otherwise hampers seedling emergence.

Table 9 Effect of wheat straw mulching on seedling emergence in cotton and soybean in a sandy loam soil

	Percent seedling emergence			
	Cotton		*Soybean*	
Treatment	*No rain*	*Simulated rain*	*No rain*	*Simulated rain*
No mulch	25	9	19	0
Mulch	31	23	30	20

Adapted from Prihar SS (1986) In: *Proceedings of Seminar Organized by Fertilizer Association of India* (Northern Region) at Dehradun, India, May 10–11, 1986, pp. 31–44.

Root growth Mulching favorably influences root growth and distribution in the upper soil layer due to favorable soil temperature and higher soil water content under mulch. Higher soil water content in the surface layer reduces the mechanical resistance to growing roots, which helps in their proliferation. Application of organic mulches to the previous standing crop after the recession of rains helps in conservation and carry-over of higher amounts of residual soil moisture at the time of sowing of the subsequent rainfed crop, leading to better root proliferation and higher crop yield in regions where the weather starts cooling in the post-rainy season (Table 10). This carry-over effect is more prominent under conservation tillage than under conventional tillage.

Improvement of Input-Use Efficiencies

The favorable effects of mulching on crop growth and yields through an improved soil environment enhance both water and nutrient-use efficiency. Increase in water availability in the upper layers due to a reduction in evaporation and runoff losses under mulch is one of the reasons for increased input-use efficiency. Furthermore, modification of the soil hydrothermal regime under mulch favors root proliferation, which can tap water and nutrients from a larger soil volume, thus improving water and nutrient-use efficiency. Mulching results in better coordination of demand and supply of water resulting in complete extraction of profile-stored water, which improves the water-use efficiency. As grain yield-water use relation is linear within the domain of limited water supplies, any increase in water supply, beyond a minimum threshold would be directly reflected in increased grain yield as well as water-use efficiency (WUE).

Higher soil water content particularly in the initial stages of crop growth results in greater total water use by the crop, leading to higher plant growth and greater yield. Another factor contributing significantly to higher yield is more efficient use of growing season precipitation when the surface is covered with mulch. The final consequence of a higher yield with a relatively small increase in total water use is higher WUE with mulching than under unmulched conditions.

Application of pine needles, a waste product from pine forests, as mulch at the time of sowing of some tuber crops like potato improves the yield and water-use efficiency and results in savings of nitrogen fertilizer and irrigation water (Table 11). Hence, within certain limits, mulches can act as a substitute for both irrigation water and nitrogen fertilizer.

Table 10 Effect of tillage and nitrogen on root growth and grain yield of rainfed wheat in a silty-clay loam soil (Typic Hapludalf)

Mulch tillage (T)	Nitrogen	Root mass density ($kg\,m^{-3}$)		Grain yield ($Mg\,ha^{-1}$)
		Tillering	Flowering	
T_1: Mulching previous standing maize with wild sage followed by its incorporation with conventional tillage immediately prior to sowing of wheat	N_{60}	0.19	0.61	2271
	N_{120}	0.21	0.67	2711
T_2: As T_1 but wheat sowing following conservation tillage and retaining the residues of wild sage as mulch	N_{60}	0.26	0.68	2532
	N_{120}	0.29	0.78	2942
T_3: As T_1 but mulching by eupatorium instead of wild sage	N_{60}	0.20	0.58	1850
	N_{120}	0.24	0.65	2330
T_4: As T_2 but mulching with eupatorium instead of wild sage	N_{60}	0.25	0.60	2252
	N_{120}	0.29	0.72	2623
T_5: Farmer's practice of conventional tillage after maize harvesting	N_{60}	0.13	0.41	1231
	N_{120}	0.17	0.53	1282
LSD ($P = 0.05$) Mulch-tillage (T)		0.05	0.06	433
Nitrogen (N)		0.03	0.04	274
Interaction (T × N)		0.07	0.08	610

Adapted from Acharya CL, Kapur OC, and Dixit SP (1998) *Soil and Tillage Research* 46: 153–163.

Table 11 Tuber yield and water-use efficiency (WUE) of potato as influenced by mulching and N application (pooled over 4 years)

Treatment[a]	Autumn potato		Spring potato	
	Yield ($Mg\,ha^{-1}$)	WUE ($kg\,ha^{-1}\ cm^{-1}$)	Yield ($Mg\,ha^{-1}$)	WUE ($kg\,ha^{-1}\ cm^{-1}$)
$60\,kg\,N\,ha^{-1}$	8.46	479.9	13.30	257.0
$120\,kg\,N\,ha^{-1}$	9.86	528.2	15.15	271.2
Pine needle mulch @ 10 Mg ha^{-1} + $60\,kg\,N\,ha^{-1}$	12.69	711.0	16.33	340.9
Pine needle mulch @ 10 Mg ha^{-1} + $120\,kg\,N\,ha^{-1}$	14.52	735.5	18.34	395.4
CD ($P = 0.05$)	1.19	-	1.46	-

[a]All the treatments received a uniform dose of FYM @ $20\,Mg\,ha^{-1}$.
Reproduced from Acharya CL and Kapur OC (2001) *Indian Journal of Agricultural Sciences* 71: 306–309.

With increased water availability under mulching, crop response to fertilizer increases, thereby improving the fertilizer-use efficiency. Enrichment of soil with plant nutrients through organic mulching materials boosts crop production. Furthermore, mulching helps in moderating the soil thermal regime, which has a profound influence on nutrient transformations in soil, root growth and nutrient uptake by crop plants. Thus it enhances the nutrient-use efficiency by crops.

Limitations of Mulching

Despite the positive aspects, the following are limitations, which must be taken into account while following the practice of mulching.

1. If mulch is applied before sowing it hinders sowing and many post-sowing operations like fertilizer applications, irrigation etc. Mulch poses special problems for the arable farmer because tillage tools become clogged with the residue, thus planting under the residue is not always successful. Some specific equipment is required for post-sowing operations in mulched plots.

2. In shallow potato planting if mulch is improperly managed it may lead to a greater percentage of stolon becoming aerial shoots and more tubers becoming sunburned. The mulch layer in potato fields should be removed once full canopy is achieved, to avoid disease and pest infestation.

3. In areas of high rainfall, thick mulching may lead to water-logging conditions and provide an environment for disease infestation. If the soil is too wet under mulch, an increase in denitrifying bacteria may cause loss of N.

4. The effectiveness of a mulch depends upon the depth and extent of its coverage. However, if the mulch is applied too thickly it may hinder seedling emergence through the deep impenetrable barrier.

Therefore, it is suggested to use a thin layer of mulch at planting so as not to impede emergence and to follow with a thicker application once the seedlings are established.

5. Plastic mulches usually require pick-up and disposal at the end of the season and their manufacture and disposal entail significant environmental costs, as these are not biodegradable. Plastic mulch can interfere with the recharge of the soil profile by rain or overhead irrigation, which can lead to water deficits in vegetables if drip irrigation is not used.

6. Numerous phytotoxic substances associated with various crop residues have been shown to limit growth of several crop species reportedly through exudation, leaching or microbial production of allelopathic chemicals.

7. There are competing demands for crop residues for other uses in the semiarid tropics of developing countries. Where most of the plant residues are used as cattle fodder, farmers may not adopt mulching under conditions of scarce fodder. In the tropics, crop residues are also used for fencing, roofing and as a source of household fuel. Mulching is thus practiced under conditions where the marginal return from this practice is higher than its other competing demands.

8. The crop residue mulches reduce the temperature of the surface soil layers. In cool climates, this reduction in soil temperature shortens the growing season of the arable crops, while in wet areas higher soil moisture may induce gleying and anaerobic conditions. The limiting problem of using residue mulch on soils in temperate regions is its inability to remove excess water to allow the soil to warm up in time for planting the next crop.

9. The mulch competes with the main crop for nitrogen as it decomposes if the C:N ratio of the material is very high due to immobilization of the soil nitrogen by the microorganisms.

10. The surface residue mulch may generate a more favorable habitat for soil and surface-dwelling insects and pathogens. The residue left at the surface reduces water loss, moderates temperature extremes and provides food resources for the pests and may increase their infestation. Besides this, for many plant pathogens, residues provide a source of food, a place to live and a place to reproduce. A significant proportion of root-infecting organisms, for example, depend upon crop residues for their survival in soil. Root diseases caused by *Pythium* species are favored by wet soils and cool temperatures, especially in corn production.

See also: **Aggregation:** Physical Aspects; **Cover Crops**; **Crop-Residue Management**; **Crusts:** Structural; **Cultivation and Tillage**; **Erosion:** Water-Induced; **Evaporation of Water from Bare Soil**; **Infiltration**; **Nutrient Management**; **Water-Use Efficiency**

Further Reading

Acharya CL, Kapur OC, and Dixit SP (1998) Moisture conservation for rainfed wheat production with alternative mulches and conservation tillage in the hills of north-west India. *Soil and Tillage Research* 46: 153–163.

Bhagat RM and Acharya CL (1987) Effect of soil management on rainfed wheat in northern India. 1. Hydrothermal regime and root growth. *Soil and Tillage Research* 9: 65–77.

Bond JJ and Willis WD (1969) Soil water evaporation: Surface residue rate and placement effects. *Soil Science Society of America Proceedings* 33: 445–448.

Hatfield JL and Stewart BA (1994) Crop residue management. *Advances in Soil Science*. Boca Raton: Lewis Publishers Inc.

Lal R (1987) Farming systems. In: Lal R (ed.) *Tropical Ecology and Physical Edaphology*, pp. 618–685. New York: Wiley-Interscience Publication.

Prihar SS (1986) Fertilizer and water use efficiency through the aid of mulching. In: *Proceedings of Seminar Organized by Fertilizer Association of India* (Northern Region) at Dehradun, India, May 10–11, 1986. pp. 31–44. Dehradun: Fertilizer Association of India.

Prihar SS, Jalota SK, and Steiner JL (1996) Residue management for reducing evaporation in relation to soil type and evaporativity. *Soil Use and Management* 12: 150–157.

Stratton ML and Rechcigl JE (1998) Organic mulches, wood products and composts as soil amendments and conditioners. In: Wallace A and Terry RE (eds) *Handbook of Soil Conditioners – Substances that Enhance the Physical Properties of Soil*, pp. 43–95. New York: Marcel Dekker Inc.

Unger PW (1990) Conservation tillage systems. *Advances in Soil Science* 13: 27–68.

Unger PW (1994) *Managing Agricultural Residues*. Boca Raton: Lewis Publishers Inc.

MYCORRHIZAL FUNGI

L M Egerton-Warburton, Chicago Botanic Garden, Glencoe, IL, USA
J I Querejeta and M F Allen, University of California – Riverside, Riverside, CA, USA
S L Finkelman, Containerless Research, Inc., Evanston, IL, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Mycorrhizae ('fungus roots') are an almost universal terrestrial mutualism between plant roots and certain soil fungi, the oldest fossil evidence of which is in the rhizomes of early vascular plants from approximately 400 My ago. Most mycorrhizae are mutualistic, i.e., the fungi provide soil resources to the plant in exchange for photosynthates. Six major types of mycorrhizae are recognized: ectomycorrhizae (EM), arbuscular (AM), monotropoid, arbutoid, orchid, and ericoid (Table 1; Figure 1a). These categories are based largely on differences in phylogenetic, structural, or ecophysiological differences among fungal–plant relationships. We focus on AM and EM, the most common types.

AM are unique in that they are the only monophyletic mycorrhizal lineage that has coevolved with land plants (Table 1; Figure 1b). Although there are relatively few species of AM fungi (approx. 175 species to date), they associate with a wide diversity of plants in grasslands, many tropical forests, and some temperate forests. AM are particularly efficient at the uptake and transfer of nutrients under conditions of low soil moisture or fertility, especially immobile nutrients with low diffusion rates or adsorbed to soil particles (phosphorus, P), and mobile mineral species (NH_4).

EM and the remaining mycorrhizal types demonstrate polyphyletic lineages that represent parallel or convergent evolution, interspersed with saprophytic relatives, and multiple lineages that have gained or lost the capacity to form mycorrhizae. These associations are formed by a vast array of fungal taxa (more than 5000 species; Table 1; Figure 1b). However, compared with AM associations, a mere 3% of plants are EM, most of which are woody trees and shrubs in boreal, temperate, and some tropical forests. These mycorrhizae also demonstrate the capacity to acquire N in a variety of forms (proteins, peptides, NH_4) and sources (weathered rock, leaf litter), and P by producing extracellular acid phosphatases that release P from organic complexes in the soil.

The remaining types are confined to specific plant families and/or habitats. Arbutoid mycorrhizae are a variant of EM and found in ericaceous plants. Ericoid and orchid mycorrhizae coexist with specific plant groups and harbor specialized fungal symbionts that promote plant growth by directly mineralizing organic substrates, often at low soil pH, by releasing enzymes into the substrates (Table 1; Figure 1b). Monotropoid mycorrhizae demonstrate the evolution toward increased host control over their fungal partners (mycoheterotrophy). Typically, the host plant is dependent on fungi for mineral nutrients and energy (carbon substrates), because they lack photosynthetic apparatus and/or substantial roots, but the fungi derive little or no benefit from these associations.

Apart from the construction of fungal structures on or in the root, mycorrhizae invest resources in producing hyphae, microscopic threads 2–10 μm in diameter, which grow out into the soil from the mycorrhizal root (Figure 1). These external hyphae ramify the soil and increase plant nutrient and water supply by extending the volume of soil accessible to plants, act as a sink for carbon, bind soil particles into soil aggregates, and provide food for animals. Mycorrhizal benefits can include greater plant yield and nutrient accumulation. Moreover, hyphal linkages shared by two or more plants may facilitate the transfer of resources in a community, or assist in seedling establishment.

For the most part, being mycorrhizal thus represents a considerable C expense to the plant. In community and ecosystem structure and functioning, a large proportion of plant photosynthate production is allocated to mycorrhizal roots, resulting in a large flux of C into the belowground detrital system. In fact, as mycorrhizal fungal tissue grows, between 3 and 20% of a plant's net primary productivity is transferred from the host plant and allocated to the pool of live fungal tissue. The obligate fungal requirement for host-derived C means that any environmental factor that affects plants also affects their fungal mutualists. Then again, mycorrhizal diversity also determines the competitive balance between plant species and the overall diversity and productivity of plant communities, because each fungal species may differentially enhance (or restrain) survival, growth, and nutrient acquisition among host plants. Accordingly, each change in the soil physical or chemical environment that influences mycorrhizal functioning may also affect the host plant. Such interdependencies or linkages typically characterize the central paradigm of complexity,

Table 1 Summary of the principal evolutionary and ecologic characters of the major mycorrhizal types

	Arbuscular	*Ectomycorrhizae*	*Ericoid–arbutoid–monotropoid*[a]	*Orchid*
Evolution				
Age of association (My)	>400	>100	±80	±100
Fungal–plant coevolution	+	+	*	*
Specificity	Low	Intermediate	Medium-high	High
Multiple lineages of fungi	No	Yes	Yes	Yes
Fungal taxa	Zygomycetes	Basidiomycete Ascomycete Zygomycete	Ascomycetes Basidiomycetes	Basidiomycetes
Plant taxa	Bryophytes Pteridophytes Gymnosperms Angiosperms	Gymnosperms Angiosperms	Ericales Monotropodeae Bryophytes	Orchidaceae
Ecology				
Fungi septate	−	+	+	+
Fungi aseptate	+	−	−	−
Intracellular colonization	+	+	+	+
Fungal sheath	−	+ or −	+ or −	−
Hartig net	−	+	+ or −	−
Vesicles	+ or −	−	−	−
Depends on host for survival	+	−	−	−
Host achlorophyllous[b]	− (?+)	−	− (Ericoid, arbutoid) + (Monotropoid)	+ (Seedling[b])
Plant depends on fungus for seed germination	−	−	+ (Monotropoid)	+ (Seedling)
Ecophysiology				
Nutrient transfer to plant (and nutrients transferred)	+ (P)	+ (N)	Yes (N, P)	Yes (N, P)
Carbon transfer to fungus	+	+	+ (Ericoid, arbutoid)	+ (Chlorophyllous spp.)
Carbon transfer to plant	−	−	+ (Monotropoid)	+ (Achlorophyllous spp., seedlings)
Specialized nutrient–carbon transfer interface	+ Arbuscule	+ Hartig net	− (Hyphal coils, peg)	− (Hyphal coils)

Character states: +, present; −, absent; *, not known.
[a]Members of the Ericales are host to arbutoid, ericoid, and monotropoid mycorrhizae. For clarity, these types are considered as one group.
[b]All orchids are achlorophyllous in the early seedling stages; however, the extent of chlorophyll development varies among adults.

from the Latin *complexus*, meaning 'entwined' or 'twisted together.'

Complexity

What is complexity? What does it mean to have a complex mycorrhizal system? Traditionally, mycorrhizal functioning has been evaluated by empirical and descriptive studies or by theoretical concepts, mathematical models, and deductive structures, i.e., analytic versus synthetic models. Most experiments are also reductionist. In other words, they successively reduce a mycorrhizal system into smaller and simpler components that are studied in isolation and in a formulated manner. In contrast, the complexity approach is rich and holistic. The individual components within a mycorrhizal system mean little on their own, but together they give rise to the collective responses or emergent patterns of the whole system, and how this system interacts with its environment (Figure 2). This tenet is also central to ecology, or the totality of relations between organisms and their environment.

Mathematically, such systems exist 'on the edge of chaos,' or somewhere between perfect order and complete randomness, where a small change in one component can lead to widely varying results overall. Moreover, complexity *per se* has been allied with terms such as 'fractals,' 'cellular automata,' 'genetic algorithms,' 'order and hierarchy,' to name just a few. Such terms, however, simply emphasize the many different approaches that may be used to understand complexity. Can one use the science of complexity to understand mycorrhizal systems better? Conventional approaches based on linear causality and stochastic properties, while illustrative of the individual components, are inappropriate, because by their very nature mycorrhizal systems are dynamic processes rather than static entities. The canon is complex interplay. Conceptual models might provide

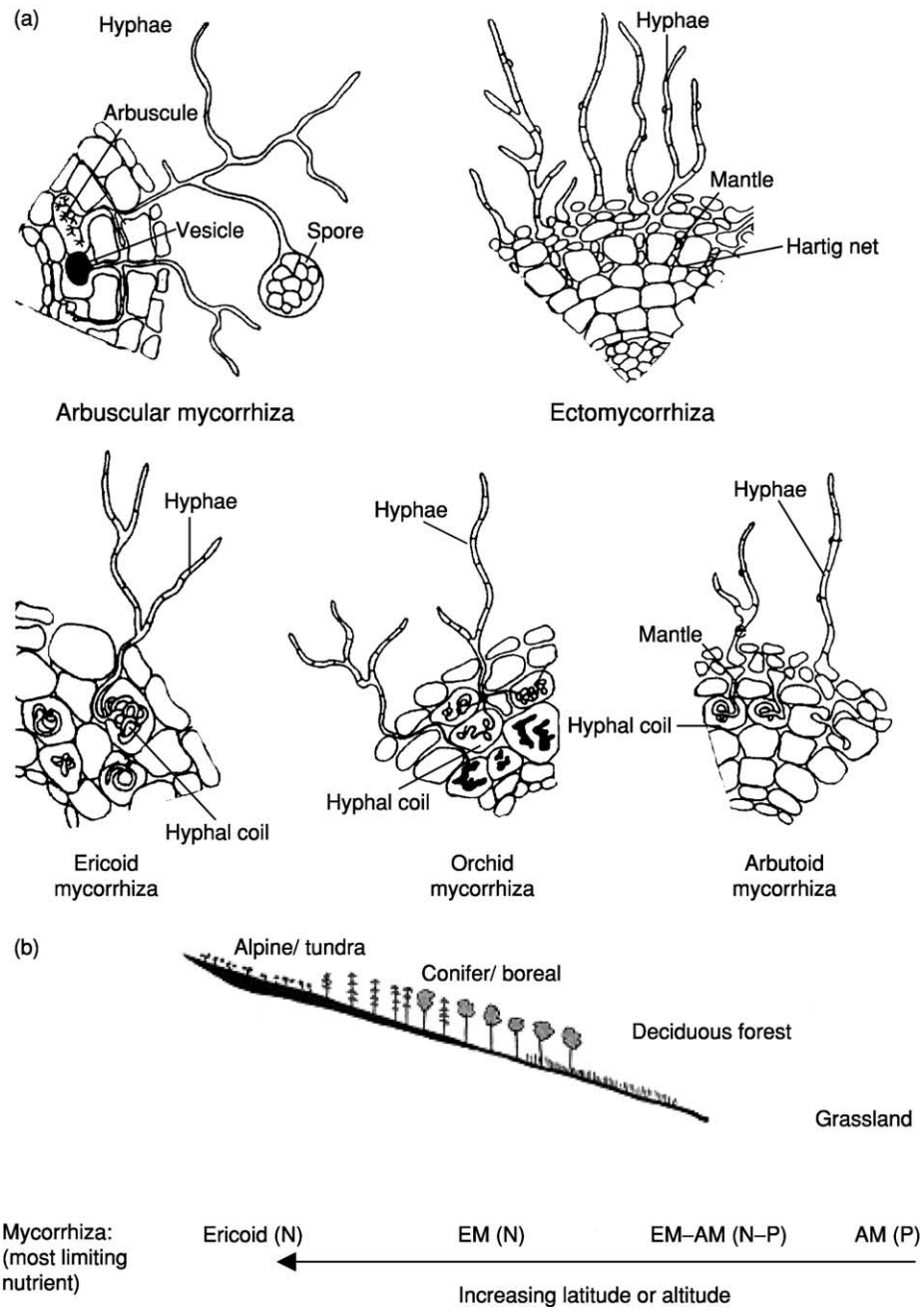


Figure 1 (a) Differential relationships between fungal and plant structures for each of the main types of mycorrhizal associations. Reprinted from Selosse *et al.* (1998) *Trends in Ecology and Evolution*, vol. 13: 15–20, with permission from Elsevier. (b) Mycorrhizal types as related to landscape position. Adapted from University of Wisconsin, Forest Ecology and Management, Department Literature.

frameworks for the identification of processes, structures, and relationships within complex systems.

Defining the Mycorrhizal System and Participants

The system is a specified region in space or a dynamic surface that changes over time (Figures 2 and 3). For example, the oak woodland bounded by grasslands and the hyphal network can both be considered as mycorrhizal systems. They are persistent and respond to the environment or repeated instances of similar or related conditions. Mycorrhizal systems repeatedly demonstrate a negative response to N fertilization and a positive response to elevated atmospheric CO_2. The environment includes organic and mineral nutrients, water, carbon, and influences such as urbanization and global change.

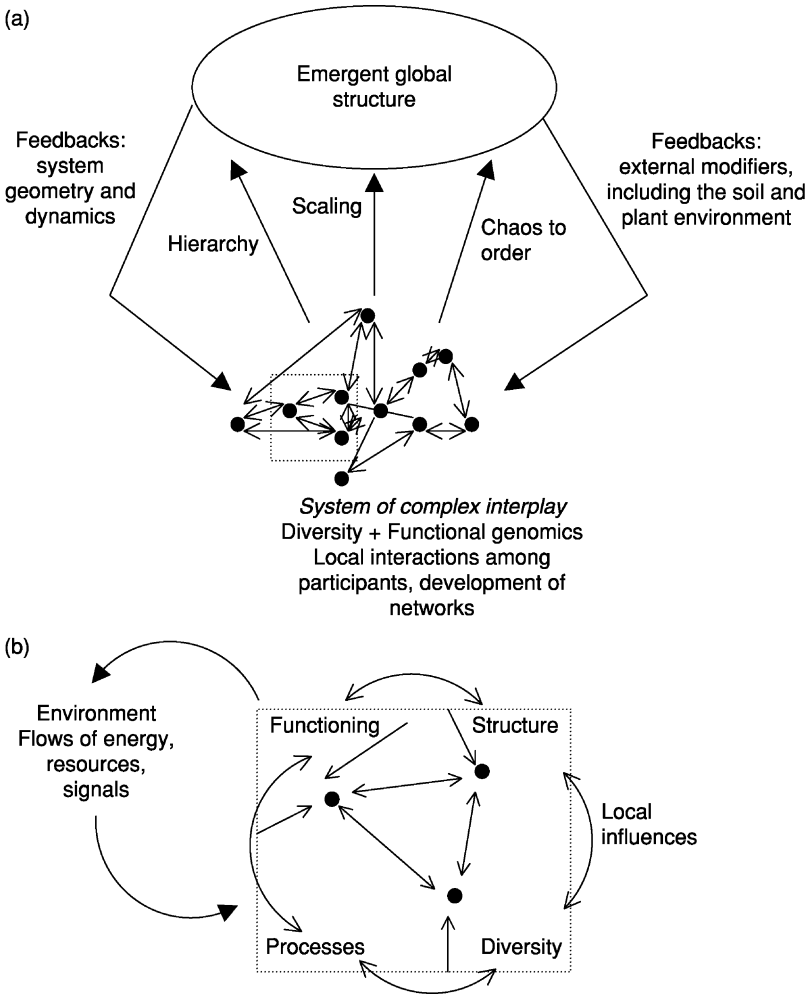


Figure 2 (a) Conceptual model of complexity as a framework for developing analyses in mycorrhizal systems; (b) components of the interactions between the mycorrhizal systems and the environment.

The system is composed of fungal and plant participants. The fungal participants may consist of single-species situations or a set of interacting species. The plant participants are networked into the system by fungi such that the system consists of a lattice (hyphal network) with a discrete agent (host plant) at each lattice site or node (Figure 3a,b). In the field, each mycorrhizal fungus attaches to as many plants as it encounters. For example, there is considerable overlap of EM associated with Ponderosa pine, Douglas fir, Tan oak, Grand fir, and Madrone. Every plant within the network interacts because any transition or change in one fungal partner leads to changes in the neighboring or connected plants. Adding new individuals progressively alters the size and degrees of connectedness within the network in a geometric fashion (Figure 3b). Connections among networks can result in increasingly complex configurations (Figure 3c). The journey into mycorrhizal complexity starts with ecologic diversity, and the structural and functional genomics of the symbiosis.

Governance of the System

Diversity and the Delivery of Ecosystem Services

The potential global taxon richness is extremely high for both plants and fungi forming mycorrhizal systems. The measures of α, β, and γ relate the number of species to area: α is the number of species per hectare (local), β, the number of species per 10–1000 ha (regional) and γ, the number of species in more than 10 000 ha (landscape). Both α and β diversity are similar among habitats and appear to reach upper limits of 5–20 species ha^{-1} for α diversity and 30–50 species for β diversity. Hence, there may be some limit to the system services provided by mycorrhizal fungi. At the γ level, patterns of diversity

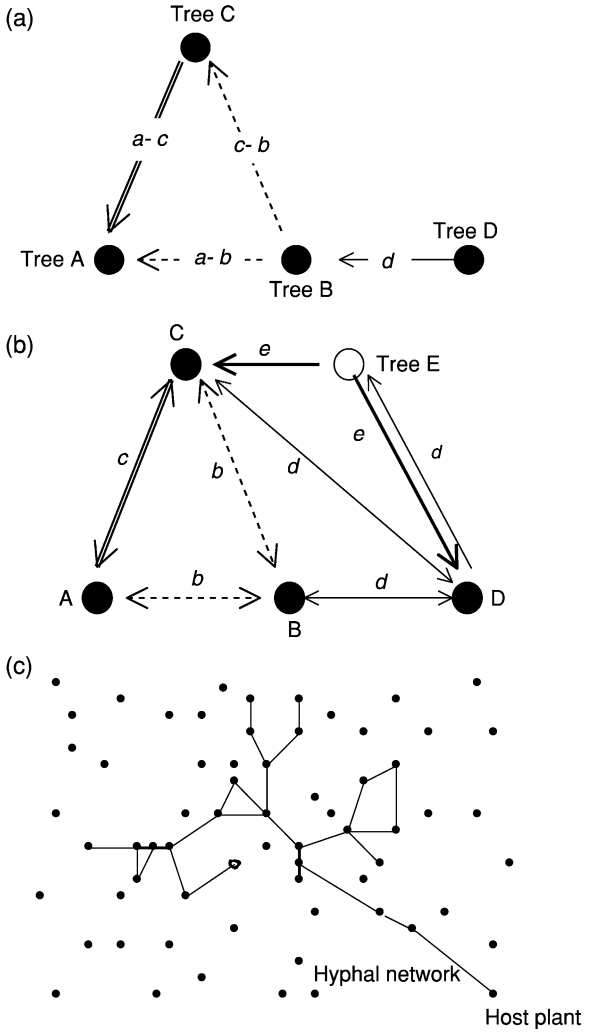


Figure 3 (a) The simplest configuration of a network among three trees; (b) the addition of another host tree and associated fungi increases the number of potential links in the network; (c) the addition of subsequent host plants further increases the complexity of links among plants and the system as a whole. Upper-case letters and filled circles denote the host tree; lower-case letters denote fungus or fungi associated with a host tree. Arrows denote the potential hyphal linkages among plants.

among mycorrhizal types diverge. Forty-eight species of AM fungi have been recovered from *Artemisia tridentata*, although only 175 species have been described globally. In contrast, more than 2000 EM species associate with Douglas fir, but worldwide the number of described EM species approaches many thousands.

What is being measured? Molecular approaches measure species and population genetic diversity, whereas sporocarp (mushroom) and morphotype estimates are responsive to host species or environment. Thus, sporocarp diversity is unlikely to reflect the diversity of EM fungi on the root tips. In *Quercus* woodlands, only three species of EM sporocarps per

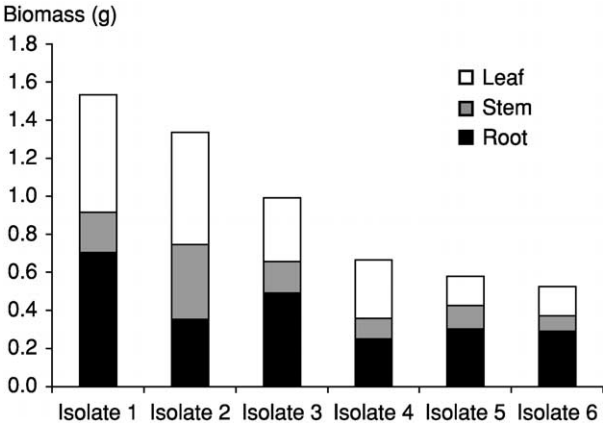


Figure 4 Biomass accumulation in the leaves, stems, and roots of *Quercus* plants grown with six different cultures of *Cortinarius* that were isolated from one clump of sporocarps.

tree have been collected compared with seven to nine morphotypes and up to 20 molecular types per tree. In addition, a few widespread species of fungi account for more than 50% of mycorrhizal abundance and the majority of species are rarely detected. In Douglas fir, 69 morphotypes of EM have been recovered, of which nine types colonize 67% of the root tips.

Fungi also separate both spatially and temporally even when associated with the same host plant. Spatial variation within a system is high because the dominant fungi vary considerably among trees. Fungi may also be spatially segregated to the undercanopy, edge of canopy, and interspace areas. From the temporal perspective, EM formed by certain species of *Laccaria*, *Amanita*, and *Hebeloma* persist in small genets for an average of 1–2 years. In contrast, large genets that last decades are formed by species of *Suillus* and *Cortinarius*. It follows that all species have the potential to contribute to system functioning, although not necessarily in synchrony.

Increasing the diversity of fungi within a system can increase the diversity, productivity, and stability of the plant community. Douglas fir grows better when supported by multiple mycorrhizal fungi, and especially in competition with Ponderosa pine. Different fungi can also undertake different ecologic roles across space and time. For example, six isolates of *Cortinarius* isolated from a single group of sporocarps produce different biomass allocation patterns in their host plant (Figure 4). Certain fungi may preferentially acquire organic nitrogen (N) (*Boletus*) and others, inorganic N (*Pisolithus*, *Laccaria*). *Cenococcum geophilum* exploits the local soil for resources, whereas *Pisolithus* forms extensive hyphal networks that run from one resource patch to another. *Rhizopogon subcaerulescens* persists deep within the soil profile, survives fire, and thus forms

mutualisms in the postfire environment, whereas the upper-profile dweller, *Thelephora*, does not. Such patterns of diversity have important implications for system functioning.

Multiple Genes Coordinate the Development of Mycorrhizae

Both the fungal and plant partner engage in a precisely orchestrated developmental sequence that results in the expression of distinct fungal or plant genes exclusively or predominantly within the symbiosis. In *Eucalyptus globulus–Pisolithus tinctorius* EM, 65 genes record the dialogue of the early development of the fungal mantle. Mycorrhizal development results in the upregulation of the translation initiation factor eIF4A gene, a high-molecular-weight (MW) protein complex, in *Pisolithus tinctorius*, sequences coding for classes of symbiosis-regulated (SR) genes, particularly the cysteine-rich hydrophobins and multigene SRAP32 families, and more recently proline-rich proteins and proteophosphoglycans. eIF4A is essential for translation and cell growth, whereas the structural proteins encoded by SR genes are usually involved in membrane synthesis and function. Similarities to genes encoding proteins involved in primary metabolism and transport (carbon, H^+-ATPase), the cell cycle, deoxyribonucleic acid (DNA) replication, cell structure (β-tubulin), fungal structure (arbuscule degradation, hyphal proliferation), signal transduction, and the suppression of plant defense genes have also been recovered from AM. In addition, mutants unable to sustain AM (Myc^-) were also unable to associate with N-fixing Rhizobium (Nod^-). Conserved signal transduction pathways may thus be common to both mutualisms during their early development

Genomic Hegemony over Nutrient and Carbon Transport

Just as many genes can be ascribed to development of the symbiosis, so can the expression of other functional genes record the dialogue of carbon and nutrient transfer in mycorrhizae. For example, the monosaccharide uptake system in *Amanita muscaria* requires a transporter, AmMST1. AmMST1 contains the open reading frame for a 56-kDa protein that is expressed when *A. muscaria* is axenically cultured at a glucose concentration of less than 5 mmol l^{-1}(the threshold expression level) for 18–24 h, and in *A. muscaria–Picea abies* and *A. muscaria–Populus tremula* × *tremuloides* mycorrhizae. The expression of AmMST1 and any subsequent increase in the import of monosaccharides by fungus triggers an increase in the supply of photoassimilates from the host. Phenylalanine ammonia lyase (AmPAL), a key enzyme of secondary metabolism (phenolics), is also involved in plant defense against pathogenic microbes. Phosphoenolpyruvate carboxylase (PEPC) is expressed in regions of intense mycorrhizal development; PEPC moderates the assimilation of NH_4 in fungi and the transport of amino acids such as glutamine from fungus to plant. However, many genes in fully sequenced genomes are of unknown function, because mycorrhizae are typically studied in a synthetic environment and outside the ecologic context in which they evolved.

Functional Genomics

Responses of a mycorrhizal system to specific environmental cues are tightly coordinated by a physiologically appropriate subset of its genes. Nutrient transport from soil to plant via the mycorrhizal fungus has been the hallmark of mycorrhizal studies for more than a century. One gene that may be involved in P transport efficiency of AM is a fungal membrane P transporter, which moves P from the soil into the hyphae. This gene is regulated by the availability of P in the soil environment, but the fungal partner controls the synthesis of polyphosphate (polyP) and thus the amount of P available in the root and shoot. In turn, *in vivo* ^{31}P nuclear magnetic resonance (NMR) illustrates that mycorrhizal plants typically accumulate P as supramolecular aggregates of polyP (chain length 5–11 units), whereas nonmycorrhizal plants accumulate solely inorganic P (P_i). In mycorrhizae, polyP represents much of the mobile P reservoir.

Mycorrhizal fungi are also capable of utilizing many forms of N and along a variety of pathways. A putative nitrate-transporter gene has been isolated from mycorrhizal *Medicago* in which the expression is downregulated by some, but not all, AM. NMR reveals that NH_4 is predominantly incorporated by glutamine synthetase/glutamate synthetase pathway (GS/GOGAT), with a small proportion by the glutamate dehydrogenase (GDH) pathway. AM may also accelerate the decomposition of organic matter and acquire N directly from these substrates. ^{15}N pulse-chase studies show the incorporation of labeled N into arginine, which is consistent with the ornithine cycle. In turn, the N is stored in relatively immobile complexes with polyP in vacuoles.

Fractionation of N during N uptake by fungi results in considerable shifts in ^{15}N abundance such that ^{15}N abundance in fungal tissues is unlikely to reflect those of the original soil N. The patterns of fractionation are also fungus and N source-specific. Intense fractionation against ^{15}N in ericoid systems results in a ^{15}N-enriched plant, whereas EM systems discriminate against ^{14}N (^{15}N-depleted plant). In

addition, δ^{15}N values for EM sporocarps in a *Quercus* woodland range from 2 to 11. Thus, organic and inorganic N compounds can be accessed differentially by fungi. ^{15}N natural abundance also demonstrates that the strict functional boundaries between mycorrhizal and decomposer fungi may be less clear-cut than previously considered (δ^{15}N 5.29% in mycorrhizal *Boletus*; 6.03% in saprobic *Bovista*).

Spectra from ^{1}H- and ^{13}C-NMR assays in AM *Gigaspora* demonstrate that fungal spore germ tubes, extraradical hyphae, and intraradical hyphae differ in their lipid (triaglycerol) and carbohydrate (hexose) metabolism. Spore germ tubes are sites of lipid breakdown and gluconeogenesis. In intraradical hyphae, hexose is transported into the cytoplasm, where lipid synthesis occurs. Extraradical hyphae act as sink for lipids, and the lipids are converted to carbohydrates (trehalose) and occasionally glycogen, and stored in spores or vesicles. Fungi convert the simple sugars to trehalose or other complex sugar to prevent or reduce reabsorption. Changes in root phytohormone composition also occur in tandem with the loading and unloading of C.

Nutrient fluxes may also be coregulated because the translocation of nonlimiting nutrients such as magnesium (Mg) depends on simultaneous long-distance translocation of P. Combined mass balance (P, potassium, K) and stable isotope analyses (^{25}Mg) show that hyphae can transfer up to 5% K and 3–4% Mg as long as P fluxes through the system.

The factors regulating the amounts of resources transported between fungi are poorly understood. Virtually all studies to date have concentrated on demonstrating the presence and form of elemental transport from soil to plant via the fungal hyphae. As a result, there is little quantitative work on the mechanisms regulating the transport of resources among fungi.

Games and Models for Mapping Causality

Knowing the structure of water yields no clues as to why water goes down a drain in a vortex. Similarly, in mycorrhizae, the genes are expressed and the metabolic machinery produces parts of the system, but they provide no information on how structure in mycorrhizal systems is generated. Mapping is a way of generalizing dynamics and establishing conceptual illustrations and models using simple input/output diagrams. The cognitive system map (Figure 5) draws a causal relationship among components and predicts how complex events might play out. Each arrow defines a causal link where 'plus' is an increase or enhanced response, 'minus' is a decrease,

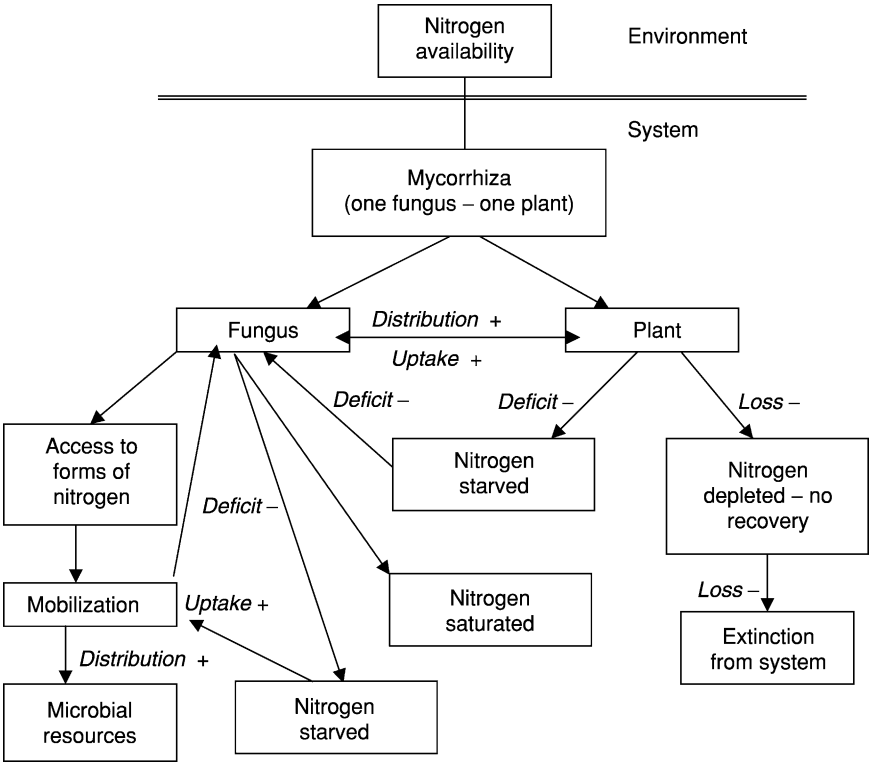


Figure 5 A single-state model of the relationship between the environment and the mycorrhizal system in response to soil nitrogen status.

or negative response, and 'zero' is stasis. The final cognitive map defines the nonlinear dynamic system.

This single-state map for a one plant–one fungus mycorrhizal system and its response to soil N availability illustrates that the fungal response is enhanced by plant nutrient stress and decreased by an increase in nutrient availability and drought. The plant response is enhanced by P availability and decreased by drought. The overall model is the 'If condition, then action,' which has the same meaning as statements in programming language. If the condition portion is satisfied, then the action is triggered and other rules are activated. This is the 'bucket brigade' algorithm (*sensu* J.H. Holland), where signals are passed down the line to reinforce a chain of effect. This is also similar to path analysis. Calculations of the potential outcomes (fungus, plant, system, or 3^n) from the model (Figure 5), indicate that if soil N is limited, then the fungus transports approximately 64% of N to plant, or strengthens the importance of the fungus to the system. However, if soil N increases then the fungus transports only approximately 34% of plant N.

Adding more participants into the system increases the number of possible outcomes exponentially. A two-fungus system responds quite differently to the environment than a one-fungus system. Fungal respiration is depressed in each of *Rhizopogon* and *Suillus* mycorrhizae by 72% and 30%, respectively, after the addition of NH_4 (Figure 6). An additive cognitive map, however, indicates that NH_4 may have a bigger negative effect on respiration in *Rhizopogon–Suillus* in concert than on a single species or the edges where they meet. Using mapping rules as patches also demonstrates the nonlinear nature of the sporocarp $\delta^{15}N$:$\delta^{13}C$ divide between mycorrhizal and saprobic fungi (Figure 7). In particular, when $\delta^{15}N$ is ± 0, $\delta^{13}C$ in mycorrhizal fungi is depleted, whereas $\delta^{13}C$ in saprobes is enriched.

As more knowledge emerges, so do more rules and a more elaborate model that is connected to actual observations. For example, the N needs of the fungus and plant are determined by the level of internal N reservoirs that are depleted at a rate that keeps the fungus and plant functioning, so that low N implies a high need for the resource. The N level in the reservoir may only be increased by providing the correct signal molecule or transporter. Adding these new parameters into Figure 6 increases the number of potential iterations and outcomes and, in turn, the complexity of the system. Such complex systems are controlled by feedback mechanisms between plants and their mycorrhizal mutualists and soil microbes, and local host-dependent preferences for certain communities of fungi.

Hierarchy and Scaling

Complex adaptive systems are hierarchically nested because they are large systems that are comprised of progressively smaller systems, etc.; these large

	Respiration (μg C mg^{-1} ergosterol h^{-1})		
Species	*Control*	*NH_4*	*% decrease*
Rhizopogon sp.	5.1	1.4	72
Suillus pungens	3.0	2.1	30

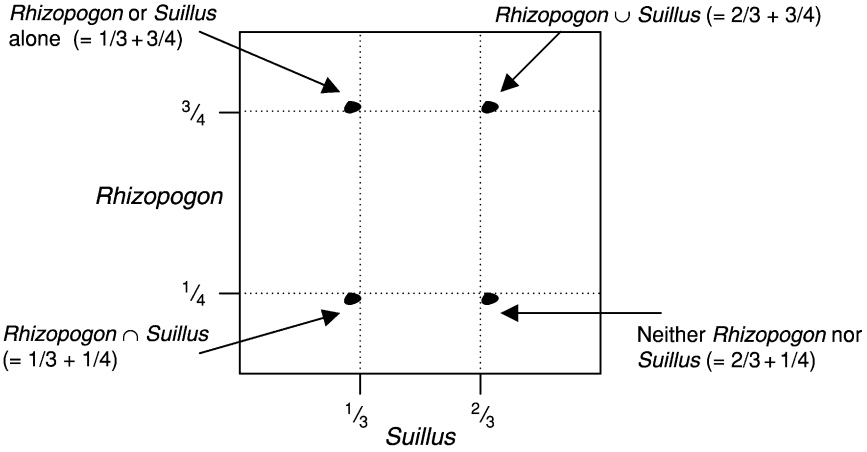


Figure 6 An additive map of the predicted interactions between the ectomycorrhizal fungi, *Rhizopogon* and *Suillus*, after media enrichment with ammonia. Ranking of analysis (highest to lowest effect) shows that NH_4 influences: two species together in culture (*Rhizopogon* ∪ *Suillus*, 17/12); single-species cultures (*Rhizopogon* or *Suillus* alone, 13/12); fungal contaminants (neither *Rhizopogon* nor *Suillus*, 11/12); patches where each species intersects (*Rhizopogon* ∩ *Suillus*, 7/12).

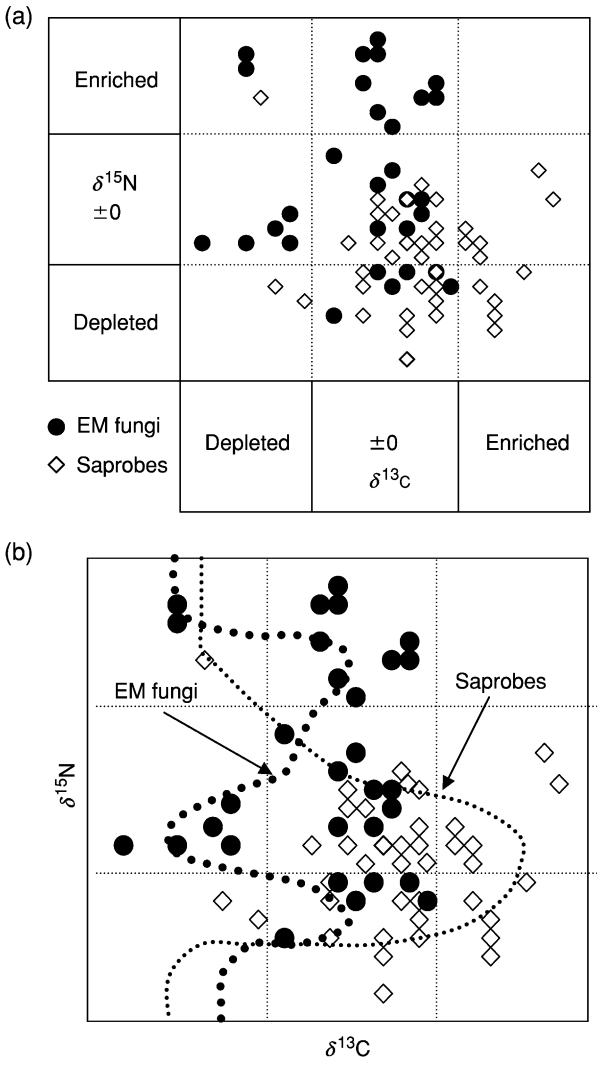


Figure 7 (a) Ratio of δ^{15}N:δ^{13}C in sporocarps of mycorrhizal and nonmycorrhizal (saprobic) fungi; (b) additive patch map of δ^{15}N:δ^{13}C in mycorrhizal and nonmycorrhizal fungi. EM, ectomycorrhizae.

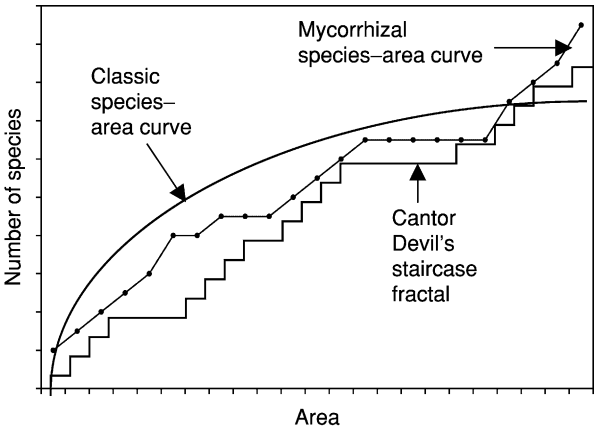


Figure 8 Overlap of the asymptotic (classic) species–area curve, the species–area relationship for morphotypes in oak woodlands, and the Devil's staircase fractal.

systems can interact with each other as well. The issue of scaling and translating information from small-scale patterns and processes to understanding of large-scale problems, such as global change, is essential. This is captured in a simple way in the study of chaos and fractals. Chaos is a rich description of a complex nonlinear system that can include the coupling of different scales over time or space and where a small, seemingly insignificant change in one of the attributes of a single agent can have a widely varying effect on the system as a whole. A fractal is a noninteger semidimension, and one of its traits can be self-similarity at different scales.

Quantifying global species' diversity from measurements of local diversity (species, transcripts, sequences) is typically summarized as a function of sample size and area, i.e., a species–area curve (Figure 8). The classic species–area curve asymptotically approaches the total diversity in the collection. On the contrary, species–area curves for EM morphotypes typically proceed in a step-wise fashion and bear a stronger resemblance to the Cantor (or Devil's staircase) fractal. An inverse relationship exists between species diversity and D, the dimension of the Cantor function. Therefore, systems with high diversity have low D ($\dim \rightarrow <0$), and conversely, mycorrhizal systems with low diversity possess a high D ($\dim \rightarrow <1$). Relating D to issues in global-change biology suggests that an increasing input of N may correspond to a progressive increase in D, while elevated atmospheric CO_2 may correspond to lower D.

Dimensionality within the Cantor function also appears to corresponds with the transition from order to chaos, because entropy (disorder) and enthalpy (energy) are inextricably linked to chaos as follows: maximum diversity: $\dim \rightarrow 0$: minimum order: maximum disorder: minimum energy. This follows from Shannon–Wiener's adaptation of Newton's second law of thermodynamics as applied to information theory. A commonly used measure of diversity, the Shannon–Wiener index, is also calculated using exactly the same formula as for statistical entropy. Thus, calculations of diversity are simply reflections of system order or disorder (chaos), where the higher the diversity index the greater the level of chaos within a system.

Autonomous Networks

Fungal and plant species differ in mutualistic functioning and tolerance to the environment, but, when networked, the benefits of each are distributed

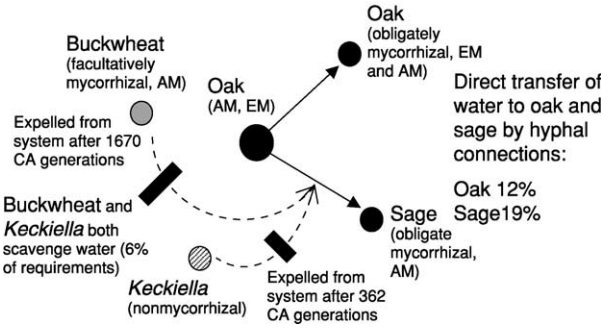


Figure 9 Model estimates of the time and sequence of plant species elimination from a fungal network based on simple cellular automata (CA) rules of water sharing and host mycorrhizal status. AM, arbuscular mycorrhizae; EM, ectomycorrhizae.

throughout the system. Fungal networks allow both plants and fungi to share resources, including carbon and water (Figure 9).

Network models based on cellular automata (CA) are appropriate in nonlinear systems. Traits or variables such as diversity (entropy) may be used for CA and, in turn, a set of cellular automata may form a Cantor set. In the simplest case, a CA consists of fungal hyphae that evolve in a series of time steps and according to simple specific rules. Even though the rules may be very simple in the initial state, the pattern produced is self-similar (or fractal). Fractal analysis derives the models of hyphal growth. EM hyphae demonstrate branching in a binary tree in which each of the nodes has two branches and operations are restricted to a two-way choice (left or right, great or less). *L*-systems describe system growth and efficiency. Because EM fungi follow the Twig model and AM fungi follow Weed models, the *L*-system can be used to describe mathematically or simulate mycorrhizal growth patterns. This is packing efficiency, or the manner in which mycorrhizal systems can squeeze a great deal of hyphal length and surface area into a finite volume and, in turn, resource acquisition. Because the production of extraradical hyphae differs significantly among mycorrhizae, packing efficiency may be a useful indicator of system functioning.

Integrating this knowledge into mycorrhizal functioning indicates how the network might proceed. Plants joined by a hyphal network may share water by either direct or indirect transfer (Figure 9). Using a CA with simple water-sharing rules, nonmycorrhizal *Keckiella* is eliminated from the group after 362 generations and facultative mycorrhizal *Eriogonum* after 1670 generations. Only plants that are directly connected by fungal hyphae remain in the system, because packing efficiency (resource transfer) increases within the networked area.

The extent of network development and the transfer resources among plants is still an open question. Spanning trees may be useful in calculating these outcomes. In addition, multiple transfers of resources suggest redundancy. Different fungal species, and possibly communities, may be capable of the same processes even though there may be different versions of the gene(s) that encode for a certain purpose, e.g., N acquisition.

Summary

Lovelock's Gaia hypothesis conceptualized biodiversity and mutualism in their most advanced and elegant integration. In mycorrhizae, diversity and mutualistic functioning unite successive systems into networks and complex systems. In order to show that complexity has increased overall, it is sufficient to show, that – all other things being equal – connections have increased in at least one dimension. What is lacking is the ability to make predictions about how complexity in mycorrhizal communities will change their function as the systems and/or their environment is altered by human impacts and global change. Combining molecular, metabolic, and biochemical data with nonlinear mathematical models might provide the foundations and rules for understanding mycorrhizal complexity. The limitations and utility of any data, however, remain in developing data-mining and complexity-modeling tools and techniques to utilize effectively the information from a local and global perspective, because data are gathered on scales from molecules to genomes, organelles, cells, tissues, and organs. Bioinformatics is the acquisition of knowledge by means of computational tools for the organization, management, and mining of genetic biological data. These analytical tools are generally applied to the oceans of data collected by molecular biology, e.g., the Human Genome project. A more appropriate term for mycorrhizal systems may be 'ecoinformatics' or the accumulation of ecologically based data sets appropriate to mycorrhizae *in situ*, followed by data integration. In doing so, it will then be appropriate to say that diversity and mutualism provide ecosystem function and what that functioning may be.

See also: **Fungi**; **Root Architecture and Growth**; **Root Exudates and Microorganisms**

Further Reading

Allen MF (1996) *The Ecology of Mycorrhizae*. Cambridge, UK: Cambridge University Press.

Smith SE and Read DJ (1997) *Mycorrhizal Symbiosis*, 2nd edn. San Diego, CA: Academic Press.

Van der Heijden MGA and Sanders IR (eds) (2002) *Mycorrhizal Ecology*. Berlin, Germany: Springer-Verlag.

ENCYCLOPEDIA OF SOILS IN THE ENVIRONMENT

EDITED BY
DANIEL HILLEL
CYNTHIA ROSENZWEIG
DAVID POWLSON
KATE SCOW
MICHAIL SINGER
DONALD SPARKS

VOLUME THREE

ENCYCLOPEDIA OF SOILS IN THE ENVIRONMENT

FOUR-VOLUME SET

by Daniel Hillel (Editor-in-Chief)

Hardcover: 2200 pages
Publisher: Academic Press; 1 edition (November 8, 2004)
Language: English
ISBN-10: 0123485304
ISBN-13: 978-0123485304

Book Description

More than ever before, a compelling need exists for an encyclopedic resource about soil the rich mix of mineral particles, organic matter, gases, and soluble compounds that foster both plant and animal growth. Civilization depends more on the soil as human populations continue to grow and increasing demands are placed upon available resources.
The Encyclopedia of Soils in the Environmentis a comprehensive and integrated consideration of a topic of vital importance to human societies in the past, present, and future.

This important work encompasses the present knowledge of the world's variegated soils, their origins, properties, classification, and roles in the biosphere. A team of outstanding, international contributors has written over 250 entries that cover a broad range of issues facing today's soil scientists, ecologists, and environmental scientists.
This four-volume set features thorough articles that survey specific aspects of soil biology, ecology, chemistry and physics. Rounding out the encyclopedia's excellent coverage, contributions cover cross-disciplinary subjects, such as the history of soil utilization for agricultural and engineering purposes and soils in relation to the remediation of pollution and the mitigation of global climate change.

This comprehensive, yet accessible source is a valuable addition to the library of scientists, researchers, students, and policy makers involved in soil science, ecology, and environmental science.

Also available online via ScienceDirect featuring extensive browsing, searching, and internal cross-referencing between articles in the work, plus dynamic linking to journal articles and abstract databases, making navigation flexible and easy. For more information, pricing options and availability visit www.info.sciencedirect.com.

* A distinguished international group of editors and contributors
* Well-organized encyclopedic format providing concise, readable entries, easy searches, and thorough cross-references
* Abundant visual resources — photographs, figures, tables, and graphs — in every entry
* Complete up-to-date coverage of many important topics — essential information for scientists, students and professionals alike

EDITOR-IN-CHIEF

Daniel Hillel
Columbia University
New York, NY
USA

EDITORS

Jerry L Hatfield
National Soil Tilth Laboratory
Ames, IA
USA

Kate M Scow
University of California
Davis, CA
USA

David S Powlson
Rothamsted Research
Harpenden
UK

Michael J Singer
University of California
Davis, CA
USA

Cynthia Rosenzweig
NASA Goddard Institute for Space Studies
New York, NY
USA

Donald L Sparks
University of Delaware
Newark, DE
USA

EDITORIAL ADVISORY BOARD

R Bardgett
Lancaster University
Lancaster
UK

J L Boettinger
Utah State University
Logan, UT
USA

G Gee
Pacific Northwest National Laboratory
Richland, WA
USA

R Keren
The Volcani Center
Bet Dagan
Israel

J Kimble
USDA Natural Resources Conservation Service
Lincoln, NE
USA

M B Kirkham
Kansas State University
Manhattan, KS
USA

M Kutilek
Prague
Czech Republic

D Martens
Southwest Watershed Research Center
Tucson, AZ
USA

K Mengel
Justus Leibig University
Giessen
Germany

K Reichardt
Center for Nuclear Energy in Agriculture
Piracicaba
Brazil

K Ritz
Cranfield University
Silsoe
UK

R Schulin
Swiss Federal Institute of Technology Zurich
Schlieren
Switzerland

N Senesi
Università di Bari
Bari
Italy

J T Sims
University of Delaware
Newark, DE
USA

K Smith
University of Edinburgh
Edinburgh
UK

R L Tate
Rutgers University
New Brunswick, NJ
USA

N van Breemen
Wageningen Agricultural University
Wageningen
The Netherlands

W H van Riemsdijk
Department of Soil Quality
Wageningen
The Netherlands

FOREWORD

The *Encyclopedia of Soils in the Environment* is a vitally important scientific publication and an equally important contribution to global public policy. The *Encyclopedia* brings together a remarkable range of cutting-edge scientific knowledge on all aspects of soil science, as well as the links of soils and soil science to environmental management, food production, biodiversity, climate change, and many other areas of significant concern. Even more than that, the *Encyclopedia* will immediately become an indispensable resource for policy makers, analysts, and students who are focusing on one of the greatest challenges of the 21st century. With 6.3 billion people, our planet is already straining to feed the world's population, and is failing to do so reliably in many parts of the world. The numbers of chronically poor in the world have been stuck at some 800 million in recent years, despite long-standing international goals and commitments to reduce that number by several hundred million. Yet the challenge of food production will intensify in coming decades, as the human population is projected to rise to around 9 billion by mid-century, with the increased population concentrated in parts of the world already suffering from widespread chronic under-nourishment.

Unless the best science is brought to these problems, the situation is likely to deteriorate sharply. Food production systems are already under stress, for reasons often related directly to soils management. In Africa, crop yields are disastrously low and falling in many places due to the rampant depletion of soil nutrients. This situation needs urgent reversal, through increasing use of agro-forestry techniques (e.g. inter-cropping cereals with leguminous nitrogen-fixing trees) and increasing the efficient applications of chemical fertilizers. In other impoverished, as well as rich, parts of the planet, decades of intensive agriculture under irrigation have led to salinization, water-logging, eutrophication of major water bodies, dangerous declines of biodiversity and other forms of environmental degradation. These enormous strains are coupled with the continuing pressures of tropical deforestation and the lack of new promising regions for expanding crop cultivation to meet the needs of growing populations. Finally, there looms the prospect of anthropogenic climate change. Global warming and associated complex and poorly understood shifts in precipitation extremes and other climate variables all threaten the world's natural ecosystems and food production systems in profound yet still imperfectly understood ways. The risks of gradual or abrupt climate change are coupled with the risks of drastic perturbations to regional and global food supplies.

The *Encyclopedia* offers state-of-the-art contributions on each of these challenges, as well as links to entries on the fundamental biophysical processes that underpin the relevant phenomena. The world-scale and world-class collaboration that stands behind this unique project signifies its importance for the world community. It is an honor and privilege for me to introduce this path-breaking endeavor.

Jeffrey D Sachs
Director
The Earth Institute at Columbia University
Quetelet Professor of Sustainable Development
Columbia University, New York, USA

PREFACE

The term 'soil' refers to the weathered and fragmented outer layer of our planet's land surfaces. Formed initially through the physical disintegration and chemical alteration of rocks and minerals by physical and biogeochemical processes, soil is influenced by the activity and accumulated residues of a myriad of diverse forms of life. As it occurs in different geologic and climatic domains, soil is an exceedingly variegated body with a wide range of attributes.

Considering the height of the atmosphere, the thickness of the earth's rock mantle, and the depth of the ocean, one observes that soil is an amazingly thin body – typically not much more than one meter thick and often less than that. Yet it is the crucible of terrestrial life, within which biological productivity is generated and sustained. It acts like a composite living entity, a home to a community of innumerable microscopic and macroscopic plants and animals. A mere fistful of soil typically contains billions of microorganisms, which perform vital interactive biochemical functions. Another intrinsic attribute of the soil is its sponge-like porosity and its enormous internal surface area. That same fistful of soil may actually consist of several hectares of active surface, upon which physicochemical processes take place continuously.

Realizing humanity's utter dependence on the soil, ancient peoples, who lived in greater intimacy with nature than many of us today, actually revered the soil. It was not only their source of livelihood, but also the material from which they built their homes and that they learned to shape, heat, and fuse into household vessels and writing tablets (ceramic, made of clayey soil, being the first synthetic material in the history of technology). In the Bible, the name assigned to the first human was Adam, derived from 'adama,' meaning soil. The name given to that first earthling's mate was Hava (Eve, in transliteration), meaning 'living' or 'life-giving.' Together, therefore, Adam and Eve signified quite literally 'Soil and Life.'

The same powerful metaphor is echoed in the Latin name for the human species – Homo, derived from humus, the material of the soil. Hence, the adjective 'human' also implies 'of the soil.' Other ancient cultures evoked equally powerful associations. To the Greeks, the earth was a manifestation of Gaea, the maternal goddess who, impregnated by Uranus (god of the sky), gave birth to all the gods of the Greek pantheon.

Our civilization depends on the soil more crucially than ever, because our numbers have grown while available soil resources have diminished and deteriorated. Paradoxically, however, even as our dependence on the soil has increased, most of us have become physically and emotionally detached from it. Many of the people in the so-called 'developed' countries spend their lives in the artificial environment of a city, insulated from direct exposure to nature, and some children may now assume as a matter of course that food originates in supermarkets.

Detachment has bred ignorance, and out of ignorance has come the delusion that our civilization has risen above nature and has set itself free of its constraints. Agriculture and food security, erosion and salination, degradation of natural ecosystems, depletion and pollution of surface waters and aquifers, and decimation of biodiversity – all of these processes, which involve the soil directly or indirectly – have become abstractions to many people. The very language we use betrays disdain for that common material underfoot, often referred to as 'dirt.' Some fastidious parents prohibit their children from playing in the mud and hurry to wash their 'soiled' hands when the children nonetheless obey an innate instinct to do so. Thus soil is devalued and treated

as unclean though it is the terrestrial realm's principal medium of purification, wherein wastes are decomposed and nature's productivity is continually rejuvenated.

Scientists who observe soil closely see it in effect as a seething foundry in which matter and energy are in constant flux. Radiant energy from the sun streams onto the field and cascades through the soil and the plants growing in it. Heat is exchanged, water percolates through the soil's intricate passages, plant roots extract water and transmit it to their leaves, which transpire it back to the atmosphere. Leaves absorb carbon dioxide from the air and synthesize it with soil-derived water to form the primary compounds of life. Oxygen emitted by the leaves makes the air breathable for animals, which consume and in turn fertilize plants.

Soil is thus a self-regulating bio-physio-chemical factory, processing its own materials, water, and solar energy. It also determines the fate of rainfall and snowfall reaching the ground surface – whether the water thus received will flow over the land as runoff, or seep downward to the subterranean reservoir called groundwater, which in turn maintains the steady flow of springs and streams. With its finite capacity to absorb and store moisture, and to release it gradually, the soil regulates all of these phenomena. Without the soil as a buffer, rain falling over the continents would run off entirely, producing violent floods rather than sustained river flow.

Soil naturally acts as a living filter, in which pathogens and toxins that might otherwise accumulate to foul the terrestrial environment are rendered harmless. Since time immemorial, humans and other animals have been dying of all manner of disease and have then been buried in the soil, yet no major disease is transmitted by it. The term *antibiotic* was coined by soil microbiologists who, as a consequence of their studies of soil bacteria and actinomycetes, discovered streptomycin (an important cure for tuberculosis and other infections). Ion exchange, a useful process of water purification, also was discovered by soil scientists studying the passage of solutes through beds of clay.

However unique in form and function, soil is not an isolated body. It is, rather, a central link in the larger chain of interconnected domains and processes comprising the terrestrial environment. The soil interacts both with the overlying atmosphere and the underlying strata, as well as with surface and underground bodies of water. Especially important is the interrelation between the soil and the climate. In addition to its function of regulating the cycle of water, it also regulates energy exchange and surface temperature.

When virgin land is cleared of vegetation and turned into a cultivated field, the native biomass above the ground is often burned and the organic matter within the soil tends to decompose. These processes release carbon dioxide into the atmosphere, thus contributing to the earth's greenhouse effect and to global warming. On the other hand, the opposite act of reforestation and soil enrichment with organic matter, such as can be achieved by means of conservation management, may serve to absorb carbon dioxide from the atmosphere. To an extent, the soil's capacity to store carbon can thus help to mitigate the greenhouse effect.

Thousands of years are required for nature to create life-giving soil out of sterile bedrock. In only a few decades, however, unknowing or uncaring humans can destroy that wondrous work of nature. In various circumstances, mismanaged soils may be subject to erosion (the sediments of which tend to clog streambeds, estuaries, lakes, and coastal waters), to leaching of nutrients with attendant loss of fertility and eutrophication of water bodies, to waterlogging and impaired aeration, or to an excessive accumulation of salts that may cause a once-productive soil to become entirely sterile. Such processes of soil degradation, sometimes called 'desertification,' already affect large areas of land.

We cannot manage effectively and sustainably that which we do not know and thoroughly understand. That is why the tasks of developing and disseminating sound knowledge of the soil and its complex processes have assumed growing urgency and importance. The global environmental crisis has created a compelling need for a concentrated, concise, and definitive source of information – accessible to students, scientists, practitioners, and the general public – about the soil in all its manifestations – in nature and in relation to the life of humans.

Daniel Hillel
Editor-in-Chief
May 2004

INTRODUCTION

The *Encyclopedia of Soils in the Environment* contains nearly 300 articles, written by the world's leading authorities. Pedologists, biologists, ecologists, earth scientists, hydrologists, climatologists, geographers, and representatives from many other disciplines have contributed to this work. Each of the articles separately, and all of them in sequence and combination, serve to summarize and encapsulate our present knowledge of the world's variegated soils, their natural functions, and their importance to humans.

Concise articles surveying specific aspects of soils (soil genesis, soil chemistry and mineralogy, soil physics and hydrology, and soil biology) are complemented by articles covering transdisciplinary aspects, such as the role of soils in ecology, the history of soil utilization for agricultural and engineering purposes, the development of soil science as a discipline, and the potential or actual contributions of soils to the generation, as well as to the mitigation, of pollution and of global climate change.

This comprehensive reference encompasses both the fundamental and the applied aspects of soil science, interfacing in general with the physical sciences and life sciences and more specifically with the earth sciences and environmental sciences.

The *Encyclopedia of Soils in the Environment* manifests the expanding scope of modern soil science, from its early sectarian focus on the utilitarian attributes of soils in agriculture and engineering, to a wider and much more inclusive view of the soil as a central link in the continuous chain of processes constituting the dynamic environment as a whole. Thus it both details and integrates a set of topics that have always been of vital importance to human societies and that are certain to be even more so in the future.

Daniel Hillel
Editor-in-Chief
May 2004

CONTENTS

Contents are given as follows: CHAPTER NAME Author(s) Page number

VOLUME 1

A

ACID RAIN AND SOIL ACIDIFICATION L Blake 1

ACIDITY N Bolan, D Curtin and D Adriano 11

AERATION D E Rolston 17

AGGREGATION

Microbial Aspects S D Frey 22

Physical Aspects J R Nimmo 28

AGROFORESTRY P K R Nair 35

AIR PHASE see AERATION; DIFFUSION

ALBEDO see ENERGY BALANCE; RADIATION BALANCE

ALLOPHANE AND IMOGOLITE see AMORPHOUS MATERIALS

ALLUVIUM AND ALLUVIAL SOILS J L Boettinger 45

ALUMINUM SPECIATION D R Parker 50

AMMONIA D E Kissel and M L Cabrera 56

AMORPHOUS MATERIALS J Harsh 64

ANAEROBIC SOILS P W Inglett, K R Reddy and R Corstanje 72

ANION EXCHANGE see CATION EXCHANGE

APPLICATIONS OF SOILS DATA P J Lawrence 78

ARCHAEA J E T McLain 88

ARCHEOLOGY IN RELATION TO SOILS J A Homburg 95

B

BACTERIA

Plant Growth-Promoting Y Bashan and L E de-Bashan 103

Soil L J Halverson 115

BACTERIOPHAGE M Radosevich, K E Williamson and K E Wommack 122

BIOCONTROL OF SOIL-BORNE PLANT DISEASES C E Pankhurst and J M Lynch 129

BIODIVERSITY D H Wall 136

BUFFERING CAPACITY B R James 142

BULK DENSITY see POROSITY AND PORE-SIZE DISTRIBUTION

C

CALCIUM AND MAGNESIUM IN SOILS N Bolan, P Loganathan and S Saggar 149

CAPILLARITY D Or and M Tuller 155

CARBON CYCLE IN SOILS

Dynamics and Management C W Rice 164

Formation and Decomposition C A Cambardella 170

CARBON EMISSIONS AND SEQUESTRATION K Paustian 175

CATION EXCHANGE L M McDonald, V P Evangelou and M A Chappell 180

CHEMICAL EQUILIBRIA A P Schwab 189

CHEMICAL SPECIATION MODELS see SURFACE COMPLEXATION MODELING

CHERNOZEMS see GRASSLAND SOILS

CHILDS, ERNEST CARR E G Youngs 195

CIVILIZATION, ROLE OF SOILS D Hillel 199

CLASSIFICATION OF LAND USE see LAND-USE CLASSIFICATION

CLASSIFICATION OF SOILS R W Arnold 204

CLASSIFICATION SYSTEMS

Australian R W Fitzpatrick 211

FAO F O Nachtergaele 216

Russian, Background and Principles M Gerasimova 223

Russian, Evolution and Examples D Konyushkov 227

USA D J Brown 235

CLAY MINERALS D G Schulze 246

CLIMATE CHANGE IMPACTS P Bullock 254

CLIMATE MODELS, ROLE OF SOIL P Smith 262

COLD-REGION SOILS C-L Ping 268

COLLOID-FACILITATED SORPTION AND TRANSPORT R Kretzschmar 276

COMPACTION J J H van den Akker and B Soane 285

COMPOST T L Richard 294

CONDITIONERS R E Sojka, J A Entry and W J Orts 301

CONSERVATION see EROSION: Water-Induced; Wind-Induced; SUSTAINABLE SOIL AND LAND MANAGEMENT; TERRACES AND TERRACING

CONSERVATION TILLAGE M R Carter 306

COVER CROPS L Edwards and J Burney 311

CROP ROTATIONS C A Francis 318

CROP WATER REQUIREMENTS L S Pereira and I Alves 322

CROP-RESIDUE MANAGEMENT D C Reicosky and A R Wilts 334

CRUSTS

Biological J Belnap 339

Structural R L Baumhardt and R C Schwartz 347

CULTIVATION AND TILLAGE M R Carter and E McKyes 356

D

DARCY'S LAW D Swartzendruber 363

DEGRADATION C J Ritsema, G W J van Lynden, V G Jetten and S M de Jong 370

DENITRIFICATION D A Martens 378

DESERTIFICATION D Hillel and C Rosenzweig 382

DIFFUSION T Addiscott and P Leeds-Harrison 389

DISINFESTATION A Gamliel and J Katan 394

DISPERSION see FLOCCULATION AND DISPERSION

DISSOLUTION PROCESSES, KINETICS K G Scheckel and C A Impellitteri 400

DRAINAGE, SURFACE AND SUBSURFACE N R Fausey 409

DRYLAND FARMING G A Peterson 414

E

EARTHWORMS see FAUNA

EDAPHOLOGY A L Ulery 419

ELECTRON PARAMAGNETIC RESONANCE see ELECTRON-SPIN RESONANCE SPECTROSCOPY

ELECTRON-SPIN RESONANCE SPECTROSCOPY N Senesi and G S Senesi 426

ELECTROSTATIC DOUBLE-LAYER see CATION EXCHANGE

ENERGY BALANCE M Fuchs 438

ENVIRONMENTAL MONITORING P J Loveland and P H Bellamy 441

ENZYMES IN SOILS R P Dick and E Kandeler 448

EROSION

Irrigation-Induced G A Lehrsch, D L Bjorneberg and R E Sojka 456

Water-Induced J E Gilley 463

Wind-Induced T M Zobeck and R S Van Pelt 470

ESSENTIAL ELEMENTS E A Kirkby 478

EUTROPHICATION A J Gold and J T Sims 486

EVAPORATION OF WATER FROM BARE SOIL C W Boast and F W Simmons 494

EVAPOTRANSPIRATION G Stanhill 502

F

FACTORS OF SOIL FORMATION

Biota A H Jahren 507

Climate O C Spaargaren and J A Deckers 512

Human Impacts J Sandor, C L Burras and M Thompson 520

Parent Material K R Olson 532

Time E F Kelly and C M Yonker 536

FAUNA T Winsome 539

VOLUME 2

FERTIGATION U Kafkafi and S Kant 1

FERTILITY J L Havlin 10

FERTILIZERS AND FERTILIZATION H W Scherer 20

FIELD CAPACITY see WATER CYCLE

FLOCCULATION AND DISPERSION I Shainberg and G J Levy 27

FLUORESCENCE SPECTROSCOPY N Senesi and V D'Orazio 35

FOLIAR APPLICATIONS OF NUTRIENTS M Tagliavini and M Toselli 53

FOOD-WEB INTERACTIONS P C de Ruiter and J C Moore 59

FORENSIC APPLICATIONS W F Rowe 67

FOREST SOILS J R Boyle 73

FOURIER TRANSFORM INFRARED SPECTROSCOPY D Peak 80

FRACTAL ANALYSIS Y Pachepsky and J W Crawford 85

FREEZING AND THAWING

Cycles B Sharratt 98

Processes G N Flerchinger, G A Lehrsch and D K McCool 104

FUNGI K Ritz 110

G

GEOGRAPHICAL INFORMATION SYSTEMS J Bo¨hner, T Selige and R Ko¨the 121

GERMINATION AND SEEDLING ESTABLISHMENT A Hadas 130

GLOBAL WARMING see CARBON EMISSIONS AND SEQUESTRATION; CLIMATE CHANGE IMPACTS; GREENHOUSE GAS EMISSIONS

GRASSLAND SOILS J A Mason and C W Zanner 138

GREEN MANURING see COVER CROPS

GREENHOUSE GAS EMISSIONS K A Smith 145

GROUNDWATER AND AQUIFERS Y Bachmat 153

GROUNDWATER POLLUTION see POLLUTION: Groundwater

H

HEAT AND MOISTURE TRANSPORT R Horton and A Globus 169

HEAT CAPACITY see THERMAL PROPERTIES AND PROCESSES

HEAT FLOW see THERMAL PROPERTIES AND PROCESSES

HEAVY METALS D C Adriano, N S Bolan, J Vangronsveld and W W Wenzel 175

HILGARD, EUGENE WOLDEMAR R Amundson 182

HOOGHOUDT, SYMEN BAREND P A C Raats and R R van der Ploeg 188

HUMIFICATION T C Balser 195

HYDRAULIC PROPERTIES, TEMPERATURE EFFECTS S A Grant 207

HYDRIC SOILS G W Hurt 212

HYDROCARBONS P Kostecki, R Morrison and J Dragun 217

HYDRODYNAMIC DISPERSION see SOLUTE TRANSPORT

HYDRODYNAMICS IN SOILS T P A Ferre´ and A W Warrick 227

HYSTERESIS J H Dane and R J Lenhard 231

I

IMMISCIBLE FLUIDS R J Lenhard, J H Dane and M Oostrom 239

INCEPTISOLS A Palmer 248

INDUSTRIAL POLLUTION see POLLUTION: Industrial

INFILTRATION T P A Ferre´ and A W Warrick 254

INFRARED SPECTROSCOPY see FOURIER TRANSFORM INFRARED SPECTROSCOPY

IRON NUTRITION K Mengel and H Kosegarten 260

IRRIGATION

Environmental Effects S Topcu and C Kirda 267

Methods D L Bjorneberg and R E Sojka 273

ISOTOPES IN SOIL AND PLANT INVESTIGATIONS K Reichardt and O O S Bacchi 280

ISOTROPY AND ANISOTROPY T-C J Yeh, P Wierenga, R Khaleel and R J Glass 285

J

JENNY, HANS R Amundson 293

K

KELLOGG, CHARLES J D Helms 301

KINETIC MODELS P M Jardine 307

KIRKHAM, DON D R Nielsen and R R van der Ploeg 315

L

LAMINAR AND TURBULENT FLOW see HYDRODYNAMICS IN SOILS

LANDFILLS see WASTE DISPOSAL ON LAND: Municipal

LAND-USE CLASSIFICATION J A LaGro Jr 321

LAWES, JOHN BENNET AND GILBERT, JOSEPH HENRY A E Johnston 328

LEACHING PROCESSES B E Clothier and S Green 336

LIEBIG, JUSTUS VON R R van der Ploeg, W Bo¨hm and M B Kirkham 343

LIMING E J Kamprath and T J Smyth 350

LIPMAN, JACOB G. J C F Tedrow 358

LOESS A J Busacca and M R Sweeney 364

LOWDERMILK, WALTER CLAY J D Helms 373

LYSIMETRY T A Howell 379

M

MACRONUTRIENTS C W Wood, J F Adams and B H Wood 387

MACROPORES AND MACROPORE FLOW, KINEMATIC WAVE APPROACH P F Germann 393

MAGNESIUM IN SOILS see CALCIUM AND MAGNESIUM IN SOILS

MANURE MANAGEMENT J T Sims and R O Maguire 402

MARBUT, CURTIS FLETCHER J P Tandarich 410

MATRIC POTENTIAL see HYDRODYNAMICS IN SOILS; WATER POTENTIAL; WATER RETENTION AND CHARACTERISTIC CURVE

MEDITERRANEAN SOILS J Torrent 418

METAL OXIDES A C Scheinost 428

METALS AND METALLOIDS, TRANSFORMATION BY MICROORGANISMS S M Glasauer, T J Beveridge, E P Burford, F A Harper and G M Gadd 438

METALS, HEAVY see HEAVY METALS

MICROBIAL PROCESSES

Environmental Factors P G Hartel 448

CommunityAnaly sis C H Nakatsu 455

Kinetics N S Panikov 463

MICRONUTRIENTS L M Shuman 479

MINERAL–ORGANIC–MICROBIAL INTERACTIONS P M Huang, M C Wang and M K Wang 486

MINERALS, PRIMARY P M Huang and M K Wang 500

MINERALS, SECONDARY see CLAY MINERALS

MINIMUM TILLAGE see CONSERVATION TILLAGE

MISCIBLE DISPLACEMENT see SOLUTE TRANSPORT

MORPHOLOGY P R Owens and E M Rutledge 511

MULCHES C L Acharya, K M Hati and K K Bandyopadhyay 521

MYCORRHIZAL FUNGI L M Egerton-Warburton, J I Querejeta, M F Allen and S L Finkelman 533

VOLUME 3

N

NEMATODES D A Neher and T O Powers 1

NEUTRON SCATTERING M J Fayer and G W Gee 6

NITROGEN IN SOILS

- Cycle M S Coyne and W W Frye 13
- Nitrates D S Powlson and T M Addiscott 21
- Nitrification J I Prosser 31
- Plant Uptake A Hodge 39
- Symbiotic Fixation J I Sprent 46

NITROGEN FERTILIZERS see FERTILIZERS AND FERTILIZATION

NUCLEAR WASTE DISPOSAL G W Gee, P D Meyer and A L Ward 56

NUTRIENT AVAILABILITY N K Fageria and V C Baligar 63

NUTRIENT MANAGEMENT G D Binford 71

O

ORGANIC FARMING C A Francis 77

ORGANIC MATTER

- Principles and Processes M Schnitzer 85
- Genesis and Formation K M Haider and G Guggenberger 93
- Interactions with Metals N Senesi and E Loffredo 101

ORGANIC RESIDUES, DECOMPOSITION A J Franzluebbers 112

ORGANIC SOILS D L Mokma 118

OVERLAND FLOW T S Steenhuis, L Agnew, P Ge´rard-Marchant and M T Walter 130

OXIDATION–REDUCTION OF CONTAMINANTS C J Matocha 133

P

PADDY SOILS C Witt and S M Haefele 141

PARENT MATERIAL see PEDOLOGY: Basic Principles; FACTORS OF SOIL FORMATION: Parent Material

PEDOLOGY

- Basic Principles M J Singer 151
- Dynamic F C Ugolini 156

PEDOMETRICS I O A Odeh and A B McBratney 166

PENMAN, HOWARD LATIMER J L Monteith 176

PENMAN–MONTEITH EQUATION R Allen 180

PERCOLATION see HYDRODYNAMICS IN SOILS

PERMAFROST see POLAR SOILS

PERMEABILITY see HYDRODYNAMICS IN SOILS

PERSISTENT ORGANIC POLLUTANTS (POPS) see POLLUTANTS: Persistent Organic (POPs)

PESTICIDES R H Bromilow 188

PETROLEUM see HYDROCARBONS

pH N Bolan and K Kandaswamy 196

PHOSPHORUS IN SOILS

Overview J T Sims and P A Vadas 202

Biological Interactions M D Mullen 210

PHYTOTOXIC SUBSTANCES IN SOILS M Qadir, S Schubert and D Steffens 216

PLANT–SOIL–WATER RELATIONS R A Feddes and J C van Dam 222

PLANT–WATER RELATIONS C Gimenez, M Gallardo and R B Thompson 231

POISEUILLE'S LAW see HYDRODYNAMICS IN SOILS

POLAR SOILS J C F Tedrow 239

POLLUTANTS

Biodegradation P B Hatzinger and J W Kelsey 250

Effects on Microorganisms M E Fuller 258

Persistent Organic (POPs) D Johnson 264

POLLUTION

Groundwater H Rubin 271

Industrial S P McGrath 282

POLYMERS AND MICROORGANISMS M C Rillig 287

POORLY CRYSTALLINE ALLUMINOSILICATES see AMORPHOUS MATERIALS

POROSITY AND PORE-SIZE DISTRIBUTION J R Nimmo 295

POTASSIUM IN SOILS P M Huang, J M Zhou, J C Xie and M K Wang 303

PRECIPITATION, WATERSHED ANALYSIS J V Bonta 314

PRECIPITATION–DISSOLUTION PROCESSES W P Robarge 322

PRECISION AGRICULTURE see SITE-SPECIFIC SOIL MANAGEMENT

PREFERENTIAL FLOW see UNSTABLE FLOW; MACROPORES AND MACROPORE FLOW, KINEMATIC WAVE APPROACH

PRODUCTIVITY D L Karlen 330

PROFILE see MORPHOLOGY

PROTOZOA W Foissner 336

Q

QUALITY OF SOIL B J Wienhold, G E Varvel and J W Doran 349

R

RADIATION BALANCE J L Hatfield, T J Sauer and J H Prueger 355

RADIONUCLIDES see ISOTOPES IN SOIL AND PLANT INVESTIGATIONS

RAINFED FARMING see DRYLAND FARMING

RANGE MANAGEMENT G L Anderson 360

RECYCLING OF ORGANIC WASTES see POLLUTANTS: Biodegradation

REDISTRIBUTION see WATER CYCLE

REDOX POTENTIAL R D DeLaune and K R Reddy 366

REDOX REACTIONS, KINETICS P S Nico and S Fendorf 372

REMEDIATION OF POLLUTED SOILS E Lombi and R E Hamon 379

REMOTE SENSING

Organic Matter D K Morris, C J Johannsen, S M Brouder and G C Steinhardt 385

Soil Moisture T J Jackson 392

RHIZOSPHERE A C Kennedy and L Z de Luna 399

RICHARDS, LORENZO A. W R Gardner 407

ROOT ARCHITECTURE AND GROWTH L E Jackson 411

ROOT EXUDATES AND MICROORGANISMS B-J Koo, D C Adriano, N S Bolan and C D Barton 421

S

SALINATION PROCESSES I Shainberg and G J Levy 429

SALINITY

Management D Hillel 435

Physical Effects D Russo 442

SALT BALANCE OF SOILS see SALINATION PROCESSES

SALT-AFFECTED SOILS, RECLAMATION R Keren 454

SAND DUNES H Tsoar 462

SATURATED AND UNSATURATED FLOW see HYDRODYNAMICS IN SOILS; VADOSE ZONE: Hydrologic Processes

SCALING

Physical Properties and Processes G Sposito 472

Transport Processes R P Ewing 477

SEPTIC SYSTEMS R L Lavigne 485

SHIFTING CULTIVATION R Lal 488

SITE-SPECIFIC SOIL MANAGEMENT C J Johannsen and P G Carter 497

SLASH AND BURN AGRICULTURE see SHIFTING CULTIVATION

SLUDGE see WASTE DISPOSAL ON LAND: Liquid; Municipal

SODIC SOILS G J Levy and I Shainberg 504

SOIL–PLANT–ATMOSPHERE CONTINUUM J M Norman and M C Anderson 513

SOLUTE TRANSPORT M C Sukop and E Perfect 521

SORPTION

Metals D L Sparks 532

Organic Chemicals B Xing and J J Pignatello 537

Oxyanions C P Schulthess, H Wijnja and W Yang 548

SORPTION–DESORPTION, KINETICS D L Sparks 556

SPATIAL PATTERNS J H Go¨rres and J A Amador 562

VOLUME 4

SPATIAL VARIATION, SOIL PROPERTIES R Webster 1

SPECIFIC SURFACE AREA K D Pennell 13

STATISTICS IN SOIL SCIENCE R Webster 19

STERILIZATION see DISINFESTATION

STOCHASTIC ANALYSIS OF SOIL PROCESSES D Russo 29

STRESS–STRAIN AND SOIL STRENGTH S K Upadhyaya 38

STRUCTURE V A Snyder and M A Va´zquez 54

SUBSOILING R L Raper 69

SULFUR IN SOILS

- Overview M A Tabatabai 76
- Biological Transformations S D Siciliano and J J Germida 85
- Nutrition M A Tabatabai 91

SURFACE COMPLEXATION MODELING S Goldberg 97

SUSTAINABLE SOIL AND LAND MANAGEMENT J L Berc 108

SWELLING AND SHRINKING D Smiles and P A C Raats 115

T

TEMPERATE REGION SOILS E A Nater 125

TEMPERATURE REGIME see THERMAL PROPERTIES AND PROCESSES

TENSIOMETRY T K Tokunaga 131

TERMITES see FAUNA

TERRA ROSSA see MEDITERRANEAN SOILS

TERRACES AND TERRACING G R Foster 135

TESTING OF SOILS A P Mallarino 143

TEXTURE G W Gee 149

THERMAL PROPERTIES AND PROCESSES D Hillel 156

THERMODYNAMICS OF SOIL WATER P H Groenevelt 163

TILLAGE see CONSERVATION TILLAGE; CULTIVATION AND TILLAGE; ZONE TILLAGE

TILTH D L Karlen 168

TIME-DOMAIN REFLECTOMETRY G C Topp and T P A Ferre´ 174

TROPICAL SOILS

Arid and Semiarid H C Monger, J J Martinez-Rios and S A Khresat 182

Humid Tropical S W Buol 187

U

UNSTABLE FLOW T S Steenhuis, J-Y Parlange, Y-J Kim, D A DiCarlo, J S Selker, P A Nektarios, D A Barry and F Stagnitti 197

URBAN SOILS J L Morel, C Schwartz, L Florentin and C de Kimpe 202

V

VADOSE ZONE

Hydrologic Processes J W Hopmans and M Th van Genuchten 209

Microbial Ecology P A Holden and N Fierer 216

VIRUSES see BACTERIOPHAGE

VOLCANIC SOILS G Uehara 225

W

WAKSMAN, SELMAN A. H B Woodruff 233

WASTE DISPOSAL ON LAND

Liquid C P Gerba 238

Municipal D A C Manning 247

WATER AVAILABILITY see PLANT–SOIL–WATER RELATIONS

WATER CONTENT AND POTENTIAL, MEASUREMENT G S Campbell and C S Campbell 253

WATER CYCLE D K Cassel and B B Thapa 258

WATER EROSION see EROSION: Water-Induced

WATER HARVESTING D Hillel 264

WATER MANAGEMENT see CROP WATER REQUIREMENTS

WATER POTENTIAL D Or, M Tuller and J M Wraith 270

WATER REQUIREMENTS see CROP WATER REQUIREMENTS

WATER RETENTION AND CHARACTERISTIC CURVE M Tuller and D Or 278

WATER TABLE see GROUNDWATER AND AQUIFERS

WATER, PROPERTIES D Hillel 290

WATER-REPELLENT SOILS J Letey 301

WATERSHED MANAGEMENT M D Tomer 306

WATER-USE EFFICIENCY M B Kirkham 315

WEED MANAGEMENT D D Buhler 323

WETLANDS, NATURALLY OCCURRING E K Hartig 328

WIDTSOE, JOHN A. AND GARDNER, WILLARD G S Campbell and W H Gardner 335

WIND EROSION see EROSION: Wind-Induced

WINDBREAKS AND SHELTERBELTS E S Takle 340

WOMEN IN SOIL SCIENCE (USA) M J Levin 345

WORLD SOIL MAP H Eswaran and P F Reich 352

Z

ZERO-CHARGE POINTS J Chorover 367

ZONE TILLAGE J L Hatfield and A T Jeffries 373

Table of Contents, Volume 3

N **1**

Nematodes **1**

Introduction - Life in a Soil Pore 1

Nematode Distribution and Abundance 3

Physiology 4

Ecology 4

Monitoring 5

Further Reading 5

Neutron Scattering **6**

Introduction 6

Theory and Instrumentation 6

Field Methodology 9

Calibration 10

Applications 11

Safety and Care 11

Future Use 12

Further Reading 12

Nitrogen In Soils **13**

Cycle 13

Introduction 13

Properties and Forms of Nitrogen in Soil 13

N-Cycling Pathways 13

Mineralization (Ammonification) 14

Assimilation (Immobilization) 14

Nitrification 14

Autotrophic nitrification 14

Heterotrophic nitrification 14

Nitrate Reduction 15

Denitrification 15

Dissimilatory nitrate reduction to ammonium 15

N2 Fixation 15

Fixed (Nonexchangeable) Ammonium 15

Regulation and Environmental Control 16

Cellular and Environmental Regulators of N Transformations 16

Chemical Regulators of N Transformations 17

Routes of N Loss From the Environment 17

Reductive Losses 17

Oxidative Losses 17

Volatilization 17

Interaction of N with Soil Organic Matter 18

Management Practices to Decrease N Losses 18

Nitrification inhibitors 18

Urease inhibitors 19

Environmental and Health Consequences of N Cycling 19

Global Warming and Ozone Destruction 19

Nutrient Losses 20

Groundwater Quality 20

Conclusion 20

List of Technical Nomenclature 20

Further Reading 20

Nitrates 21

Chemical Structure and Physical Properties 23

Biological Transformations in Soil 23

Forms of Nitrogen Fertilizer 24

Management of Nitrate in Agriculture 24
Loss of Nitrate from Fertilizer During the Crop-Growing Period 24
Losses of Nitrate Derived from Soil or Fertilizer Outside the Main Crop Growth Period 26
Improving the Management of Inorganic Fertilizer 27
Regulations Affecting N Use in Agriculture 29
Nitrate and Health 29
Stomach Cancer 29
Methemoglobinemia (Blue-Baby Syndrome) 30
Beneficial Effects of Nitrate on Health 30
Impacts of Nitrate on Ecology of Surface Waters 30
Conclusions 30
Further Reading 31
Nitrification 31
Introduction 31
Autotrophic Ammonia-Oxidizing Bacteria 32
Taxonomy 32
Biochemistry 32
Carbon Metabolism 32
Ammonia Oxidation under Anaerobic Conditions 33
Autotrophic Nitrite-Oxidizing Bacteria 33
Heterotrophic Nitrification 33
Analysis of Nitrifiers and Nitrification 34
Nitrification Activity and Rates 34
Enumeration 34
Enrichment and Isolation 35
Soil Communities and Diversity 35
Influence of Soil Characteristics on Ammonia-Oxidizer Communities 36
Soil pH 36
Nitrogen Fertilization 37
Inhibition of Soil Nitrification 38
Summary 38
Further Reading 38
Plant Uptake 39
Introduction 39
Plant-Microbial Competition for N 39
Inorganic N 40
Plant Inorganic N Uptake and Rhizosphere pH 41
Organic N 41
Root Responses 42
Mycorrhizal Contribution to Plant N Uptake 45
Further Reading 46
Symbiotic Fixation 46
Introduction 46
Cyanobacterial Symbioses 46
Lichens 46
Bryophytes 48
Ferns 48
Cycads 49
Angiosperms 49
Nodulated Angiosperms 50
Actinorhizal Plants 52
Legumes 53
The Importance of Nitrogen-Fixing Symbioses to Soil Fertility 55
Further Reading 55
Nuclear Waste Disposal 56

Introduction 56
Categories 56
High-Level Waste 56
Transuranic Waste 57
Low-Level Waste 57
Uranium Mill Tailings 57
Mixed Waste 57
Waste Inventories and Environmental Discharges 58
Radiation Exposures 58
Waste Storage and Disposition 58
Environmental Uncertainties 60
Summary 62
Further Reading 63
Nutrient Availability 63
Introduction 63
Essential Nutrients for Plant Growth and Their Functions 64
Diagnostic Techniques for Nutrient Availability 65
Visual Symptoms 66
Soil Testing 67
Plant Analysis 68
Crop Growth Response 70
Summary 70
Further Reading 71
Nutrient Management 71
Introduction 71
Nutrient-Management Planning 72
Step 1: Creating Maps 72
Step 2: Nutrients Required by Crops or Plants 72
Step 3: Inventory of Available Nutrients on the Operation 73
Step 4: Nutrient Balance 74
Step 5: Emergency Action Plan 75
Step 6: Record Keeping 75
Summary 75
Further Reading 75
O 77
Organic Farming 77
Introduction 77
Building Soil Fertility 78
Protecting Plants against Pests 80
Conversion to Organic Production and Certification 81
Biodynamic Farming, Biological Farming, and Integrated Systems 82
Marketing and Consumption of Organic Products 82
Future Perspectives for Organic Farming and Foods 83
Further Reading 84
Organic Matter 85
Principles and Processes 85
Introduction 85
Extraction and Fractionation of Humic Substances 85
Extraction of Humic Substances from Soils 85
Fractionation of Extracted Humic Substances 85
In Situ Analysis of SOM 85
Analytical Characteristics of HAs and FAs 86
Elemental Composition and Functional Group Content 86
Infrared and Fourier Transform Infrared Spectrophotometry 86
Chemical Structure of SOM 86
13C NMR Spectroscopy 86

Electron-Spin Resonance Spectroscopy 87
Oxidative Degradation 88
Reductive Degradation 88
Pyrolysis-Field Ionization Mass Spectrometry of HAs, FAs, and Humins..... 88
Curie-Point Pyrolysis-Gas Chromatography/Mass Spectrometry of HAs and FAs 88
Other Characteristics 89
Model Structures of HA 89
Two- and Three-Dimensional Model Structures of HA 89
Reactions of SOM with Metals and Organics..... 90
Formation of Water-Soluble Metal-Organic Complexes 90
Other Types of Metal SOM Interactions 91
Nitrogen in Soils and SOM 91
Phosphorus and Sulfur in Soils and SOM 92
Interactions of SOM with Pesticides and Herbicides 92
List of Technical Nomenclature..... 92
Further Reading 92
Genesis and Formation 93
Introduction 93
Soil Fabric and its Impact on SOM Stabilization 93
Transformation and Stabilization Processes of Free and Occluded Particulate Organic Matter 93
Stabilization of SOM by Sequestration in Microporous Structures or on Hydrophobic Surfaces 94
Chemical Information About Organic Matter in Aggregates and Organomineral Complexes 96
Stabilization of SOM by Humification 97
Biochemical Stabilization Mechanisms of Main Components from Plant Residues in SOM..... 98
Polysaccharide-related materials..... 98
Lignin-related materials 98
Impact and Function of the Soil Microflora in Relation to SOM Stability..... 99
Further Reading 100
Interactions with Metals 101
Introduction 101
Molecular and Mechanistic Aspects.....101
Metal Reactivity, SOM Binding Sites, and Complexation Stoichiometry 101
Experimental Evidence.....102
IR, UV-VIS, and fluorescence spectroscopies.....102
ESR, XAS, Mossbauer, and NMR spectroscopies 104
Quantitative Aspects 107
Metal-Binding Capacity of HS.....107
Stability Constants and Related Parameters of Metal-HS Complexes.....108
Modeling and Models.....110
Experimental Methods.....110
List of Technical Nomenclature.....111
Further Reading 112
Organic Residues, Decomposition.....112
Introduction.....112
Types of Organic Residues 112
Environmental Factors.....114
Organism Interactions.....116
Nutrient Cycling 116
Future Advances.....117
List of Technical Nomenclature 117
Further Reading.....118

Organic Soils **118**
Introduction 118
Formation 118
Parent Materials 119
Topography 119
Organisms 119
Climate 119
Time 119
Distribution 120
Utilization 121
Classification 128
Further Reading 129
Overland Flow **130**
Introduction 130
Saturation-Excess Overland Flow 130
Calculation Method 131
Water Quality Impacts of Variable Source Area Hydrology 132
Further Reading 132
Oxidation-Reduction of Contaminants **133**
Introduction 133
Redox Status of Reactive Fe and Mn Species 133
Kinetics and Mechanism 134
Inorganic Contaminants 135
Manganese as an Oxidant 135
Fe(II) as a Reductant 136
Organic Contaminants 137
Background 137
Oxidation and Reduction in Soil Organic Matter 137
Manganese and Iron as Oxidants 138
Adsorbed Fe(II) as a Reductant 139
Further Reading 139
P **141**
Paddy Soils **141**
Introduction 141
Cropping Environments in Asia's Lowlands 141
Distribution and Important Characteristics of Paddy Soils 142
Paddy Soil Fertility and Nutrient Management 144
Lowland Rice Systems and Declining Water Availability 147
Sustainability and Environmental Issues 147
Summary 149
List of Technical Nomenclature 149
Further Reading 150
Pedology **151**
Basic Principles 151
Introduction 151
History 151
Overview 152
Soil Classification 152
Soil Genesis 153
Soil Morphology 154
Soil Mapping 155
Summary 155
List of Technical Nomenclature 155
Acknowledgments 156
Further Reading 156
Dynamic 156

Early Stages of Soil Development on Recent Volcanic Ejecta.....157
Migration of Particles in the Process of Podzolization.....158
Evidence of Catastrophic Events and Horizon Inversion 159
Effects of Species Substitution.....160
Impact of Burial by Volcanic Ash on Preeruption Soils 161
Testing New Theories of Soil Formation 162
Summary.....164
List of Technical Nomenclature.....165
Further Reading 165
Pedometrics.....166
Introduction 166
Methods and Problems.....166
Pedometrics: Principles and Hypotheses.....166
Soil Measurements and Properties 167
Soil color and spectroradiometry 167
Field properties: hard and fuzzy descriptors?.....167
Pedotransfer functions.....168
Quantitative Soil Geography - Classification Models.....168
Soil Classification and Soil Map Units.....168
Fuzzy classes and the breakdown of the classification paradigm 170
Spatial Prediction of Soil Classes.....170
The corpt approach.....170
Spatial Prediction of Soil Attributes 171
The scorpan approach.....171
Geostatistics 171
The hybrid of scorpan and geostatistics 172
Soil Utility and Quality.....173
Land Evaluation, Soil Quality, and Soil Utility.....173
Recent Advances: Quantitative Pedogenesis.....173
Process Models.....173
Further Reading.....175
Penman, Howard Latimer 176
Further Reading.....179
Penman-Monteith Equation.....180
Derivation of the Penman-Monteith Equation 180
Other Parameters in the Penman-Monteith Equation.....183
Aerodynamic and Surface Parameters 183
Parameterization of Surface Resistance 184
Multilayer ET Models 185
Net Radiation.....185
Soil Heat Flux 186
Application of the Penman-Monteith Equation.....186
Further Reading.....187
Pesticides 188
Introduction 188
Physicochemical Properties of Pesticides 189
Persistence in Soil 190
Loss Processes and Factors of Influence 190
Methods for Studying Pesticide Breakdown.....192
Sorption of Pesticides by Soil 192
Movement of Pesticides in Soil.....193
Uptake of Pesticides from Soil by Plants 194
Effects on Soil Processes.....194
List of Technical Nomenclature 195
Further Reading.....195
pH 196

Introduction 196
Theoretical Basis of pH....196
Measurement of pH 196
Potentiometric Method 197
Determination of pH Using pH-Sensitive Dyes....198
Measurement of Soil pH 199
Factors Affecting Soil pH 200
Carbon Dioxide 200
Salt Effect....200
Suspension Effect 200
Importance of Soil pH 201
Summary 201
Further Reading....201
Phosphorus in Soils 202
Overview....202
Introduction 202
The Soil Phosphorus Cycle....202
Major Forms of Phosphorus in Soils....202
Overview of Soil Phosphorus Cycling and Transformations....203
Phosphorus Additions to Soils....203
Phosphorus Losses from Soils 204
Soil Phosphorus Management 204
Soil Testing for Phosphorus....204
Characteristics of Phosphorus Fertilizer Sources....204
Nutrient Management Practices for Phosphorus....205
Soil and Water Conservation Practices for Phosphorus....207
Plant Cover, Crop Rotation, and Strip Cropping....207
Companion Crops, Cover Crops, Green Manures, and Crop Residues....207
Conservation Tillage, Contour Cultivation, and Deep Plowing 208
Terraces, Vegetated Waterways, and Buffer Strips....208
Innovations in Soil Phosphorus Management....208
Animal Feed Formulations....208
Soil and Manure Testing....208
Soil and Manure Amendments 208
Water-Quality Modeling and the Phosphorus Index 209
Further Reading 209
Biological Interactions....210
Introduction 210
Forms of Soil Phosphorus....210
The Phosphorus Cycle....212
Mineralization and Immobilization of P....213
Microbial Solubilization of Inorganic P....213
P-Solubilizing Microorganisms and Plant Growth 214
Phosphorus and the Environment....215
Further Reading 215
Phytotoxic Substances in Soils....216
Introduction....216
Pesticides 216
Metals and Metalloids....217
Soil Acidity and Aluminum and Manganese Toxicity 218
Soil Flooding and Manganese and Iron Toxicity....219
Soil Salinity and Sodium Toxicity....220
Remediation Strategies for Phytotoxic Substances....220
Further Reading....222
Plant-Soil-Water Relations 222
Introduction....222

Liquid Water Flow Inside the Plant 223
Measurement of Root Water Uptake 225
Modeling Root Water Uptake in Ecologic, Hydrologic, and Atmospheric Communities 227
Ecologic Communities: Detailed Radial Root Model 227
Hydrologic Communities: General Root System Model 228
Atmospheric Communities: Large-Scale Root System Model 229
Further Reading 230
Plant-Water Relations 231
Introduction 231
Water in Plants 231
Plant Water Status Indicators 232
Direct Indicators of Plant Water Status 232
Relative water content 232
Water potential and components 232
Indirect Indicators of Plant Water Status 233
Stomatal conductance 233
Stem diameter variations 233
Sap flow 233
Leaf or canopy temperature 234
Other Methods 234
Water Deficits in Plants 234
Expansive Growth 235
Stomatal Responses 235
Root Signals 236
Applications of Plant-Water Relations to Irrigation Scheduling 236
Further Reading 238
Polar Soils 239
Introduction 239
Climate 239
Biotic Factors 239
Permafrost 239
Soil Development and Classification 240
Systems of Classification 240
Tundra Zone 241
Polar Desert Zone 242
Subpolar Desert Zone 244
Patterned Ground 244
Cold Desert Zone 245
Further Reading 249
Pollutants 250
Biodegradation 250
Introduction 250
Biodegradation 250
Aerobic Transformation Processes 251
Anaerobic Transformation Processes 252
Chemical Properties Affecting Biodegradation 253
Environmental Factors Affecting Biodegradation 254
Bioavailability 254
Sorption 254
Entrapment within the soil matrix 255
Nonaqueous-phase liquid 256
Nutrient Availability 256
pH 256
Temperature 256
Water Content 256

Presence of Appropriate Microorganisms....257
Predation....257
Presence of Other Compounds....257
Summary....257
Further Reading....258
Effects on Microorganisms....258
Introduction....258
Types of Effects....258
Microbial Community Composition....258
General....258
Specific....259
Microbial Community Function....260
General....260
Specific....260
Effects of Different Classes of Pollutants....262
Petroleum Products....262
Halogenated Compounds....262
Pesticides....262
Explosives....262
Metals....264
Further Reading....264
Persistent Organic (POPs)....264
Introduction....264
Persistent Organic Pollutants....265
Physicochemical Properties of Persistent Organic Pollutants in Soils....265
Persistence and Aging Behavior of Organic Chemicals in Soils....265
The 'Aging' of Persistent Organic Pollutants Within Soils....266
Mechanisms Contributing to the Persistence of Organic Chemicals in Soils....267
Sorption-Related Mechanisms....267
Sorption Retarded Intraparticle Diffusion....267
Intraorganic Matter Diffusion....269
The (Bio)availability of Organic Chemicals in Soils....269
Availability to Plants....270
Availability to Soil Microflora....270
Grazing Animals....270
Availability to Earthworms....270
Further Reading....271
Pollution....271
Groundwater....271
Introduction....271
Association of Pollutant Transport with the Water Cycle....272
Simplified Mass Budget and Flux Balances in Phreatic Aquifers....274
Penetration of Solutes into the Phreatic Aquifer....277
Pollutant Transport in the Aquifer....277
Contamination of Groundwater by NAPLS....278
Summary....280
List of Technical Nomenclature....281
Further Reading....281
Industrial....282
Introduction....282
Origins of Industrial Pollution....283
Types of Pollutant....283
Inorganic Pollutants....284
Organic Pollutants....285
Examples of Pollutants....285

Legislative Controls on Pollution 285
Source-Pathway-Receptor Principle 285
Persistent, Bioaccumulative, and Toxic Pollutants 286
Further Reading 287
Polymers and Microorganisms 287
Introduction 287
Extracellular Polymers: Definition and Overview 287
Hydrophobins: Secreted Fungal Proteins 289
Structure and Properties 289
Functions 289
Hydrophobins allow fungi to escape their aqueous environment 289
Glomalin: An Abundant Glycoproteinaceous Substance Produced by Arbuscular Mycorrhizal Fungi 291
Properties 291
Functions 291
Soil aggregation 291
Role in the biology of AMF and symbiosis 292
Exopolysaccharides 292
Structure 292
Functions 292
EPS mediate attachment to roots 292
EPS as symbiotic signals: Rhizobium 292
Desiccation resistance 292
Soil aggregation 293
Alginate: a model system for EPS biology 293
Extracellular Lipid Compounds 294
Further Reading 294
Porosity and Pore-Size Distribution 295
Porosity 295
Porosity in Natural Soils 295
Measurement of Porosity 296
Pores and Pore-Size Distribution 296
The Nature of a Pore 296
Measurement 298
Representation 299
Typical Features of a Pore-Size Distribution 300
Significance to Soil and Water Behavior 300
Significance of Porosity 300
Significance of Pore-Size Distribution 300
Pore Dynamics 301
Applications to Soil Transport Properties 301
Summary 302
Further Reading 302
Potassium in Soils 303
Introduction 303
Forms of Soil Potassium 303
Solution K 303
Exchangeable K 303
Fixed K 304
Structural K 304
Ion Selectivity and Fixation of Potassium in Soils 304
Ion Selectivity 304
Mineralogical properties 304
Chemical properties 306
Fixation of Potassium 306
Potassium Release from Soils 307

Primary Minerals 307
Micas 307
Potassium-bearing feldspars 308
Natural Soils 309
Availability of Soil Potassium 309
Quantity/Intensity Relationships 309
The Rhizosphere Chemistry and Labile Pool of Soil Potassium 311
Assessing Potassium Availability by Soil Tests 312
Plant Analysis for Potassium 312
Further Reading 313
Precipitation, Watershed Analysis 314
Introduction 314
Measurement of Precipitation 314
Precipitation Sensing and Resolution 314
Precipitation-Measurement Representativeness 315
Characterization of Precipitation 316
Dry Times Between Storms 316
Storm Depths and Durations 317
Within-Storm Intensities and Mass Curves 317
Spatial Variability 318
Watershed Response to Precipitation due to Physical Characteristics 319
Precipitation and Soil Variability 319
Watershed Processes 319
Watershed and Weather Modeling 320
Watershed-Precipitation Interactions 320
Watershed Models 320
Weather and Precipitation Modeling 320
Spatial Variability of Weather and Precipitation 321
Summary 321
Further Reading 321
Precipitation-Dissolution Processes 322
Introduction 322
Macroscale Processes 322
Ion Activity Products 323
Activity Diagrams 323
Microscale Processes 326
Precipitation 326
Dissolution 327
Rate law 328
Summary 328
Further Reading 329
Productivity 330
Introduction 330
Productivity Effects of the Inherent Soil-Forming Factors 330
Humankind's Impact on Productivity 331
Soil Productivity Comparisons 333
Productivity Ratings 334
Processes Reducing Soil Productivity and Practices for Restoring It 334
Summary 336
Further Reading 336
Protozoa 336
Introduction 336
A Brief History of Soil Protozoology 337
Diversity and Biology 337
Diversity and Geographic Distribution 337
Morphological and Physiological Adaptations 337

Morphological adaptations....337
Physiological adaptations....338
Food and feeding....338
r/K-selection....338
Cysts....339
Anabiosis....341
Number and Biomass....341
Vertical and Horizontal Distribution....342
The 'Water Paradox'....343
Ecological Significance....343
Standing Crop, Respiration, and Annual Production....343
Mineralization and Plant Growth....344
Earthworm Growth....344
Protozoa as Bioindicators in Terrestrial Environments....345
Soil Protozoa as Indicators in Natural Ecosystems....345
Protozoa as Indicators in Reclaimed Open-Cast Coal Mining Areas....346
Protozoa as Indicators of Soil Decontamination....346
Effects of Biocides....346
Further Reading....347
Q....349
Quality of Soil....349
Introduction....349
Management Systems....349
Soil Quality Assessment....350
Soil Function-Soil Indicator Relationship....351
Summary....352
List of Technical Nomenclature....353
Further Reading....353
R....355
Radiation Balance....355
Incoming Solar Radiation....355
Albedo....356
Incoming Longwave Radiation....358
Terrestrial Longwave Radiation....358
Radiation Balance....358
Implications....359
Further Reading....359
Range Management....360
Introduction....360
What are Rangelands?....360
How Much Area is Covered by Rangelands?....360
What Determines How Rangelands are Managed?....360
Biological Diversity....362
Introduction of Undesirable Plant Species....362
The Role of Disturbance in Rangeland and Soil Health....364
Livestock Grazing....364
No Domestic Grazing....364
Domestic Grazing....364
Overgrazing....365
Summary....365
Further Reading....365
Redox Potential....366
Introduction....366
Theoretical Relationships....366
Redox-pH Relationships....366
Normal Limits of pH and Redox Potential (Eh) in the Environment....367

Aerobic and/or Anaerobic Redox Potential Range in the Environment 368
Redox Couples in Wetlands 368
Predicting Mineral Stability 369
Intensity and Capacity of Reduction 369
Intensity 369
Capacity 370
Redox Condition and Metal/Metalloid Chemistry 370
Measurement of Redox Potential 371
Further Reading 371
Redox Reactions, Kinetics 372
Kinetic Rate Laws 372
Integrated Rate Laws 373
Initial Rates 374
Pseudo nth-Order Reactions 375
Second-Order, Two-Reactant Rate Laws 375
Reaction Diagrams and Activation Energies 375
Reaction Mechanism (Rate-Limiting Step and Preequilibria) 376
Biologically Controlled Redox Reactions 377
Conclusion 378
Further Reading 378
Remediation of Polluted Soils 379
Introduction 379
Containment Versus Clean-Up Strategies 379
In Situ Versus Ex Situ Strategies 379
Physical Treatment Technologies 380
Excavation and Off-site Disposal (Ex Situ, Requires Ongoing Monitoring) 380
Dilution (In Situ or Ex Situ) 380
Capping and Encapsulation (In Situ, Requires Ongoing Monitoring) 380
Separation-Fractionation (Ex Situ) 380
Electrokinetic Separation (In Situ) 380
Soil-Flushing (In Situ) 381
Vitrification (In Situ or Ex Situ, Requires Ongoing Monitoring) 381
Solidification (In Situ or Ex Situ, Requires Ongoing Monitoring) 381
Incineration (Ex Situ) 381
Soil Vapor Extraction (In situ or Ex situ) 381
Photo-oxidation (Ex Situ) 381
Chemical Treatment Technologies 381
Monitored Natural Attenuation (In Situ, Requires Ongoing Monitoring) .381
Fixation-Stabilization (In Situ or Ex Situ, Requires Ongoing Monitoring) 382
Chemical Oxidation-Reduction (Ex Situ, may Require Monitoring (Inorganics)) 382
Aqueous and Solvent Extraction (Ex Situ) 382
Dehalogenation (Ex Situ) 382
Biological Treatment Technologies 383
Biodegradation Techniques (In Situ) 383
Composting and Biopiles, and Slurry-Phase Biological Treatment (Ex Situ) 383
Phytoremediation (In Situ) 383
Selection of Remediation Technologies and Future Directions 384
Further Reading 385
Remote Sensing 385
Organic Matter 385
Introduction 385
Mapping of Soils 386

Remote Sensing for Soil Mapping....387
Organic Matter Detection....388
The Future....391
Further Reading....392
Soil Moisture....392
Microwave Remote Sensing of Surface Soil Moisture....392
Reflectivity and Soil Moisture....393
Passive Microwave Techniques....394
Active Microwave Methods....394
Aircraft Platforms and Application Example....395
Passive Microwave Satellite Observing Systems and Products....396
Special Satellite Microwave/Imager....396
Tropical Rainfall Measurement Mission Microwave Imager....397
Advanced Microwave Scanning Radiometer....397
Future Passive Systems....398
Active Microwave Satellite Observing Systems....398
Current Active Systems....398
Future Active Systems....398
List of Technical Nomenclature....398
Further Reading....398
Rhizosphere....399
Introduction....399
Organisms in the Rhizosphere....400
Bacteria....400
Fungi....401
Fauna....401
Microfauna (protozoa)....401
Mesofauna (nematodes and microarthropods)....402
Interactions in the Rhizosphere....402
Dinitrogen-Fixation Symbiosis....402
Mycorrhizal Associations....402
Plant Growth-Promoting Organisms....403
Plant Growth-Inhibiting Organisms....403
Bioremediation....403
Biological Control....405
Summary....406
Further Reading....406
Richards, Lorenzo A....407
Further Reading....410
Root Architecture and Growth....411
Introduction....411
Description of Roots and Root Architecture....412
Tradeoffs in Root Form....413
Root Traits for Stress Tolerance....414
Genetic Analysis of Root Traits....418
Root Architectural Effects on Plant Communities and Ecosystem Processes....418
Further Reading....420
Root Exudates and Microorganisms....421
Introduction....421
Rhizosphere Microzone....421
Microorganisms and Structure of Microbial Communities in the Rhizosphere....422
Composition of Root Exudates....423
Collection and Quantitation of Root Exudates....424
Factors Affecting Root Exudates....424

Role of Root Exudates on Plant Nutrient and Metal Dynamics 425
Root Exudates and Biodegradation of Organic Contaminants 426
Summary 428
Acknowledgments.....428
Further Reading.....428
S 429
Salination Processes 429
Origin and Distribution of Saline Soils 429
Properties of Saline and Sodic Soils.....429
Total Salt Concentration.....430
Sodicity and Sodium Adsorption Ratio 430
Anionic Composition.....431
Concentration of Elements Toxic to Plants 431
Response of Crops to Salinity 432
Response of Soils to Salinity and Sodicity.....432
Reclamation of Salt-Affected Soils 433
Reclamation of Sodic Soils 433
Summary 434
Further Reading.....434
Salinity 435
Management.....435
Natural and Induced Soil Salinity 435
Irrigation Water Quality.....436
Groundwater Drainage 438
Leaching Processes.....439
Soil Amendments 440
Early-Warning Systems.....441
Further Reading 442
Physical Effects 442
Introduction 442
Water Retention and Swelling in Salt-Affected Soils.....442
Effects of Salts on Soil Hydraulic Conductivity.....444
Modeling Hydraulic Conductivity and Water Retentivity in Salt-Affected Soils 445
Calculations of the b(h, R, C0) and N(h, R, C0) Functions 445
Calculations of the theta(h, R, C0) and K(h, R, C0) Functions 446
Implications for Flow and Transport 450
Summary.....451
List of Technical Nomenclature.....452
Further Reading 453
Salt-Affected Soils, Reclamation 454
Introduction 454
Effect of Salts on Soil Properties 454
Hydraulic Conductivity.....454
Infiltration Rate 454
Reclamation of Salt- and Sodium-Affected Soils 454
Reclamation of Salt-Affected Soils 454
Leaching Methods 455
Continuous and intermittent ponding 455
Sprinkling 455
Postleaching operation 456
Reclamation of Sodium-Affected Soils 456
Reclamation by Adding Soil Amendments 456
Gypsum 456
Acids and sulfur 460
Reclamation without Adding Soil Amendments.....460

Calcium carbonate....460
Successive dilutions of highly saline water....461
Further Reading....461
Sand Dunes....462
Introduction....462
Dune Sand Texture....462
The Effect of Dune Sand Texture on Water Infiltration....463
The Effect of Sand Texture on the Thermal Properties of Sand Dunes....465
The Advantage and Disadvantage of Sand Texture on Vegetation Growth in Deserts....466
Mobility and Stability of Sand Dunes....466
Dune Shape and Dynamics....467
Further Reading....471
Scaling....472
Physical Properties and Processes....472
Introduction....472
Scale Invariance....472
Scaling Relationships....473
Example: Hamra Soil Field Experiment....474
Physical Significance of Scaling....475
Further Reading....476
Transport Processes....477
Introduction....477
Diffusive-Dispersive Transport....477
Sir Geoffrey Taylor and the Spreading of Solute in a Tube....478
Solute Spreading in Soils....480
Scale....481
Conceptual Models....481
Observations and Mechanisms....483
Model Selection....484
Further Reading....484
Septic Systems....485
Introduction....485
Technical and Historical Overview....485
Outhouse - Structure and Function....485
Cesspool - Structure and Function....486
The Modern Septic System....486
Septic Tank....486
Soil Absorption Systems....486
SAS Size and Soil Infiltration Rates....486
Advanced Treatment Systems....487
Artificial Wetlands....487
Summary....487
Further Reading....487
Shifting Cultivation....488
Introduction....488
Shifting Cultivation....489
Productivity of Complex Farming Systems....490
Ecosystem-Related Constraints to Agricultural Intensification....491
Sustainable Management of Soils Under Shifting Cultivation....491
Characteristics of an Improved/Sustainable Soil Management System....492
Examples of Sustainable Soil Management Systems....493
Ecological Approach to Management of Soils under Shifting Cultivation....494
Research and Development Priorities....495
Conclusions....496
Further Reading....497

Site-Specific Soil Management 497
Introduction 497
Technologies 498
Soil Management 499
Proper Seeding 500
Harvesting 500
Adoption Considerations 501
The Future 501
Further Reading 503
Sodic Soils 504
Introduction 504
Definition of Sodic Soils 504
Origin and Distribution of Sodic Soils 504
Mechanisms Responsible for Sodic Behavior 505
Structural Stability of Sodic Soils 505
Hydraulic Conductivity 505
Sodicity and electrolyte concentration 506
Soil properties affecting sodic behavior 507
Sodicity and conditions prevailing in the soil 507
Infiltration Rate 507
Sodicity and Soil Productivity 509
Reclamation of Sodic Soils 510
Chemical Reclamation 510
Chemical amendments 510
Successive dilutions of high-salt water 511
Management of Sodic Soils 511
Soil Amendments 511
Tillage 512
Sodicity-Related Environmental Hazards 512
Conclusions 512
Further Reading 512
Soil-Plant-Atmosphere Continuum 513
Introduction 513
Fundamental Principles 514
Water in the Soil-Plant-Atmosphere-Continuum 515
Small-Scale Observations and Model Building 516
Application to Regional and Global Scales 519
Institutional Issues 520
Further Reading 521
Solute Transport 521
Introduction 521
Transport Processes 522
Advection 522
Diffusion 522
Dispersion 522
Dimensionality 523
Longitudinal Dispersivity 524
Breakthrough Curves 524
Adsorption and Exclusion 525
Advection-Dispersion Equation 527
Two-Region Advection-Dispersion Equation 528
Fractional Advection Dispersion and Continuous Time Random Walk Models 529
List of Technical Nomenclature 530
Further Reading 531
Sorption 532

Metals 532
Introduction 532
Terminology and Role of Sorption Processes in Soils 532
Surface Functional Groups....532
Surface Complexation....532
Sorption of Metal Cations....534
Metal Surface Precipitation 534
Further Reading 537
Organic Chemicals 537
Introduction 537
Physisorption of Nonionic Compounds 538
General Considerations....538
Forces involved....538
Categories 538
Thermodynamic considerations....538
Physisorption of Nonionic Compounds to Mineral Solids 539
Physisorption of Nonionic Compounds to Natural Organic Matter 540
General considerations....540
Natural organic materials in soils....541
Physisorption to humic substances and kerogen 541
Physisorption to black carbon....543
Kinetics of Physisorption....544
Physisorption Hysteresis 545
Ion-Exchange Sorption....546
Chemisorption 547
Further Reading 547
Oxyanions....548
Introduction 548
Techniques for Detection and Quantification of Oxyanions in Solution....548
Techniques for Investigation of Oxyanions Adsorbed at Mineral-Water Interfaces 549
Vibrational Spectroscopies 550
X-Ray Absorption Spectroscopy....550
Sorption Strength of Oxyanions 551
Impact of Ions on Adsorption of Oxyanions....553
Sorption Mechanisms of Oxyanions....554
Site Preference and Site Reactivity....555
Further Reading 555
Sorption-Desorption, Kinetics 556
Introduction....556
Rate-Limiting Steps and Time Scales for Sorption Processes 556
Sorption/Desorption Kinetics of Metals....557
Kinetics of Metal Hydroxide Surface Precipitation/Dissolution 558
Kinetics of Organic Contaminant Sorption/Desorption 560
Conclusions and Future Directions....561
Further Reading....561
Spatial Patterns 562
Introduction....562
Background....562
Sources of Variability....562
Pedological Factors of Soil Biological Variability....562
Spatial Scales Related to Organisms....563
Quantification of Spatial Properties 564
First-Order Statistics and Spatial Independence....564
Second-Order Statistics and Spatial Dependence....565
Sampling to Characterize Spatial Variability....566

Models of Spatial Variability....568
Linear Statistical Models....568
Physical Mechanistic Process Models....568
Biological Mechanistic Process Models....569
Applications of Spatial Analysis....569
Summary....570
Further Reading....570

N

NEMATODES

D A Neher, University of Toledo, Toledo, OH, USA
T O Powers, University of Nebraska, Lincoln, NE, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction – Life in a Soil Pore

Few people are aware of the most numerous of all soil-dwelling animals, the nematodes. These minute, unsegmented roundworms are actually aquatic organisms, inhabiting water films (1–5 μm thick) that coat and surround soil particles. (This chapter focuses on nematodes that typically complete their entire lifecycle within terrestrial soils or within plant roots. Numerous parasites of vertebrates have soil-dwelling stages, but their development is dependent on entering an appropriate host species. Vertebrate nematode parasites include hookworms, intestinal roundworms, pinworms, whipworms, and other serious pests of humans and domesticated animals. Information about vertebrate parasites is readily found in textbooks of parasitology.) Their size, typically 0.40–1.0 mm in length, is ideal for navigating this porous matrix. Although fully functional within a water film, they are ultimately bounded by soil structure, restricted in movement within soil pores less than 30 μm diameter.

Nematodes are simple organisms (Figure 1). Comprised of approximately 1000 somatic cells in the adult stage, these worm-like organisms are an example of functional and anatomical economy. Lacking eyes, appendages, and true segmentation, nematodes use mechanosensory and chemosensory neurons embedded in the cuticle to orient and respond to a wide range of environmental stimuli. At the nematode's anterior end is a circlet of sensilla arranged around an oral opening. These sensilla, including two larger, laterally placed chemosensory organs called 'amphids,' detect subtle chemical gradients, providing directional information to the nematode (Figure 2). Plant-feeding nematodes respond to slight variations in CO_2 and root exudates. A positive chemosensory signal initiates a snake-like sinuous movement characteristic of most nematodes, created by longitudinal body muscles working in apposition to a hydrostatic skeleton. Directional movement more than a few centimeters per day is unlikely in the soil environment. Once they contact the root surface, plant-feeding nematodes will probe with their stylet, a hollow, protrusible, hypodermic needle-like feeding tube (Figure 3a). Some plant-feeding nematodes will feed externally to the root surface, using this stylet to puncture cells, withdrawing the cytoplasmic contents. Others will penetrate the root and establish permanent feeding sites within the root cortex, or migrate cell-to-cell, leaving a trail of damaged necrotic tissue.

Free-living nematodes also occupy the interstitial spaces in soil. Unlike the plant-feeders, free-living nematodes are seldom sedentary, continually moving to feed on a diverse array of food, including bacteria, algae, fungi, protozoa, small invertebrates, and other nematodes. They use the same set of sensilla to track their food source; however, their feeding structures are modified to suit their meal. Bacterial-feeders graze using a relatively simple tubular mouthpart,

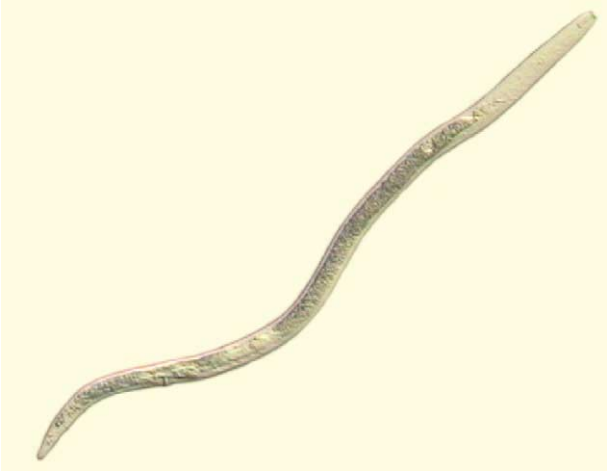


Figure 1 Entire body view of a female *Pratylenchus agilis* (×100 magnification), collected on the Konza Prairie (96° W 35′ 39° N 05′) beneath Scribner's panicum (*Panicum scribnerianum*) and bluegrass (*Poa pratensis*) near Manhattan, Kansas. Photograph is provided courtesy of Peter Mullin (2000).

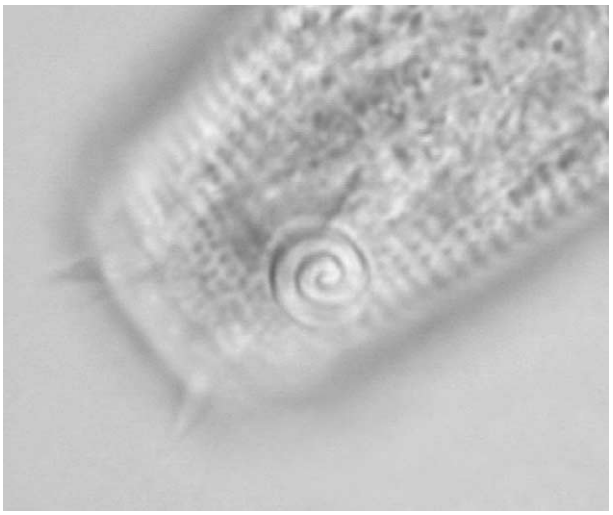


Figure 2 Spiral-shaped chemosensory organs called 'amphids' in an anterior position of *Achromadora* sp. collected from soil of Jumbo Valley fen in Cherry County, Nebraska.

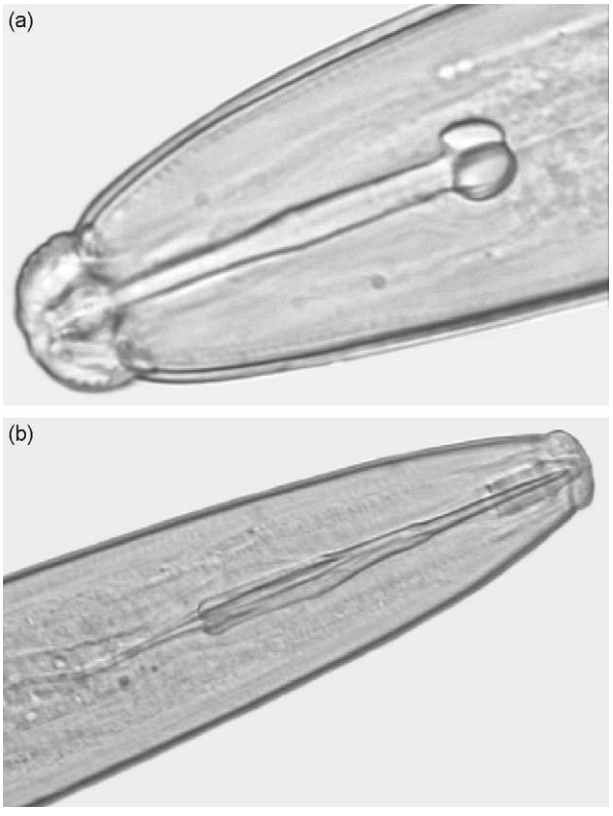


Figure 3 Variation in morphology of spear-like structure in oral opening: (a) male plant-parasite *Hoplolaimus galeatus* (×1000 magnification), collected from soil with big bluestem (*Andropogon gerardii* Vitman) in the Konza Prairie (96° W 35′ 39° N 05′), near Manhattan, Kansas; and (b) female fungivore *Enchodelus hopedorus* (× 400 magnification), collected from the summit of Long's Peak, Colorado (105° W 35′ 40° N16′). Photographs are provided courtesy of Peter Mullin (2000).

although the cuticle surrounding the oral opening may be modified elaborately to direct food toward the stoma (Figure 4). Predators may be adorned with 'teeth,' used to puncture or shred the integument of various invertebrates (Figure 5). Fungal-feeders and omnivores have a stylet similar in appearance to that of plant-feeders, distinguished by the lack of stylet 'knobs,' points of muscle attachment for stylet protrusion (Figure 3b).

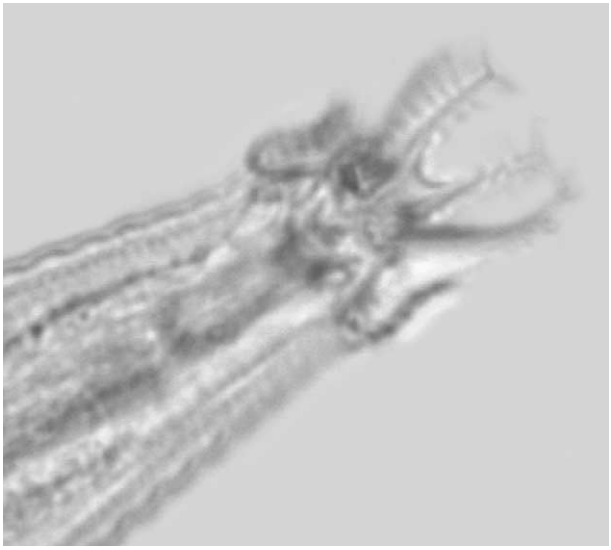


Figure 4 Cuticle ornamentation of oral opening of *Acrobeles ctenocephalus* (×1000 magnification), collected in soil with little bluestem (*Andropogon scoparius*) in the Konza Prairie (96° W 35′ 39° N 05′) near Manhattan, Kansas. Photograph is provided courtesy of Peter Mullin (2000).

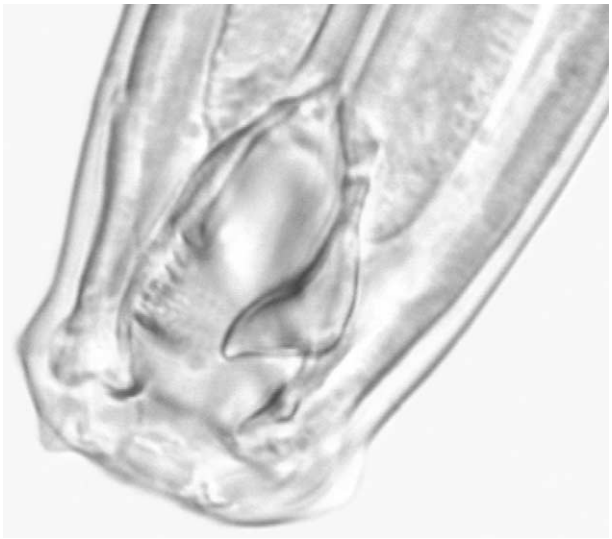


Figure 5 Teeth of oral opening of predator *Mylonchulus montanus* (×1000 magnification), collected in soil with big bluestem in the Konza Prairie (96° W 35′ 39° N 05′) near Manhattan, Kansas. Photograph is provided courtesy of Peter Mullin (2000).

Figure 6 Ovary in reproductive tract of female *Axonchium micans* (× 400 magnification), collected in soil with big bluestem/Scribner's panicum in the Konza Prairie, near Manhattan, Kansas. Photograph is provided courtesy of Peter Mullin (2000).

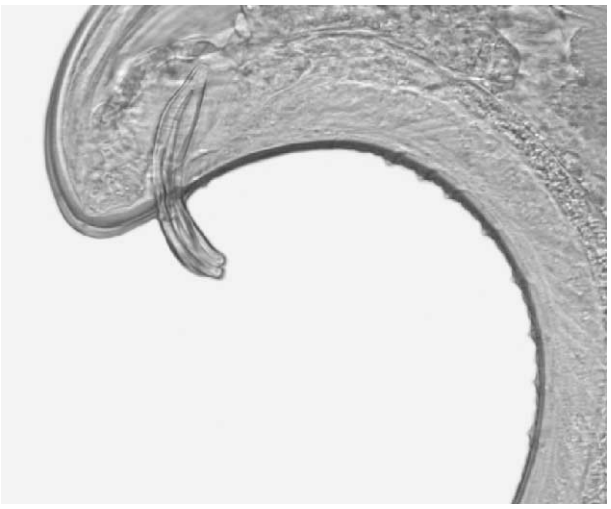


Figure 7 Protrusible modification of male genitalia of *Aporcelaimellus obscuroides* (×400 magnification), collected in soil with big bluestem/Scribner's panicum in the Konza Prairie, near Manhattan, Kansas. Photograph is provided courtesy of Peter Mullin (2000).

Early observations of nematodes by light microscopy led scientists to refer to nematodes as a 'tube within a tube.' Food entering the nematode is channeled quickly into a tubular esophagus and then passed into an intestine that comprises the bulk of the body cavity. Waste products are eliminated through a posterior ventral anus. There is no circulatory or respiratory system within the nematode. In addition to the alimentary system, the reproductive system is a conspicuous feature in the adult nematode (Figure 6). Females and hermaphrodites are characterized by a branched genital system, which produces eggs that exit the body via a ventral genital opening in the body wall. Males possess a spicule, a cuticularized, protrusible modification of the genital system that guides sperm into the female for internal fertilization (Figure 7). Many nematodes are strictly bisexual, but parthenogenesis and self-fertilizing hermaphroditism are also common among species. Regardless of mode of reproduction, nematodes are prolific animals. Some species complete a life-cycle of 3 or 4 days. Others require several months to go from egg to egg-laying female. *Meloidogyne* spp., an economically important group of plant parasites, will typically begin to produce eggs within 3 weeks of hatching; each female is capable of producing hundreds of eggs during an adult life.

Nematode Distribution and Abundance

The hallmark of Phylum Nemata is the exceptional abundance and ubiquitous presence of nematodes (Table 1). It is a challenging exercise to identify a habitat not occupied by a diverse community of nematodes. Consider the dry valleys of Antarctica, an environment considered so extreme that it is thought to provide the bare minimum for sustaining life on earth. There, in McMurdo Valley, three nematode species occupy the top levels in the soil food chain, two bacterivores and a third species, *Eudorylaimus antarcticus*, that feeds on the two bacterivorous species. Soils of the high Arctic are populated so heavily with nematodes that biologists have searched for explanations for an Arctic nematode biodiversity that exceeds that of the humid tropics. Temperate grasslands appear to sustain the greatest levels of diversity and abundance in the phylum. On Konza Prairie, one of the last remaining large stands of true tallgrass prairie in North America, each handful of soil is likely to contain 50–100 different nematode species. An estimated 1.5 million nematodes can be extracted from a square meter of soil, just within the surface 10 cm. Some researchers have considered the plant-feeding or herbivorous species as 'belowground buffalo,' referring to their impact as grazers of plant roots. Nearly all researchers agree that the approximately 25 000 described species of nematodes represent a small percentage of the species that actually exist in nature.

Finding agreement on those factors responsible for extraordinary nematode abundance and diversity is not so simple. Clearly, the evolutionary age of nematodes has provided considerable time for nematode diversification. Recent molecular studies indicate that the roots of the nematode evolutionary tree are hundreds of millions of years older than the corresponding tree for insects. Marine nematodes, which inhabit even the deepest ocean floors, have a nucleotide diversity that suggests nematodes flourished before life

Table 1 Nematode communities in selected studies

Ecosystem type	Abundance ($10^6\ m^{-2}$)	Genera (n)	Species (n)	Bacterivore (%)	Fungivore (%)	Herbivore (%)	Predator/ omnivore (%)
Grassland	2.3–20	30–124	71–384	24–38	5–20	16–41	26–40
Deciduous forest	2.3–3.7	22–81	34–175	27–46	15–25	13–31	5–32
Tropical rainforest	NA	107	204	35	5	29	30
Tundra	–	58	162	55	3	22	19
Heath	1.2	–		57	–	2	–
Agroecosystems	3.5–5.0	19–44	33–100	36–46	13–19	33–41	5–7
Temperate desert	–	18	23	33	22	17	28
Antarctic coastal	–	15	27	47	27	0	27
Antarctic dry valley	–	3	3	67	0	0	33

NA, not available.

colonized land. The same nematode survival mechanisms that allowed them to be among the first colonizers of land most likely contribute to a competitive advantage in current-day harsh environments.

Physiology

Although nematodes depend on free water for normal functioning, under harsh conditions such as freezing or drying many nematodes are capable of entering a cryptobiotic state, essentially a reversible state of suspended animation until favorable conditions return. Anhydrobiotic success, the ability to withstand the lack of moisture, depends on a slow rate of dehydration, rapid decline in cuticle permeability, and accumulation of compounds such as trehalose and glycerol to stabilize phospholipids and proteins. Species including *Acrobeloides* spp., *Aphelenchus avenae*, and *Scutellonema brachyurum* start to enter anhydrobiosis by coiling at gravimetric water contents of 3.7% (−300 kPa), 9% (−30 kPa), and 15% (−10 kPa), respectively. It is unknown how long nematodes may persist in this state and still survive hydration. Specimens of the wheat gall nematode *Anguina tritici* have been revived after more than 30 years in anhydrobiosis.

Nematodes are not alone among invertebrates in their ability to enter a cryptobiotic state. Therefore, additional clues to their successful colonization of soil are found in their physiological versatility. Physiological adaptations have allowed nematodes to invade habitats in which few other animals can survive and avoid interspecific competition and many environmental selection pressures. Nematodes are capable of regulating their uptake of oxygen over a wide range of partial pressures from 100 to 5%. When oxygen falls below 5%, nematodes convert their use of energy reserves. For example, *Aphelenchus avenae* uses reserves of neutral lipids under aerobic conditions and glycogen under anaerobic conditions. Not only does a hydrostatic skeleton facilitate movement, but also an ability to expand or contract body size in response to gradual changes in concentration of solutions containing sodium, calcium, magnesium, and/or potassium ions. The adjustment process, osmoregulation, is possible by modifying the permeability of their cuticle. Furthermore, nematodes can tolerate a wide pH range, with some species capable of withstanding strong acids or bases (pH 1.6–11.0). Nematode tolerance for temperature is equally remarkable. In hot springs, nematodes live at temperatures as high as or higher than any other Metazoa. *Aphelenchoides parietinus*, a cosmopolitan species, has been reported from 58°C springs in Chile and 61°C springs in New Zealand.

Ecology

Although nematodes have adapted mechanisms to survive extremities of climate, their activity is stimulated by the return of more moderate conditions. For example, communities of nematodes are revived after rain in desert soils or after a relatively warm period in soils of polar regions. Species are found interacting with other organisms in a variety of occupations including competitor, opportunist, parasite, host, predator, or prey. As a competitor, nematodes rival protozoa for access to bacterial food sources. However, differences in size and generation time can alleviate competition by specializing on microhabitats fine-tuned to minute spatial and temporal scales. As opportunists, nematode species are phoretically transported by insects, adhering to their bodies, providing an opportunity to reach food sources at a longer distance than they would be capable of reaching alone. For example, Psychodidae flies carry nematodes to fermenting organic matter; dung beetles carry nematodes to fresh dung; and Scolytidae bark beetles transport Tylenchida and Rhabditida to their tunnels, where they feed on bacteria and fungi. Nematodes that are parasites of plants include both specialists and generalists.

Host ranges may extend to hundreds of plant species or be restricted to a single plant variety. Some nematodes specialize in precise feeding sites along the root. Plant-parasitic nematodes affect primary productivity of plants by altering uptake of water and nutrients. These abnormalities may result from changes in root morphology and/or physiology resulting in reduced productivity. Other nematodes infect insects, killing them within 48 h by releasing insect-pathogenic bacteria into the insect. The insect host dies from the bacterial infection, and nematodes feed on the bacteria and develop and reproduce inside the insect cadaver. In this case, the nematode and bacteria have a mutualistic association. Sometimes, a nematode finds itself in a role reversal, becoming the host or target of fungi specialized for trapping nematodes, using constricting rings, sticky knobs, or hyphal nets. Being in the middle of a food chain, nematodes eventually become prey to higher-order consumers. Nematodes provide a portion of the diet for many kinds of small soil arthropods. For example, a symphylan can hold seven large nematodes in its gut at one time. In 1 day, a mite may consume two large nematodes or several smaller nematodes.

Genetic diversity, nematode abundance, and the variety of niches occupied by nematodes has led to the assessment of nematode communities as a means of making predictions about past or present soil conditions. The species composition and relative nematode abundance of species comprising a nematode community allow nematodes to serve as biological indicators for soil disturbance as well as 'soil health.'

Monitoring

Nematodes contribute directly to biogeochemistry of soils by regulating processes such as decomposition and nutrient cycling. Nematodes do not feed on decaying organic matter directly, but on bacteria, fungi, algae, and actinomycetes that colonize and decompose decaying plant and animal debris. Indirectly, nematodes control availability of nutrients by regulating the abundance or activity of these organisms, releasing nitrogen and phosphorus from microbes they digest, immobilizing nutrients in their live tissues, and excreting excess nitrogen as ammonium. Under field conditions, bacterial-feeding and predatory nematodes contribute 8–19% of nitrogen mineralization in conventional and integrated farming systems, respectively.

Nematode community structure and function are known to change in response to land-management practices such as nutrient enrichment through fertilization by organic or inorganic nitrogen, cultivation, liming, drainage, plant community composition and age, and toxic substances such as heavy metals, pesticides, and polycyclic aromatic hydrocarbons. Disturbance can affect the survival of individuals directly, or indirectly by changing resource levels. For example, a larger ratio of fungal-feeding to bacterial-feeding nematodes suggests less effect by cultivation than a smaller ratio. Any disturbance that results in compaction reduces soil porosity, and the numbers of relatively large-sized nematodes. General biocides such as methyl bromide can nearly eliminate nematodes, returning the ecological succession of soil communities to that assembling a depauperate soil matrix. Although recovery occurs eventually, abundance and diversity of nematode communities may take years to achieve prefumigation levels. Alternatively, some herbicides affect nematodes indirectly by reducing vegetation and smaller additions of organic matter to soil.

The composition of nematode communities may reflect the frequency of disturbance to soil, whether the disturbance is primarily physical and/or chemical in origin. This concept is based on the principle that different species have contrasting levels of sensitivity or tolerance to stress because of unique survival and/or reproductive traits. Smaller values of an index that integrates relative abundance of species within each sensitivity/tolerance class are indicative of a more disturbed environment, and larger values may indicate a less-disturbed environment. Ideally, indicators of soil health would correspond with ecological processes occurring in soil. Because initial experiments have been correlative in nature, it is too soon to claim such an association with great confidence.

The concept of nematodes as indicators of soil condition and their predictive use in soil management represents a tremendous shift in emphasis in the science of nematology. Formerly a science focused on the control of parasitic and harmful species, now nematodes are seen as an integral and potentially useful component of soil systems. In light of this new perspective, it is likely that the study of nematodes will contribute significantly to our understanding of biological processes in the soil.

Further Reading

Fitter AH (1985) *Ecological Interactions in Soil*. Oxford, UK: Blackwell Press.

Freckman DW (1982) *Nematodes in Soil Ecosystems*. Austin, TX: University of Texas Press.

Freckman DW (1988) Bacterivorous nematodes and organic-matter decomposition. *Agriculture, Ecosystems and Environment* 24: 195–217.

Freckman DW and Baldwin JG (1990) Nematoda. In: Dindal DL (ed.) *Soil Biology Guide*, pp. 155–200. New York: John Wiley.

Ingham RE, Trofymow JA, Ingham ER, and Coleman DC (1985) Interaction of bacteria, fungi, and their nematode grazers: effect on nutrient cycling and plant growth. *Ecological Monographs* 55: 119–140.

Maggenti A (1981) *General Nematology*. New York: Springer-Verlag.

Nicholas WL (1984) *The Biology of Free-living Nematodes*. Oxford, UK: Clarendon Press.

Perry RN and Wright DJ (1998) *The Physiology and Biochemistry of Free-living and Plant-parasitic Nematodes*. New York: CABI Publishing.

Yeates GW (1999) Effects of plants on nematode community structure. *Annual Review of Phytopathology* 37: 127–149.

Zunke U and Perry RN (1997) Nematodes: harmful and beneficial organisms. In: Benckiser G (ed.) *Fauna in Soil Ecosystems: Recycling Processes, Nutrient Fluxes, and Agricultural Production*, pp. 85–133. New York: Marcel Dekker.

NEUTRON SCATTERING

M J Fayer and G W Gee, Pacific Northwest National Laboratory, Richland, WA, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Soil water makes land-based life possible by satisfying plant water requirements (e.g., for crops and forests), by serving as a medium for nutrient movement to plant roots and nutrient cycling (e.g., fungal and microbial decay, mineralization, fertilization), and by controlling the fate and transport of contaminants in the soil environment. Therefore, understanding the dynamics of plant growth, nutrient cycling, and contaminant behavior in the soil requires knowledge of the soil-moisture status as well as its spatial and temporal variability.

The neutron probe (also known as the neutron moisture meter) is a popular instrument for measuring soil-water content. The probe emits high-energy neutrons and measures the low-energy neutrons that result from interactions with the soil constituents. What makes a neutron probe viable for measuring soil water is that hydrogen nuclei, which are part of the water molecule, play a dominant role in normal soils in reducing the energy level of neutrons.

The neutron probe measures soil-moisture status without the need to collect innumerable soil samples. This feature alone is one of the main reasons the technique is popular, because it allows researchers and farm managers to monitor the water content at multiple soil depths at the same location repeatedly without disturbing the soil. Thus, neutron probes make it possible to conduct long-term studies 'on the same plot of land' without fear of altering the plot by repetitive sampling. Neutron probes are easy to deploy for studies of spatial variability, thus overcoming the cost and destructiveness of sampling or the cost, disturbance, and inflexibility of weighing lysimeters. Another reason for the rise in popularity of the neutron probe is the large sensing volume, which nominally approximates a sphere with a 15- to 30-cm radius. The large sensing volume reduces the influence of small-scale variability in the soil, the number of measurements needed, and the sensitivity to soil disturbance caused by installation.

The idea of using neutrons to measure soil conditions arose in the 1940s, and the first use of a neutron device in a borehole was reported in 1947. By the 1950s, researchers around the world were extending neutron probe technology to agricultural conditions and reporting success in measuring water content in laboratory and field experiments. Since then, the neutron probe has become a standard field method for measuring soil-water content. As shown in **Figure 1**, the agriculture community uses neutron probes to monitor field water status in croplands and identify when irrigation is needed to meet plant water requirements. The environmental community uses neutron probes to monitor water conditions in and around waste disposal sites. The academic research community uses neutron probes to study and understand the hydrologic cycle.

Theory and Instrumentation

The neutron is an elementary particle with a mass similar to that of a proton. Unlike the proton, the neutron has no charge, which allows it to interact primarily with atomic nuclei. A free neutron eventually decays according to:

$$ {}^{1}_{0}\mathrm{n} = {}^{1}_{1}\mathrm{H} + \beta \qquad [1] $$

Figure 1 Neutron probe in use in a field containing soybean stubble. Photo courtesy of Steve Evett, USDA–ARS, Soil and Water Management Research Unit, Bushland, Texas.

with a half-life of 10.4 min and a β energy of 0.782 MeV.

Neutrons are produced by the transformation from one radionuclide to another. Three modes of transformation suitable for soil and environmental studies are spontaneous fission, alpha-neutron (α-n) interactions, and neutron generation. Of these, the α-n mode has proven to be most effective for agricultural soils.

The α-n method for generating neutrons involves mixing two materials: one material to provide a source of alpha particles, the other material to provide a source of neutrons when exposed to alpha particles. Potential sources of alpha particles include americium, californium, polonium, plutonium, and radium. The americium decay reaction is:

$$^{241}_{95}\mathrm{Am} = ^{237}_{93}\mathrm{Np} + ^{4}_{2}\alpha + \gamma \qquad [2]$$

which generates an α with an energy of 5.48 MeV. The material typically chosen to provide the neutrons is beryllium, although lithium has also been used. For the americium–beryllium (Am-Be) combination, the equation is:

$$^{4}_{2}\alpha + ^{9}_{4}\mathrm{Be} = ^{12}_{6}C + ^{1}_{0}\mathrm{n} + \gamma \qquad [3]$$

The optimum material combination from the standpoint of commercial application is americium–beryllium, which has the desirable characteristics of a long half-life (458 years), a moderately high neutron-emission rate of 67.6 n s^{-1} MBq^{-1} (2500 n s^{-1} mCi^{-1}), and a low gamma dose rate of 2.78 nSv s^{-1} at 1 m (1 mrem h^{-1} at 1 m) under a fast neutron flux of 10^6 n s^{-1}. The long half-life ensures that a sealed source will last many years with negligible decay. The high emission rate ensures that enough neutrons will be present to make detection feasible. The low gamma dose rates and moderate fast neutron fluxes make it relatively safe to use.

The energy of neutrons generated by the Am-Be source ranges from 1 to 14 MeV and averages approximately 4.5 MeV. Neutrons at these energy levels are classified as fast neutrons. As they permeate the surrounding medium, fast neutrons interact with nuclei in scattering (both elastic and inelastic) and absorption events. The probability that a nucleus will interact with a free neutron can be statistically determined and represented using the microscopic cross-section (σ); the standard unit of measurement is the barn, which is 10^{-28} m^2 per nucleus. The ability of a soil to interact with neutrons will be a function of the soil's elemental composition, the σ values associated with each element, and the element density (atoms per volume).

In elastic scattering events, the neutron strikes the nucleus and rebounds; energy and momentum are conserved. This type of scattering is colloquially referred to as 'billiard ball scattering' and is the dominant process for slowing down neutrons under normal soil conditions. In inelastic scattering events, the neutron strikes the nucleus and rebounds with a loss of energy via gamma-ray emission. In absorption events, neutrons strike and react with the nucleus, resulting in the release of additional radiation (e.g., α, β, γ, n). Both inelastic scattering and absorption of fast neutrons are minor neutron-slowing processes relative to elastic scattering.

Table 1 shows that soil constituents differ in their efficiency at slowing down neutrons by elastic scattering. Hydrogen, which is present primarily in soil water, requires the least number of collisions to thermalize a 14 MeV neutron. Table 1 shows that, although the slowing-down power of the elements is a function of the energy range, hydrogen consistently has the greatest potential.

After repeated scattering events, the energy level of the neutron decreases to less than 1000 eV, at which point the neutron is classified as a slow or thermalized neutron. The petroleum industry uses slow neutrons in the energy range from 1000 to nearly 0.025 eV for geophysical measurements (e.g., petroleum reservoir

Table 1 Fast and thermal neutron information for the most common soil constituents

Element	Nominal soil fraction (wt %)	Estimated no. collisions to thermalize neutron from 14 MeV to 0.01 eV	Estimated fast neutron slowing-down cross section 10–0.183 MeV (barn)	Estimated fast neutron slowing-down cross section 0.183 MeV to 1.44 eV (barn)	Thermal capture cross section at 0.0253 eV (barn)
H	–	19	2.76	19.2	0.332
O	49	154	0.079	0.48	0.0002
Si	33	297	0.164	0.16	0.16
Al	7	290	0.668	0.13	0.23
Fe	4	539	0.132	0.32	2.62
C	2	112	0.234	0.77	0.00373
Ca	1	380	0.172	0.12	0.44
K	1	418	0.151	0.11	2.07
Na	0.7	190	0.204	0.28	0.505
Mg	0.6	235	0.153	0.27	0.069
Ti	0.5	512	–	–	6.09
N	0.1	155	–	–	1.88
P	0.08	333	–	–	0.166
Mn	0.08	586	–	–	13.4
S	0.05	344	–	–	0.51
Cl	0.01	343	–	–	33.8
B	0.001	89	–	–	759

Adapted from: IAEA (1970) *Neutron Moisture Gauges*. Technical Reports Series 12. Vienna, Austria: International Atomic Energy Agency; Bohn HL, McNeal BL, and O'Connor GA (1979) *Soil Chemistry*. Chichester: John Wiley & Sons, Inc.; Hearst JR, Nelson PH, and Paillet FL (2000) *Well Logging for Physical Properties*, 2nd edn. New York: John Wiley & Sons, Inc.; Chang J *Table of Nuclides*, KAERI (Korea Atomic Energy Research Institute).

characterization); these slow neutrons are called epithermal neutrons. Slow neutrons near or below the energy level of 0.025 eV are called thermal neutrons because they are in an energy condition similar to thermal equilibrium with their environment.

The density of thermal neutrons that forms around the probe is balanced between thermal neutron creation (via thermalization of fast neutrons) and the processes of thermal neutron capture, decay, and diffusion away from the probe. Once formed, thermal neutrons do not travel far; the diffusion length is highly variable but is on the order of centimeters. If a fast neutron encounters many nuclei close to the source, the neutron will be reduced to thermal conditions while it is still relatively close to the source. If a fast neutron encounters few nuclei close to the source, the neutron will tend to migrate further from the source before it is reduced to a thermal energy level. In other words, the nature of the medium dictates how far from the source the thermal neutrons are created. Because hydrogen (and thus water) has such a high macroscopic cross-section relative to all other normal soil constituents, water content is the primary controller of the density of thermal neutrons near the probe and thus determines how much of the medium is sensed. This phenomenon is called the volume of influence. Figure 2 shows how the volume of soil sensed by the neutron probe decreases as the water content increases. The radius of influence is typically 15 cm for wet soil and as much as 30 cm for dry soil.

Thermal neutrons can be detected using the ionization products that form when the neutron is captured by an element. Early detectors used gas-filled tubes containing boron trifluoride (BF_3), but modern detectors tend to use the helium isotope ^{3}He, which reacts as follows:

$$^3_2\mathrm{He} + {}^1_0\mathrm{n} = {}^3_1\mathrm{H} + \mathrm{p} \qquad [4]$$

The ^{3}He isotope has a higher thermal cross-section (5400 barns) than ^{10}B (3840 barns) and can be stored in the detector tube to a greater density. For both the B and He detectors, the ionization products within the gas in the tube alter the current through the tube, thus allowing the thermal neutrons to be detected and counted. These detectors are insensitive to fast neutrons.

Radioactive decay is a random process that affects the rate of fast neutrons emanating from the probe. To overcome the randomness of the source, neutron probes are designed to count the number of thermal neutrons for a specific duration. Repeated measurements under the same conditions, such as in a standard medium, produce a sample population of counts that resembles a Poisson distribution, which has the property that the standard deviation of the finite sample population can be approximated by the square root of

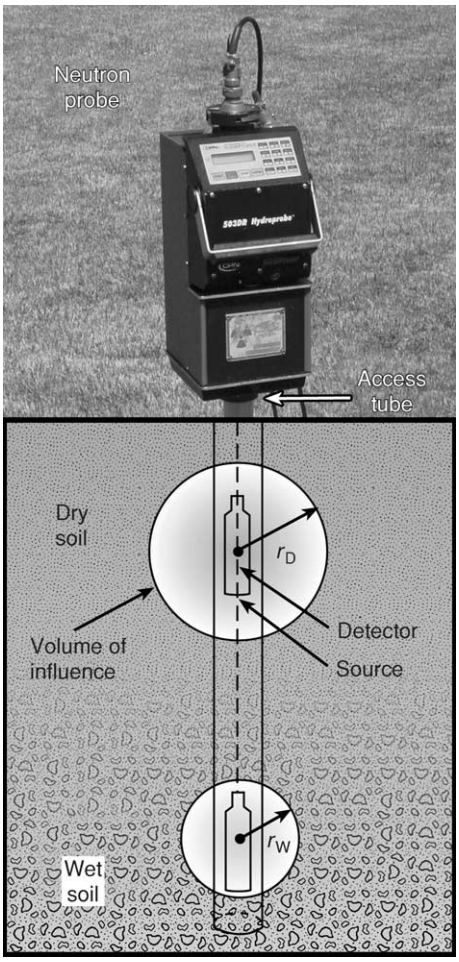


Figure 2 A neutron probe volume of influence in wet soil and dry soil.

the mean count. Because a Poisson distribution can be approximated with a normal distribution (given sufficient counts), the mean count and standard deviation can be used to describe the sample distribution. Thus, they can be used to control measurement precision by estimating the counting time needed to achieve a specific precision. They can also be used for quality control by comparing the results with a normal distribution to detect abnormal amounts of fluctuation that would indicate a malfunctioning probe.

Neutrons emitted by a neutron probe have the potential to activate (make radioactive) some of the elements (e.g., Na, Mg, Cl, Si, Al, Fe, Mn) in the soil and access tube. For example, the aluminum in an aluminum access tube can become activated as follows:

$$^{27}_{13}\mathrm{Al} + ^{1}_{0}\mathrm{n} = ^{28}_{13}\mathrm{Al} \qquad [5]$$

The resulting aluminum isotope has a short half-life (2.3 min) and yields a beta particle during decay. Other activation product half-lives range from seconds to hours. The probability of significant activation is very low because the concentrations in the soil are generally so small and the neutron flux is relatively low (more than an order of magnitude lower than required to cause measurable activity). Because activation products are belowground and of short duration, the risk of exposure is extremely small.

Field Methodology

To use a neutron probe to measure soil moisture requires an access hole in the soil into which the probe can be inserted. Ideally, the hole should be just wide enough to permit entry of the probe while minimizing excess air gaps. In reality, drilling a uniform-width hole and keeping it open and unchanging is highly unlikely, even under the best of field conditions. Therefore, the commonly accepted practice is to install a rigid-wall tube (or pipe, conduit, casing) that allows the probe to access the soil repeatedly without degrading the hole geometry.

The access-tube material should be chosen based on ease of installation, durability, availability, and minimal impact on neutron flux. Thin-walled tubing made of aluminum, schedule-10 galvanized steel, or schedule-40 polyvinyl chloride (PVC) is the preferred choice, but thicker iron well casing has been shown to be acceptable for specific applications. Successful installation requires that air gaps and packing material (e.g., grout) be minimized or, even better, eliminated and that the installation procedure be used consistently for all tubes. The key to choosing the correct access-tube material and installation method is to demonstrate, through calibration, that the probe and access-tube combination retains sufficient sensitivity to be an effective soil-moisture measurement system.

The most common orientation for access tubes is vertical so that the operator can monitor the depth distribution of soil water. However, access tubes can also be installed at an angle, or even horizontally, to allow the operator to monitor beneath an object such as a buried storage tank. Access tubes can become damaged during installation; once installed, the tube should be tested with a dummy probe to determine its integrity. Dummy probes are inert objects, usually made of metal, that are shaped like a real probe but slightly larger. If the dummy probe can be passed through the access tube, the operator can be assured that the neutron probe will not become lodged.

In addition to protecting the access tube, the operator must protect the access-tube site from compaction by foot and vehicular traffic. Relative to the tube surroundings, compaction increases the density of

neutron-affecting minerals, alters the soil hydraulic properties, and affects plant growth and behavior. The seriousness of compaction can increase if the site is used for multiyear tests involving many site visits.

Calibration

A neutron probe detects and measures thermal neutrons to indicate soil-moisture status. In addition to soil water, the formation and maintenance of a cloud of thermal neutrons within the sensing volume of the probe depends on the mineral composition of the soil matrix, bulk density, bound water, soil-water constituents, and organic matter. Table 1 shows how the major soil elements vary in their ability to slow down fast neutrons as well as capture thermal neutrons. Clearly, hydrogen has the greatest potential to thermalize neutrons and capture them, and thus control the thermal neutron density around the probe. As noted above, increases in bulk density translate directly into increases in the density of neutron-affecting minerals. Bound (or structural) water is part of the soil matrix (e.g., between clay lamella). Bound water affects neutron thermalization just like free water, but it is not available to plants and does not interact with free-flowing soil water.

Natural soil-water constituents are generally present in concentrations too low to affect a neutron probe and are normally not considered during probe calibration and use. However, operators should be cognizant when soils are subject to significant fertilization, salts (e.g., carbonate and sulfate deposits in semiarid and arid soils), sludge amendments, or waste disposal. Under these conditions, the soil water could contain constituents with significant thermal neutron-capture cross sections (e.g., B, Cl, Fe, Mn) that would render the probe calibration less representative unless modified.

Another soil component normally not considered during probe calibration and use is organic material, which includes living plant roots and associated subsurface plant components, dead and decaying plant parts, and fully developed organic matter. The key feature of organic material that could affect probe sensitivity is the presence of water, either as part of living plant tissue or in response to changes in the water-holding capacity of the soil. These factors should be assessed seasonally and in special zones in the soil where plant root changes are greatest (e.g., soil surface and soil-layer interfaces with contrasting soil texture).

Because they are affected by conditions such as mineralogy, bulk density, and sediment discontinuities, neutron probes should be calibrated for the specific conditions under which they will be used in the field. In a typical field calibration, multiple access tubes are emplaced to allow access to the entire thickness of soil. The neutron probe is lowered to different depths in each access tube and readings performed. Immediately subsequent to collecting neutron readings, the soil is excavated to collect multiple undisturbed soil samples directly adjacent to the tubes and at the same probe-measurement depths. The undisturbed samples are then analyzed in the laboratory to determine the measured volumetric water contents and bulk densities for those depths. These values are then related to the neutron-probe counts taken at the respective depths. The key to a successful calibration is to collect data under wet and dry conditions in each soil layer to cover the complete range of expected water contents.

The thermal neutron count rate is converted to soil moisture using a calibration equation. Separate equations may be required for different soil horizons. The most common calibration equation is a linear function of the form:

$$\theta = m(\mathrm{CR}) + b \qquad [6]$$

where θ is the volumetric moisture content, CR is the thermal neutron count ratio, and m and b are calibration parameters. The count ratio is the thermal-neutron count in the soil relative to the thermal-neutron count in a standard medium. The rationale for referencing to a standard medium is to minimize the impacts caused by changes in the probe source and electronics due to aging and decay that might not be readily apparent to the probe operator. The standard medium must have neutron-slowing properties and the properties must remain constant. Typical standards are the probe shield or a sealed barrel containing nonchanging material such as water or wax. If the probe shield is used, it ought to be located 0.8 m from the ground surface and up to 3 m from structures, vehicles, and people to minimize interferences. The distance recommendations are less stringent if the larger standards such as water- or wax-filled barrels are used. Figure 3 shows a typical linear calibration equation and the supporting field measurements.

Equation [6] has been shown to be appropriate for a wide variety of soils and moisture conditions, but other forms of calibration equations have been proposed and used for special cases. These other equations include multiple regression equations that use other soil information such as bulk density and quadratic equations that allow for a better fit of data for specific cases (e.g., very low water contents; large-diameter, iron-cased wells).

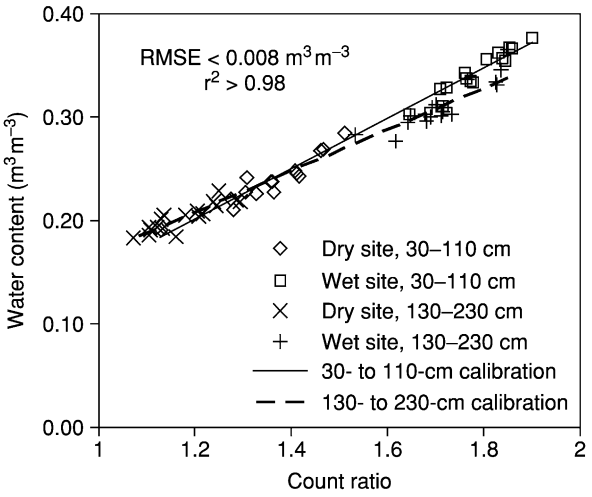


Figure 3 Calibration of a neutron probe in Pullman soil at Bushland, Texas. RMSE, root mean square error. Courtesy of Dr. Steve Evett, USDA–ARS, Soil and Water Management Research Unit, Bushland, Texas.

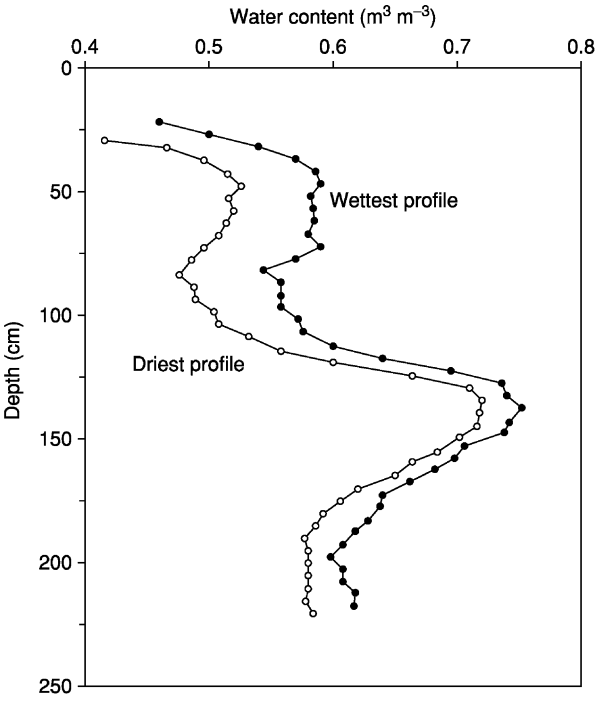


Figure 4 Water-content profiles in a Tasmanian soil on two different dates. Adapted from Greacen EL (1981) *Soil Water Assessment by the Neutron Method*. Victoria, Australia: CSIRO.

Applications

The output of the neutron probe is converted to soil moisture data. These data are used to assess the depth distribution of water, the overall soil-water status, and the movement of water in the soil profile. When collected on multiple occasions, the temporal data can reveal long-term changes such as withdrawal of water by plants, loss of water by evaporation, and responses to weather and irrigation conditions. Figure 4 shows a typical variation in water-content profile determined with a neutron probe.

Safety and Care

Human health is affected by exposure to radiation. The nature and degree of health effects depend on the type of radiation and the length of time one spends in the radiation field. Although the radioactive source in a neutron probe is sealed and unreachable by the operator (barring the use of force to overcome the seal), the probe operator is subject to the radiation emitted through the seal and the probe housing. Emissions from neutron probes consist of neutrons and gamma rays at various energy levels. Measurements adjacent to the probe shield have shown that dose rates are dominated by the fast neutron flux. Thus, most probe monitoring efforts focus on measuring the fast neutron flux. The impact to an operator can be seen in the following example. The radiation field at a distance of 5 cm from the shield of a 3.7-GBq (100 mCi) Am-Be source was measured and found to be $0.12\,\mathrm{mSv\,h^{-1}}$. At that dose rate, a neutron probe operator could be 5 cm from the shield for 8 h every week of the year and not exceed the International Standard worker limit of 50 mSv in any single year.

Radiation from a neutron probe is generally not a health hazard to someone within the radiation field if ALARA (as low as reasonably achievable) principles are followed. For example, the radiation field strength can be reduced by limiting the size of the source and surrounding it with a shield. Typical shields are composed of hydrogenous material to reduce neutrons. The dose from a 3.7-GBq (100 mCi) Am-Be source is reduced by approximately 20% when behind a 7-cm-thick plastic shield; the resulting dose is primarily from fast neutrons ($0.1\,\mathrm{mSv\,h^{-1}}$ ($10\,\mathrm{mrem\,h^{-1}}$) versus $0.01\,\mathrm{mSv\,h^{-1}}$ ($1\,\mathrm{mrem\,h^{-1}}$) from gammas and thermal neutrons). The dose from more modern probes is less because the source activities have been reduced, e.g., 0.37 GBq (10 mCi) and 1.8 GBq (50 mCi). Operators can reduce exposure further by using distance and time to advantage. For example, the radiation filed is inversely proportional to the distance squared. By doubling the distance to the probe, the operator can reduce the radiation field by a factor of four. The operator can also reduce the radiation field to some extent by using intervening material (e.g., shielding between a truck cab and the truck bed when the probe is transported). In addition to increasing distance to the

probe, the operator can also reduce exposure by minimizing the time spent within the radiation field. Prior planning, quality-assurance procedures, and efficient work habits can reduce the amount of unnecessary time spent lingering near the probe.

Modern neutron probes are robust but, because of the radioactive source and the electronics, these probes require administrative oversight and care in handling. The International Atomic Energy Agency (IAEA) promulgates safety and regulatory requirements designed to ensure safe use of neutron probes. More than 125 nations belong to the IAEA and they use national organizations to address the IAEA requirements. These requirements include the need for routine periodic leak tests of sealed sources to look for particles of the source (thus indicating a failure of the source seal) and transportation and storage controls for safety, theft protection, and disposal.

Manufacturers provide guidance to operators for probe care, cleaning, and battery replacement. For maintenance and repair work, manufacturer recommendations differ. Some require sending the probe back to the factory; others allow the electronics to be disengaged from the source and sent to the factory while the source remains stored locally. A probe operator can prolong the life of the probe, cable, and electronics by taking care to avoid dropping them, subjecting them to extremes of temperature, pressure, and acceleration, and by minimizing their exposure to solar radiation and water. The operator should also prevent undue stress on the cable.

Access tubes can become damaged during use. Caps or covers on the access tubes help to limit accidental bumps as well as to keep out water, soil, and organic debris. Even so, the tubes should be periodically checked for dents, decay, and debris. Tube cleanliness can be maintained by using a cleaning rag on a pole. The radial uniformity of the tube should be checked occasionally using a dummy probe to ensure that the actual neutron probe will pass easily through the tube.

Future Use

After more than 50 years, neutron probes remain the most reliable tool available for field monitoring of soil-water content. Neutron probes provide integrated measurements over relatively large volumes of soil and, with proper access, allow for repeated sampling of the subsurface at the same locations. The limitations of neutron probes include costly and time-consuming manual operation, lack of data automation, and costly regulatory requirements. As more nonradioactive systems for soil-water monitoring are developed to provide automated profiling capabilities, neutron-probe usage will probably decline. Until then, neutron probes will continue to be a standard for reliable measurements of field water contents in soils around the globe.

See also: **Time-Domain Reflectometry**; **Water Content and Potential, Measurement**; **Water Cycle**; **Water Potential**

Further Reading

American Society for Testing Materials (2002) *Standard Test Method for Water Content of Soil and Rock In-Place by the Neutron Depth Probe Method,* D 5220-92 (97), vol. 04.08, pp. 1089–1094. West Conshohocken, PA: ASTM International.

Bohn HL, McNeal BL, and O'Connor GA (1979) *Soil Chemistry*. New York: John Wiley & Sons, Inc.

Evett SR (2002) Exploits and endeavours in soil water management and conservation using nuclear techniques. In: Eaglesham A (ed.) *International Symposium on Nuclear Techniques in Integrated Plant Nutrient, Water and Soil Management*, pp. 151–177. Vienna, Austria: International Atomic Energy Agency.

Gardner W and Kirkham D (1952) Determination of soil moisture by neutron scattering. *Soil Science* 73: 391–401.

Gee GW, Stiver JF, and Borchert HR (1976) Radiation hazard from americium-beryllium neutron moisture probes. *Soil Science Society of America Journal* 40: 492–494.

Greacen EL (1981) *Soil Water Assessment by the Neutron Method*. Victoria, Australia: CSIRO.

Hearst JR, Nelson PH, and Paillet FL (2000) *Well Logging for Physical Properties*, 2nd edn. New York: John Wiley & Sons, Inc.

Hignett C and Evett SR (2002) Neutron thermalization. In: Dane JH and Topp GC (eds) *Method of Soils Analysis, part 4, Physical Methods*. Madison, WI: Soil Science Society of America.

IAEA (1970) *Neutron Moisture Gauges*. Technical Reports Series 112. Vienna, Austria: International Atomic Energy Agency.

IAEA (1999) *Nuclear Geophysics and Its Applications*. Technical Reports Series 393. Vienna, Austria: International Atomic Energy Agency.

IAEA (2002) *Neutron and Gamma Probes: Their Use in Agronomy*. Training Course Series No. 16. Vienna, Austria: International Atomic Energy Agency.

Tyler SW (1988) Neutron moisture meter calibration in large diameter boreholes. *Soil Science Society of America Journal* 52: 890–893.

NITROGEN IN SOILS

Contents

Cycle
Nitrates
Nitrification
Plant Uptake
Symbiotic Fixation

Cycle

M S Coyne and W W Frye, University of Kentucky, Lexington, KY, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Nitrogen (N) is the element most often limiting plant nutrition in terrestrial ecosystems. The greatest source of available N is atmospheric dinitrogen (N_2), which is relatively inert and can only be used by symbiotic and free-living prokaryotic bacteria with the capacity for N_2 fixation (Table 1). For other plants and soil organisms, the slow release of N from rocks and minerals and cycling between organic and inorganic forms in soil is crucial to life. Unfortunately, some aspects of N cycling can be problematic. Nitrogen can be readily lost from terrestrial soils, leading to reduced fertility and surface- or groundwater contamination, and several transformations give rise to intermediate or final products that can have negative environmental consequences.

Properties and Forms of Nitrogen in Soil

There are two overriding chemical properties of the N atom worth noting. First, the dinitrogen molecule, N_2, is very stable. Once formed and lost from the soil environment, N_2 is relatively inert and requires considerable energy ($946\,kJ\,mol^{-1}$) to break the triple bond between N molecules and reintroduce N to biological systems. The second significant feature of N, owing to its atomic structure and electron configuration, is its ability to undergo changes in oxidation state from $+5$ to -3 depending on its environment. This property is important for soil transformations, because N, in its oxidized forms, can be an electron acceptor in anaerobic respiratory metabolism. N, in its reduced forms, can also be an energy source for various chemolithotrophic bacteria. The N cycle in soil is driven by these transformations between oxidized and reduced states and between organic and inorganic forms (Figure 1).

The principal soluble inorganic N forms in soil are NO_3^-, NO_2^-, and NH_4^+. The inorganic gaseous N forms in soil are N_2, N_2O, NO, and NH_3. Organic N takes many forms, such as amino acids (e.g., glycine, glutamine), amino sugars (e.g., glucosamine, galactosamine), nucleosides (e.g., adenine, guanine), peptides, phospholipids (e.g., phosphatidylethanolamine, phosphatidylserine), vitamins (e.g., niacin), and other compounds such as creatine, cyanide, allantoin, various alkyl amines, and urea. Unidentifiable forms of N are typically polymerized with soil organic matter fractions such as humic and fulvic acids.

N-Cycling Pathways

The major organic and inorganic N transformations that occur in soil can be grouped into five pathways: mineralization (ammonification), assimilation (immobilization), nitrification, nitrate reduction, and N_2 fixation.

Table 1 Global sources of available N

Source	*N content ($\times 10^{16}$ kg)*[a]	*Total (%)*
Atmosphere	386.000	87.0000
Oceans and lakes		
Dissolved N_2	2.190	0.5000
Soluble inorganic N	0.110	0.0200
Sediments	35.000–55.000	12.0000
Sea bottom organic N	0.054	0.0100
Terrestrial biosphere	0.028–0.065	0.0100
Terrestrial soils		
Organic N	0.022	0.0050
Clay-fixed NH_4^+	0.002	0.0005

[a]The data exclude N in the lithosphere such as igneous rocks, sediments, and coal that do not actively participate in global N cycling.
Source: Stevenson FJ and Cole MA (1999) *Cycles of Soil*, 2nd edn. New York: John Wiley and Sons, Inc.

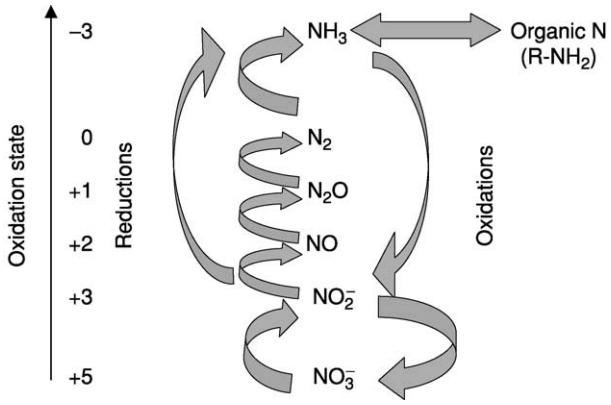


Figure 1 Oxidation–reduction transformations of nitrogen.

Mineralization (Ammonification)

During mineralization, organic N is hydrolyzed to release NH_3, which is subsequently protonated in the soil solution to become the cation NH_4^+ (eqn [1a]). Consequently, mineralization is frequently referred to as ammonification:

$$\text{Organic N (R-NH}_2) \longrightarrow \text{NH}_3 \longrightarrow \text{NH}_4^+ \quad [1a]$$

$$\text{Glutamate} + \text{NAD}^+ + \text{H}_2\text{O} \xleftrightarrow{\text{Glutamate dehydrogenase}} \alpha\text{-Ketoglutarate} + \text{NH}_4^+ + \text{NADH} + \text{H}^+ \quad [1b]$$

Another example of mineralization that has great importance in soil is urea hydrolysis by the seemingly ubiquitous enzyme urease, which is shown in eqn [2]. Urea is one of the dominant organic N forms excreted by animals, and it is also one of the most popular forms of solid fertilizer used on a global basis. Urea hydrolysis in soil is rapid, and the subsequent NH_3 formation can cause considerable N loss through volatilization. One of the immediate effects of urea hydrolysis is also a rapid increase in the soil pH:

$$\text{H}_2\text{O} + \text{NH}_2\text{CONH}_2 \xrightarrow{\text{Urease}} \text{CO}_2 + 2\text{NH}_3 \quad [2]$$

Assimilation (Immobilization)

Assimilation (immobilization) is the incorporation of inorganic N as NH_4^+, NO_3^-, or NO_2^- into biomass (eqns [3] and [4]). Immobilization can also refer to the binding of NH_4^+ to certain clays in soil and to the interaction of inorganic N with soil organic matter:

$$\text{NH}_4^+ \longrightarrow \text{Organic N (R-NH}_2) \quad [3]$$

$$\text{NO}_3^- \text{ or } \text{NO}_2^- \longrightarrow \text{NH}_4^+ \longrightarrow \text{Organic N (R-NH}_2) \quad [4]$$

Nitrification

Nitrification is the general term that describes oxidation of reduced inorganic and organic N and formation of NO_2^- or NO_3^-. It is an aerobic process that occurs by one of two pathways: autotrophic nitrification, carried out by chemolithotrophic bacteria, and heterotrophic nitrification, carried out by chemoheterotrophic bacteria and fungi.

Autotrophic nitrification Autotrophic nitrification is a two-step process that is carried out by chemolithotrophic bacteria, which obtain energy by oxidizing NH_3 to NO_3^- (eqn [5]).

$$\text{NH}_4^+ \underset{+\text{H}^+}{\overset{-\text{H}^+}{\rightleftharpoons}} \text{NH}_3 \longrightarrow \text{NO}_2^- \longrightarrow \text{NO}_3^- \quad [5]$$

The first step in the pathway, which is performed by NH_3-oxidizing bacteria such as *Nitrosomonas*, involves the sequential oxidation of NH_3 to hydroxylamine (NH_2OH) and then to NO_2^- (eqns [6a] and [6b]):

$$\text{NH}_3 + \text{O}_2 + 2\text{H}^+ \xrightarrow{\text{Ammonia monooxygenase (AMO)}} \text{NH}_2\text{OH} + \text{H}_2\text{O} \quad [6a]$$

$$\text{NH}_2\text{OH} + \text{H}_2\text{O} \xrightarrow{\text{Hydroxylamine oxidoreductase (HAO)}} \text{NO}_2^- + 5\text{H}^+ \quad [6b]$$

The second step in the pathway is performed by NO_2^- oxidizing bacteria typified by bacteria in the genus *Nitrobacter*. Nitrite oxidation is generally so rapid that NO_2^- rarely accumulates to any significant extent in soil:

$$\text{H}^+ + \text{NO}_2^- + \text{H}_2\text{O} \xrightarrow{\text{Nitrite dehydrogenase}} \text{NO}_3^- + 3\text{H}^+ \quad [7]$$

Autotrophic nitrification is believed to be the dominant type of nitrification occurring in agricultural soils. Nitrification is an acidifying process and use of reduced organic and inorganic N for plant nutrition will lower the soil pH with time because of nitrification.

Heterotrophic nitrification Heterotrophic nitrification is the oxidation of organic N and NH_3 by various chemoheterotrophic bacteria and fungi such as *Arthrobacter*, *Streptomyces*, and *Aspergillus*. Heterotrophic nitrification produces NO_2^- and NO_3^- in soils typically too acid to permit autotrophic nitrification (eqn [8a]).

$$\text{Organic N (R-NH}_2) \longrightarrow \text{NO}_2^- \text{ or } \text{NO}_3^- \quad [8a]$$

Products of heterotrophic nitrification also include nitroso-compounds, which have been identified as potential carcinogens (eqn [8b]):

$$\text{R-NH}_2 \longrightarrow \text{RNHOH} \longrightarrow \underset{\text{Nitroso-compound}}{\text{R-NO}} \longrightarrow \text{R-NO}_3 \longrightarrow \text{NO}_3^- \quad [8b]$$

Nitrate Reduction

In anaerobic or waterlogged environments, denitrification and dissimilatory nitrate reduction to ammonium (DNRA) are the two major pathways by which the NO_2^- and NO_3^- formed by nitrification can be subsequently reduced. Facultatively anaerobic bacteria perform these N transformations, although it has been suggested that some fungi are also capable of denitrification.

Denitrification Denitrification is the major reductive fate of NO_3^- in most soil environments. It is a respiratory process in which nitrogen oxides are used as terminal electron acceptors in place of O_2 and produces gas as the terminal reduction product (eqn [9]). Denitrification is a multistep process catalyzed by distinct enzymes for each step:

$$\text{NO}_3^- \xrightarrow{\text{Nitrate reductase}} \text{NO}_2^- \xrightarrow{\text{Nitrite reductase}} \text{NO} \xrightarrow{\text{Nitric oxide reductase}} \text{N}_2\text{O} \xrightarrow{\text{Nitrous oxide reductase}} \text{N}_2 \quad [9]$$

Dissimilatory nitrate reduction to ammonium The NH_4^+ produced during DNRA is not assimilated but is excreted into the environment (eqn [10]). The bacteria that perform DNRA are distinct from denitrifiers. They predominate in environments that are continuously saturated, anaerobic, and have high concentrations of available C relative to NO_3^-:

$$\text{NO}_3^- \longrightarrow \text{NO}_2^- \longrightarrow \text{NH}_4^+ \quad [10]$$

N_2 Fixation

Nitrogen fixation is an energy-demanding process by which prokaryotic bacteria reduce N_2 gas to NH_3 either independently or in symbiosis with various leguminous and nonleguminous plant species (eqn [11]). The most significant types of terrestrial nitrogen fixation are the legume/rhizobia symbiosis and the actinorhizal symbioses, which occur between woody shrubs and bacteria of the genus *Frankia*. In aquatic systems, photosynthetic cyanobacteria such as *Anabaena* and *Nostoc* are responsible for N_2 fixation.

$$\text{N}_2 + 16\text{Mg-ATP} + 8\text{H}^+ \xrightarrow{\text{Nitrogenase}} 2\text{NH}_3 + \text{H}_2 + 16\text{ADP} + 16\text{Pi} + 16\text{Mg}^{2+} \quad [11]$$

Nitrogenase is actually an enzyme complex consisting of dinitrogenase reductase and dinitrogenase, which are highly conserved, O_2-sensitive enzymes working in tandem to reduce N_2. Nitrogenase is somewhat nonspecific with respect to substrate, so it will also reduce H^+ to H_2, N_2O to N_2, and C_2H_2 to C_2H_4. The latter reduction is the basis for the acetylene reduction assay (ARA), which is used to detect and quantify N_2 fixation in different environments.

Fixed (Nonexchangeable) Ammonium

An important part of the inorganic soil N of some soils is fixed or nonexchangeable NH_4^+. Ammonium ions in soil exist in a chemical equilibrium as follows:

$$\text{NH}_4^+(\text{solution}) \leftrightarrow \text{NH}_4^+(\text{exchangeable}) \leftrightarrow \text{NH}_4^+(\text{fixed})$$

Figure 2 shows the crystalline structure of a weathered, micaceous silicate clay mineral, such as illite, showing the locations of solution, exchangeable, and fixed NH_4^+. Much of the fixed NH_4^+ can be slowly available to growing plants and soil

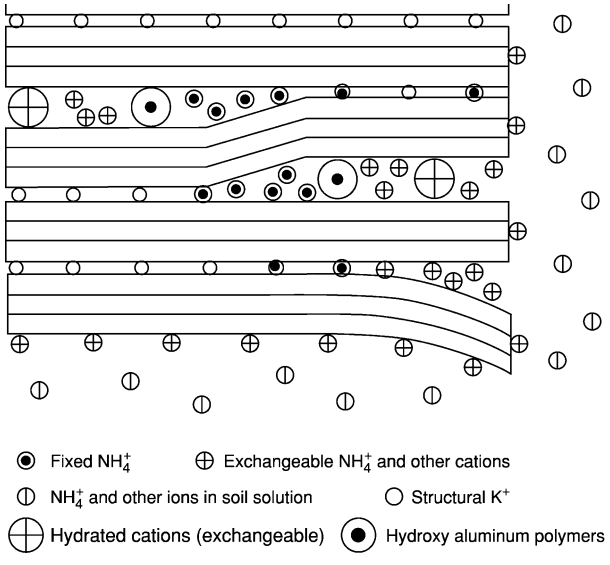


Figure 2 Schematic of micaceous-type soil clay (vermiculite, hydrous mica) showing wedge-shaped zones and weathered edges where NH_4^+ can be fixed. Adapted from Dixon JB and Weed SB (eds) (1989) *Minerals in Soil Environments*, 2nd edn. Soil Science Society of America Book Series 1. Madison, WI: Soil Science Society of America with permission.

microorganisms. As plant uptake or biomass immobilization depletes the readily available NH_4^+, i.e., solution NH_4^+ and exchangeable NH_4^+, fixed NH_4^+ is released by cation replacement, especially with K^+. Conversely, when NH_4^+ is added to the soil from any source, more NH_4^+ is fixed to establish a new equilibrium. Therefore, the fixed NH_4^+ sites serve as a reserve source of N and as a sink for excess NH_4^+.

Regulation and Environmental Control

Nitrogen transformations between organic and inorganic forms in soil depend on the type and availability of inorganic N and on the relative availability of inorganic N to mineralizable C.

Cellular and Environmental Regulators of N Transformations

Microorganisms in soil will preferentially assimilate NH_4^+ rather than NO_3^- or NO_2^-. Even low NH_4^+ concentrations in soil solutions will quickly stop NO_3^- uptake and, soon after, inhibit enzymes associated with assimilatory NO_3^- reduction (Table 2). If NH_4^+ concentrations in soil solutions exceed $1\,mmol\,l^{-1}$, assimilation by glutamate dehydrogenase will occur in which the first recognizable product is glutamate. If soil N is limiting, microorganisms will utilize an enzyme system referred to as GS/GOGAT (glutamine synthetase/glutamine oxoglutarylamino transferase) in which the first identifiable form of organic N is glutamine. Nitrate, if it is assimilated, will first be reduced to NO_2^- by assimilatory NO_3^- reductase and the NO_2^- subsequently reduced to NH_4^+ by assimilatory NO_2^- reductase. Thereafter, the pathways are the same as for NH_4^+ assimilation.

Organic N will be mineralized in soil in the absence of available inorganic soil N, although, as a general rule, N-containing compounds such as amino acids are not mineralized only to be subsequently resynthesized. Most N cycling in soil and inorganic N release is actually driven by the mineralization of soil C for biomass and energy; inorganic N accumulates in soil only when the microbial demand for N is met. When the C/N ratio of decomposing material exceeds approximately 30, net immobilization of inorganic N from soil occurs because there is insufficient N in the mineralizing material to sustain microbial biomass formation. In contrast, when the C/N ratio is less than 30, the N needs for microbial growth are met, and inorganic N is released into the soil environment. The decomposition of organic materials with very low C/N ratios can therefore cause inorganic N to accumulate briefly in excess of microbial and plant needs, with the potentially adverse effects of volatilization, runoff loss, or plant toxicity in the case of NH_4^+.

The three environmental factors that regulate nitrification in soil are NH_4^+ availability, aeration, and pH (Figure 3). As long as the oxidation–reduction potential (E_h) of soil exceeds approximately 400 mV, NO_3^- will remain oxidized and nitrification of NH_4^+ will proceed. In contrast to most other N transformation processes, autotrophic nitrification is extremely pH-sensitive. Below pH 5.5, NH_4^+ oxidation will not occur, and above pH 8 NO_2^- oxidation slows, because NH_3 inhibits NO_2^--oxidizing bacteria. Ammonification and urease activity can rapidly increase the NH_3 concentration in soil sufficiently to prevent NO_2^- oxidation and thereby cause brief NO_2^- accumulation. Plant toxicity can result.

In grassland and prairie environments, autotrophic nitrification is often minimal. Although it has been suggested that this is due to allelopathic effects of the native vegetation, the most likely explanation is

Table 2 Inhibition of NO_3^- assimilation by NH_4^+

NO_3^- concentration ($\mu g\ NO_3^-$-N ml^{-1})	*N source*	*Inhibition (%)*	*Response time (min)*
2	NH_4^+	55	1
	Arginine	51	6
	Asparagine	50	1
	Glutamine	48	4
	Glutamate	22	nd
	Aspartate	21	nd
	Alanine	8	nd
	Glycine	1	nd
10	NH_4^+	79	5
	Glutamine	72	3

nd, not detectable.

Source: Rice CW and Tiedje JM (1989) Regulation of nitrate assimilation by ammonium in soils and in isolated soil microorganisms. *Soil Biology and Biochemistry* 21: 597–602.

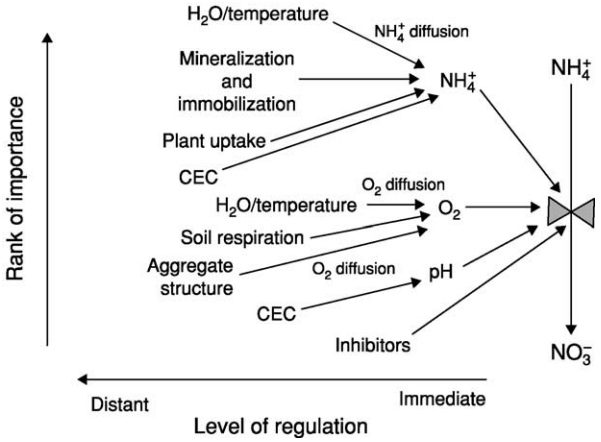


Figure 3 Regulators of nitrification in the soil environment. CEC, cation exchange capacity.

that intense competition from plant roots and heterotrophic microorganisms deprives nitrifiers of the NH_4^+ that they require for growth.

Nitrification slows and ceases as soils become increasingly saturated. The dominant fate of NO_3^- in most waterlogged soil environments is denitrification. Oxygen concentration is the overwhelming environmental factor controlling denitrification and DNRA (Figure 4). Limited C and NO_3^- availability will also slow denitrification and affect whether incompletely reduced intermediates such as NO and N_2O accumulate.

Environmental control over whether denitrification or DNRA predominates in an environment appears to be governed by the extent and duration of anoxia and by the relative proportions of soluble C and NO_3^- (Figure 5). Denitrifiers are more competitive when C is limiting because they obtain more energy through anaerobic respiration with NO_3^- than do NH_4^+ dissimilators (1121 versus 600 kJ mol^{-1} of NO_3^- reduced).

Nitrogen fixation is an energy-demanding process. Consequently, when N is available in the form of inorganic N or readily mineralizable organic N, the synthesis and activity of nitrogenase are suppressed. High soil N concentrations will also suppress the establishment of symbiotic N_2-fixing relationships between bacteria and plants.

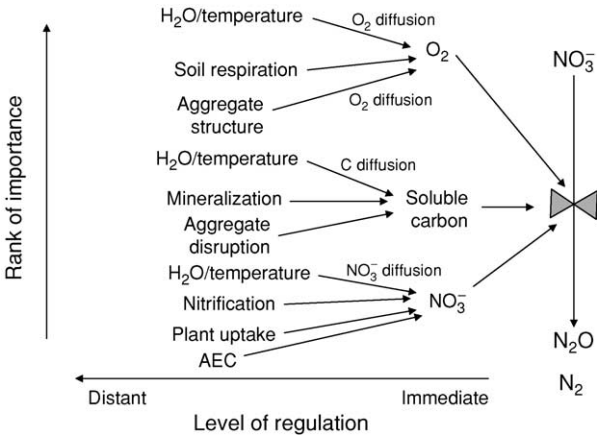


Figure 4 Regulators of NO_3^- reduction in the soil environment. AEC, anion exchange capacity.

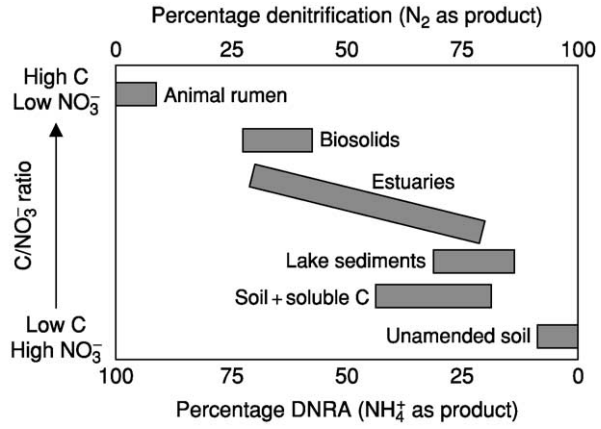


Figure 5 Environmental control of the partitioning between denitrification and dissimilatory nitrate reduction to ammonium (DNRA). Adapted from Tiedje JM, Sexstone AJ, Myrold DD, and Robinson AJ (1982) Denitrification: ecological niches, competition, and survival. *Antonie van Leeuwenhoek* 48: 569–583.

Chemical Regulators of N Transformations

Chemicals have been developed that can effectively regulate (inhibit) two important N transformations on an agricultural field scale – nitrification and urea hydrolysis. A nitrification inhibitor slows the rate of autotrophic nitrification by selectively suppressing *Nitrosomonas* bacteria in the soil (eqn [5]). A urease inhibitor functions by interfering with the enzymatic activity of urease in catalyzing the hydrolysis of urea (eqn [2]). The use of these chemicals will be discussed later as part of management practices to decrease N losses.

Routes of N Loss From the Environment

The N transformations in soil inevitably result in intermediate or final products that can be lost from the soil environment through gaseous diffusion, leaching, or surface erosion (Figure 6). The route of N loss greatly depends on the type of transformation – whether oxidative or reductive.

Reductive Losses

The major reductive losses of N occur during denitrification, which produces N oxides such as NO and N_2O as intermediates and N_2 as a final product. All three compounds are gases that can be lost to the atmosphere. Nitrate reduction to NH_4^+ seemingly preserves N from loss; however, the NH_4^+ formed can still be leached in soils with low cation exchange capacities or lost through erosion.

Oxidative Losses

Autotrophic nitrification frequently results in the production of NO and N_2O as gaseous intermediates. The major concern with nitrification is the production of NO_3^-, which is readily leached from soil environments.

Volatilization

Ammonium exists in chemical equilibrium with NH_3. The pK for this equilibrium is approximately pH 9.5. Consequently, when the soil pH is sufficiently high, or

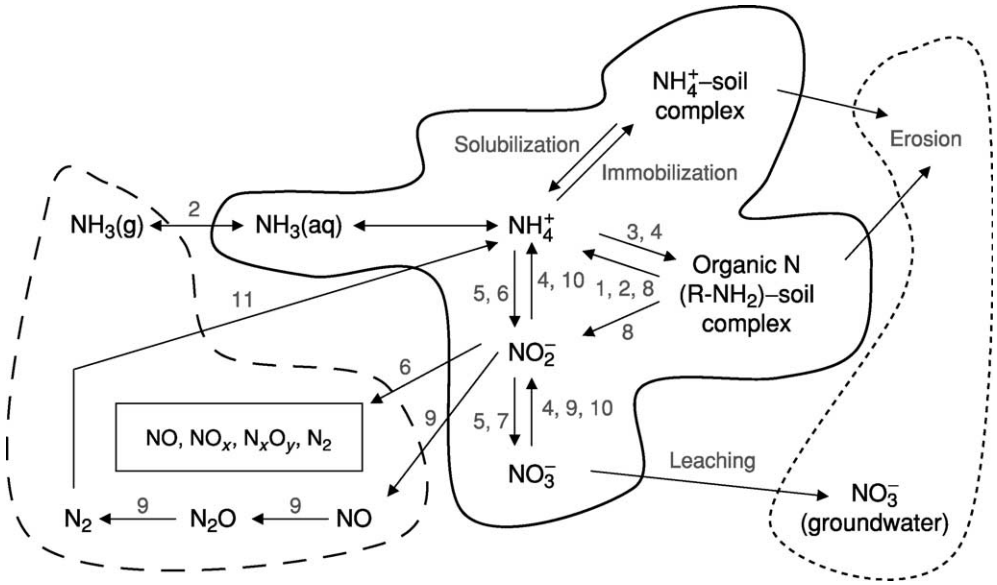


Figure 6 Paths of nitrogen transformation and loss from soil. Numbered pathways correspond to the chemical equations in the text.

Table 3 Some patented nitrification inhibitors

Chemical	*Common name*	*Developer*	*Inhibition by day 14 (%)*
2-Chloro-6-(trichloromethyl)pyridine	N-serve	Dow Chemical	82
4-Amino-1,2,4-6-triazole-HCl	ATC	Ishihada Industries	78
2,4-Diamino-6-trichloromethyltriazine	CL-1580	American Cyanamid	65
Dicyandiamide	DCD	Showa Denko	53
Thiourea	TU	Nitto Ryuso	41
1-Mercapto-1,2,4-triazole	MT	Nippon	32
2-Amino-4-chloro-6-methylpyrimidine	AM	Mitsui Toatsu	31

Adapted from Paul EA and Clark FE (1989) *Soil Microbiology and Biochemistry*. San Diego, CA: Academic Press with permission.

if NH_4^+ production is sufficiently rapid, aqueous NH_3 will be volatilized from soil and lost to the atmosphere as a gas. Animal manures, which contain readily mineralizable forms of organic N such as urea, can produce considerable amounts of volatile NH_3 due to this process.

Interaction of N with Soil Organic Matter

A significant fate of inorganic N is its interaction with soil organic matter. Although this can retard N loss from the environment, it also temporarily renders part of the inorganic N unavailable. Studies in California and Iowa have demonstrated that much of the inorganic NO_3^- and NH_4^+ added to soil can almost immediately become immobilized. Nitrification-mediated immobilization involves the reaction of NO_2^-, formed during the first step of nitrification, with soil organic matter at acidic and neutral pH to form nitrosophenols. Nitrite can also react with amino acids and other forms of reduced organic N to liberate gases as illustrated in eqns [12] and [13]:

$$NH_3 + HNO_2 \longrightarrow NH_4NO_2 \longrightarrow 2H_2O + N_2 \quad [12]$$

$$RNH_2 + HNO_2 \longrightarrow ROH + H_2O + N_2 \quad [13]$$

Management Practices to Decrease N Losses

Figure 6 shows the transformation pathways that potentially result in losses from the N cycle. Some of these losses, including soil erosion, NO_3^- leaching, denitrification, and NH_3 volatilization, can be managed to varying extents. Inhibiting nitrification and urease activity affects all of these pathways except erosion.

Nitrification inhibitors Several chemicals have been shown to inhibit nitrification (Table 3), but few have been proven agronomically and economically effective. The first nitrification inhibitor to be approved by the US Environmental Protection Agency was nitrapyrin, manufactured by Dow Chemical Company

and marketed as N-Serve. Nitrapyrin is 2-chloro-6-(trichloromethyl) pyridine. It may be used with any NH_4^+-N fertilizer, including anhydrous ammonia, urea, urea–ammonium nitrate solutions (UAN), ammonium nitrate, ammonium sulfate, and animal manures.

Nitrapyrin has a relatively high vapor pressure (2.8×10^{-3} mmHg at 23°C). Because of its tendency to volatilize, the manufacturer initially recommended incorporation into soil immediately after application. Research in Kentucky showed that nitrapyrin was effective when sprayed directly on to N fertilizer particles, which were then surface-applied for no-tillage corn. The researchers believed that the immediate proximity of nitrapyrin and NH_4^+ ions together at the sites of nitrification facilitated its effectiveness, even when surface-applied.

Dicyandiamide (DCD), also known as cyanoquanidine, is a water-soluble, organic amide that is both a slow-release N source and a nitrification inhibitor. The composition of DCD is $C_2N_4H_4$, and it contains at least 65% N. It is a fine-grained, nonhygroscopic, crystalline powder, which is easily incorporated into solid, solution, or suspension fertilizers.

Widespread research on nitrapyrin and DCD shows generally that both are very effective in suppressing nitrification in soil for several weeks after application. The ratio of NH_4^+ to NO_3^- is usually much higher in soils, and NO_3^- is lower in drainage water from soils following application of ammonical fertilizers treated with nitrapyrin or DCD than the same applications without a nitrification inhibitor. A crop yield response to a nitrification inhibitor, however, occurs only if N was lost by leaching or denitrification and then only if those losses resulted in N deficiency sufficient to reduce yields without the nitrification inhibitor. Rarely are yield increases observed if nitrification inhibitors are used with much higher than optimum application rates of N. The environmental benefits of decreasing NO_3^- leaching with a nitrification inhibitor would none the less be present and even more important where an excess of N was applied.

Urease inhibitors As in the case of nitrification inhibitors, several chemicals inhibit urease activity. Because mechanical incorporation is usually not practical, a urease inhibitor must suppress hydrolysis of broadcast-applied urea to the surfaces of soil and plant residues for no-tillage agriculture, or growing plants in pastures, forage fields, or lawns, and the effect must last long enough to allow rainfall to leach the urea molecules into the soil, where hydrolysis can occur with little risk of NH_3 volatilization loss from most soils.

Table 4 Corn yield response to urease inhibitors with fertilizers surface applied for no-tillage corn

	Grain yields (Mg ha^{-1})	
Fertilizer[a]	*Pennsylvania*[b]	*Kentucky*[c]
UAN	6.96	5.14
UAN + ATS	7.03 NS	–
UAN + NBPT	7.40*	5.52*
Urea	6.65	5.27
Urea + NBPT	7.53**	6.15**
Ammonium nitrate	7.59	5.96

[a]UAN, urea ammonium nitrate solution; ATS, ammonium thiosulfate; NBPT, *N*-(*n*-butyl) thiophosphoric triamide.
[b]From Fox and Piekielek (1993).
[c]Unpublished data, Department of Agronomy, University of Kentucky.
NS, not significantly different ($P \geq 0.10$) by LSD.
Significantly different by LSD: *$P \leq 0.10$, **$P \leq 0.01$, respectively.

In many research studies (examples shown in Table 4), one urease inhibitor has been shown to be particularly efficacious for urea fertilizer management. Allied Chemical Corporation patented *N*-(*n*-butyl)thiophosphoric triamide (NBPT) as a urease inhibitor in 1985. Today, NBPT is marketed by Agrotain International as a technology entity that includes a practical means of treating solid urea or UAN solution with NBPT, or NBPT and DCD, at a fertilizer-blending facility and delivering the product in the field as a fertilizer application. The combination of NBPT and DCD provides urease and nitrification inhibition for a urea application.

Environmental and Health Consequences of N Cycling

The principal environmental and health concerns associated with N cycling in soil are the potential for trace gas production and its effect on global warming and ozone depletion, the potential for nutrient loss, and the movement of N to environments where it can cause eutrophication and potential health problems.

Global Warming and Ozone Destruction

Nitrous oxide and nitric oxide are trace gases formed during nitrification and denitrification transformations of N in soil. Molecule for molecule, N_2O is 150 times more effective at adsorbing thermal radiation than is CO_2, the major greenhouse gas. Nitric oxide plays a well-known role in the destruction of stratospheric ozone as illustrated in the following reactions:

$$O_3 + h\nu \rightarrow O_2 + O\cdot \quad [14]$$

$$N_2O + O\cdot \rightarrow 2NO \quad [15]$$

$$NO + O_3 \rightarrow NO_2 + O_2 \quad [16]$$

$$NO_2 + O\cdot \rightarrow NO + O_2 \quad [17]$$

Nutrient Losses

Denitrification and NO_3^- leaching have been associated with the loss of up to 70% of the inorganic N in some fertilized fields. Ammonia volatilization occurring during hydrolysis of urea (eqn [2]) can result in losses of 10–50% of the N from surface-applied urea fertilizers. Ammonia volatilization due to urease activity is given as one explanation for the loss of 3–50% of the N in unincorporated manures such as poultry litter applied to neutral soils. Ammonia in high enough concentrations is a toxic gas, and its production in confined animal feeding operations (CAFOs) is a major concern. Unfortunately, venting such gases into the atmosphere is problematic for two reasons. First, the volatile N gases are odoriferous and generate nuisance complaints. Second, the volatile gases eventually return to soil or water as wet and dry deposition and can contribute to eutrophication in pristine environments.

Groundwater Quality

Groundwater is the source of drinking water for more than half of the USA, and in rural areas it represents more than 95% of the water used for domestic purposes. Nitrate-N concentrations in streams and groundwater are typically less than $2\,mg\,l^{-1}$ but have been steadily increasing in the USA and Europe. In the USA, for example, well-drained soils receiving high N inputs have median NO_3^--N concentrations approaching $5\,mg\,l^{-1}$ and up to 26% of sample wells have levels exceeding $10\,mg\,NO_3^-N\,l^{-1}$, which is the US Environmental Protection Agency (USEPA) and Public Health Service drinking water standard. Many drinking water systems in Europe currently exceed the EC limit of $11.3\,mg\,l^{-1}$.

The consequences of chronic consumption of NO_3^--contaminated drinking water and vegetation can be severe and include methemoglobinemia (blue-baby syndrome) and increased potential for stomach cancer due to nitrosamine formation.

The consequences of nitrification and subsequent NO_3^- leaching also have severe ecological consequences. The inability of growing plants to capture completely the N applied in fertilizer, and natural mineralization and nitrification in undisturbed soils opened to agriculture, grazing, and irrigation, has caused N pollution problems from the Chesapeake Bay in the eastern USA to the Baltic. The Gulf of Mexico, for example, receives drainage from approximately 41% of the continental USA via the Mississippi River and its tributaries. Due to nutrient leaching from agricultural watersheds, a hypoxic zone some 7000 square miles in size (approximately the same area as the state of New Jersey), which has significantly affected the shrimp and shellfish industries, has developed in the Gulf of Mexico off the Louisiana coast.

Conclusion

The N cycle in soil is dynamic and inherently leaky. Off-site losses of N gases, soluble ions, and organic compounds are virtually impossible to eliminate completely. Nevertheless, there are several management practices for erosion control and fertilizer use that can minimize N losses from soil, and the cycle itself can be used to preserve environmental quality. For example, biological N_2-fixation rather than commercial fertilization will generally retain more soil N. Nitrates and nitrites in groundwater have been remediated through denitrification on increasingly larger scales to preserve groundwater quality and public health.

List of Technical Nomenclature

Anoxia	Conditions in which O_2 is absent. Synonymous with 'anaerobic'
Autotrophy	Growth using inorganic elements as C and energy sources
Biomass	Living organisms in soil, whether plants, animals, or microorganisms
Eutrophic	Excessive growth typically associated with aquatic systems in which one or more limiting nutrients such as N or P is provided
Heterotrophy	Growth using organic C as the source of C for biomass production and for energy
Hydrolysis	Addition of water to chemical bonds during chemical reactions
Inhibition	Suppression or slowing of a process for an indefinite time, usually a few weeks

See also: **Enzymes in Soils**; **Greenhouse Gas Emissions**; **Microbial Processes:** Environmental Factors; **Nutrient Management**

Further Reading

Dixon JB and Weed SB (eds) (1989) *Minerals in Soil Environments*, 2nd edn. Soil Science Society of America Book Series 1. Madison, WI: Soil Science Society of America.

Fenchel T, King GM, and Blackburn TH (1998) *Bacterial Biogeochemistry*, 2nd edn. San Diego, CA: Academic Press.
Follett RF, Keeney DR, and Cruse RM (1991) *Managing Nitrogen for Groundwater Quality and Farm Profitability*. Madison, WI: Soil Science Society of America.
Fox RH and Piekielek WP (1993) Management and urease inhibitor effects on nitrogen use efficiency in no-till corn. *Journal of Production Agriculture* 6: 195–200.
Paul EA and Clark FE (1989) *Soil Microbiology and Biochemistry*. San Diego, CA: Academic Press.
Pearce RC, Zhan Y, and Coyne MS (1998) Nitrogen transformations in the tobacco float system. *Tobacco Science* 42: 82–88.
Rice CW and Tiedje JM (1989) Regulation of nitrate assimilation by ammonium in soils and in isolated soil microorganisms. *Soil Biology and Biochemistry* 21: 597–602.
Schlesinger WH (1997) *Biogeochemistry: an Analysis of Global Change*, 2nd edn. San Diego, CA: Academic Press.
Spalding RF and Exner ME (1993) Occurrence of nitrate in groundwater – a review. *Journal of Environmental Quality* 22: 392–402.
Stevenson FJ (ed.) (1982) *Nitrogen in Agricultural Soils*. Agronomy Monograph 22. Madison, WI: American Society of Agronomy.
Stevenson FJ and Cole MA (1999) *Cycles of Soil*, 2nd edn. New York: John Wiley.
Thorn KA and Mikita MA (2000) Nitrite fixation by humic substances: nitrogen-15 nuclear magnetic resonance evidence for potential intermediates in chemodenitrification. *Soil Science Society of America Journal* 64: 568–582.
Tiedje JM, Sexstone AJ, Myrold DD, and Robinson AJ (1982) Denitrification: ecological niches, competition, and survival. *Antonie van Leeuwenhoek* 48: 569–583.
US Environmental Protection Agency (1995) *Drinking Water Regulations and Health Advisories*. Washington, DC: Office of Water, USEPA.
Zumft WG (1997) Cell biology and molecular basis of denitrification. *Microbiology and Molecular Biology Reviews* 61: 533–616.

Nitrates

D S Powlson and T M Addiscott, Rothamsted Research, Harpenden, UK

© 2005, Elsevier Ltd. All Rights Reserved.

Under aerobic conditions nitrate (NO_3^-) is the final form of nitrogen produced in soil following the decomposition of soil organic matter or any added plant or animal material. It is the main chemical form in which nitrogen is taken up by plants, though most plants can also absorb N in the ammonium form (NH_4^+). Figure 1 shows the central position of soil nitrate within the nitrogen cycle. Nitrate is a natural product that is almost always present in soil. In natural or seminatural ecosystems nitrogen is usually cycled very efficiently between soil microbes, soil solution, and plants so the concentration of nitrate in soil at any time is normally low and losses are small. By contrast, in order to grow large yields of agricultural crops, nitrogen is required in large quantities. For example, winter wheat growing under UK conditions would typically take up about 200 kg N ha^{-1} and most agricultural crops, other than those grown under subsistence farming conditions, take up between 100 and 200 kg N ha^{-1}. Various practices are adopted in order to provide sufficient N and the use of inorganic fertilizers is particularly important, with almost 80 million tonnes of N being used annually worldwide (Table 1). Inorganic fertilizers either contain nitrate or forms of N such as ammonium or urea that are rapidly converted to nitrate by soil microorganisms. Alternatively N may be supplied by adding animal manure or by incorporating residues of plants rich in nitrogen, such as N-fixing legumes, which may have been grown as the previous crop, as a green manure or as part of a pasture which is then plowed up. In all cases, except in anaerobic conditions such as flooded rice, much of the N from these organic sources will be converted to nitrate before being absorbed by the crop. It is highly desirable to utilize N from organic materials but it is mistaken to assume that losses of nitrate are necessarily less than with inorganic fertilizers as the production of nitrate from organic sources is determined by the activity of soil microorganisms and not by crop demand – some nitrate is produced at times when crop uptake is small and the risk of nitrate loss is large.

The nitrate ion is extremely soluble in water and can easily be washed out of soil to natural waters. There are two main pathways by which nitrate ions dissolved in soil water are lost through the movement of water: (1) leaching – the vertical movement of water and its solutes through the soil profile to either groundwater or field drains; and (2) surface runoff – the horizontal movement of water and solutes over the soil surface until it reaches a river or lake. Surface runoff occurs where the soil is either compacted or is naturally poorly draining, e.g., due to high clay content. A variation on surface runoff is sometimes termed subsurface flow, in which water infiltrates the soil to some extent but then moves horizontally due to impeded drainage, for example due to compaction at the base of the plow layer in cultivated soil or because of a change in soil physical properties with depth.

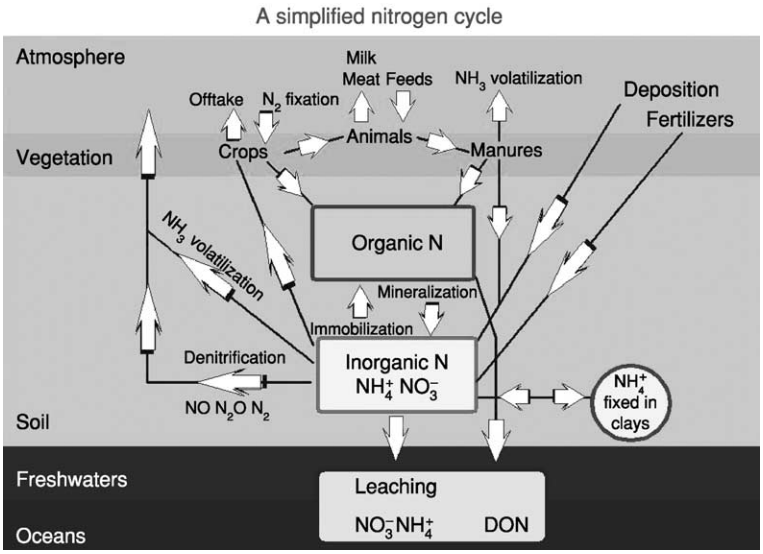


Figure 1 Soil nitrate within the nitrogen cycle. Reproduced with permission from Davies DB (2000) *Soil Use and Management* 16: 142–144.

Table 1 World consumption in 1994/1995 of nitrogen in various types of fertilizer

Type of fertilizer	*N consumed globally (million tons)*	*% of world total N usage*
Ammonium nitrate (AN)	6.58	9.0
Calcium ammonium nitrate (CAN)	3.69	5.1
Urea	31.57	43.3
Ammonium sulfate	2.37	3.2
Injected ammonia gas (anhydrous ammonia)	4.20	5.8
N solutions	3.79	5.2
Other 'straights' (i.e., N not mixed with other nutrients)	9.05	12.4
N in compound fertilizers	11.68	16.0
Total	72.93	100.0

Source: *World Fertilizer Consumption Statistics* (1997), No. 28, p. 17. Paris: International Fertilizer Industry Association.

Compared to natural ecosystems, the risk of nitrate loss through leaching, surface runoff, or denitrification is much greater in agricultural soils because: (1) nitrate concentration in soil is high for some periods of the year; and (2) at least with arable crops, soil is bare for part of the year so there is no removal of nitrate by plants. Any loss of N from agriculture is an economic loss for the farmer and a waste of a valuable resource for society, so much research over the last century has been directed at minimizing losses. Since about the 1970s issues of water pollution (some genuine and some not) have led to regulations and policies to limit the concentration of nitrate in water and this has become a main driver of research and a major influence on farming practice. In the European Union a limit of $50\,mg\,l^{-1}$ of nitrate (equivalent to $11.3\,mg\,l^{-1}$ of nitrate-N) has been imposed on any ground- or surface-water to be used as a source of drinking water and any surface water where eutrophication is considered to be a risk. In the USA the limit is $45\,mg\,l^{-1}$ of nitrate, equivalent to $10\,mg\,l^{-1}$ of nitrate-N. These limits have a significant influence on agricultural practices: for example, under the climatic and agronomic conditions of the UK, annual leaching losses from arable soils are rarely less than $20\,kg\,N\,ha^{-1}$ and often twice this. But in the drier eastern regions, which are important for arable cropping, excess winter rainfall is often only 200 mm or less: under these conditions a nitrate loss of only $20\,kg\,N\,ha^{-1}$ will cause the nitrate concentration limit in water to be exceeded.

In addition to issues of water pollution, nitrate can also be involved in the emission of environmentally damaging gases to the atmosphere. Under anaerobic conditions nitrate can be microbially reduced to the gas nitrous oxide (N_2O) which is a powerful greenhouse gas and also has a role in the destruction of ozone in the stratosphere; this provides a good reason for minimizing the amount of excess nitrate in the environment.

Chemical Structure and Physical Properties

The nitrate ion, NO_3^-, has a symmetrical planar trigonal structure. It is negatively charged and so is termed an anion. Salts formed by the nitrate ion with cations (e.g., Ca^{2+}, K^+, Na^+) are generally soluble and, in soil, are always present in solution. These salts are dissociated so it is the nitrate ion that is the entity of interest. The practice of referring to nitrates in soil or water is incorrect – nitrate is a unique species, not a plural.

Being an anion, nitrate is attracted to positively charged surfaces. Some acid soils contain positive surface charge that can retain nitrate ions, but this is the exception. In the vast majority of soil worldwide, especially those used for agriculture, nitrate is not retained in this way and remains in soil solution and moves through the processes of mass flow or diffusion.

Biological Transformations in Soil

Numerous organisms in soil are involved in the breakdown of organic matter to inorganic (mineral) materials, the process termed mineralization. During aerobic mineralization, C in organic materials is converted to CO_2 and N is initially converted to ammonium; this step of N transformation is termed ammonification and is carried out by a wide range of heterotrophic organisms including various fungi and bacteria. A much smaller range of autotrophic bacteria are responsible for the conversion of ammonium to nitrate via a two-step process, together termed nitrification. The first step is conversion of ammonium to nitrite; classically this is described as being mediated by bacteria of the *Nitrosomonas* species, but it is now known that the less-studied species *Nitrosolobus*, *Nitrosospira*, and *Nitrosococcus* are at least as important:

$$2NH_4^+ + 3O_2 \rightarrow 2NO_2^- + 4H^+ + 2H_2O + \text{energy}$$

Nitrite rarely accumulates in soil, which is fortunate as it is toxic to plants and many microbes. It is rapidly oxidized in aerobic soils to nitrate by *Nitrobacter* species:

$$2NO_2^- + O_2 \rightarrow 2NO_3^- + \text{energy}$$

Although the nitrifying bacteria in soil cover a narrow group of genera, they are ubiquitous, so it can be assumed that any ammonium (whether from added fertilizer or from mineralization) will be converted to nitrate within a period of days to weeks provided the soil is aerobic and conditions of temperature, moisture, and pH are conducive for general microbial activity. The first step of nitrification produces H^+ ions and is thus an acidifying reaction. This is illustrated in Figure 2, which shows changes in soil pH over a period of 150 years in a long-term grassland experiment at Rothamsted, UK. In the treatment where N fertilizer is applied as ammonium sulfate, pH decreased from 5.7 to 3.7. In the treatment receiving the same quantity of N in the nitrate form there was little change in soil pH.

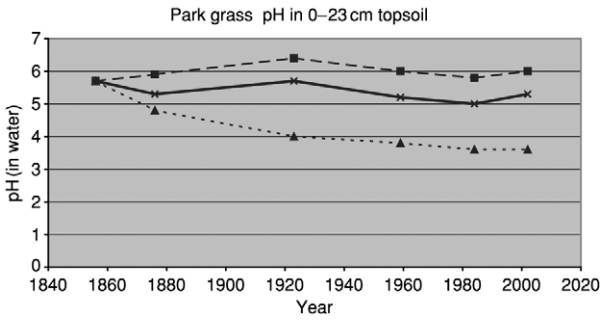


Figure 2 Influence of the long-term application of ammonium sulfate fertilizer on soil pH in the Park Grass grassland experiment at Rothamsted, UK. Plots received N as either ammonium sulfate or sodium nitrate every year since 1856. Nitrification of ammonium in the ammonium sulfate treatment caused an acidification of soil but application of N already in the nitrate form did not. × Control; ■ sodium nitrate; ▲ ammonium sulfate.

Nitrate in soil can be destroyed by the action of denitrifying bacteria when oxygen is in short supply. Under anaerobic conditions the nitrate ion can be used instead of O_2 as the acceptor of free electrons produced during respiration. The products are a mixture of nitrous oxide (N_2O) and nitrogen gas (N_2) – the sequence can be represented as:

$$NO_3^- \rightarrow NO_2^- \rightarrow N_2O \rightarrow N_2$$

The extent to which the reduction proceeds is strongly influenced by the pH and temperature of the soil. At a low temperature and a pH less than 5, the ratio of N_2O to N_2 is about 1, but at a temperature of 25°C or more and a pH greater than 6, most of the N_2O is reduced to N_2. Soils are not usually uniform and even in mainly aerobic soils there are often small localized anaerobic zones in which denitrification occurs. This spatial variability, together with the variable ratio of N_2O to N_2, makes denitrification very difficult to measure. It also implies that nitrification and denitrification may be occurring simultaneously close to each other in the soil. Nitrous oxide is also produced to some extent as a by-product of nitrification (Figure 3) and it is often difficult to distinguish between the sources. The environmental impacts of nitrous oxide are mentioned later.

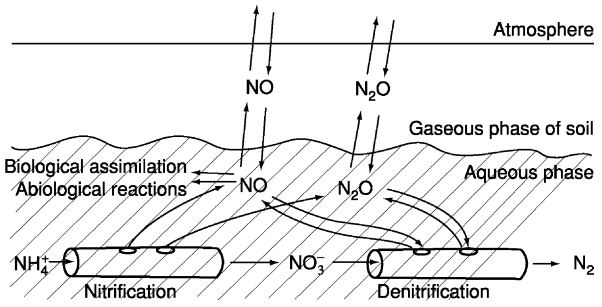


Figure 3 Formation of nitrous oxide (N_2O) and nitric oxide (NO) through denitrification and nitrification. Reproduced with permission from Davidson EA (1991) Fluxes of nitrous oxide and nitric oxide from terrestrial ecosystems. In: Rogers JE and Whitman WB (eds) *Microbial Production and Consumption of Greenhouse Gases: Methane, Nitrogen Oxides, Halomethanes*, pp. 219–235. American Society for Microbiology.

Forms of Nitrogen Fertilizer

In Europe ammonium nitrate or calcium ammonium nitrate (CAN) is widely used. These forms supply 50% of their N as pure nitrate and so instantaneously increase the quantity of nitrate in topsoil from a very low value, typically about 10 kg nitrate-N ha^{-1}, to 50 kg nitrate-N ha^{-1} or more depending on the amount applied. A typical application of N to an arable crop in spring, at the beginning of the rapid crop growth period, might be 100–200 kg N ha^{-1}, but usually divided into two or more portions. The ammonium N is typically nitrified to nitrate over a period of 1–4 weeks. Worldwide, urea is by far the most commonly used form of nitrogen fertilizer. Before becoming available for plant uptake, the N in urea first has to be converted to ammonium (the process of urea hydrolysis) but this occurs within a few days. The ammonium is then nitrified within 2–3 weeks. Thus, in practice, the difference between urea and ammonium nitrate or CAN in terms of impact on soil nitrate is fairly small. These transformations can be studied by labeling the N in urea or either the ammonium or nitrate parts of ammonium nitrate with ^{15}N, the heavy isotope of nitrogen. In some experiments in northern France in which ^{15}N-labeled fertilizers were applied to field plots of winter wheat in April, intact urea molecules were virtually undetectable in soil within a week of application. All of the N derived from urea was either converted to ammonium or nitrate, absorbed by the crop, or immobilized by soil microorganisms within 2 weeks of application. Where ammonium nitrate was applied, the majority of the ammonium had been converted to nitrate or absorbed by plants or microorganisms within 2 weeks. For applications in March, when soil temperatures were lower, the transformations were slower but the trends the same.

A small but potentially significant difference between urea and nitrate-containing fertilizers is that the peak concentration of nitrate in soil is likely to be less with urea. Thus urea may give a decreased risk of nitrate loss by leaching or denitrification shortly after application. However, with urea some N may be lost as ammonia gas under some circumstances and slightly more may be absorbed by microorganisms as they tend to take up ammonium rather than nitrate if both are available. Overall agronomic experiments show only small differences in crop utilization of N from urea compared with nitrate-containing fertilizers.

A number of additives are available commercially that retard either urea hydrolysis or nitrification, so retaining a greater proportion of N from fertilizer in the ammonium form and decreasing the amount as nitrate. In principle these additives (urease inhibitors or nitrification inhibitors) should decrease the losses associated with nitrate by decreasing the quantity in soil. Whilst this can be useful for nitrate coming from fertilizer, the impact of inhibitors on overall nitrate loss is likely to be small because it is nitrate derived from the mineralization of soil organic matter that is usually the main source. Also, until now, their high cost has precluded their use from normal agricultural crops; they are used to some extent to make fertilizer N slowly available to certain high-value horticultural crops.

Management of Nitrate in Agriculture

Loss of Nitrate from Fertilizer During the Crop-Growing Period

The period immediately after N fertilizer application is a time of high risk of nitrate loss. During this period nitrate concentration in soil is high as a result of the added fertilizer and the mineralization and nitrification of N derived from soil organic matter. The concentration rapidly declines as crop uptake proceeds. One method of measuring the total loss of N from fertilizer during the growing season of a crop is to apply N fertilizer labeled with ^{15}N, the heavy isotope of nitrogen. In this way N derived from fertilizer can be distinguished from N derived from soil organic matter or other sources in soil as these are predominantly composed of ^{14}N. ^{15}N is a stable isotope, so is not radioactive and can be used in field or laboratory experiments without the need for special safety procedures. Figure 4 shows the results of such an experiment in which about 170 kg N ha^{-1} was applied in spring to autumn-sown wheat in south-east England. The N was applied as ammonium nitrate with both the ammonium and nitrate parts labeled with ^{15}N, i.e. $^{15}NH_4$ $^{15}NO_2$. About two-thirds of the applied ^{15}N-labeled N derived from fertilizer was recovered in the above-ground parts of the crop when it was harvested. A further 20% was retained in the soil, almost entirely as organic forms as this N had

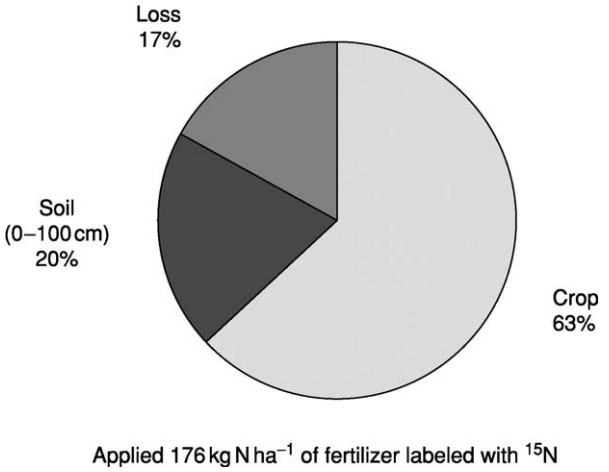


Figure 4 Fate of ^{15}N-labeled fertilizer (double-labelled ammonium nitrate, $^{15}NH_4$ $^{15}NH_3$) applied to winter wheat in south-east England. N was applied as a single dressing in spring (April) to a healthy crop of winter wheat, sown the previous autumn. The crop and soil were sampled at the time of harvest in August.

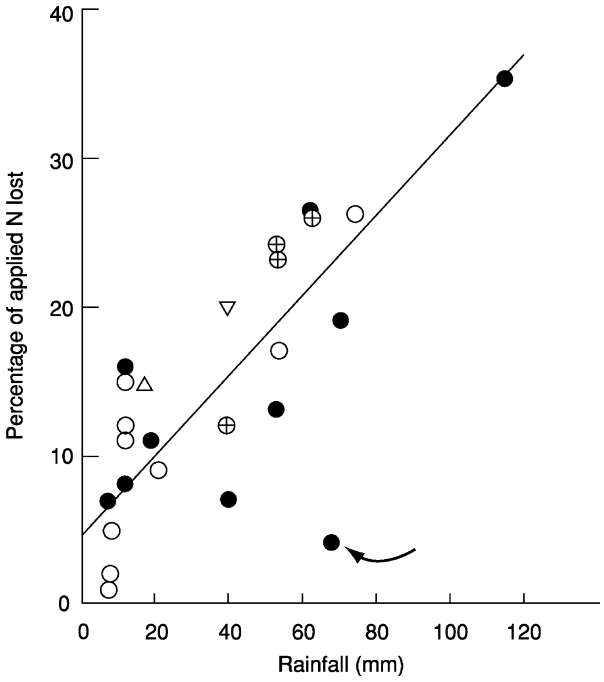


Figure 5 Relationship between loss of ^{15}N-labeled fertilizer applied to crops in spring and rainfall in the 3-week period following application. All circles (●, ⊕, ○) refer to winter wheat crops. Different symbols indicate crops receiving rates of fertilizer N known to be above (⊕) or below (○) that giving maximum grain yield. All other wheat crops indicated by ●. Arrowed result excluded from regression. Results included for experiments with forage maize (▽) and canola (oilseed rape, △) but used in calculating regression: Percentage of applied N lost = 5.0 + 0.264 (rainfall in 3 weeks after N application, mm). Reproduced with permission from Powlson DS, Hart PBS, Poulton PR, Johnston AE and Jenkinson DS (1992) Influence of soil type, crop management and weather on the recovery of 15N-labelled fertilizer applied to winter wheat in spring. *Journal of Agricultural Science, Cambridge* 118: 83–100.

been incorporated into wheat roots, soil organisms, and soil organic matter – this can also be seen in Figure 6, later. About 17% of the applied N had been lost – it could not be found in the crop or in soil to a depth of 1 m. This result is typical for healthy well-managed winter wheat growing in a temperate climate. In other experiments in similar environments losses ranged from less than 5% of applied fertilizer N to a maximum of 35%. Other studies indicated that in this environment dentrification was the major and most common cause of loss; but in years when rainfall was particularly heavy soon after N fertilizer application nitrate leaching became more important, as expected, especially in sandy soils. Taking together results from a series of similar field experiments in south-east England, there was a correlation between the percentage of the applied fertilizer N lost during the growing season and the total quantity of rain falling in the 3-week period following fertilizer application (Figure 5). This emphasizes the importance of the vulnerable period for loss when much nitrate is present in soil because crop uptake is not yet complete. It has also been found that any factors such as disease or weeds that inhibit crop growth (and uptake of N) tend to increase N losses from soil.

Although losses of fertilizer N can be relatively low in some situations, as described above, this is not universally the case. There are many examples of much lower recoveries, implying losses of 50% or sometimes even more. In some cases this is because of high rainfall in the period soon after N application – if this is an occasional event there is little that can be done to avoid it other than for the farmer to take note of local weather forecasts and for meteorologists to improve these as much as possible. There are, however, some measures that can be taken to decrease the risk of loss. For some crops and growing conditions it is common to apply N fertilizer at the time of sowing the crop, or even before. This means that nitrate is present in soil for a prolonged period as uptake by the crop will not start until seeds have germinated and the plants reached a certain size; depending on the crop and climatic region this could be as long as 2 months, allowing considerable opportunity for losses by leaching or denitrification unless it is a region where rainfall is very low at that time. Eliminating N applications at a time when the crop is not growing rapidly is one means of decreasing N losses and increasing the efficiency of use of nitrate derived from fertilizers.

An example of inappropriate timing of N application is shown in Table 2. In these experiments ^{15}N-labeled potassium nitrate was applied to winter wheat at the time of sowing the crop in autumn. Under the climatic conditions of north-west Europe

Table 2 Fate of ^{15}N-labeled fertilizer applied to winter wheat at sowing time in autumn in south-east England

Soil type	*Loss of autumn-applied ^{15}N at the time of harvest (%)*
Silty clay loam	46
Silty clay loam	63
Silty clay loam	60
Sandy clay loam	39
Sandy loam	78

wheat grows to a limited extent in autumn and early winter, typically taking up about 20 kg N ha^{-1}, and then no more until the main growth period in spring. Thus, in most conditions, little of the applied nitrate is taken up during the winter period and is exposed to loss, mainly by leaching or surface runoff as low temperature in winter limits denitrification. In maritime climates rainfall in winter exceeds evapotranspiration so water can move through the soil and cause leaching of nitrate, either continuously or intermittently. In continental climates soil is frozen during much of the winter so leaching does not occur. When the soil thaws in spring, a large amount of leaching (and/or surface runoff) occurs in a short period but the overall effect on nitrate loss is similar. Table 2 shows that between 39 and 78% of the autumn-applied N was lost from the crop/soil system, almost certainly by nitrate leaching. In the past it was thought that applying fertilizer N at the time of sowing was beneficial for early growth of the crop and 'seedbed' applications were recommended. It is now known that there is no benefit and it simply causes an unnecessary loss of N, and is a waste of money. A similar inappropriate timing of N fertilizer is the situation where N is applied at about sowing time of a crop and is immediately followed by irrigation. In this situation some of the nitrate in soil, not yet taken up by the newly established crop, can be leached below rooting depth or denitrified depending on soil and climatic conditions.

Losses of Nitrate Derived from Soil or Fertilizer Outside the Main Crop Growth Period

Although some losses of fertilizer-derived nitrate do occur during the growing period of agricultural crops, in most environments the largest losses by nitrate leaching or surface runoff occur when crops are not actively growing or are not present at all. The examples discussed above of fertilizer being applied at the time of sowing rather then just before the main growing period illustrate this. It is obviously important to know the source of nitrate in soil during these periods of high risk. It is commonly assumed that it is unused fertilizer but much research has shown this is not usually the case for the major arable crops in

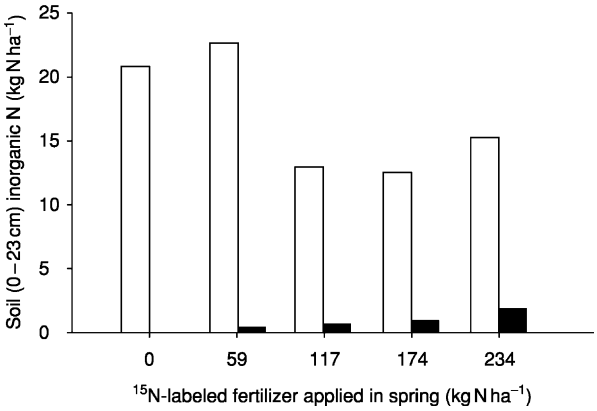


Figure 6 Labeled and unlabeled nitrate in plow-layer soil (0–23 cm) at the time of harvesting winter wheat. A range of rates of fertilizer N, labeled with ^{15}N, was applied in spring. Almost all the nitrate present in soil at the time of harvest was unlabeled, indicating that it was derived from the mineralization of soil organic N, not from fertilizer. Additional nitrate was present in deeper soil layers, almost all unlabeled. This soil-derived nitrate, plus additional nitrate from later mineralization, is the main source of nitrate leaching to water during the subsequent winter. □ unlabeled N; ■ labeled N. Reproduced with permission from Macdonald AL, Powlson DS, Poulton PR, and Jenkinson DS (1989) Unused fertilizer nitrogen in arable soils–its contribution to nitrate leaching. *Journal of the Science of Food and Agriculture* 46: 407–419. © Society of Chemical Industry.

the world or for pasture. Short-season horticultural crops can be exceptions to this generalization as they are often given very large rates of fertilizer N and this can be used inefficiently, leaving significant residues unused in the soil.

Experiments using fertilizer labeled with ^{15}N have been informative in identifying the sources of nitrate in soil in autumn and winter. Figure 6 shows the quantities of nitrate in soil at the time of harvesting a winter wheat crop in England. The fertilizer applied in spring had been labeled with ^{15}N, but the results show that very little of the nitrate left in soil after harvest in late summer was derived from this – the vast majority of the nitrate present at this time, and subject to leaching during the forthcoming winter, was unlabeled because it was derived from the mineralization of organic matter in the soil. Similar results have been obtained from numerous other studies with crops such as wheat, barley, and maize in Europe, North America, and elsewhere.

Winter losses of nitrate by leaching arise because the soil microbes responsible for mineralization and nitrification are active when conditions of temperature and moisture are suitable, which does not always coincide with crop uptake – this is illustrated in Figure 7. N mineralized and nitrified in spring does coincide with the major period of N uptake for most crops but that mineralized in autumn or early winter, before low temperature stops the process, is likely to remain in soil and be exposed to a high risk

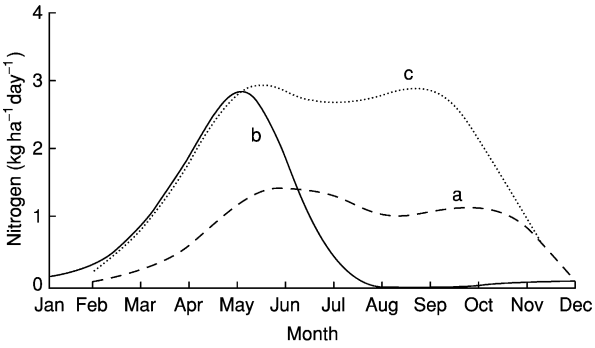


Figure 7 Representation of the time course of (a) mineralization of soil organic N to produce nitrate, (b) N uptake by an autumn sown arable crop and (c) N uptake by grass. — wheat; ····· grass; - - - mineralization. Reproduced from Addiscott TM, Whitmore AP, and Powlson DS (1991) *Farming, Fertilizers and the Nitrate Problem*. CAB International.

of leaching. Leaching may occur intermittently during winter (as in maritime climates where soil is not frozen during winter) or in a shorter period after the spring thaw in continental climates, as described above for autumn-applied fertilizer. Unfortunately, many agricultural practices that are beneficial for crop nutrition and soil quality tend to worsen the problem of winter nitrate loss. For example, additions of organic manures lead to increased production of nitrate at all periods of the year – partly when it will be absorbed by the growing crop (spring and summer) but partly when it will be at high risk of loss due to low crop uptake and the movement of water through soil (autumn and winter). Similarly, plowing up pasture in a mixed grassland/arable rotation such as organic farming brings both benefits and problems. Nitrate produced during the growing season of the subsequent crop substitutes for inorganic fertilizer that would otherwise be needed, but that produced too late for uptake in summer or too early for spring uptake will be subject to leaching.

Winter leaching losses from grass can be less than from arable crops as the plants are present throughout the year and able to absorb nitrate later in the autumn than an arable crop and start earlier in the spring, as the grass will already have a well-established root system. In this sense, grassland is more like a natural ecosystem with plants present throughout the year. However, low leaching is generally only the case with cut grass. In grassland that is grazed by animals, or has slurry applied, mineralization of these organic inputs can lead to nitrate formation in late autumn that is not taken up by the grass and is at risk of being leached.

As a general principle, arable crops that are sown in autumn rather than spring minimize the opportunity for leaching losses during winter – some nitrate is taken up in autumn before leaching begins and roots are already present in spring. With a spring-sown crop, the soil is bare during winter and there is a period of several weeks after sowing in the spring before the crop takes up nitrate in significant quantity. This trend is generally shown when measurements of drainage are made from arable fields or nitrate losses are estimated from measurements of nitrate concentration in soil water made using porous ceramic cups. However, it cannot be assumed that the presence of an autumn-sown crop solves the problem as the extent to which the crop is an effective sink for nitrate is highly dependent on the time of sowing and on weather conditions. A crop sown rather late in autumn and slow in establishing may only absorb $10\,kg\,ha^{-1}$ of nitrate before winter leaching begins and thus have little impact on the amount of nitrate lost. By contrast, a vigorously growing early-sown crop may absorb 30–$50\,kg\,ha^{-1}$ of nitrate and greatly decrease losses. Where spring crops are grown, a management practice sometimes used to decrease nitrate loss is a winter cover crop or 'catch crop.' This is a crop grown specifically to absorb nitrate during the period from late autumn to early spring that is incorporated into the soil before sowing the spring crop. Cover crops can sometimes be very effective at decreasing nitrate leaching, although they do introduce additional agronomic difficulties. Also, as the incorporated plant material is mineralized it is common for rather little of it to become available to the subsequent spring-sown crop – mineralization is usually fairly slow, occurring over several years, and some contributing to nitrate leaching in the next winter.

In a series of experiments using hydrologically isolated plots in central England (the Brimstone Farm experiment), varied crop rotations and management practices were compared for their influence on winter leaching of nitrate. Drains were installed under each plot so that water could be collected and analyzed for nitrate. Losses were greatest when soil was bare prior to a spring-sown crop. They could be decreased in that year by growing a cover crop but this led to increased leaching in later years. The only way to decrease nitrate loss greatly was to abandon arable cropping and convert the land to unfertilized grass. Overall it was concluded that, under the soil and environmental conditions of the site, continuous autumn cereals were the best compromise between decreased nitrate movement to water and profitable agriculture.

Improving the Management of Inorganic Fertilizer

Although it is generally true that nitrate from mineralization of organic matter in soil, or added to soil, is the major source of nitrate losses, there are

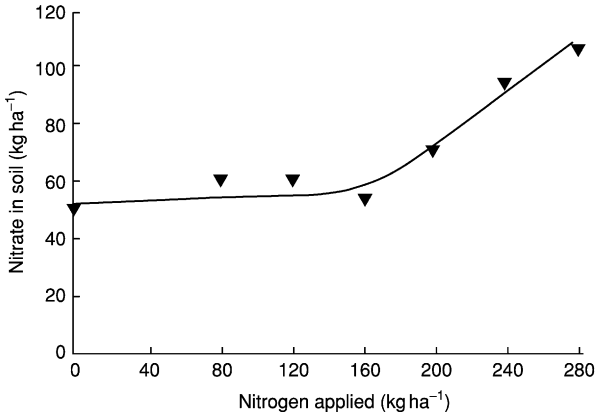


Figure 8 Mean quantities of nitrate in soil to a depth of 90 cm after harvesting winter wheat at eight sites in central and eastern England. Reproduced with permission from Cheney K (1990) Effect of nitrogen fertilizer rate on soil nitrate nitrogen content after harvesting winter wheat. *Journal of Agricultural Science, Cambridge* 114: 171–176.

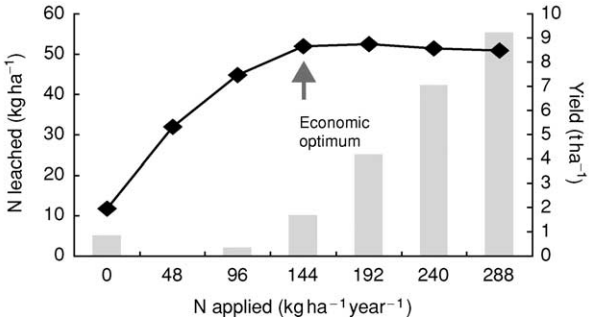


Figure 9 Nitrate lost in drainage water from plots of the long-term Broadbalk Experiment at Rothamsted that have received uniform applications of inorganic N fertilizer for many years. ◆, wheat grain yield; ▮, N leached. Reproduced from Goulding K (2000) Nitrate leaching from arable and horticultural land. *Soil Use and Management* 16: 145–151.

exceptions. If more inorganic fertilizer is applied to a crop than is required, some of the excess may be left in the soil and contribute to the pool of nitrate liable to be lost in the period after crop harvesting – an example is shown in Figure 8. The amount of nitrate in the soil profile (to a depth of 90 cm) was measured after harvesting winter wheat at eight sites in central and eastern England. A range of fertilizer N rates, including zero, had been applied in spring. Figure 8 shows the mean nitrate quantities for the eight sites. For rates of fertilizer N between zero and 160 kg N ha^{-1} (the average rate giving maximum crop yield), about 60 kg ha^{-1} of nitrate-N remained in soil but when this application was exceeded, residual nitrate rapidly increased to over 100 kg ha^{-1} of nitrate-N. It should be noted that, even where no fertilizer N had been applied, 50 kg ha^{-1} of nitrate-N remained in the soil, so small decreases in applications below a rate likely to give approximately maximum yield will have little influence on residual nitrate and nitrate leaching.

Figure 9 shows the measured amounts of nitrate lost in drainage water during winter from plots of the long-term Broadbalk wheat experiment at Rothamsted, UK. Each plot has received a specific rate of N fertilizer annually for many years. Up to the rate of N fertilizer corresponding to maximum grain yield (and also the calculated economic optimum rate), nitrate losses were about 10 kg N ha^{-1} or less but after this fertilizer N rate was exceeded, losses increased greatly. Other studies have shown that a plot receiving organic manure annually loses the largest amount of nitrate. To avoid excessive amounts of fertilizer being applied it is necessary to estimate in advance two quantities, both of which are difficult to predict: (1) the likely requirement of N by the specific crop; and (2) the amount of N likely to become available to the crop from mineralization or other sources. Estimating crop N requirement is best achieved by referring to the yields achieved on that field or region in the past. Of course, the value obtained will only be an estimate as weather in the coming growing season will be unknown at the time. For estimating N supply from sources other than fertilizer the following approaches are currently used or being developed:

1. Using computer models of N cycle processes to estimate the production of inorganic N from mineralization based on soil type, past cropping, past inputs of organic manures and either average or expected climatic conditions for the site. Such models also include estimates of likely N losses based on soil type, timing of agronomic operations, and weather. Modeling systems may also include a model for crop growth or this may be estimated separately. Decision support systems based on computer models have been developed in many regions of the world over the last 20 years and, in principle, can be highly site-specific. The availability of home computers and computerized farm record-keeping has made this approach realistic in many countries. The models used range from simple nutrient balances to complex descriptions of the main N cycle processes. The more complex the model, the more detailed data are required to run it – a judgment has to be made between simplicity and complexity.
2. Using tables of typical N supply values based on runs of models for a range of conditions or measurements from specific experiments. This system can never be as site-specific as a modeling system but is often very effective, low-cost, and requires a minimum of specialist personnel.
3. Measurements of inorganic N in the soil profile before the main crop-growing period. This is often termed the N_{min} method. Assumptions have to be made regarding the rooting depth of the crop and amount of inorganic N that will be lost and so not

be available to the crop. In regions with fairly stable weather conditions (e.g., continental areas within Europe) and soils that contain few stones and so are easy to sample (e.g., loess), this method has proved successful. The measurements can also be used to adjust predictions based on models or tables.

4. Variable N applications. In some situations, where fields are large and soils are not uniform, it may be helpful to adjust N applications according to the variations in soil type and crop growth – giving larger applications where growth can reasonably be expected to be most vigorous and small applications where it is constrained. The technology for variable applications is now readily available but the limitations to its use are: (1) insufficient understanding of the factors influencing crop growth in a particular year; and (2) insufficient data on soil variability. Maps of yield variation from previous years are easy to obtain with modern combine harvesters but it is often difficult to interpret their significance for fertilizer use in future years.

Regulations Affecting N Use in Agriculture

In many parts of Europe and North America concern about nitrate entering water has led to regulations being imposed on agricultural practices. In the European Union areas subject to regulation are termed 'nitrate-vulnerable zones' (NVZs). The regulations are based on current understanding of the causes of nitrate loss, especially the need to minimize nitrate in soil during the vulnerable winter period. The measures adopted in NVZ regulations include the following:

1. Closed periods for N applications. No applications of organic or inorganic forms of N during specified periods in winter appropriate for climate and growing conditions in the area (about 5 months, e.g., 1 September to 1 February for inorganic fertilizer to arable land in the UK). The closed season is shorter for grassland, recognizing that N uptake is possible over a longer period than for arable crops.

2. Limits on quantities of N applied. The farmer is required to use a recognized method for assessing crop requirement and this must not be exceeded. For organic manures the maximum allowed application decreases after 4 years, recognizing that the continued use of organic manure will increase the amount of mineralizable N in soil. For example, in the UK the maximum for grassland is initially $250\,kg\,ha^{-1}$ of total N in manure but decreases to $170\,kg\,ha^{-1}$.

3. Application controls. Limitations on applications when weather conditions could lead to runoff to adjacent surface waters, e.g., because soil is frozen or waterlogged. Also ensuring that applications are made accurately and evenly.

4. Slurry storage. Ensuring that a farm with animals has sufficient capacity to store slurry during the closed period.

5. Record-keeping. Keeping adequate records on N applications, cropping and animal numbers for at least 5 years.

Nitrate and Health

One of the main drivers for agricultural measures to decrease losses of nitrate from soil to water has been the claim that nitrate in drinking water is a cause of two serious diseases. The evidence for the link has never been strong and medical research now indicates that it does not exist at all. There is also evidence that some intake of nitrate has a positive effect on health.

Stomach Cancer

In the past it was claimed that nitrate (NO_3^-) from drinking water or vegetables was reduced to nitrite (NO_2^-) by bacteria in the mouth and that when nitrite entered the stomach it could react with amines from the breakdown of proteins to form N-nitrosamines, which are known to be carcinogenic. One early epidemiological study appeared to support this as it showed a higher incidence of stomach cancer in an area of the UK having high concentrations of nitrate in drinking water. However, a later and much larger study, published in the mid-1980s, contradicted this result. Nitrate concentration in saliva was measured in people from two areas in the UK, one with an unusually high incidence of stomach cancer and one with an unusually low incidence. The results were the opposite of what would be predicted if nitrate was associated with stomach cancer – nitrate concentrations in saliva were about 50% higher in the low-cancer-risk area. Such evidence in itself can neither prove nor disprove a link but it clearly indicates that any link must be weak and outweighed by other factors. Several other studies have shown a negative correlation between nitrate and incidence of stomach cancer. One study was based on the number of deaths of workers in a factory manufacturing ammonium nitrate fertilizer over a 35-year period. Because of the ammonium nitrate dust in the factory, the workers had a larger intake of nitrate than the general population and this was shown by increased concentrations of nitrate in their saliva. However, the number of deaths from stomach cancer among these workers between 1946 and 1981 was exactly the number expected on the basis of incidence in the general population. It can be concluded that there is now no medical evidence to suggest any link between nitrate intake and stomach cancer.

Methemoglobinemia (Blue-Baby Syndrome)

Methemoglobinemia or blue-baby syndrome is a serious disease that can kill infants less than 1 year old. Nitrite derived from the reduction of nitrate can react with hemoglobin and block the site that normally transports oxygen around the body. The first report of the problem in the USA in 1945 referred to 'well-water methemoglobinemia,' and all the cases identified in a 1991 review were caused by water from wells, 98% of which were described as 'privately dug.' It is now thought that the condition only occurs when water is contaminated by gastroenteritic bacteria from sewage or manure heaps, as can occur if a shallow well is dug on farm premises. In such cases nitrate and bacteria both contaminate the water. The condition is currently known in parts of Africa where surface waters are used for human consumption but are contaminated by animal excreta, introducing both bacteria and nitrate. The publisher of the original report was well aware that both bacteria and nitrate were involved in the condition, but the possible role of the bacteria seems to have been played down subsequently. It is now seems likely that it is the bacterial contamination that is the main factor. There are no reports of blue-baby syndrome associated with bacteriologically clean water from a mains supply.

Beneficial Effects of Nitrate on Health

Recent medical research has indicated that nitrate can protect humans and animals from gastroenteritis. It has shown that nitrate is reduced to nitrite by denitrifying bacteria in the mouth, similar to those in soil. It also now appears that the human metabolic system actively secretes nitrate into saliva for this purpose if the intake of nitrate from water or vegetables is insufficient. The proposed reaction between nitrite and amines in the stomach (mentioned above in relation to the proposed link with stomach cancer) probably does not occur because the acid conditions in the stomach destroy nitrite and nitric oxide, NO, is formed. A great deal is now known about the physiological role of NO, one of which is the killing of bacteria such as *Salmonella* and *Escherichia coli* which can cause gastroenteritis. In this way it is now thought that nitrate is a key precursor for NO and thus plays an important role in the body's defense against stomach disease.

Impacts of Nitrate on Ecology of Surface Waters

Any nitrate escaping from agricultural soil to streams, rivers, or lakes can increase the growth of water plants and thus alter the ecology of aquatic systems. This is undesirable in itself and can also interfere with fisheries and navigation for shipping by clogging the waterway with large quantities of water weed. Nitrate can also contribute to the growth of algae in surface waters which are unsightly and some of which are toxic. When algae die the bacteria decomposing them use oxygen dissolved in the water, thus deoxygenating the water with adverse consequences for fish and other organisms. In freshwaters phosphate is usually the limiting factor controlling the growth of algae, but nitrate is also required. In river estuaries and sheltered seas (e.g., the Baltic Sea and the Gulf of Mexico), nitrate is probably the controlling factor. Thus the impact of nitrate on the ecology of natural waters is a genuine and significant reason for controlling losses from agriculture.

Conclusions

Nitrate is the most common form of inorganic nitrogen in soil, being the final product of the decomposition of organic materials under aerobic conditions. It is the form normally taken up by plants, though most plants can also take up ammonium. If sufficient food is to be grown for the current world population of over 6 billion and its expected growth during the twenty-first century, greatly increased amounts of nitrogen will be needed. It will be necessary to use increasing amounts of inorganic fertilizers in addition to better utilization of organic manures and other wastes and improvement in the use of legumes to supply nitrogen to crops. This continued intensification of agriculture will inevitably lead to further losses of nitrate to natural waters. The considerable amount of research conducted over recent decades can help to decrease these losses, both by increasing the efficiency of use of inorganic fertilizers and by better retaining nitrate from organic matter mineralization within the soil/plant system. However, losses can never be entirely eliminated as the nitrate ion is soluble and is always at some risk of being washed out of soil. It can also be converted to nitrous oxide, a powerful greenhouse gas. The ecological impacts of nitrate on aquatic ecosystems can be serious so, for a range of environmental reasons, it is important to minimize nitrate losses from agricultural soils as much as possible. This is in addition to the economic and sustainability reasons for avoiding waste of a vital resource. However, the proposed adverse health impacts of nitrate in water that have been the main driver for limiting nitrate loss now appear to be unfounded. Indeed, the opposite appears to be the case – there is evidence from medical research that some intake of nitrate has beneficial effects on health through the production of nitric oxide in the stomach and its role in the control of gastroenteritis.

See also: **Eutrophication**; **Greenhouse Gas Emissions**; **Isotopes in Soil and Plant Investigations**; **Nitrogen in Soils:** Cycle; Nitrification; Plant Uptake; Symbiotic Fixation; **Pollution:** Groundwater

Further Reading

Addiscott TM, Whitmore AP, and Powlson DS (1991) *Farming, Fertilizers and the Nitrate Problem*. Wallingford: CAB International.

Benjamin N (2000) Nitrates in the human diet – good or bad? *Annales de Zootechnologie* 49: 207–216.

Catt JA *et al.* (1998) Strategies to decrease nitrate leaching in the Brimstone Farm experiment, Oxfordshire, UK, 1988–1993: the effects of winter cover crops and unfertilized grass leys. *Plant and Soil* 203: 57–69.

Cheney K (1990) Effect of nitrogen fertilizer rate on soil nitrate nitrogen content after harvesting winter wheat. *Journal of Agricultural Science, Cambridge* 114: 171–176.

Davidson EA (1991) Fluxes of nitrous oxide and nitric oxide from terrestrial ecosystems. In: Rogers JE and Whitman WB (eds) *Microbial Production and Consumption of Greenhouse Gases: Methane, Nitrogen Oxides, Halomethanes*, pp. 219–235. Washington, DC: American Society for Microbiology.

Dykhuizen RS, Fraser A, Duncan C, Golden M, Benjamin N, and Leifert C (1999) Antimicrobial effect of acidified nitrite on gut pathogens: the importance of dietary nitrate in host defences. In: Wilson WS, Ball AS, and Hinton RH (eds) *Managing Risks of Nitrates to Humans and the Environment*, pp. 295–316. Cambridge: Royal Society of Chemistry.

Goss MJ, Howse KR, Christian DG, Catt JA, and Pepper TJ (1998) Nitrate leaching: modifying the loss from mineralized organic matter. *European Journal of Soil Science* 49: 649–659.

Goulding K (2000) Nitrate leaching from arable and horticultural land. *Soil Use and Management* 16: 145–151.

Goulding KWT, Poulton PR, Webster CP, and Howe MT (2000) Nitrate leaching from the Broadbalk wheat experiment, Rothamsted, UK, as influenced by fertilizer and manure inputs and weather. *Soil Use and Management* 16: 244–250.

Hatch D, Goulding K, and Murphy D (2002) Nitrogen. In: Haygarth PM and Jarvis SC (eds) *Agriculture, Hydrology and Water Quality*, pp. 7–27. Wallingford: CAB International.

Jarvis SC (2000) Progress in studies of nitrate leaching from grassland. *Soil Use and Management* 16: 152–156.

Macdonald AL, Powlson DS, Poulton PR, and Jenkinson DS (1989) Unused fertilizer nitrogen in arable soils – its contribution to nitrate leaching. *Journal of the Science of Food and Agriculture* 46: 407–419.

Macdonald AJ, Poulton PR, Powlson DS, and Jenkinson DS (1997) Effects of season, soil type and cropping on recoveries, residues and losses of 15N-labelled fertilizer applied to arable crops in spring. *Journal of Agricultural Science, Cambridge* 129: 125–154.

Powlson DS (1993) Understanding the soil nitrogen cycle. *Soil Use and Management* 9: 86–94.

Powlson DS (1999) Fate of nitrogen from manufactured fertilizers in agriculture. In: Wilson WS, Ball AS, and Hinton RH (eds) *Managing Risks of Nitrates to Humans and the Environment*, pp. 42–57. Cambridge: Royal Society of Chemistry.

Powlson DS (1997) Integrating agricultural nutrient management with environmental objectives – current state and future prospects. *Proceedings No. 402*. York: The Fertiliser Society.

Powlson DS, Hart PBS, Poulton PR, Johnston AE, and Jenkinson DS (1986) Recovery of ^{15}N-labelled fertilizer applied in autumn to winter wheat at four sites in eastern England. *Journal of Agricultural Science, Cambridge* 107: 611–620.

Recous S, Fresnau C, Faurie G, and Mary B (1988) The fate of labelled ^{15}N urea and ammonium nitrate applied to a winter wheat crop. *Plant and Soil* 112: 205–214.

Wilson WS, Ball AS, and Hinton RH (eds) (1999) *Managing Risks of Nitrates to Humans and the Environment*. Cambridge: Royal Society of Chemistry.

Nitrification

J I Prosser, University of Aberdeen, Aberdeen, UK

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Nitrification is the oxidation of reduced forms of nitrogen, ultimately to nitrate. It is carried out by three microbial groups: (1) autotrophic ammonia oxidizers, (2) autotrophic nitrite oxidizers, and (3) heterotrophic nitrifiers. Autotrophic ammonia and nitrite oxidizers are characterized by their ability to oxidize ammonia sequentially to nitrate under aerobic conditions, but these organisms possess a wider metabolic diversity. This has significant consequences for their ecology, for the occurrence and kinetics of nitrification, and for interactions with other organisms. The major sources of reduced forms of soil nitrogen are excretion and decomposition of organic nitrogen, derived from animals and plants, and application of ammonia-based fertilizers. Many microorganisms and plants require ammonium for growth, while others assimilate nitrate. For both groups, nitrification is important in regulating the supply or loss of nitrogen from the environment. While ammonia has a propensity to bind to soil particles, its conversion to nitrate leads to significant losses of soil nitrogen through leaching and conversion to gaseous forms through denitrification. Nitrification also contributes to soil acidification, with consequent increases in

mobilization of toxic metals, particularly in heavily fertilized and poorly buffered soils. In addition, nitrifiers are involved in production and conversion of a number of greenhouse gases.

Autotrophic Ammonia-Oxidizing Bacteria

Taxonomy

Autotrophic ammonia oxidizers are Gram-negative bacteria, traditionally placed within the *Nitrobacteriaceae* and characterized by their ability to oxidize ammonia. Further classification is based on a limited number of characteristics, including cell shape, internal membrane structures, flagella, inclusion bodies, salt tolerance, urease activity, and substrate affinity and inhibition. Confirmation of taxonomic groupings and species descriptions is by DNA:DNA hybridization of pure cultures. The type strain, *Nitrosomonas europaea*, has been the subject of the majority of biochemical, physiological, and genomic studies.

More recently, analysis of sequences of 16S ribosomal ribonucleic acid (rRNA) and *amo* A genes (encoding the active subunit of ammonia monooxygenase) has enabled phylogenetic classification of cultured ammonia oxidizers. This places *Nitrosococcus oceanus* in the γ-proteobacteria and other known autotrophic ammonia oxidizers in seven lineages within a monophyletic group within the β-proteobacteria: *Nitrosospira*, *Nitrosomonas europaea/Nitrosococcus mobilis*, *Nitrosomonas communis*, *Nitrosomonas marina*, *Nitrosomonas oligotropha*, *Nitrosomonas cryotolerans*, and *Nitrosomonas* sp. Nm143. Difficulties in cultivating ammonia oxidizers limit available physiological data, but links exist between these phylogenetic groups, their physiological characteristics, and their environmental origin. Analysis of 16S rRNA gene sequences also indicates evolutionary relationships with photosynthetic bacteria and other groups possessing intracytoplasmic membrane structures, suggesting that ammonia oxidizers evolved from photosynthetic organisms. 16S rRNA sequence analysis suggests only peripheral associations with methane oxidizers, but similarities exist between *amo*A genes of γ-proteobacterial ammonia oxidizers (*N. oceanus*), and *pMMO* (particulate methane monooxygenase), suggesting a common ancestral origin.

Biochemistry

Ammonia oxidation by β-proteobacteria involves two steps:

$$NH_3 + O_2 + 2H^+ + 2e^- \rightarrow NH_2OH + H_2O$$
$$NH_2OH + H_2O \rightarrow NO_2^- + 5H^+ + 4e^-$$

In the first step, ammonia is oxidized to hydroxylamine by a membrane-bound ammonia monooxygenase. The reaction is endergonic and requires oxygen and a source of reducing equivalents. Hydroxylamine is then oxidized to nitrite by hydroxylamine oxidoreductase, using oxygen from water in an exergonic reaction. Two electrons are required for ammonia oxidation, while the remainder pass down the electron transport chain. The NO_2^-/NH_3 redox potential is -340 mV and reduced nicotinamide adenine dinucleotide (NADH) is synthesized by reverse electron flow using adenosine triphosphate (ATP). Hydrogen ions generated by the second step lead to acidification of unbuffered growth medium, limiting sustained growth.

An alternative mechanism for the first step has recently been proposed, involving nitrogen dioxide and tetroxide, based on studies of *N. eutropha*:

$$NH_3 + N_2O_4 + 2H^+ + 2e^- \rightarrow NH_2OH + 2NO + H_2O$$
$$2NO + O_2 \rightarrow 2NO_2(N_2O_4)$$

This mechanism does not appear to involve ammonia monooxygenase, and studies have also demonstrated reduction in lag phase and the stimulation of growth, activity, and cell yield in cultures supplied with nitric oxide. Any reduction in lag phase will provide ammonia oxidizers with advantages in soil where ammonia supply is intermittent.

Carbon Metabolism

Organic carbon is obtained by nitrifiers via carbon dioxide fixation, and all strains studied as of 2003 possess type I Rubisco genes, but with sequence variation between strains. Rubisco enzymes in some strains are located in carboxysomes. Energy generated by ammonia oxidation is only just sufficient for growth, particularly as carbon dioxide fixation requires significant reducing equivalents, and generation of reducing equivalents consumes ATP.

Ammonia monooxygenase has broad substrate specificity, and ammonia oxidizers can oxidize and assimilate, but not grow on, a number of organic compounds, including methanol, bromoethane, ethylene, propylene, cyclohexane, benzene, and phenol. Acetate, formate, and pyruvate can increase biomass yield by as much as 130%, but higher concentrations inhibit growth. The ecological implications for this are not clear, but assimilation of soil organic carbon may provide a selective advantage, while ammonia oxidizers may be important in biodegradation of organic pollutants in oligotrophic environments, where alternative carbon sources are scarce.

Ammonia monooxygenase is similar to the membrane-bound methane monooxygenase of methanotrophic bacteria. Ammonia oxidizers can oxidize but

not grow on methane and carbon monoxide, and the enzymes have common inhibitors. Incorporation of methane is inhibited by ammonia, and vice versa, and complete methane oxidation may be prevented by inhibition of hydroxylamine oxidation by formaldehyde, produced during methane oxidation. In general, ammonia oxidizers have higher oxidation rates and lower K_m values for ammonia than methanotrophs, while the converse is true for oxidation and affinity for methane. Competition between the two groups for ammonia and methane therefore appears unlikely, but high levels of ammonia-based fertilizer may suppress methanotroph activity. Despite these considerations, ammonia oxidizer 16S rRNA and *amo*A sequences have been found in ^{13}C-labeled nucleic acids extracted from grassland soils treated with ^{13}C-labeled methanol and methane, suggesting their role in soil methane oxidation. Ammonia oxidizers may also have a role in the oxidation of another greenhouse gas, carbon monoxide. They have greater affinity for carbon monoxide than carboxydobacteria, and may provide an important carbon monoxide sink. Carbon monoxide oxidation only occurs during ammonia oxidation, which is required as an energy source, even though carbon monoxide and ammonia compete for the ammonia monooxygenase enzyme.

Ammonia Oxidation under Anaerobic Conditions

Although traditionally considered to be aerobic, autotrophic ammonia oxidizers maintain activity at low oxygen concentrations and under anaerobic conditions. *N. europaea* possesses a nitrite reductase and can denitrify nitrite under anaerobic conditions using hydrogen or organic compounds as electron donors. (They can also denitrify under oxic conditions, using ammonia as electron donor.) This results in the production of nitric and nitrous oxides, and dinitrogen gas. At atmospheric oxygen concentrations, nitrogen oxides constitute 0.1–0.5% of the total N oxidized, but proportions rise to 2.5–10% at low oxygen concentrations.

Anaerobic ammonia oxidation by *N. eutropha* has recently been demonstrated in which nitrogen dioxide or nitrogen tetroxide acts as an electron acceptor, generating nitric oxide and hydroxylamine. The latter is converted to nitrite with release of dinitrogen gas. In addition, anaerobic ammonia oxidation is carried out by autotrophic planctomycetes, which utilize nitrite as an electron acceptor, generating nitrate and nitrogen, with hydrazine and hydroxylamine as intermediates. The process is sensitive to oxygen and nitrite, and the calculated mass balance is:

$$NH_4^+ + 1.32NO_2^- + 0.066HCO_3^- + 0.13H^+ \rightarrow 0.26NO_3^- + 1.02N_2 + 0.066CH_2O + 0.5N_{0.15} + 2.03H_2O$$

However, neither of these anaerobic processes has been detected in soil, where their significance is unclear. Both require substrates which are produced under oxic conditions (nitrogen dioxide and nitrite) but the processes may occur at oxic–anoxic interfaces, such as in biofilms formed on soil particulate material.

Autotrophic Nitrite-Oxidizing Bacteria

Nitrite oxidizers are classified into four genera within the proteobacteria: *Nitrobacter* (the most studied; α-proteobacteria), *Nitrococcus* (γ-proteobacteria), *Nitrospira* (δ-proteobacteria), and *Nitrospina* (forming its own subdivision). *Nitrobacter* strains are related to several denitrifying organisms, while *Nitrobacter* and *Nitrococcus* possess intracytoplasmic membranes and are related to photosynthetic bacteria. δ-Proteobacterial nitrite oxidizers do not contain intracytoplasmic membranes and are not related to photosynthetic organisms. Only *Nitrobacter* and *Nitrosospira* have been detected in soil. Nitrite oxidation is catalyzed by nitrite oxidoreductase, the NO_3^-/NO_3^- redox potential is 430 mV, and reducing equivalents are generated by reverse electron flow, consuming ATP:

$$NO_2^- + H_2O \rightarrow NO_3^- + 2H^+ + 2e^-$$

Nitrite oxidizers fix CO_2 for growth but, unlike ammonia oxidizers, many are capable of mixotrophic and heterotrophic growth on a range of simple organic substrates. Heterotrophic growth is slow (doubling times of 30–150 h) but leads to greater biomass yields. Nitrite oxidoreductase enables heterotrophic growth under anaerobic conditions, with nitrate as the electron acceptor and production of nitrite and ammonia, and nitrogen oxides. This metabolic versatility may explain greater abundance of nitrite oxidizers than ammonia oxidizers in soil.

Heterotrophic Nitrification

Heterotrophic nitrification is the oxidation of inorganic and organic reduced forms of N, to nitrate, by a wide range of fungi and heterotrophic bacteria. In some organisms, the mechanism is similar to that in autotrophic ammonia oxidizers and is linked, in some strains, to aerobic denitrification. The second mechanism, termed fungal nitrification, is linked to lignin degradation and involves reaction of reduced organic compounds with hydroxyl radicals produced in the

presence of hydrogen peroxide and superoxide. There is little evidence that heterotrophs gain energy or other benefit from nitrification, and cellular rates of heterotrophic nitrification activity are significantly lower than for autotrophs. Nevertheless, it may be important in acid soils or where C:N ratios and heterotroph biomass are high.

Analysis of Nitrifiers and Nitrification

Nitrification Activity and Rates

Nitrification rates in soil are traditionally measured as ammonia consumption or nitrite plus nitrate production during incubation of soil or soil slurry. Zero-order kinetics result from short-term measurements at nonlimiting ammonia concentrations and measure potential nitrification, which can be used to estimate nitrifier abundance. Short-term measurements at limiting substrate concentrations generate Michaelis–Menten kinetics. Long-term incubation of nonenriched soil leads to exponential increases in nitrite plus nitrate owing to unrestricted growth, at nonlimiting substrate concentrations, or growth at submaximal and decreasing rates at limiting substrate concentrations. Nitrate-production kinetics are frequently fitted to the logistic equation. Kinetics can be interpreted differently, however, if nitrifiers are considered to be distributed heterogeneously, as clusters growing on discrete sources of ammonium within the soil. If nitrification is measured in individual soil aggregates, the observed nitrate production curves fit well with the cumulative distribution curves for lag periods. Bulk nitrate concentrations therefore reflect not cumulative product of a uniformly growing population but the continued activation of clusters of nitrifying bacteria. The increase in nitrate concentration then depends on the distribution of lag periods prior to the onset of nitrification. Further evidence for heterogeneous distribution of ammonia oxidizers in soil is provided by molecular analysis of communities in grassland soils. Regular application of inorganic nitrogen fertilizer reduces heterogeneity in ammonia oxidizer communities, in association with similar reduction in the heterogeneity of soil pH and ammonia concentration. Analysis of nitrification kinetics during incubation is therefore important in providing information on soil population levels and substrate concentrations. Growth parameters are more easily obtained from laboratory cultures and typical values are given in Table 1.

Other nitrogen cycle processes, e.g., denitrification, immobilization, and ammonification, will influence concentrations of ammonium, nitrite, and nitrate. Rigorous measurement of soil nitrification rates, and distinction between autotrophic and heterotrophic nitrification therefore necessitate the use of metabolic blocks and ^{15}N-based techniques. The former involve measurement of nitrate production in the presence of specific nitrification inhibitors e.g., nitrapyrin and acetylene. The second involves applying ^{15}N as NH_4^+ or NO_3^- and following changes in the concentration of label in the respective pools.

Enumeration

Dilution-plate enumeration of viable nitrifiers is not feasible because of poor growth on solid media and enumeration is traditionally achieved using the most probable number (MPN) method. This involves multiple inoculation from dilutions of a soil suspension into mineral-salts medium containing ammonium and a pH indicator. After incubation for several (typically 4) weeks, growth is assessed by nitrite production, ammonium loss, and/or acidification, and cell concentrations are calculated using statistical tables. The requirement for laboratory cultivation leads to underestimation of cell concentrations. In addition, growth may be biased by ammonium concentrations in the medium, with inhibition of some strains at high concentrations (Figure 1). MPN counts for typical soils range between 10^3 and 10^5 cells g^{-1}

Table 1 Reported ranges for kinetic constants derived from pure cultures of ammonia and nitrite oxidizers

	Ammonia oxidizers	*Nitrite oxidizers*
Maximum specific growth rate (h^{-1})	0.02–0.088	0.018–0.058
Doubling time (h)	8–35	12–39
Biomass yield (g biomass mol^{-1} substrate)	0.42–1.72	1.11–1.51
Cell yield (cells mol^{-1} substrate)	1.85–10.6	
Carbon yield (ratio of CO_2 fixed: NO_2^- or NO_3^- produced)	0.014–0.096	0.013–0.031
Cell activity (fmol NO_2^- or NO_3^- $cell^{-1}$ h^{-1})	0.9–31.3	5.1–42
Biomass activity (fmol NO_2^- or NO_3^- g $biomass^{-1}$ h^{-1})	4–200	15.7–25.2
Saturation constant for growth (mmol l^{-1} NH_4^+ or NO_2^-)	0.12–14	0.045–0.178
Saturation constant for activity (mmol l^{-1} NH_4^+ or NO_2^-)	0.05–0.07	1.6–3.6

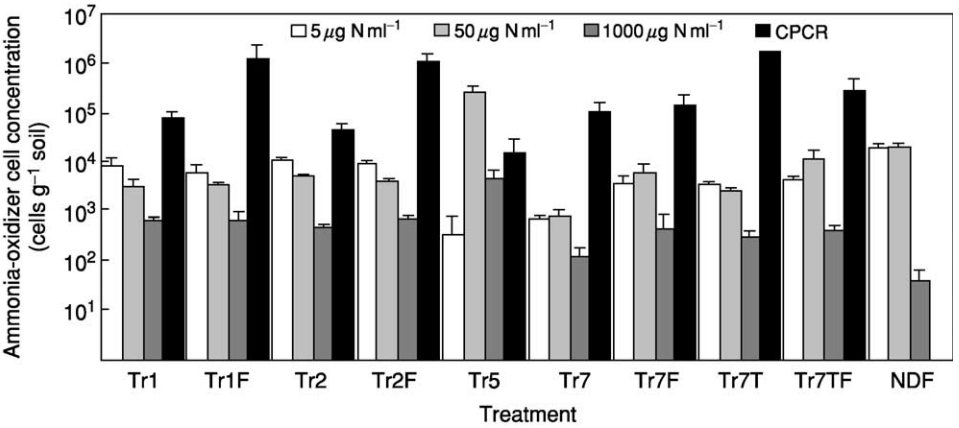


Figure 1 Enumeration of ammonia-oxidizer cells in long-term ecological research (LTER) soils using the most probable number (MPN) method with medium containing 5, 50, or 1000 mg of NH_4^+-N ml^{-1} and competitive polymerase chain reaction (C-PCR). Treatments are: Tr1, conventional tilling; Tr2, no tilling; Tr5, *Populus* perennial cover crop; Tr7, historically tilled; NDF, natural deciduous forest. Error bars represent the standard errors. Suffixes 'T' and 'F' indicate tillage and fertilization, respectively. (Reproduced with permission from Phillips CJ, Harris D, Dollhopf SL *et al.* (2000) Effects of agronomic treatments on the structure and function of ammonia oxidising communities. *Applied and Environmental Microbiology* 66: 5410–5418.)

soil for ammonia oxidizers, with nitrite oxidizers slightly more abundant. Quantitative molecular techniques (e.g., competitive PCR (cPCR), real-time PCR) avoid cultivation bias and detect the total nitrifier populations, giving cell concentrations 1–2 orders of magnitude greater than MPN counts (Figure 1).

Enrichment and Isolation

Autotrophic ammonia-oxidizing bacteria are readily enriched by inoculation of an inorganic medium, containing ammonium with soil, and incubation for several weeks. Isolation of pure cultures is more difficult, due to faster growth of heterotrophs on organic carbon contaminants or by-products of nitrifier growth. This typically takes several months and cultures frequently do not survive continued subculturing. Laboratory selection also reduces the likelihood of obtaining dominant or representative environmental strains, but enrichment of a strain is necessary for unequivocal evidence of its presence and viability in a particular soil.

Soil Communities and Diversity

Information on 16S rRNA and *amo*A gene sequences from cultured ammonia oxidizers has led to the design of oligonucleotide polymerase chain reaction (PCR) primers specific for β-proteobacterial autotrophic ammonia oxidizers. These have been used to amplify ammonia-oxidizer gene fragments from nucleic acids extracted directly from soil. Gene fragment sequences may then be analyzed by cloning, sequencing of clone library representatives, and phylogenetic analysis. Alternatively, fingerprinting techniques such as denaturing or temperature gradient gel electrophoresis (DGGE, TGGE) and terminal restriction fragment length polymorphism (T-RFLP) analysis may be used. These approaches do not require laboratory cultivation and thereby eliminate many of the disadvantages of cultivation-based techniques, and difficulties in purifying and identifying ammonia-oxidizing bacteria.

The application of molecular techniques has greatly expanded our knowledge of nitrifier communities in soils. Phylogenetic analysis of amplified sequences reveals considerable diversity within soil ammonia oxidizers. Duplicate sequences and sequences identical to pure cultures are rarely obtained, although this partly reflects the few pure culture sequences in databases and, potentially, sequencing and cloning errors. Sequences are sometimes homologous to those of laboratory cultures enriched from the same environment and the majority fall within the lineages obtained with sequences from cultivated organisms. One exception is the possible subdivision of *Nitrosospira* into four clusters, and there is evidence of differences in physiological characteristics of pure and enriched cultures representative of these different *Nitrosospira* clusters.

Soils are dominated by β-proteobacterial ammonia oxidizers, with γ-proteobacterial sequences apparently restricted to marine environments. The links between 16S rRNA-defined lineages and physiological characteristics aid in the interpretation of the distribution of these organisms under different soil conditions. This approach also indicates that *Nitrosospira* may be more abundant than *Nitrosomonas* in

soil. If so, more studies are required of *Nitrosospira* physiology, rather than basing interpretation on information on *Nitrosomonas europaea*.

Fingerprinting techniques, particularly DGGE, have been used to study the effects of environmental factors and soil treatments on ammonia-oxidizer communities. This provides more rapid assessment and greater replication than cloning and/or sequencing approaches. In addition, bands of interest on DGGE gels may be excised, cloned, and sequenced and information on identity obtained by comparison with database sequences. These approaches can also be used to assess quantitative or semiquantitative changes in ammonia-oxidizer community structure.

Comparison of 16S rRNA gene sequences in enrichment cultures with those amplified from nucleic acids extracted from the same soil provides an indication of selection during laboratory cultivation. The relative abundances obtained using these two approaches differ but, in one study, 16% of sequences were common to both enrichment cultures and clones analyzed. Molecular analysis of cultures obtained during MPN enumeration also indicates selection, with frequent domination by *Nitrosomonas* strains in samples from environments which appear, from molecular analysis, to be dominated by *Nitrosospira*. Molecular- and cultivation-based techniques therefore provide different views of diversity and species composition within soil ammonia-oxidizer communities. However, molecular techniques are also subject to bias and additional techniques are required. In this respect, direct detection of cells using fluorescent *in situ* hybridization (FISH) with specific ammonia-oxidizer probes would be useful. This approach, however, is limited by autofluorescence and the relatively low cell concentrations of ammonia oxidizers in soil.

Influence of Soil Characteristics on Ammonia-Oxidizer Communities

Soil pH

Ionization of ammonia at low pH values reduces substrate availability for ammonia oxidizers, and laboratory batch growth of ammonia oxidizers does not occur at pH values of less than 6.5. However, autotrophic ammonia oxidizers with neutral pH optima are frequently isolated from acid soils, and autotrophic nitrification has been demonstrated in soils with pH values as low as 3.5. In some situations, nitrification in acid soils is due to heterotrophic nitrification. Coupled denitrification–nitrification by heterotrophic nitrifiers is unlikely, due to the requirement for high substrate concentrations, and heterotrophic nitrification involving ammonia monooxygenase-based reactions suffers from reduced ammonia availability at low pH. However, acidophilic fungal heterotrophic nitrifiers have been characterized and fungal nitrification is significant in some acid soils. Nevertheless, use of specific inhibitors of autotrophic ammonia oxidation and ^{15}N-based studies have demonstrated active, autotrophic ammonia oxidation in acid soils.

One explanation for this phenomenon is the possession by some ammonia oxidizers of urease activity, enabling growth on urea as a sole nitrogen source. Urease activity and growth on urea are independent of pH in the range 4–8, possibly through uptake by passive diffusion. Laboratory studies show that, under these conditions, ammonia oxidation can proceed at pH values as low as 4, until urea is exhausted, after which it ceases, even if ammonium is present in solution. This has been demonstrated in a strain of *Nitrosospira* that cannot oxidize ammonium at pH values less than 7 (Figure 2). Urease activity therefore provides

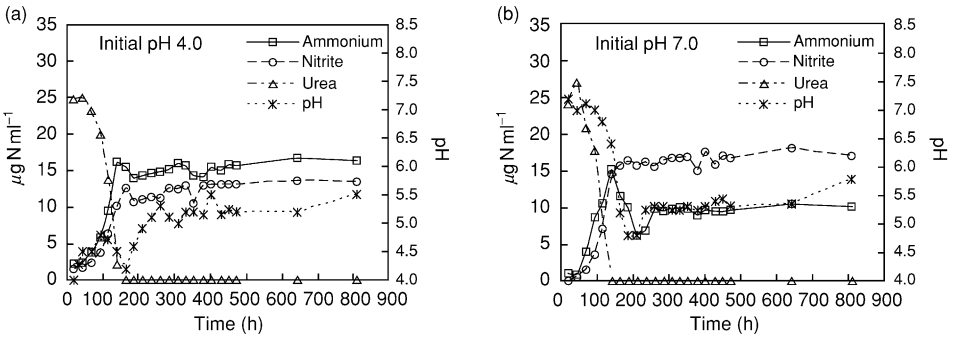


Figure 2 Changes in pH and in concentrations of urea, ammonium, and nitrite during growth of *Nitrosospira* strain NPAV in liquid batch culture on poorly buffered medium containing urea, at initial pH values of (a) 4 and (b) 7. (Reproduced with permission from Burton SAQ and Prosser JI (2001) Autotrophic ammonia oxidation at low pH through urea hydrolysis. *Applied and Environmental Microbiology* 67: 2952–2957.)

the potential for acidophilic ammonia oxidation, but is not a prerequisite for ammonia oxidizers isolated from acid soils.

Autotrophic ammonia oxidizers appear to be protected from low pH in soil when occurring at high cell concentrations and associated with particulate matter. High cell concentrations can be achieved in aggregates, which when inoculated into liquid medium can nitrify at low pH, and nitrification at pH 4 has been demonstrated for cells immobilized in alginate. Biofilm formation on surfaces such as glass slides, glass beads, ion-exchange resin beads, and clay minerals can reduce the pH minimum for growth and activity of *N. europaea* by up to 2 pH units. Surface growth protection increases with the cation exchange capacity of the particulate material and lag phases prior to growth of stationary-phase or starved cells are reduced for attached cells. Protection from low pH may result from preferential growth and nitrification at the clay surface, due to adsorption of ammonium. It is greatest when the only source of ammonium is that bound to clay minerals. This has been demonstrated by growth on ammonia-treated vermiculite (ATV), in which ammonia is fixed to the vermiculite by high-temperature exposure to ammonia. In the presence of ATV, *N. europaea* growth occurs without a lag phase and with greater nitrite yield than for suspended cells, but with no decrease in the pH of the medium. This, and the other clay effects, suggests localized buffering; ammonium released from the clay surface is utilized by adsorbed cells and H^+ ions are effectively inactivated by exchange for ammonium ions. It is not clear why biofilm growth should reduce the lag phase, but this presents a significant ecologic advantage in competing for transient supplies of ammonium. A possible mechanism involves cell–cell signaling molecules, acyl homoserine lactones, which will accumulate at high cell concentrations. When *N. europaea* biofilms established on glass beads are starved and then resupplied with ammonia, oxidation occurs immediately. However, supply of ammonia to suspended, nonbiofilm cells in liquid medium results in a lag period, prior to ammonia oxidation. The lag period increases with the length of the preceding starvation period but is significantly reduced by addition of acyl homoserine lactones. Signaling compounds are produced by a wide range of soil microorganisms and affect a number of ecologically important processes. These compounds may therefore mediate interactions between nitrifiers and other soil microorganisms, particularly in biofilms and in the rhizosphere, where biomass concentrations are sufficient to lead to high concentrations of signaling molecules.

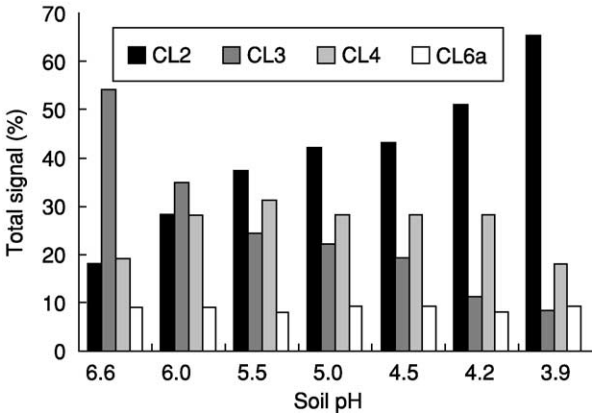


Figure 3 The influence of soil pH on the relative abundances of *Nitrosospira* clusters 2 (CL2), 3 (CL3), and 4 (CL4), and *Nitrosomonas* cluster 6a. Abundances are calculated as relative intensities of denaturing gradient gel electrophoresis (DGGE) analysis of ribosomal ribonucleic acid (rRNA) genes amplified from DNA extracted from soil maintained at a range of pH values, using ammonia oxidizer-specific primers. (Reproduced with permission from Stephen JR, Kowalchuk GA, Bruns MA *et al.* (1998) Analysis of β-subgroup ammonia oxidiser populations in soil by DGGE analysis and hierarchical phylogenetic probing. *Applied and Environmental Microbiology* 64: 2958–2965.)

The development of molecular techniques has enabled assessment of the influence of soil pH on ammonia-oxidizer communities. In one study, DGGE analysis was carried out on β-proteobacterial 16S rRNA gene fragments amplified from DNA extracted from soils maintained for 36 years at pH values in the range 3.9–6.6. Relative abundances of different clusters were assessed as relative intensities of DGGE bands and showed a greater proportion of *Nitrosospira* cluster-2 sequences in acid soils and a greater proportion of *Nitrosospira* cluster 3 at higher pH (Figure 3). Other clusters detected were *Nitrosospira* cluster 4, which may have increased in relative abundance at pH 5.5, and *Nitrosomonas* cluster 6a, which was independent of pH. Molecular studies of agricultural, grassland, and forest soils provide further evidence for the association of *Nitrosospira* cluster-2 sequences with acid soils. However, lack of information on distinctive physiological characteristics specific to representatives of this group, and of physiological diversity within ammonia oxidizers in general, prevents establishment of links between species diversity, functional diversity, and mechanisms for nitrification at low pH.

Nitrogen Fertilization

Sustained nitrogen fertilization and liming increase nitrifier population size, as indicated by cell counts (MPN and cPCR) and nitrification potential

measurements. In addition, evidence from molecular studies of nitrifiers in long-term ecological research (LTER) sites subjected to a number of different fertilization and cultivation conditions indicates an influence on the ammonia-oxidizer communities. Tilled soils were dominated by *Nitrosospira* cluster-3 sequences, while never-tilled soils contained *Nitrosospira* clusters 3 and 4, and *Nitrosomonas* cluster 6. Similar effects of fertilization on the relative proportions of *Nitrosospira* cluster-3 and -4 sequences have been found in chalk grassland soil. Little is known about the persistence of changes after perturbations.

Nitrogen supplied as inorganic fertilizer or, in grazing systems, as animal urine leads to localized, high ammonia concentrations. These concentrations are sufficient to inhibit ammonia-sensitive groups of ammonia oxidizers, while permitting growth of more tolerant strains. Soil microcosm experiments have demonstrated that the kinetics of nitrification of ammonia derived from synthetic sheep urine applied to grassland soil depends on the relative abundances of ammonia-sensitive and ammonia-tolerant strains. Low numbers of the latter result in significant lag periods before nitrification, impact significantly on maintenance of high levels of ammonia for plant growth, and reduce nitrogen losses associated with leaching and denitrification of nitrate.

Inhibition of Soil Nitrification

Specific inhibitors of autotrophic nitrification (e.g., nitrapyrin, dicyandiamide, etridiazole) have been used in the field to reduce losses of fertilizer nitrogen and to reduce nitrate pollution. Inhibitors, particularly acetylene, are also used as metabolic blocks when measuring nitrification rates. Higher concentrations of these compounds are required for inhibition in soil than in laboratory culture because of inactivation and degradation and because attachment of nitrifiers to soil surfaces protects cells from inhibition. Protection is greater with expanding clays with high cation exchange capacity. In many mature coniferous soils, nitrogen is present as ammonium rather than nitrate and net nitrification rates are low. This has led to suggestions of allelopathic inhibition of nitrification by polyphenolic compounds released by roots, which can inhibit laboratory growth of nitrifiers. However, use of ^{15}N-based techniques has demonstrated high gross nitrification rates in these soils and low nitrate levels are more likely due to greater immobilization of nitrate. There is little direct evidence for specific effects of plant diversity and community structure on ammonia-oxidizer communities, although plants can exert indirect effects through their influence on soil ammonium concentrations.

Summary

Nitrifying bacteria convert the most reduced form of soil nitrogen, ammonia, into its most oxidized form, nitrate. In itself, this is important for soil ecosystem function, in controlling losses of soil nitrogen through leaching and denitrification of nitrate. Nitrifiers also contribute to other important processes, including nitrous oxide production, methane oxidation, degradation of organic compounds, and carbon monoxide oxidation. The development of ^{15}N-based techniques has increased significantly our ability to dissect soil nitrogen transformations and their rates, while molecular techniques now enable characterization of soil nitrifier community structure and changes in species composition. These approaches are increasing our understanding of the ecology of soil-nitrifying bacteria and provide the potential for determining relationships between diversity, community structure, and ecosystem function in this important group of organisms.

See also: **Nitrogen in Soils:** Cycle; Nitrates; Plant Uptake

Further Reading

Belser LW (1979) Population ecology of nitrifying bacteria. *Annual Review of Microbiology* 33: 309–333.

De Boer W and Kowalchuk GA (2000) Nitrification in acid soils: microorganisms and mechanisms. *Soil Biology and Biochemistry* 33: 853–866.

Koops H-P and Pommerening-Röser A (2001) Distribution and ecophysiology of the nitrifying bacteria emphasizing cultured species. *FEMS Microbiology Ecology* 37: 1–9.

Koops H-P, Purkhold U, Pommerening-Röser A, Timmermann G, and Wagner M (2003) The lithoautotrophic ammonia-oxidizing bacteria. In: Dworkin D, Falkow S, Rosenberg E, Schleifer K-H, and Stackebrandt E (eds) *The Prokaryotes: An Evolving Electronic Resource for the Microbiological Community.* 3rd edn, release 3.13. New York: Springer-Verlag.

Kowalchuk GA and Stephen JR (2001) Ammonia-oxidizing bacteria: a model for molecular microbial ecology. *Annual Review of Microbiology* 55: 485–529.

Laanbroek HJ and Woldendorp JW (1995) Activity of chemolithotrophic nitrifying bacteria under stress in natural soils. *Advances in Microbial Ecology* 1: 275–304.

Prosser JI (ed.) (1986) *Nitrification*. Oxford, UK: IRL Press.

Prosser JI (1989) Autotrophic nitrification in bacteria. *Advances in Microbial Physiology* 30: 125–181.

Prosser JI and Embley TM (2002) Cultivation-based and molecular approaches to characterisation of terrestrial and aquatic nitrifiers. *Antonie van Leeuwenhoek* 81: 165–179.

Prosser JI, Embley TM, and Webster G (2002) The influence of selection pressures on species diversity, functional gene diversity and activity of ammonia oxidising bacteria. In: Hails R, Beringer JE, and Godfray HCJ (eds) *Genes in the Environment*, pp. 187–202. Oxford, UK: Blackwell Publishing.

Schmidt I, Slickers O, Schimid M *et al.* (2002) Aerobic and anaerobic ammonia oxidizing bacteria – competitors or natural partners? *FEMS Microbiology Ecology* 39: 175–181.

Plant Uptake

A Hodge, University of York, York, UK

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Nitrogen (N) is essential for life because it is a component of nucleic acids and amino acids and thus peptides and proteins. In addition, plants also require N for chlorophyll. In most terrestrial ecosystems, N is the primary limiting nutrient. In soils, N is present predominantly in organic form, and inorganic N is made available by the mineralization of organic N to ammonium (NH_4^+) and subsequent oxidation to nitrate (NO_3^-). Both these inorganic N sources can be taken up by plant roots, but roots have to compete with soil microorganisms as well as other plant roots for this N. Recent evidence suggests that plants may be able to access simple organic N sources directly without the need for the mineralization process to release inorganic N, but these N sources also have to be captured in a competitive environment. The distribution of N sources in soil, particularly organic N forms, is both spatially and temporally heterogeneous. Plant roots respond to this heterogeneity in a number of ways in order to capture N from the soil system.

Plant–Microbial Competition for N

Mineralization of organic material to NH_4^+ is carried out by a wide range of soil animals and microorganisms. However, mineralization and the subsequent conversion of NH_4^+ to NO_3^- are generally considered to be the rate-limiting steps in the terrestrial N-cycle. Due to their rapid turnover times and rapid adaptation rate, microorganisms are generally thought to utilize inorganic N sources first, with the plant only being able to access what is left over after microbial uptake. Exceptions are plants such as legumes and alder which form symbiotic associations with rhizobium and *Frankia* bacteria, respectively. These symbionts form nodules on the roots which are capable of fixing dinitrogen and thus supply the plant with N directly. The majority of plants, however, do not form N_2-fixing symbiotic associations and must rely on N uptake via their roots or mycorrhizas.

Most transformations of N also depend on the supply of carbon (C), and thus these two nutrient cycles are closely linked. The C:N ratio of the substrate determines the rates at which N (and C) are released and mineral N made available for plant uptake. Also influential are the composition of the decomposer microbial community and its requirements for N. Fungi tend to have a higher substrate assimilation efficiency and a lower N requirement per unit of C assimilated than bacteria. Thus, if the decomposer population is predominantly fungal, more N may be released than if it was predominantly bacterial. However, fungi also tend to turn over less rapidly than bacterial populations, so the N contained in their structural components will be locked up for a longer time than bacterial N. Generally net mineralization, resulting in the release of N, occurs if the C:N ratio of the substrate being decomposed is lower than approx. 25:1. At higher C:N ratios, net immobilization occurs and the microbial decomposer community requires additional N from the soil system to decompose the substrate.

Net immobilization of N reduces the N available for the plant and can reduce overall plant growth. Although other factors influence the release of N, for example lignin content of the substrate, there is a strong inverse relationship between the amount of plant N captured from organic substrates and the C:N ratio of the substrate, as shown in Table 1.

In addition to microorganisms soil animals also play an important role in the mineralization process through physical, chemical, and biological processes. Soil animals can physically break up and redistribute organic materials in the soil. They also ingest organic materials which are subsequently broken down by a range of enzymes and/or by microflora present in their guts. Much of the C they consume is used for metabolic processes and the excreted waste products, in the form of NH_4^+, urea, or amino acids, are a much more readily available form of N for plant uptake than the organic material originally ingested. Soil animals feed on microorganisms as well as other soil animals. The plant benefits from the regulation of the size and activity of the microbial population, and from the release of more N into the soil–plant system. An increase in plant N capture of up to 75% has been reported as a result of protozoa grazing on soil microorganisms. Thus, although microorganisms may

Table 1 The percentage N captured by plants from organic material (*Lolium perenne* shoots) of varying C:N ratio

Plant species	*C:N ratio*	*Duration (days)*	*N from organic material captured by plant (%)*
Lolium perenne[a]	12:1	30	22
Lolium perenne[b]	12:1	70	26
Poa pratensis[b]	12:1	70	26
Lolium perenne[c]	21:1	49	11
Lolium perenne[d]	31:1	39	5
Poa pratensis[d]	31:1	39	3
Lolium perenne[e]	31:1	56	9
Poa pratensis[e]	31:1	56	5

[a]Data from Hodge A, Stewart J, Robinson D, Griffiths BS, and Fitter AH (2000) Plant N capture and microfaunal dynamics from decomposing grass and earthworm residues in soil. *Soil Biology and Biochemistry* 32: 1763–1772.
[b]Hodge A, Stewart J, Robinson D, Griffiths BS, and Fitter AH (2000) Spatial and physical heterogeneity of N supply from soil does not influence N capture by two grass species. *Functional Ecology* 14: 645–653.
[c]Hodge A, Stewart J, Robinson D, Griffiths BS, and Fitter AH (2000) Competition between roots and soil micro-organisms for nutrients from nitrogen-rich patches of varying complexity. *Journal of Ecology* 88: 150–164.
[d]Hodge A, Stewart J, Robinson D, Griffiths BS, and Fitter AH (1998) Root proliferation, soil fauna and plant nitrogen capture from nutrient-rich patches in soil. *New Phytologist* 139: 479–494.
[e]Hodge A, Robinson D, Griffiths BS, and Fitter AH (1999) Why Plants bother: root proliferation results in increased nitrogen capture from an organic patch when two grasses compete. *Plant, Cell and Environment* 22: 811–820.

initially sequester inorganic N, protozoa grazing ensure that this N is released back into the root-soil environment and becomes available for root uptake.

Inorganic N

The fate of N into different pools in the soil can be measured by addition of materials labeled (or enriched) with the stable isotope ^{15}N. Short time scales (i.e., a few hours to days) are used to measure flow rates between pools, because recycling of labeled material over longer incubations may occur and lead to underestimations of the total flux through the pool. Such short-term studies on partitioning of labeled inorganic N sources added to soils have been used to measure the competition between microorganisms and plants (Table 2). Generally, a larger percentage of the ^{15}N-labeled inorganic N material is detected in the microbial biomass than plants, indicating that, at least in the short term, microorganisms are the superior competitors. These types of studies also indicate how rapid the flux is through the inorganic pools in soil. Thus the size of the pool, which is typically measured to estimate available N, does not reflect the N that is actually available to the plants and soil microorganisms if that pool has a high flux rate.

The studies summarized in Table 2 added N compounds enriched with ^{15}N in order to trace the uptake of N by the different groups of organisms (i.e., microorganisms or plants). Another method used by some investigators is to measure differences in the natural abundance of ^{15}N among different pools of N in soil to determine the source of N that plants or microbes are utilizing. This technique, however, is complicated by fractionation of ^{15}N by biochemical processes and mixing of N among the various soil pools. In addition, there is evidence that the $\delta^{15}N$ of plants is affected by their mycorrhizal status. Thus, some investigators have concluded that plant $\delta^{15}N$ values cannot be used to identify the N source that a plant is accessing in the soil. Although natural abundance techniques have considerable potential, many technical difficulties need to be resolved before the technique can be effectively used in a wide range of environments.

Some plants, such as members of the Ericaceae, prefer to take up NH_4^+ over NO_3^-. There is some evidence to suggest that these plants may lack the capability to reduce large quantities of NO_3^- in their tissues. These plants generally grow in acidic soils, where nitrification of NH_4^+ to NO_3^- is inhibited and NH_4^+ accumulates, and this may explain their inorganic N source preference. If NH_4^+ accumulates inside the plant it becomes phytotoxic, thus NH_4^+ is rapidly transformed to amide or amino acid form after absorption. For most plants, NO_3^- is the major form of N absorbed and nitrate is then reduced to nitrite (NO_2^-) and to NH_4^+ by the linked enzymes nitrate and nitrite reductase. When NO_3^- is mixed with a small amount of NH_4^+ (rather than just NO_3^- alone), plant growth generally increases. These conditions reflect the availability of inorganic N in most fertile soils but of course this will also depend on soil type. In soil solution, NH_4^+ concentrations are generally lower than NO_3^- concentrations because of rapid transformation to NO_3^- or because NH_4^+ is strongly held on to negatively charged particles such as soil organic materials and clays, and so unavailable for plant uptake. The uptake of a particular inorganic N source may also vary depending on the time of

Table 2 Data from field studies following the partitioning of inorganic N between plants and the microbial biomass after addition of ^{15}N-labeled sources

Plant species	Added material	Duration	^{15}N detected (%)	
			Plants	Microorganisms
Grassland[a]	$^{15}NH_4^+$	1 d	9–11	46–61
Temperate forest[b]	$^{15}NO_3^-$	1 d	20–26	38–50
Well drained	$NH_4^{15}NO_3$	2 d	0.19	20
	$^{15}(NH_4)_2SO_4$	2 d	0.04	21
Poorly drained	$NH_4^{15}NO_3$	2 d	0.01	0
	$^{15}(NH_4)_2SO_4$	2 d	0.01	0
Allium tricoccum[c] in a hardwood forest floor	$^{15}(NH_4)_2SO_4$	2 d	1	22
	$K^{15}NO_3$	2 d	3	32

[a]Data from Jackson LE, Schimel JP, and Firestone MK (1989) Short-term partitioning of ammonium and nitrate between plants and microbes in an annual grassland. *Soil Biology and Biochemistry* 21: 409–415.
[b]Groffman PM, Zak DR, Christensen S, Mosier A, and Tiedje JM (1993) Early spring nitrogen dynamics in a temperature forest landscape. *Ecology* 74: 1579–1585.
[c]Zak DR, Groffman PM, Pregitzer KS, Christensen S, and Tiedje JM (1990) The vernal dam: plant–microbe competition for nitrogen in northern hardwood forests. *Ecology* 71: 651–656.

year, as it is well established that NO_3^- uptake is more sensitive to low temperature than is NH_4^+.

In addition, plants can absorb N through their foliage. Foliar additions of N in the form of urea, ammonium nitrate, and potassium nitrate are added to some crops, although this practice is not widespread. Plant foliage can also absorb nitrogenous gases such as NO, N_2O, and particularly ammonia (NH_3) from the atmosphere or volatilized from the soil surface. Losses of N via NH_3 volatilization can be considerable, i.e., approx. 50% of the N in animal slurries and manures can be lost in this way. Ammonia volatilization is increased under dry conditions, high temperatures, high air movement and in soils of alkaline pH and low organic matter or clay content. The presence of plants helps to reduce NH_3 losses because their roots compete for NH_4^+ in the soil, preventing its transformation and subsequent loss as NH_3. Also their foliage directly absorbs volatilized NH_3 from the soil surface. Foliar absorption of NH_3 is of considerable importance to the plant's N uptake in some terrestrial N cycles. However, cell membranes are highly permeable to NH_3 and the local, rapid increase in pH within the plant cell may irreversibly damage it. Thus although this mechanism can contribute to the plant's total N uptake, at high absorption levels NH_3 becomes toxic to the plant.

Another input of N is emission of NH_3 and other gases including nitrogen oxides (NO_x; i.e., NO and NO_2) and N_2O from intensive agricultural and industrial sources, and their subsequent deposition elsewhere via wet (largely in the form of nitrate and ammonium) or dry (largely in the form of NO_x and ammonia) deposition. This redistribution of N can have a positive effect on plant productivity in N-limited systems. However, excessive levels can cause soil acidification and adversely affect plant diversity.

Plant Inorganic N Uptake and Rhizosphere pH

The form of inorganic N taken up by the plant can strongly influence the pH of the rhizosphere. To maintain the electrochemical gradient between root cells and soil (cells typically have electrochemical potentials of less than −100 mV) which drives cation uptake, plants taking up N as nitrate will release bicarbonate (HCO_3^-) and/or hydroxyl (OH^-) ions, causing the rhizosphere pH to increase. In contrast, plants taking up ammonium as the main form of N will tend to release protons, causing the rhizosphere pH to decrease. However, even when grown on nitrate as the sole source of N, the rhizosphere pH of some plants (such as chickpea) can decrease. This decrease is almost certainly due to a higher ratio of cation to anion uptake. Leguminous plants with active N_2-fixing nodules also tend to have a large cation to anion uptake ratio and so a tendency to excrete protons and so decrease the pH of their rhizosphere soil. Thus the consequences for the rhizosphere pH will depend not only on the form of N (i.e., NH_4^+, NO_3^- or N_2 fixation) the plant is taking up but also on the plant species present and the buffering capacity of the soil.

Organic N

In some ecosystems such as the arctic, alpine, and boreal forest, annual plant N uptake is higher than should be available through inorganic N capture

alone. This has led to the conclusion that many of the plant species in these ecosystems are accessing organic N sources directly without the prior need for microbial mineralization. Forms of organic N in these soils range from the complex (e.g., humus) to the simple (i.e., amino acids and urea), although most of the evidence for plant uptake of intact organic N comes from studies on amino acid N. Some amino acids are present in the dissolved organic nitrogen (DON) fraction, but the majority are protected from uptake by adsorption on to soil particles or humic substances. Environmental fluctuations such as freeze–thaw and drying–rewetting events can result in the disruption of these soil particles and some of this bound amino N may subsequently be released. These environmental events can also disrupt microbial cells, leading to the death of a large fraction of the microbial community and a competitive advantage for the plant providing the plant itself is not also adversely affected by the same environmental events. The advantage to the plant is due not only to a decline in the number of microbial competitors but an increase in N supply through the release of bound amino N forms from soil particles and from the lysis of the microbial cell contents.

Studies using plants grown in solution culture or excised roots have demonstrated that uptake of organic N can occur at levels comparable with, or even in excess of, N uptake from inorganic sources. However, in such studies the uptake of organic N is maximized because, unlike in soil, these N compounds are freely available rather than physically or chemically bound and the plant does not have to compete with microorganisms for these N sources. Although a number of amino acid transporters have been identified on plant roots, amino acids in soil have to move across the rhizosphere, a zone of high microbial activity. Many of these microbes also possess amino acid transporters. The exception is with plants that form mycorrhizal associations, in which case uptake by the fungal hyphae may occur. In these cases, transport across the hyphosphere (the volume of soil influenced by the external mycelium of the mycorrhizal fungus) instead of the rhizosphere occurs. How significant organic N is as a source of N to the nonmycorrhizal plant is questionable. Nevertheless studies conducted in soil microcosms or under field conditions have demonstrated that uptake of intact amino acids can occur regardless of the mycorrhizal status of the plant. The amino acid is generally labeled with both ^{13}C and ^{15}N (i.e., 'dual'-labeled) so that uptake of intact organic N can be distinguished from uptake of mineralized N. If N was being taken up by the plant in mineralized form, ^{15}N levels would be enriched in the plant tissue but there would be no ^{13}C enrichment; whereas, if N was being taken up in organic form, both ^{13}C and ^{15}N would be enriched in the plant tissue. The uptake of the amino acid glycine has been used widely to demonstrate such an uptake mechanism (Table 3). However, glycine is a relatively poor substrate for microbial growth and the amounts of glycine added in such experiments generally far exceed levels normally found in the soil solution. Thus, what has been observed in these studies may not be applicable to ambient conditions. Moreover, although the uptake of glycine in intact organic form is high in plants harvested a few hours or days after glycine addition, in the longer term, plants probably capture more total N from glycine in the form of mineralized N (Table 3). Further progress in understanding plant uptake of organic N sources requires advances in analytical techniques and also measurement of amino acids at more realistic concentrations.

Root Responses

Root systems are modular which allows them to be very plastic in their response to prevailing environmental factors, including the distribution of nutrient resources. When N availability is low, plants generally increase their root weight ratio (RWR; ratio of root weight to total plant weight) because more biomass is allocated to the root system. However, the RWR is not necessarily the best root trait to measure in order to assess the ability of plants to capture water, N, and nutrients in general. The RWR will include old lignified roots not actively involved in nutrient uptake *per se*, and also changes in the architecture of the root system in response to nutrient availability can occur independent of changes in root biomass. Uptake of nutrients is more closely related to root length than root mass. Roots can only absorb ions from the soil that are transported to them by diffusion and mass flow. Beneficial physiological changes in ion uptake rates by roots can enhance uptake of mobile ions such as nitrates. In contrast, more immobile ions (e.g., phosphate) are limited more by diffusion to the root rather than uptake at the root surface so would be less affected by changes in root uptake rates. The diffusion coefficient for NO_3^- in the soil environment is approx. $10^{-10}\,m^2\,s^{-1}$ and, in contrast, the diffusion coefficient for phosphate is approx. 10^{-13}–$10^{-15}\,m^2\,s^{-1}$. The lower diffusion coefficient of phosphate is due to insoluble complexes formed by phosphate ions with cations such as Al^{3+}, Fe^{3+}, and Ca^{2+}. Diffusion coefficients for NH_4^+ range between approx. 10^{-10} and $10^{-12}\,m^2\,s^{-1}$, but the availability of NH_4^+ is further reduced by sorption in soil. Thus, the distribution of NO_3^- in soils should be more uniform than

Table 3 Uptake of intact glycine from various studies

Plant species	*Harvest time after addition of glycine*	*Glycine added taken up in intact form (%)*	*Mycorrhizal status*	*Comments*
Vaccinium myrtillus[a]	Few hours to 7 days	91	Ericoid mycorrhizal	Glycine injected into the mor layer of a boreal forest. Rates of glycine uptake were similar to those of [^{15}N]ammonium
Deschampsia flexuosa[a]		64	Arbuscular mycorrhizal	
Pinus sylvestris[a] and *Picea abies*[a]		42	Ectomycorrhizal	
Trifolium hybridum[b]	21 h	23	Arbuscular mycorrhizal	Plants grown in pots in the field. None of the plant species took up glycine at greater rates compared with inorganic N
Trifolium pratense[b]		23	Arbuscular mycorrhizal	
Phleum pratense[b]		19	Arbuscular mycorrhizal	
Ranunculus acris[b]		20	Not tested	
Triticum aestivum[c]	4–8 h	20	Nonmycorrhizal	Plants grown in the field
Lolium perenne[d]	3 h	80	Nonmycorrhizal	Plants grown in solution culture
Plantago lancelota[e]	41 days	Negligible	Nonmycorrhizal and arbuscular mycorrhizal	Plants grown in pots. Uptake of intact glycine was negligible, but 72% of N from glycine was detected in the plants

[a]Data from Näsholm T, Ekblad A, Nordin A, Giesler R, Högberg M, and Högberg P (1998) Boreal forest plants take up organic nitrogen. *Nature* 392: 914–916.
[b]Näsholm T, Huss-Danell K, and Högberg P (2000) Uptake of organic nitrogen in the field by four agriculturally important plant species. *Ecology* 81: 1155–1161.
[c]Näsholm T, Huss-Danell K, and Högberg P (2001) Uptake of glycine by field grown wheat. *New Phytologist* 150: 59–63.
[d]Thornton B (2001) Uptake of glycine by non-mycorrhizal *Lolium perenne*. *Journal of Experimental Botany* 52: 1315–1322.
[e]Hodge A (2001) Arbuscular mycorrhizal fungi influence decomposition of, but not plant nutrient capture from, glycince patches in soil. *New Phytologist* 151: 725–734.

that of phosphate and, in principle, even a relatively low root length density (RLD) should be enough to capture sufficient N from the soil when present in the form of NO_3^-. In actuality, plants have to capture N (and indeed other nutrients) in competition with other plant roots as well as soil microorganisms, also a large number of inputs in natural systems occur in organic form via leaf litter, dead roots, and animals, leading to a heterogeneous distribution of resource supply both spatially and temporally. Thus, the distribution of resources in soil, even of NO_3^-, is not uniform but patchy.

Root systems can respond to nutrient-rich zones or patches through proliferation of roots as demonstrated in the classic experiments conducted by Drew and colleagues in the 1970s. Barley roots exposed to a localized supply of nitrate, ammonium, and phosphate increased both the initiation (i.e., number) and extension (i.e., length) of first- and second-order laterals within that localized zone. Moreover shoot weight was only slightly depressed when these nutrients were supplied locally compared with controls receiving a uniformly high supply of these nutrients throughout their entire root system. However Drew's experimental system involved plants grown in sand units supplied with nutrient solutions and did not reflect the complexity of the soil environment. Nevertheless, the study demonstrated that roots respond, often spectacularly, to inorganic N as well as phosphate-rich zones (**Figure 1**).

Although root proliferation has been widely reported, it has only recently been demonstrated to have a benefit in terms of N capture in soil. When two different plant species are grown in competition for resources from an organic N-rich patch (grass shoots dual-labeled with the stable isotopes ^{15}N and ^{13}C) added to soil, plant ^{15}N capture was related to RLD produced in the patch. No ^{13}C enrichments were detected in the plant tissue, thus N must have been taken up as mineralized N. Thus, when in competition for N, even mineralized N, which can be highly mobile in soil, RLD is important, because the supply of N from the organic patch is finite. In contrast, in the absence of plant–plant competition, the supply of N can be thought of as invariant and RLD is less important.

The timescales to construct new roots in these nutrient-rich zones are generally weeks rather than days (**Figure 2**). In contrast, microbial populations can respond much more rapidly in time scales of hours to days. Thus, root proliferation responses are considered to be more important in terms of plant–plant competition rather than plant–microbial competitive interactions. High-order laterals produced upon root proliferation also tend to have higher N concentrations and thus are more expensive to

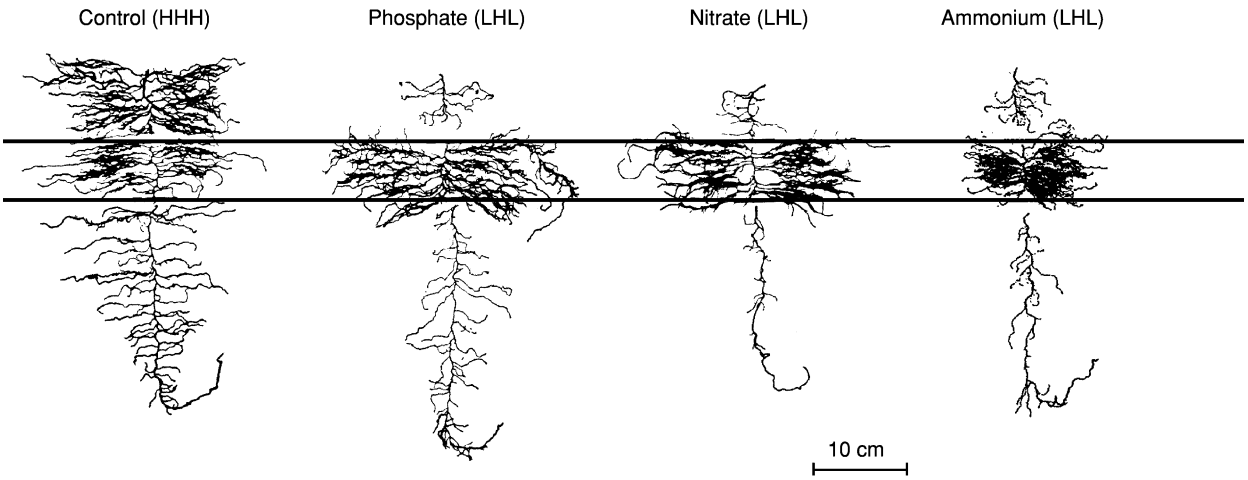


Figure 1 The proliferation of primary and secondary laterals of barley (*Hordeum vulgare*) root systems when supplied with a 100-fold greater concentration of phosphate, nitrate, or ammonium to their middle part of their root system (indicated by the lines). H, high; L, low, which refer to the nutrient concentrations of phosphate, nitrate, or ammonium received by the top, middle, and bottom of the root system, respectively. The control which received uniformly high levels of nutrients to all parts of its root system is shown as a comparison. Reproduced with permission from Drew MC (1975) Comparison of the effects of a localized supply of phosphate, nitrate, ammonium and potassium on the growth of the seminal root system, and the shoot, in barley. *New Phytologist* 75: 479–490, with kind permission of the trustees of the *New Phytologist*.

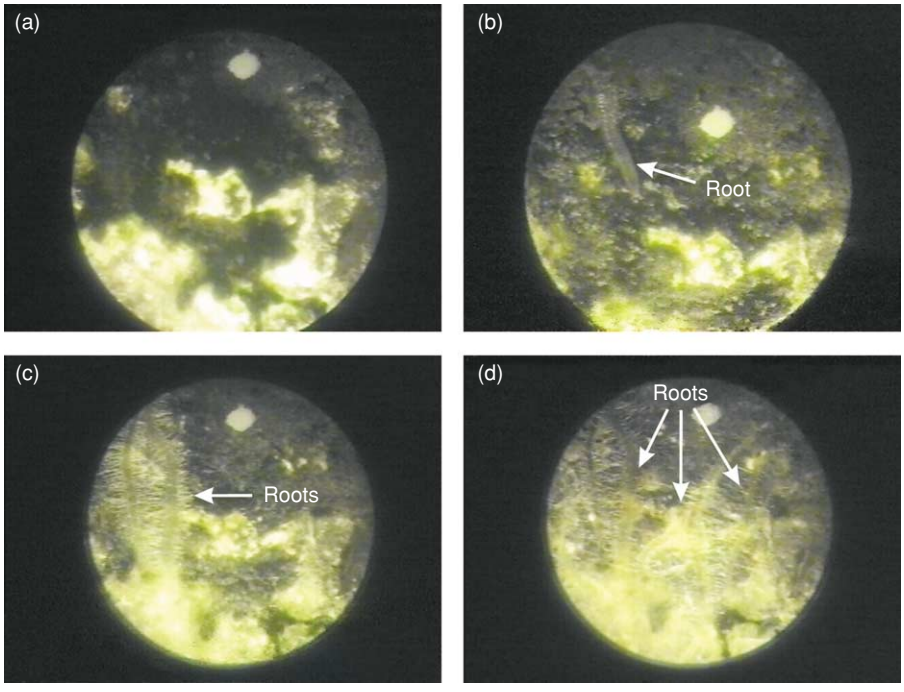


Figure 2 Root growth and proliferation in an N-rich organic material added as a spatially discrete patch and comprising lyophilized algal cell organic material patch (C:N ratio of 3.2:1). (a) One day after the patch was added, no roots in the patch zone were observed; (b) 21 days after the patch was added, a single root was seen growing into the N-rich zone. (c) The same image 25 days and (d) 48 days after the patch was added: increased root production and root-hair development in the N-rich zone was clearly observed. After 49 days, the plants were harvested and found to contain 54% of the N orginally added in the algal cell patch in their tissues. Adapted from *Trends in Plant Science*, vol. 5, Hodge A, Robinson D, and Fitter A. Are microorganisms more effective than plants at competing for nitrogen? pp. 304–308. © 2000, with kind permission from Elsevier Science.

construct but presumably the benefits of exploiting a nutrient-rich patch and the capture of N from it outweigh the disadvantages of these roots' construction costs. Moreover, increased root turnover, i.e., a decrease in root longevity, is often observed in both fertile soils and fertile patches, but again presumably the advantages of having a young population of roots actively involved in nutrient-uptake processes outweigh the disadvantages of high root turnover rates.

Mycorrhizal Contribution to Plant N Uptake

Most plants form mycorrhizal associations under natural conditions. Mycorrhizal symbiosis is discussed in detail elsewhere so only the benefit of mycorrhizas for plant N uptake will be considered here. Mycorrhizal associations may be expected to enhance N uptake from inorganic N sources because, by virtue of the extension of the fungal mycelium in the soil, they should be better able to compete with the soil microbial community for these N sources compared with plant roots. Many of the fungi involved in the ectomycorrhizal and ericoid mycorrhizal associations have been demonstrated to have a preference for NH_4^+ rather than NO_3^- when grown in pure culture. The reasons for this are probably the same as the preference for inorganic shown by plant roots in that it reflects the dominant inorganic N source under which these fungi usually form their associations. In addition, the arbuscular mycorrhiza (AM), ericoid, and ectomycorrhizal associations have all enhanced N capture for their host plant when supplied with NH_4^+. Mycorrhizas are probably more effective at enhancing N capture when supplied with NH_4^+ than NO_3^- because of the differences in mobilities of these ions. As NH_4^+ is much less mobile than NO_3^-, NH_4^+ uptake can be enhanced in the same manner that phosphate (which is even more immobile) is enhanced by mycorrhizal colonization, whereas the very mobile NO_3^- diffuses readily to the root surface. However, under certain circumstances, such as drought conditions, when ion mobility is restricted, mycorrhizas have also been demonstrated to enhance NO_3^- uptake for their host plant. In addition, AM colonization of legume roots has been demonstrated to have a positive effect on N nutrition for these plants when their hosts are in a tripartite symbiosis with N_2-fixing *Rhizobium* spp. but in this case the enhanced N nutrition of the plant is believed to be a consequence of increased plant P status and uptake of other micronutrients which may have indirectly enhanced the N_2-fixation process in addition to improving overall plant growth. AM mycorrhizal fungi have also been suggested to play a role in the redistribution of N within a plant community via fungal connections from legume to nonlegume plants, allowing transfer of N-containing products to the nonlegume host. However, it is unlikely that such a mechanism alone would be sufficient to meet the N demand of the nonlegume host.

The ectomycorrhiza and ericoid mycorrhizas also enhance N nutrition for the host plant by accessing organic N sources. These fungi produce hydrolytic enzymes which mineralize complex N forms, such as chitin and proteins, and then absorb the NH_4^+ liberated. Uptake of N from simple organic sources such as amino acids and peptides has also been demonstrated. However, there is considerable variability among species and even among different strains of the same species in their ability to acquire N from organic sources. For example ectomycorrhizal fungi such as *Suillus bovinus*, *Paxillus involutus*, and *Hebeloma crustuliniforme* are classified as 'protein fungi' because their growth yields on organic N are the same, or in excess of, yields on ammonium, whereas the so-called 'nonprotein fungi' *Laccaria laccata* and *Lactarius rufus* grow poorly and have low yields on organic N (gliadin and bovine serum albumin, BSA) sources. Once the more-complex organic N sources are cleaved by the mycorrhizal fungi, this may aid capture of N in more simple organic form by either different mycorrhizal fungi or plant roots. The ability of AM associations to use organic N sources to any great extent is controversial, although it has been shown that AM fungi can access such N sources and transfer the N to their host plant.

There is much current interest in the uptake of intact organic N sources by plants; however, its significance as a major N-acquiring system in most ecosystems is still unclear. For plants, the availability of N whether in inorganic or organic form will be determined by the activities of the microbial community in soil and, thus, plant uptake of N cannot be studied independently of these microorganisms. Plants are effective at competing for N with microorganisms because of their much slower tissue turnover times. Also, by remobilization of N from senescing tissues, plants can reduce their losses of N back into the soil system.

See also: **Acid Rain and Soil Acidification**; **Buffering Capacity**; **Carbon Cycle in Soils:** Dynamics and Management; **Foliar Applications of Nutrients**; **Isotopes in Soil and Plant Investigations**; **Mycorrhizal Fungi**; **Nitrogen in Soils:** Cycle; Symbiotic Fixation; **Organic Residues, Decomposition**; **Rhizosphere**; **Root Architecture and Growth**

Further Reading

Drew MC (1975) Comparison of the effects of a localized supply of phosphate, nitrate, ammonium and potassium on the growth of the seminal root system, and the shoot, in barley. *New Phytologist* 75: 479–490.

Fitter A, Hodge A, and Robinson D (2000) Plant response to patchy soils. In: Hutchings MJ, John EA, and Stewart AJA (eds) *The Ecological Consequences of Environmental Heterogeneity*, pp. 71–90. Oxford, UK: Blackwell Science.

Hodge A, Robinson D, Griffiths BS, and Fitter AH (1999) Why plants bother: root proliferation results in increased nitrogen capture from an organic patch when two grasses compete. *Plant, Cell and Environment* 22: 811–820.

Hodge A, Robinson D, and Fitter A (2000) Are microorganisms more effective than plants at competing for nitrogen? *Trends in Plant Science* 5: 304–308.

Hutchings MJ and de Kroon H (1994) Foraging in plants: the role of morphological plasticity in resource acquistion. *Advances in Ecological Research* 25: 159–238.

Lipson D and Näsholm T (2001) The unexpected versatility of plants: organic nitrogen use and availability in terrestrial ecosystems. *Oecologia* 128: 305–316.

Marschner H (1995) *Mineral Nutrition of Higher Plants*, 2nd edn. London, UK: Academic Press.

Robinson D (1994) The responses of plants to non-uniform supplies of nutrients. *New Phytologist* 127: 635–674.

Smith SE and Read DJ (1997) *Mycorrhizal Symbiosis*. San Diego, CA: Academic Press.

Symbiotic Fixation

J I Sprent, University of Dundee, Dundee, UK

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Nitrogen-fixing bacteria are known to form symbiotic associations with some members of all major groups of plants, as well as with some fungi. Although there are numerous reports of nitrogen-fixing bacteria occurring in animals, for example termite guts, the significance to their hosts remains to be proven. In global terms, nodulated plants (both legume and actinorhizal) fix most nitrogen, but many of the other symbioses are very important within their own ecosystems. All nonnodulated nitrogen-fixing symbioses have cyanobacteria as their endosymbiont.

Common features to most symbiotic systems that have been studied in detail are: first, that given a supply of combined nitrogen (nitrate, ammonium, organic N) from the soil, they will tend to use this in preference to fixing N_2; in most cases this partially or completely excludes the development of the symbiosis, a situation frequently described in agricultural systems, but which is likely to occur also in natural ecosystems; second, that the host plant assimilates the ammonium produced by the nitrogen-fixing endophyte. Thus a series of control mechanisms is invoked whereby the plant host supplies organic carbon, amino compounds, and mineral elements in exchange for ammonium. The only exception is in lichens, where the fungal component is probably responsible for supplying minerals from the soil or other substratum, and carbon fixation is carried out by either a green alga or the nitrogen-fixing endophyte (see Table 1).

Where estimates of fixed nitrogen are given, these are subject to large errors because of limitations of available methods for use on a field scale.

Cyanobacterial Symbioses

Table 1 lists the major types of symbiosis between cyanobacteria and plants. Nitrogen-fixing cyanobacteria usually have nitrogenase located in specialized cells known as heterocysts, and this is true of almost all symbiotic forms studied.) Exceptions to this observation are those listed in Table 1 as having 'coccoid forms' of endophyte. Shortly after infection, heterocyst frequency may be similar to that of free-living forms, but subsequently it increases markedly as the endophyte fixes nitrogen using carbon fixed by the host, making photosynthesis by vegetative cells unnecessary. Further, in many systems, the cyanobacteria do not receive light, although they continue to produce photosynthetic pigments. Only in two membered lichens (Table 1) does the cyanobacterium need to fix both CO_2 and N_2.

Another common feature to all except some diatoms is that water is necessary for infection by motile filaments (hormogonia) of the cyanobacterium. In bryophytes *Azolla* and *Gunnera*, these invade cavities in the host plant: *Azolla* also has a means of dispersing its endophyte during spore production. Within cavities, host cells make close contact with their symbionts and often undergo structural modifications enabling the efficient interchange of nutrients. The need for water may be one of the reasons why cyanobacterial symbioses are limited in number. Only lichens are found in very dry areas.

Lichens

Identification of the fungi involved in lichen formation is difficult, but it is likely that about 20% of known fungi can be found in lichens. Of these, the vast majority are ascomycetes. No phycomycetes are known to form lichens, although one coenocytic species, *Geosiphon pyriforme*, produces bladders that can house *Nostoc punctiforme*. This interesting association may

Table 1 Major categories of nitrogen-fixing symbioses involving cyanobacteria

Category	*Genus of host*	*Genus of cyanobacterium*	*Comments*
Diatoms	*Bacteriastrum*	*Richelia*	Includes both marine and freshwater genera. Marine genera may be important in vertical translocation of organic carbon in oceans
	Chaetoceros	*Richelia*	
	Denticulata	Coccoid form	
	Epithemia	Coccoid form	
	Hemiaulus	*Richelia*	
	Neostreptotheca	Coccoid form	
	Rhopalodia	Coccoid form	
	Rhizosolenia	*Richelia*	
	Rupertia	*Nostoc*	
	Streptotheca	Coccoid form	
Lichens	*Collema*	*Nostoc*	Lichens are important in boreal and temperate forests, in arid areas, and on rocks, often in coastal areas. They may have only two components (fungal, cyanobacterial), or three, when a green alga is also present. In the former the cyanobacterial member fixes both CO_2 and N_2, in the latter, only N_2. Some lichens have unicellular endophytes
	Dendriscocaulon	*Scytonema*	
	Dictyonema	*Scytonema*	
	Ephebe	*Fischerella*	
	Leptogium	*Nostoc*	
	Lichina	*Calothrix*	
	Lobaria	*Nostoc*	
	Massalongia	*Scytonema*	
	Nephroma	*Nostoc*	
	Pannaria	*Nostoc*	
	Parmelia	*Nostoc*	
	Peltigera	*Nostoc*	
	Placopsis	Uncertain	
	Placynthium	*Dichothrix*	
	Polychidium	*Scytonema*	
	Pseudocyphellaria	*Nostoc*	
	Solorina	*Nostoc*	
	Stereocaulon	*Nostoc*	
	Sticta	*Nostoc*	
Hornworts	*Anthoceros*	*Nostoc*	The cyanobacteria are located in pockets in the thallus
	Phaeoceros	*Nostoc*	
	Notothylas	*Nostoc*	
	Dendroceros	*Nostoc*	
Thalloid liverworts	*Blasia*	*Nostoc*	The cyanobacteria are located in pockets in the thallus
	Cavicularia	*Nostoc*	
Ferns	*Azolla*	*Anabaena*	Although '*Anabaena*' is in common use, it is really a *Nostoc* and is found in leaf pockets
Cycads	*Bowenia*	*Nostoc* is very common, other reports include *Calothrix*	All extant cycads are symbiotic. Cyanobacteria are found in specially modified roots, usually referred to as 'coralloid,' but sometimes called 'nodules'
	Ceratozamia		
	Cycas		
	Dioon		
	Encephalartos		
	Lepidozamia		
	Macrozamia		
	Microcycas		
	Stangeria		
	Zamia		
Angiosperms	*Gunnera*	*Nostoc*	All extant species are symbiotic. Unlike all the above, except some diatoms, cyanobacteria are intracellular

well represent another type of symbiosis for which more examples remain to be discovered.

In some areas, lichens are responsible for significant inputs of fixed nitrogen into soils, either directly or indirectly. Two of the most widely studied systems are the boreal forests and arctic tundra, where nitrogen-fixing lichens (often thalloid *Sterocaulon* spp.) may be the dominant component of the ground flora, as in parts of Scandinavia and northern Canada. Although the total amount of nitrogen fixed may seem low by agricultural standards (1–2 compared with up to $400\,\mathrm{kg\,N\,ha^{-1}}$ per year), when set against the requirements of growth at low temperatures, it may be sufficient not only for the lichens, but, following

mineralization, for other organisms in the ecosystem. Similar arguments can be applied to the epiphytic lichen *Lobaria oregana* found on old-growth Douglas fir in the Pacific Northwest of the USA, where it contributes 3–4 kg N ha^{-1} per year, which may enter the soil via stem flow. In dry areas, both cold and hot, ranging from Antarctica to the tropics, lichens and free-living cyanobacteria may be among the few photosynthetic organisms found. Lichens can survive years of desiccation then resume growth and both CO_2- and N_2-fixing activity. However, little fixed nitrogen may be released into the soil, as the lichens are part of a tightly controlled ecosystem where the presence of bacteria and other organisms suggest that the entire nitrogen cycle occurs within it. Similar arguments apply to crustose lichens occurring on rocks (and buildings).

Bryophytes

This category includes the hornworts and thalloid liverworts of Table 1. They are found in all parts of the world on bare, nitrogen-deficient soils that are very moist but not flooded. Botanically they are interesting because of their life cycle and the means of transport between the gametophyte and sporophyte generations. Figure 1a shows the main features of a hornwort, *Anthoceros*. A relevant question relating to this association is: Why are there so few bryophytes with nitrogen-fixing symbionts? Current molecular evidence suggests that extant symbiotic forms arose from an ancient branch of the bryophyte lineage. Many modern bryophytes, particularly mosses, have adapted to live in harsh environments, including dry areas. As water is essential for the cyanobacteria to enter pockets of the developing thallus, there may have been a selective disadvantage to symbiotic forms.

Ferns

Azolla is an unusual fern in being both aquatic and having a nitrogen-fixing endophyte (Figure 1b). Being aquatic, it is not strictly relevant to a discussion on soils, but, as it has been and still is widely used as a nitrogen fertilizer for wetland rice, it will be mentioned. Its agricultural use has been the subject of many reviews. These have taken different standpoints

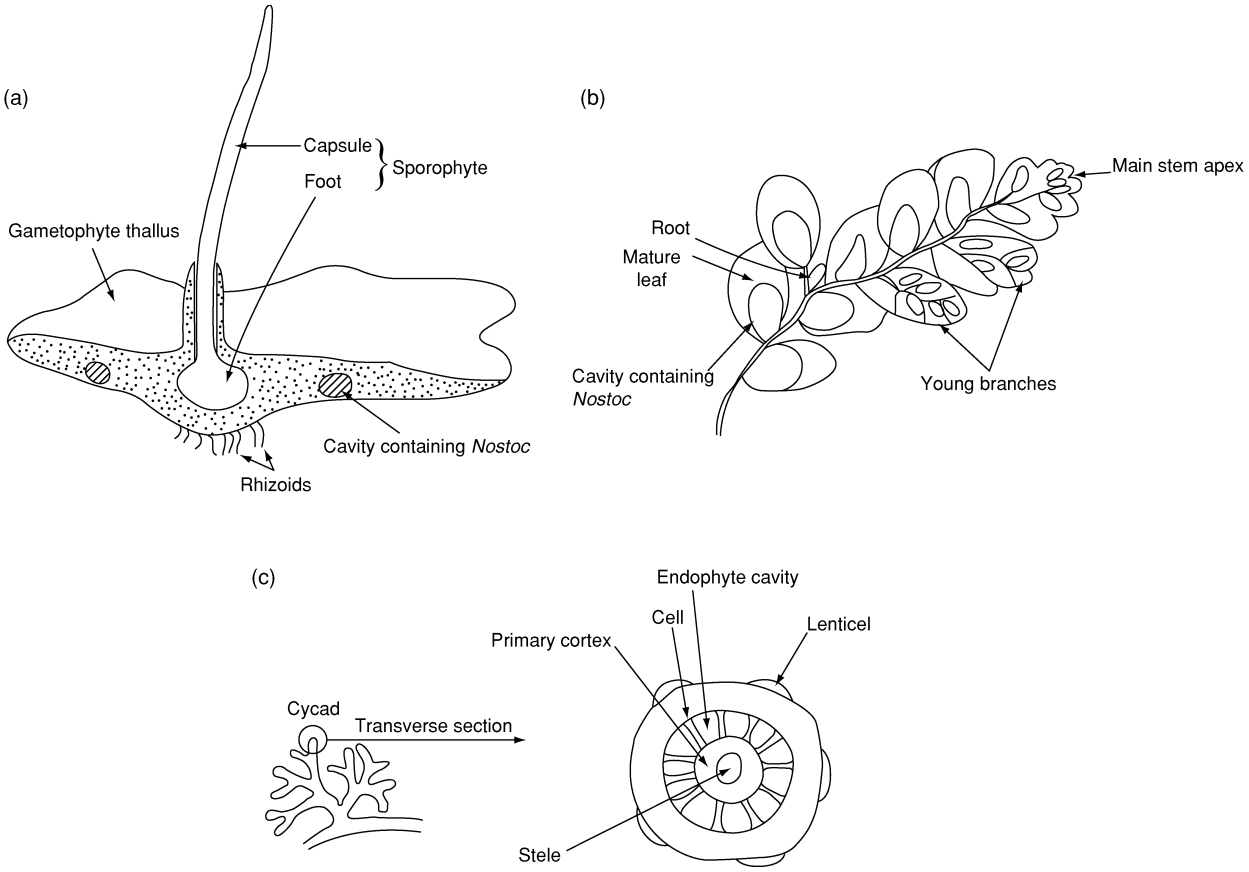


Figure 1 Some cyanobacterial symbioses with plants: (a) the hornwort *Anthoceros*, approximately ×10 natural size; (b) *Azolla*, ×10 natural size; (c) generalized cycad, approximately natural size; and magnified transverse section. © J Sprent.

as to the practicality of the use of *Azolla* in modern rice-growing, although there is general agreement that it can yield significant amounts of easily mineralized N. However, in many areas of the world where *Azolla* has been introduced, for example South Africa, it can also be a nuisance by blocking water channels. Recently attention has been given to control of *Azolla* by insects and other agents, and in its use for feeding fish, poultry, and pigs.

Why only one genus of fern is known to have a nitrogen-fixing symbiosis is not clear. Current fossil evidence is that *Azolla* evolved comparatively recently compared with other ferns, suggesting that it may have been a 'one-off' event to fill a particular need.

Cycads

Many texts state that cycads are only found naturally in the southern hemisphere, but this is not true. Although their frequency may be greater in the southern hemisphere, cycads occur in tropical and warm temperate areas of both hemispheres. This distribution can be traced through the fossil record, which shows that cycads were around at least 230 million years ago, before Pangaea broke up into Gondwana (which moved south) and Laurasia (which moved north). It is possible that the comparative rarity of natural stands of cycads in the northern hemisphere relates to destruction of their habitat. Certainly several of them are listed as endangered. Biologically, cycads are particularly interesting in that all extant and fossil forms appear to be symbiotic and yet no evolutionarily related groups of plants (for example gymnosperms) have nitrogen-fixing members. There must have been pressure toward symbiosis at an early stage in the evolution of cycads. This may have been a shortage of combined nitrogen, but evidence supporting this hypothesis is not yet clear. Many of the most abundant cycads today live in dry environments with low-nutrient soils, but others are found in rainforests and seasonally dry areas.

Hormogonia apparently invade gaps in the cortex of cycad root and stimulate the formation of a much-branched structure, the coralloid root (**Figure 1c**). Within this, *Nostoc* is found in cavities between the primary and secondary cortex. It is very easy to find coralloid roots in cycads grown as pot plants, as they are near to and often protrude from the surface of the soil. However, in other cycads, for example *Macrozamia reidlii*, coralloid roots are found well below ground level (down to 1 m). This species is common in areas of Western Australia that are prone to fire. *M. reidlii* can regenerate after fire from subterranean tubers and the deep-seated coralloid roots remain functional. In some eucalypt forests, cycads apparently provide an important source of nitrogen, equivalent to that of understorey legumes in other eucalypt forests. The largest genus of cycads is *Encephalartos*, widely distributed in a variety of habitats in east and southern Africa. Some species are fire-resistant, others frost-tolerant, and one at least is deciduous. Many standard textbooks describe cycads as 'living fossils.' Whilst they may not be as dominant as in former times, in areas such as parts of Australia and South Africa and in the everglades of Florida they are still a prominent component of natural vegetation. However, as they are widely collected for use in horticulture, some species are now threatened.

The list of cycad genera given in **Table 1** is conservative. Other genera have been proposed, of which the most often cited is *Chigua*, with two species native to Colombia; however, some workers believe that these should be included in *Zamia*. The division of cycads into families is also an area of current uncertainty. However, as far as soils are concerned, it is probably safe to assume that, at any site where cycads grow, there will be an input of symbiotically fixed nitrogen.

Angiosperms

Gunnera is the only genus of angiosperm known to have a symbiotic association with a cyanobacterium, in this case *Nostoc punctiforme*. Taxonomically, *Gunnera* appears to be isolated, placed in its own family, and far removed from other angiosperms with nitrogen-fixing symbioses. There are approximately 40 species and all contain *Nostoc* in modified glands (sometimes referred to as nodules) at the base of petioles. Plants are herbs that vary in size from those with leaves over 1 m in height to small creeping forms. In common with other angiosperms, the endophyte of *Gunnera* is intracellular, in that it is contained within the volume surrounded by a plant cell wall, but extracellular in that it is confined within a membranous sac (now often called a symbiosome). Transport of material into and out of the glands is via phloem, unlike nodulated plants in which outward transport is via the xylem. *Gunnera* appears to lack nitrate reductase, so is unable to use nitrate. All studies suggest that *Nostoc* is capable of fixing enough nitrogen for the needs of its host and hence, in the longer term, is likely to be a good source of N for soil mineralization. There is abundant fossil pollen evidence that *Gunnera* is of ancient origin and its present distribution, largely in the southern hemisphere, is restricted to areas with high humidity. This may have been its downfall if it was unable to adapt to the increasingly dry conditions that spread over much of the world since angiosperms first evolved. However, *Gunnera* appears to be very important in some of its present locations, where nodulated legumes and actinorhizal plants are

scarce or absent. An extreme case of this is in the Malvinas (Falkland Islands) off the coast of Argentina, where *Gunnera magellanica* is the only native angiosperm with a nitrogen-fixing symbiont. Here it is common and widespread, being found from sea level to all elevations, and clearly has a vital role to play.

Nodulated Angiosperms

These fall into two categories, based on the nature of the endophyte. *Frankia*, a filamentous bacterium belonging to the actinomycetes, nodulates a variety of plants (Table 2). Legumes and the Ulmaceous genus *Parasponia* are nodulated by unicellular bacteria from the α- and β-proteobacteria, collectively known as rhizobia (Table 3). Recent molecular evidence has placed all nodulated plants in a distinct group, known as the Rosid I clade. However, the clade also contains many genera that lack the ability to nodulate. If, as is generally assumed, all members of the clade have some predisposition to form nodules, why is it only expressed in certain genera? The implications of this for the evolution of nodulation are an area of intense current interest, especially as the internal structure of nodules in the two groups is different.

A problem common to all aerobic nitrogen-fixing organisms, including the symbiotic *Frankia* and rhizobia, is to reconcile the need for oxidative phosphorylation to fuel the nitrogen fixation with the fact that the enzyme is inactivated by oxygen. Nodules have a number of ways of controlling the rate of oxygen diffusion to the active enzyme so that nitrogen fixation is balanced with the supply of fixed carbon. In some cases this means reducing the concentration of oxygen from the 20% found in air, by means of fixed and/or variable diffusion barriers. Once a sufficiently low level (about 1%) is reached, there may be a need to maintain a constant flux of oxygen; this is achieved by plant forms of hemoglobin. It has been known for many years that hemoglobin gives most active

Table 2 Genera of plants known to be nodulated by *Frankia*. These are grouped according to family (numbers in parentheses indicate the total number of genera in the family); families are arranged in four groups, separated by an additional line space, according to their position in the Rosid 1 clade of angiosperms. The first group also contains the Ulmaceae, in which *Parasponia*, the only genus outside the legumes known to nodulate with rhizobia, is placed. It should be noted that there are six groups in this clade. Of the other three, two contain no nitrogen-fixing members and the other, the legumes. All groups have nonnodulating members

Family	*Nodulating genera*	*Comments*
Elaeagnaceae (3)	*Elaeagnus* *Hippophae* *Comptonia*	Where known in this group, infection is intercellular and no nodule rootlets are formed. Vesicles in the first two families are spherical and septate, with the exception of *Ceanothus*, where they lack septa. In the Rosaceae, vesicles are elliptical and aseptate. Within the whole group, Rosaceae is sister to the other two families
Rhamnaceae (55)	*Ceanothus* *Colletia* *Discaria* *Kentrothamnus* *Retanilla* *Talguenea* *Trevoa*	
Rosaceae (100)	*Cercococarpus* *Chamaebatia* *Cowania* *Dryas* *Purshia*	
Betulaceae (6) Myricaeae (3)	*Alnus* *Comptonia* *Myrica*	In this group, infection is through root hairs. Nodule rootlets are present in all genera except *Alnus*
Casuarinaceae (5)	*Allocasuarina* *Casuarina* *Ceuthostoma* *Quadrangula* *Gymnostoma*	Vesicles are absent from *Allocasuarina* and *Casuarina* nodules. *Quadrangula* has recently been proposed for two species previously included in *Casuarina*. In all other genera vesicles are septate, but vary in shape from spherical to club-shaped (Myricaeae)
Coriariaceae (1) Datiscaceae (3)	*Coriaria* *Datisca*	Details of infection in this group are still unclear. Nodule rootlets occur in *Datisca*. Vesicles are filamentous and aseptate. Nodules on these genera have several other distinct features

legume nodules their pink internal color, but it is now apparent that the pigment is also a necessary part of the functioning of some actinorhizal nodules. Species living in waterlogged conditions may also need an anatomical pathway that allows air to pass through a series of spaces from above the water level to the nodule. This may be achieved by negatively geotropic nodule rootlets, as in *Myrica*, or by a continuous layer

Table 3 Currently recognized bacterial genera and species known to form nitrogen-fixing root nodules in association with legume roots (sometimes stems). All except those indicated are in the α-proteobacteria, but not all are in the same branch. The genera *Agrobacterium* and *Allorhizobium* may be incorporated into *Rhizobium*

Genus	*Species*	*Examples of legumes nodulated and comments*
Agrobacterium	?	Some nodulating bacteria appear to belong to this genus, but their exact nature has not been confirmed
Allorhizobium	*undicola*	*Neptunia natans*
Azorhizobium	*caulinodans*	*Sesbania rostrata* stems
Blastobacter	Not determined	*Pterocarpus erinaceous*, *P. lucens*
Bradyrhizobium	*elkanii*	*Glycine max*
	japonicum	*Glycine max*
	liaoningense	*Glycine max*
	sp. (host genus)	Several potential species e.g., sp. (*Lupinus*) are currently defined by their hosts
Burkholderia	Two spp.?	*Aspalathus carnosa*
		Machaerium lunatum
		In the β-proteobacteria
Devosia	*neptuniae*	*Neptunia natans*
Mesorhizobium	*amorphae*	*Amorpha fruticosa*
	chacoense	*Prosopis alba*
	ciceri	*Cicer arietinum*
	huakuii	*Astragalus sinicus*
	loti	*Lotus corniculatus*
	mediterraneum	*Cicer arietinum*
	plurifarium	*Acacia senegal*
	tianshanense	*Glycyrrhiza pallidiflora*
Methylobacterium	*nodulans*	Some *Crotalaria* spp.
Ralstonia	*taiwanensis*	*Mimosa* spp. In the β-proteobacteria. Now renamed *Wautersia taiwanensis*
Rhizobium	*etli*	
	bv. phaseoli	*Phaseolus vulgaris*
	bv. mimosae	*Mimosa affinis*
	gallicum	
	bv. gallicum	*Phaseolus vulgaris*
	bv. phaseoli	*Phaseolus vulgaris*
	galegae	
	bv. officinalis	*Galega officinalis*
	bv. orientalis	*Galega orientalis*
	giardinii	
	bv. giardinii	*Phaseolus vulgaris*
	bv. phaseoli	*Phaseolus vulgaris*
	hainanense	*Desmodium sinuatum*
	huautlense	*Sesbania herbacea*
	leguminosarum	
	bv. phaseoli	*Phaseolus vulgaris*
	bv. trifolii	*Trifolium pratense*
	bv. viciae	*Pisum sativum*
	mongolense	*Medicago ruthenica*
	tropici	*Phaseolus vulgaris*. This is a complex species, with several types within it
	yanglingense	*Gueldenstaedtia multiflora*
Sinorhizobium	*arboris*	*Acacia senegal*
	fredii	*Glycine max*
	kostiense	*Acacia senegal*
	medicae	*Medicago truncatula*
	meliloti	*Medicago sativa*
	saheli	*Sesbania pachycarpa*
	terangae	
	bv. acaciae	*Acacia laeta*
	bv. sesbaniae	*Sesbania rostrata* roots

of aerenchyma, as in a number of wetland legume genera.

Within the two groups of nodulated angiosperms, individual genera and species are nodulated by different strains of either *Frankia* or rhizobia, in other words there is specificity between the partners. In order for nodulation to proceed, the endophytes must recognize and be recognized by their hosts, gain entry into roots (occasionally stems in legumes), by-pass the host's defense reaction, and stimulate production of nodule tissue. The recognition process involves a series of exchanges of key molecules produced by the host and the endophyte, following which various nodulation genes in each partner are induced. Some cell division in host roots often starts toward the end of this recognition pathway. Entry may be gained via root hairs, as in many legumes and actinorhizal genera such as *Alnus* or by penetration between epidermal cells or at the point of emergence of lateral roots (adventitious roots in stem-nodulating legumes). Subsequent to the initial infection, nodule development in legumes and actinorhizal plants proceeds in different ways. The common feature of all actinorhizal nodules and those induced by rhizobia on *Parasponia* is that the vascular system is central and the infected tissue cortical. In legume nodules the vascular system is cortical and the infected tissue is central. Even so, there is increasing evidence that similar nodulation genes may be involved in both types, consistent with a common predisposition for nodulation.

Actinorhizal Plants

Plants nodulated by *Frankia* are collectively referred to as actinorhizal. Technical difficulties, coupled with a lack of funding for nonagricultural species has resulted in research on these plants lagging behind that on nodulated legumes. However, progress is now rapid and the important role of these plants is increasingly being recognized. Table 2 lists the currently agreed actinorhizal genera, and this list is unlikely to be radically altered, as close relatives of the nodulating genera have been extensively examined in almost all areas of the world. However, there have been isolated (unconfirmed) reports of nodules on members of the familes Cucurbitaceae and Zygophyllaceae, both of which are in the Rosid I clade. Molecular evidence has grouped actinorhizal plants in a way that fits well with the known variations in infection pathways and nodule characters, as can be seen in Table 2. All nodules have indeterminate growth (Figure 2), most are perennial, and older parts may become very woody, with the active regions confined to the tips of nodule lobes. It is now generally agreed that nodulation with *Frankia* has evolved on a number of different occasions. One question that is now being addressed is whether or not there is a relationship between particular host plants and the type of *Frankia* with which they nodulate. Although host–endophyte specificity has been shown in a number of cases, it varies widely. On balance it is probably true to say that there is more evidence for coevolution between host plants and *Frankia* than there is for legumes and rhizobia.

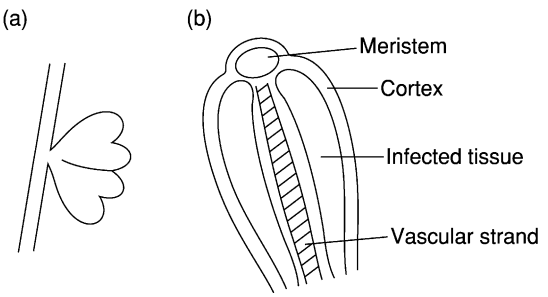


Figure 2 Actinorhizal nodules. (a) A young nodule as found in *Alnus*. Nodules branch repeatedly, forming a woody coralloid structure up to 8–10 cm in diameter; (b) longitudinal section through the tip of a lobe of a generalized nodule; in some genera this also produces a nodule rootlet (Table 2). The detailed structure of the infected tissue varies greatly with genus, and not all nodules have lobes which are symmetrical in longitudinal section. © J Sprent.

Growing *Frankia* in pure culture from crushed nodules has proved a difficult task and one that has still not been achieved for all host genera. Molecular methods have enabled more detailed studies of *Frankia* in soil to be undertaken and it now seems likely that, as with rhizobia, it has a saprophytic stage that may tide it over during periods when no compatible host is present. Being filamentous, *Frankia* has the potential for differentiation. This may occur in the production of vesicles and of spores. Vesicles are usually swollen structures formed at the ends of filaments under conditions of low oxygen concentration. Where present they are the site of nitrogenase activity and the thickness of their lipid-based walls is adjusted to prevent oxygen reaching the enzyme. The location of vesicles within the infected host cells varies considerably between genera. Some strains of *Frankia* form spores within nodules; others do not. Evidence that spores are effective perennating structures is not strong and spore-positive nodules often show lower nitrogenase activity.

Actinorhizal plants are mainly woody in nature, most being shrubs or trees. They are found worldwide from the arctic (*Dryas*, an unusual genus in that it nodulates in some areas but not others) to the tropics (members of the Casuarinaceae), on high mountains (Himalayan alder) and in lowland bogs (*Myrica gale*). The major type of vegetation from which they seem to

be absent is tropical rain forest. Many are drought-resistant (*Cercocarpus* and other shrubs from the rain shadow of the US Rocky Mountains), a factor often coupled with saline tolerance (some species of *Casuarina*). They are used in a variety of ways, including forestry (both as a nitrogen input, e.g., *Ceanothus* and as a source of timber, e.g., *Alnus*, *Casuarina*), for land reclamation (*Casuarina*) for animal browse (e.g., *Purshia*), as an agricultural crop (*Hippophäe*), and for pharmaceutical purposes (*Myrica*). The amounts of nitrogen fixed when they are growing in their natural environment or when planted is related to soil and climatic conditions, but is similar in magnitude to that of legumes in similar environments (up to approximately $100\,kg\,N\,ha^{-1}$ per year). As with legumes, optimum nitrogen fixation may need formation of other structures such as endo- and/or ectomycorrhizas and cluster (proteoid) roots.

Legumes

The Leguminosae, also known as the Fabaceae, is the third-largest family of flowering plants. However, only a small number have been exploited by humans, and our knowledge of nitrogen fixation (if any) by most of them ranges from nonexistent to sparse. A great deal of effort has gone into describing legumes from around the world, but not many plant taxonomists look beneath their feet, generally using only characters such as flowers, fruit, and habit that are visible above ground. Most taxonomists agree that the family should be divided into three subfamilies, Caesalpinioideae, Mimosoideae, and Papilionoideae, with a possible fourth currently being considered. It is also widely understood that nodulation is rare in the first, common in the second, and virtually always present in the third. Recent work has not altered this understanding (Table 4), but has identified some closely related groups, mostly trees, from each subfamily that lack the ability to nodulate. All known naturally nonnodulating legumes are from tropical and subtropical regions, although some are cultivated at higher latitudes. Their distribution within the family is consistent with the suggestion that nodulation in legumes arose at least twice, possibly three or four times.

Legume nodules are found in a variety of shapes, sizes, and structures (Figure 3). As with actinorhizal plants, nodule features tend to be uniform with taxonomic groupings. Some have determinate growth, with only a transient meristem; these are relatively short-lived. Most are indeterminate, with a distinct apical meristem, often showing branching. These are usually longer-lived and may be perennial. They are generally better able to withstand adverse conditions, because they can resume growth and produce new nitrogen-fixing tissue after periods of stress such as drought or low temperature. Studies with agricultural species such as clovers, alfalfa (lucerne), and beans generally show that nodulated plants are less able to withstand environmental extremes than the same species grown in soil containing combined nitrogen. Many commercial cultivars respond well to 'starter nitrogen,' i.e., a small dose of nitrogen fertilizer at planting. To some extent this reflects the fact that until recently most plant breeders raised crops on very fertile soils and thus may have been inadvertently selecting against efficient nitrogen fixation. Around the beginning of the twenty-first century, there was a major effort to raise legume cultivars that would produce economic yields on less-fertile soils and under environmental stresses such as drought and salinity. However, it is also important to go back to the field and look for wild legumes that grow well in extreme environments (Table 5). The only conditions where nodulated legumes do not seem to occur naturally is in sea water, although some may grow well in saline areas and others submerged in fresh water. In human terms, perhaps even more importantly, it is now being recognized that there are many nodulated legumes that have been used by native peoples for millennia that can provide food often under very adverse conditions. Early settlers in the USA made great use of tubers of *Apios americana*, a species used by American Indians. In dryland Africa, many tubers are used for food and even water under dry conditions. Whilst most of these

Table 4 Genera in legume subfamilies which can and cannot nodulate and for which there is no information. Since legume taxonomy is currently undergoing many rearrangements, all numbers must be regarded as very approximate. Many of the genera for which there are no reports are monotypic. Only three genera are known so far in which there is good evidence for both nodulating and nonnodulating species, *Acacia* and *Pentaclethra* (Mimosoideae), and *Pterocarpus* (Papilionoideae)

	Number of genera			
Subfamily	*Total*	*Nodulated*	*Nonnodulated*	*Unknown*
Caesalpinioideae	157	8	72	76
Mimosoideae	78	38	9	31
Papilionoideae	479	353	27	99

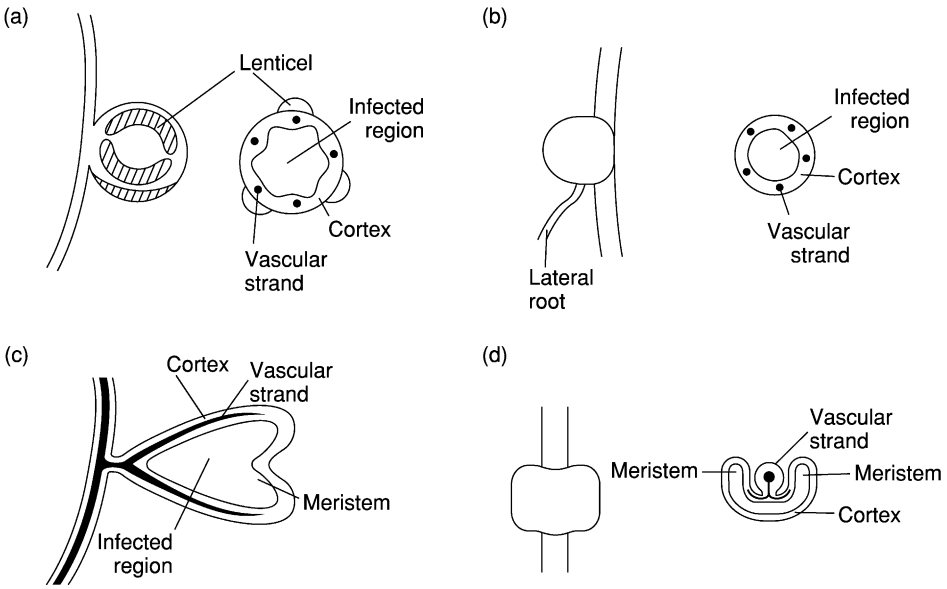


Figure 3 Legume nodules, drawn from field material, and showing the major types of nodule formed. (a, b, d) Left, whole nodules; right, transverse sections through these nodules; (c) a longitudinal section. (a) Determinate nodule of the desmodioid type, as found in soybean and related genera. Size varies from 2–7 mm in diameter. Lenticels on the surface facilitate gaseous exchange; (b) determinate nodule of the aeschynomenoid type, always associated with lateral or adventitious roots. They are found in many genera, including peanut and important tree genera such as *Dalbergia* and *Pterocarpus*, and on stems of some legumes, such as certain species of *Aeschynomene*. Size is usually between 2 and 6 mm in diameter. The infected tissue lacks the uninfected cells present in desmodioid nodules; (c) indeterminate nodule, drawn from a species of *Acacia*. This is the commonest type of nodule and occurs in all three subfamilies. It is characterized by an apical meristem, and the infected tissue has a gradation from young, inactive cells through active, nitrogen-fixing cells to (in older nodules) an inactive senescent region. They vary greatly in ratio of length to breadth and in the degree of branching (if any). Size varies from 2–3 mm in length to complex coralloid structures which may be more than 5 cm in diameter; (d) lupinoid nodules. These are not common, but are typical of many species of *Lupinus*. They have a lateral meristem which gradually encircles the root, forming a collar. As with other indeterminate nodules, there is a gradient of cells in the infected region. © J Sprent.

Table 5 Some environmental constraints to nitrogen fixation by legumes used in agriculture and examples of wild legumes which appear adapted to such conditions

Constraint	*Wild legumes showing tolerance*
Low temperature	*Astragalus alpinus*, *Hedysarum arcticum*, *Oxytropis sordida*
Shade	Some *Inga* spp., *Ormosia countinho*, *Poecilanthe hostmannii*
Flooding	*Discolobium pulchellum* (requires flooding for nodulation), some *Astragalus* spp., some *Neptunia* spp.
Salinity	*Halimodendron halodendron*, *Prosopis tamarugo*, *Vigna marina*
Drought	Some *Acacia* spp., *Faidherbia albida*, some *Vigna* spp.
Low available P	Some *Acacia* spp., *Crotalaria grahamiana*, *Cajanus cajan*
Low soil pH	Some *Astragalus* spp., some *Lupinus* spp.
High soil Al	*Dioclea guianensis*, some *Stylosanthes* spp.

can nodulate, there are exceptions, such as the marama bean, *Tylosema esculentum*. If 'old' new crops are to be sustainably exploited, it is essential to know: First, can it nodulate; and, second, will it nodulate under the conditions under test?

The soil contains a large number of bacterial genera known to nodulate legumes (Table 3). What is not clear from Table 3 is that one strain of rhizobia may nodulate one or many hosts, and that some hosts are nodulated by one strain, others by many. These variations raise a number of theoretical and practical questions. Whereas early studies placed great emphasis on host–rhizobial specificity, assuming that the two partners had coevolved, this is now much less clear. Most of the early work on legume nitrogen fixation was carried out on temperate crop plants and in these the 'rules' usually apply. The main exception is the common bean *Phaseolus vulgaris*, now known to nodulate with several different rhizobial genera (Table 3). Since about 1989, there has been a major thrust toward examining rhizobia from tropical areas. This is when the rules start to break down. New phenomena such as stem-nodulating legumes, photosynthetic rhizobia, new combinations of growth rate and acid/alkali production, and many more have been discovered. Many examples of broad host-range rhizobia have been discovered, adding to the classic one, originally referred to as the 'cowpea miscellany.' At that stage, broad host

range and slow growth were thought to be associated, but there is now a fast-growing strain, NGR 234, isolated from Papua New Guinea, that can nodulate at least 70 genera, including some from each of the three legume subfamilies. This and other findings have challenged the idea of coevolution. We now know that clusters of bacterial genes, called 'symbiotic islands,' are capable of lateral transfer in the soil between genera, and this may account for the growing number of new rhizobial genera which are being discovered. It may also help to explain the close relationship between some rhizobia and plant and animal pathogens.

If it is possible to make any generalizations about nodulated legumes, it may be that, when radiating into higher latitudes, legumes tend to be herbaceous, always to nodulate, and to have a recognizable level of host–rhizobial specificity. Nodules are also usually indeterminate, a benefit in areas with a markedly seasonal climate. In low latitudes, there are many nonnodulating legumes and, of the nodulating ones, many are nodulated by a wide range of rhizobia, producing nodules that vary greatly in their ability to fix nitrogen (effectivity). From the bacterial perspective, infectivity may be more important than effectivity.

The Importance of Nitrogen-Fixing Symbioses to Soil Fertility

One of the problems with trying to estimate biological nitrogen fixation is the lack of suitable methods for field use, especially where trees are involved. As a result, published estimates for global nitrogen fixation vary widely. The true value may lie in the range 100–200 Tg N per year. In recent years, the availability of cheap fertilizer N has reduced both the amount of biologically fixed nitrogen in agriculture and also the amount of field research on the subject in the developed world. However, the rise in the organic movement and fears about nitrate pollution have provided a stimulus for renewed study. Developing countries have no doubt about the significance of nitrogen fixation for the well-being of their peoples and their economies. It is now very clear that optimization of nitrogen fixation in these countries requires much more research, as many of the procedures used in developed countries are inappropriate. To take a simple example, the diversity of rhizobia found in tropical soils poses major problems for the production of effective, competitive inoculant strains.

What is the fate of nitrogen fixed by symbiotic systems? Initially it is to supply the partners, but in the real world plants do not live on their own. We have little idea of how much, if any, of nitrogen fixed by plants that are also mycorrhizal (which most, including bryophytes, are) is exchanged for other nutrients (phosphorus, fixed carbon, trace elements) with nonfixing plants via hyphal connections. If this occurs, then selective harvesting of, for example, large nodulated trees could have a major effect on other plants in the ecosystem. Harvesting crops can also lead to loss of soil N if the amount removed in the crop exceeds that fixed. Even the benefits of crop rotation have been reassessed. A concomitant of nitrogen reduction by nitrogenase is the simultaneous production of hydrogen.) Although the energy so wasted may be partly recovered in some systems by an uptake of hydrogenase, there is little evidence that plants lacking such a system are at a disadvantage. A recent suggestion is that hydrogen evolution in the soil alters the microbial flora in such a way that plants may benefit from improved nutrition. This could help to explain some of the benefits of crop rotation which cannot be accounted for by improved soil fertility and/or reduction of pest and pathogen load.

See also: **Nitrogen in Soils:** Plant Uptake

Further Reading

Allen ON and Allen EK (1981) *The Leguminosae: A Source Book of Characteristics, Uses and Nodulation*. Madison: University of Wisconsin Press/and London, UK: Macmillan.

Benson DR and Silvester WB (1993) Biology of *Frankia* strains, actinomycete symbionts of actinorhizal plants. *Microbiological Reviews* 57: 293–319.

Berg RH, Langwenstein B, and Silvester WB (1999) Development in the *Datisca–Coriaria* nodule type. *Canadian Journal of Botany* 77: 1334–1350.

Bergman B, Johansson C, and Söderbäck E (1992) The *Nostoc–Gunnera* symbiosis. *New Phytologist* 122: 379–400.

Cleveland CC, Townsend AR, Schimel DS *et al.* (1999) Global patterns of terrestrial biological nitrogen (N_2) fixation in natural ecosystems. *Global Biogeochemical Cycles* 13: 623–645.

Giller KE (2001) *Nitrogen Fixation in Tropical Cropping Systems*, 2nd edn. Wallingford, UK: CABI.

Huss-Danell K (1997) Actinorhizal symbioses and their N_2 fixation. *New Phytologist* 136: 375–405.

Jones DL (1993) *Cycads of the World*. Washington, DC: Smithsonian Institution Press.

Rai AN, Söderbäck E, and Bergman B (2000) Cyanobacterium-plant symbioses. *New Phytologist* 147: 449–481.

Soltis DE, Soltis PS, Chase ME *et al.* (2000) Angiosperm phylogeny inferred from 18S rDNA, *rbc*L, and

*atp*B sequences. *Botanical Journal of the Linnean Society* 133: 381–461.
Spaink HP, Kondorosi A, and Hooykaas PJJ (eds) (1998) *The Rhizobiaceae – Molecular Biology of the Model Plant-Associated Bacteria*. Dordrecht, the Netherlands: Kluwer Academic.
Sprent JI (2001) *Nodulation in Legumes*. Kew, UK: Royal Botanic Gardens.
Sprent JI (2002) Knobs, knots and nodules. *New Phytologist* 153: 2–7.
Sprent JI and Sprent P (1990) *Nitrogen-Fixing Organisms: Pure and Applied Aspects*. London, UK: Chapman and Hall.
Wagner GM (1997) *Azolla*: a review of its biology and utilization. *Botanical Review* 63: 1–26.
Wall LG (2000) The actinorhizal symbiosis. *Journal of Plant Growth Regulation* 19: 167–182.

Nitrogen Fertilizers *See* **Fertilizers and Fertilization**

NUCLEAR WASTE DISPOSAL

G W Gee, P D Meyer, and A L Ward, Pacific Northwest National Laboratory, Richland, WA, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Nuclear wastes are by-products of nuclear weapons production and nuclear power generation, plus residuals of radioactive materials used by industry, medicine, agriculture, and academia. Their distinctive nature and potential hazard make nuclear wastes not only the most dangerous waste ever created by humanity, but also one of the most controversial and regulated with respect to disposal. Nuclear waste issues, related to uncertainties in geologic disposal and long-term protection, combined with potential misuse by terrorist groups, have created uneasiness and fear in the general public and remain stumbling blocks for further development of a nuclear industry in a world that may soon be facing a global energy crisis.

Categories

High-Level Waste

The category 'high-level waste' (HLW) consists of either used (spent) nuclear fuel or the waste generated by reprocessing spent nuclear fuel. HLW generates much heat and radiation and requires remote handling and significant shielding, using materials such as lead, concrete, soil, or water, to prevent excessive radiation exposure to workers. Spent-fuel waste consists of fuel elements removed primarily from commercial reactors, after 3 or 4 years, when their useful life is over and they no longer generate electricity efficiently. The spent fuel assemblies from nuclear power plants are stored on-site, where they are typically submerged in large, water-filled pools, awaiting final disposal. In contrast, reprocessing wastes are stored mainly at nuclear weapons production facilities in liquid form in large underground storage tanks. Figure 1 shows construction of large (1 million gallons, approx. 3.6 million liters) storage tanks buried at the Hanford Site in Washington state, USA. In some cases, the liquid wastes have been calcined into a dry granular form. In the USA, all HLW liquids will be solidified (vitrified) and shipped in secure canisters to a deep geologic repository. Similar plans exist for other countries.

Figure 1 Underground million-gallon (3.6 million liter) storage tanks for nuclear waste at the Hanford site in Washington State, USA.

Transuranic Waste

Transuranic waste (TRU) is derived from the reprocessing of spent fuel and the use of plutonium in fabricating nuclear weapons. TRU is waste contaminated with alpha-emitting radionuclides of atomic number greater than 92 (i.e., greater than uranium, hence the term 'transuranic') and half-lives greater than 20 years in concentrations greater than 3700 $\mathrm{Bq\,g^{-1}}$. TRU typically generates less heat and emits less radiation than HLW, but it remains toxic for centuries and thus requires the same long-term isolation as HLW. TRU in the USA is stored at the Waste Isolation Pilot Plant near Carlsbad, New Mexico.

Low-Level Waste

The category 'low-level waste' (LLW) includes all radioactive waste other than HLW and TRU. Uranium mill tailings, a special form of LLW, are discussed separately. Most LLW contains relatively short-lived nuclides and has less radioactivity than found in HLW. LLW comes from a wide range of institutions, including nuclear power plants, government laboratories, test reactors, hospitals, laboratories, and industrial plants. LLW takes a wide variety of forms, including materials from medical research and treatments; contaminated rags and protective clothing; paper towels, filters and filter residues; and hand tools and equipment. Radiation from LLW can be high enough to require shielding for handling and shipping. Figure 2 show disposal of LLW in concrete (shielded) containers at the Hanford LLW disposal site, USA.

LLW is classified into four groups according to degree of hazard. LLWs that can be disposed of by near-surface burial are classed as A, B, and C types, ranging from the lowest hazard (A) to the highest (C). The US Nuclear Regulatory Commission (NRC) provides requirements for packaging and the form of burial for each of these classes. LLW too hazardous to bury in near-surface facilities is termed 'Greater-Than-Class-C' waste and requires disposal in a deep geologic repository. LLW also includes radioactive materials from natural sources, including radium-226, radium-228, radon-222, plus other radioactive elements from terrestrial sources. Discrete and diffuse sources are part of the LLW inventory. Discrete waste sources include spent industrial gauges, radium needles (used in medical equipment), and exchange resins (e.g., those that separate radium and other radionuclides from contaminated waters). Diffuse LLW can include such things as coal ash, phosphate waste from fertilizer production, and contaminated water and drilling equipment from oil and gas production. Some LLW has been reused for construction backfill and road-building activities and thus spread over long distances. While these radiation levels are relatively low, they produce conditions that can be a detriment to human health. As an example, uranium tailings high in radium can produce significant levels of elevated radon-222 gas. People in homes with basements surrounded by tailings or other radium sources can be exposed to radon gas levels that exceed permissible limits (e.g., 0.74 $\mathrm{Bq\,m^{-2}\,s^{-1}}$).

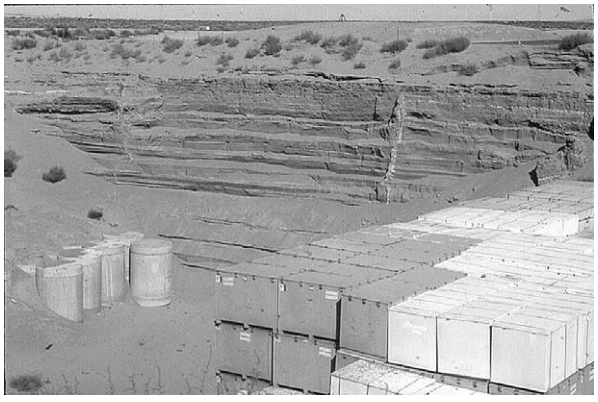


Figure 2 Low-level waste storage at the Hanford low-level waste facility in Washington State, USA.

Uranium Mill Tailings

This special category of LLW consists of residues in the form of fine sands and silts generated from the mining and milling of uranium ore. Large volumes of uranium tailings exist (more than 118 million cubic meters in the USA alone) containing low concentrations of naturally occurring radioactive materials, including thorium-232, radium-226, and its decay product, radon-222. Radon gas is the largest single radiation hazard from uranium mill tailings. Soil covers placed over uranium tailings piles have generally reduced the probability of radon escaping from them. Other environmental concerns include elevated uranium, sulfate, nitrate, and ammonia concentrations in groundwater beneath tailings piles.

Mixed Waste

Waste that contains both hazardous chemicals and radioactive wastes (generally those in the LLW category) is termed 'mixed waste.' In the USA, mixed waste is subject to both NRC and Environmental Protection Agency (EPA) regulations. Sites where chemical contaminants such as dense, nonaqueous-phase liquids (DNAPLs, e.g., carbon tetrachloride), or heavy metals (e.g., chromium, lead, mercury) are mixed with radionuclides are more difficult to clean up than sites where radionuclides are the

sole contaminant, because of the complex nature of the mixed waste coupled with the heterogeneous nature of the subsurface soils and sediments in which the wastes reside.

Waste Inventories and Environmental Discharges

Since the end of World War II, large amounts of nuclear wastes have accumulated in temporary waste repositories around the world. Table 1 shows the volumes of HLW, TRU, and LLW that have accumulated in the past and are projected to accumulate in the USA by the year 2030. Uranium mill tailings are also shown for comparison. In terms of shear volume, uranium mill tailings contribute by far the largest volume of waste (over 157 million cubic meters), with LLW contributing approximately one-fifth of the total volume (or nearly 38 million cubic meters of LLW). Nearly double the USA volumes have been generated in the former Soviet Union (FSU), with lesser volumes in other countries. Decay of radionuclides reduces the massive amounts of nuclear waste but does not eliminate the hazard, because long-lived radionuclides dominate in later years. As shown, the hazard of radioactive waste remains in perpetuity, so treatment or indefinite containment strategies are needed.

In addition to stored wastes, considerable volumes of wastes have been previously discharged directly to the environment. These environmental insults range from approximately 1.4×10^{19} Bq (390 million curies [where 1 Ci $= 3.7 \times 10^{10}$ radioactive disintegrations per second]) in the USA (primarily from liquid discharges during weapons production) to more than 6.3×10^{19} Bq (1700 million curies) in the FSU. At nuclear facilities, the discharges have been directly to the soils and adjacent waterways, burdening them with radioactive decay products whose half-lives range from a few years to millions of years. Long-term environmental consequences from these waste discharges focus primarily on inventories of carbon-14 (half-life of nearly 6000 years), technetium-99 (Tc-99) (half-life of 213 000 years) and iodine-129 (I-129) (half-life of nearly 16 million years). The chemical forms in which Tc-99 and I-129 are present are not strongly sorbed in most sediments so they tend to move to groundwater. Concentrations in excess of 7400 Bq l^{-1} of Tc-99 (more than 100 times permissible limits) have been found in the deep groundwater under waste tanks at the US Department of Energy (DOE) Hanford Site in Richland, WA, USA.

Table 1 Current and projected volumes of various types of nuclear waste in the USA

	Volume of waste ($\times 10^3$ m^3)		
Types of waste	*Year 2000*	*Year 2010*	*Year 2030*
High-level waste			
Commercial-liquid wastes	0.24	0.24	0.24
Commercial-spent fuel	17	24	35
Defense (total)	395	335	346
Transuranic			
Defense/environmental cleanup	275	297	191
Low-level waste			
Commercial	1722	2055	2508
Defense/decommissioning/ cleanup	4707	22 777	36 524
Uranium mill tailings			
Commercial and defense wastes	152 400	155 400	157 400

Radiation Exposures

Figure 3 shows the major pathways through which radioactivity from nuclear sources can affect living organisms. While the impact on the environment has been significant, immediate health effects to the general population have been small. Exposure to materials from the nuclear fuel cycle (including nuclear waste) is less than 0.2% of the total annual radiation dose to the general population. In contrast, dental and medical diagnosis and therapy account for more than 90% of all doses from nonnatural sources. Some occupations and geographic settings expose an individual to a greater-than-average radiation dose. A person living in Denver, CO, USA (1.5 km above sea level) is exposed to nearly twice as much cosmic radiation as a person living at sea level in San Diego, CA, USA. Individuals living in Spokane, WA, USA are exposed to higher concentrations of radon from the soil than those living in Midland, TX, USA. The question for many is that, while radiation doses to the general public are currently small, as nuclear waste discharges continue and dispersal to the environment proceeds, what future risks to the human population can be expected? What are the long-term risks? What can be done to ensure the risks are minimal and that, in the immediate and foreseeable future, the human race is protected against undue and unnecessary health risk from nuclear waste?

Waste Storage and Disposition

Most HLW, including spent fuel rods, and reprocessing wastes are in temporary storage (Table 2). No permanent repository for HLW has been licenced anywhere in the world to date (2004). Spent fuel rods are largely contained at sites near nuclear reactors. These

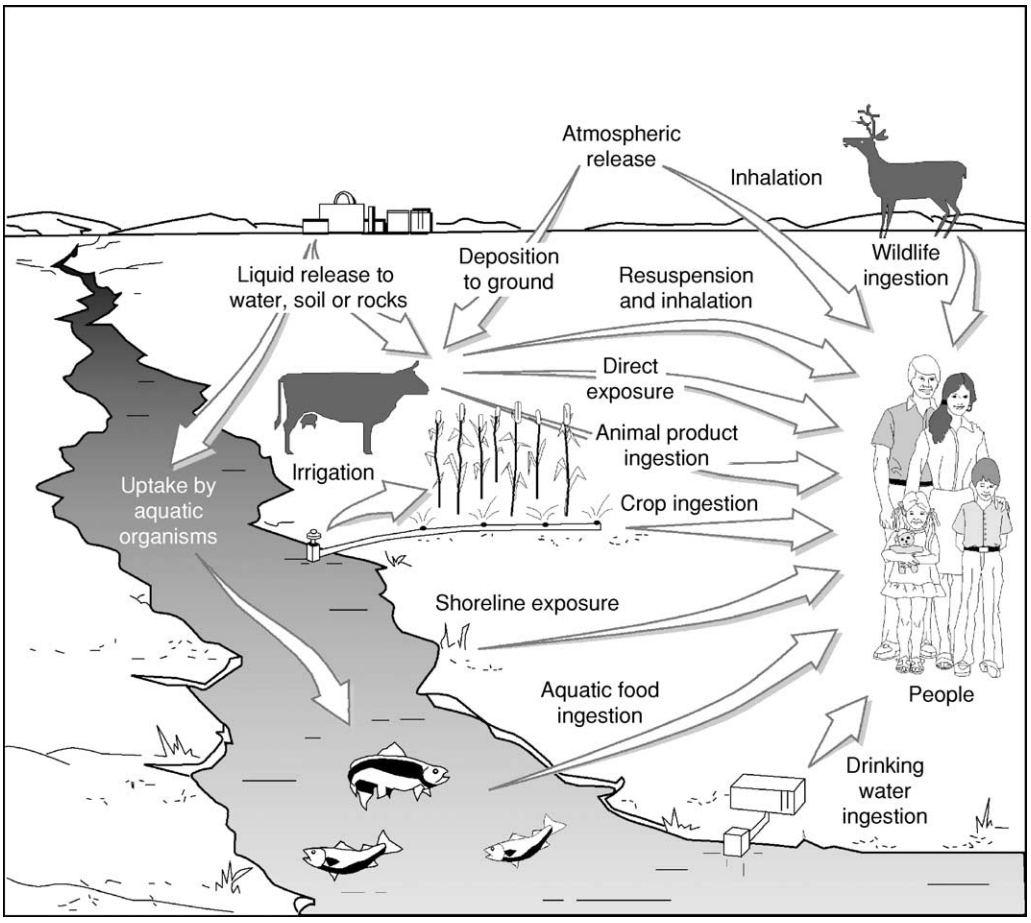


Figure 3 Potential radionuclide pathways from nuclear wastes.

Table 2 High-level waste burial programs in various countries

Country	*Earliest planned year*	*State of program*
Belgium	2020	Underground disposal in clay at Mol, Belgium
Canada	2025	Bury irradiated fuel in granite at yet-to-be identified site
China	Undetermined	Fuel to be reprocessed. Burial planned for Gobi Desert
Finland	2020	Field studies conducted – final site selection not made
France	2010	Two sites selected and studied with final selection in 2006
Germany	2008	Gorleben salt dome site only site studied
India	2010	Fuel to be reprocessed – granite repository to be selected
Italy	2040	Fuel to be reprocessed – storage for 50 years before burial in clay or granite
Japan	2020	Cooperative study at a Chinese research facility
Netherlands	2040	Interim storage for 50 – 100 years before burial sub-seabed or exterior
Russia	Undetermined	Uncertain
Spain	2020	Burial in clay, granite, or salt formation
Sweden	2020	Aspo granite site selected near Oskarshamn nuclear facility
Switzerland	2020	Burial in granite or sedimentary formation to be selected
United States of America	2010	Yucca Mountain – deep vadose zone, federally recommended site
United Kingdom (Great Britain)	2030	Interim storage approved through 2022. Options include sub-seabed burial

fuel rods pose a threat to humanity because they can be reprocessed for plutonium, a key component of nuclear weapons, so permanent disposal is highly desirable. Many billions of dollars have been spent by the US government on characterizing Yucca Mountain as a potential repository for more than $70\,000\,m^{-3}$ of nuclear wastes, including spent fuel rods. The US Department of Energy is currently

proceeding with a license application to store nuclear wastes at the Yucca Mountain site. Should a license be approved, no nuclear waste will be stored there until well after 2010. Issues of transportation, monitoring, and final waste configurations remain to be resolved. In the interim, most of the spent fuel remains in relatively vulnerable locations, often near large urban centers where nuclear terrorism may be difficult to prevent and where consequences of such acts of terrorism could be deadly and have long-range consequences. These issues suggest that action be taken soon to finalize waste repositories at locations that effectively control access to waste, particularly HLW that can be converted into weapons.

After reprocessing wastes are removed from underground storage tanks, some residual contaminants will remain. Several options are being considered for containing the waste and leaving the tanks, contaminated soil, piping, and support structures in place. The following options are being considered:

- Stabilization: tanks would be stabilized to prevent dome collapse by filling with inert material such as bentonite, sand, gravel, cement, or a combination of these materials;
- Immobilization: subsurface injections of either reactive chemicals, to immobilize contaminants and prevent them from moving, or physical barriers (low-permeability grouts, including bentonite and mineral wax), to retard the migration of the contaminants to the water table;
- Decontamination: chemical or mechanical decontamination by using high-pressure water or frozen carbon dioxide blasting, or in-place solvent washes and extractions;
- Flushing: a combination of a subsurface barrier and a soil-flushing treatment where chemicals such as carbonate solutions would extract contaminants. Subsurface barriers made of grout or polymers would minimize the chance of flushing the contaminants and mixing with groundwater;
- *In-situ* vitrification: this process uses heated electrodes to melt the waste in place. The waste and surrounding soil are vitrified (turned into glass) at temperatures exceeding 1500°C and when cooled resemble natural glass obsidian. Volatile organics would be destroyed and heavy-metal contaminants and radionuclides would be physically bound in the glass. This process has been successfully demonstrated for LLW and mixed wastes at a number of locations throughout the world;
- Surface barriers: after the empty tanks are treated in place, aboveground barriers could be placed over the tanks to protect the area against intrusion, wind and water erosion, and net water infiltration. If successful, such efforts would ensure that any residual contaminants not fully treated would tend to remain in the subsurface for thousands of years. Long-term surface barriers (intended to have a lifetime of at least 1000 years) consisting of specially designed soil layers have been tested and show promise for deployment at the Hanford Site in the USA. Figure 4 shows a schematic picture of a backfilled storage tank covered by a 1000-year soil barrier.

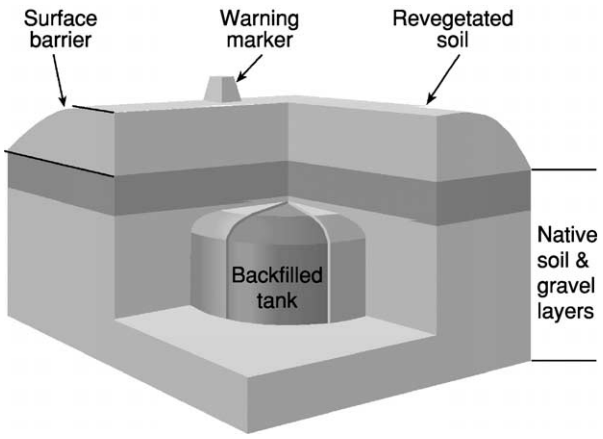


Figure 4 Schematic of a surface barrier consisting of natural earthen materials placed over an evacuated and stabilized high-level waste tank. The surface barrier is designed to last at least 1000 years and protect against plant and animal intrusion, wind and water erosion, and water infiltration.

LLW disposal actions remain site-specific. Under present regulatory guidance, there will be minimal waste treatment, with site closures depending on cover systems that can provide control of erosion, intrusion, and minimization of water invasion into the waste. Longevity for such cover systems remains largely unproven, since most conventional landfill designs are built to last decades not millennia. Efforts to develop innovative and scientifically sound, long-term performance tests are still in their infancy at the beginning of the twenty-first century.

Environmental Uncertainties

One of the characteristics of nuclear wastes is their potential to cause adverse health and environmental effects for many thousands of years. For nuclear waste disposal facilities and sites contaminated with radioactive materials, there is consequently a need to assess safety far into the future. In the USA, regulatory limits for the disposal of HLW and TRU are required to be met over a 10 000-year time period; a 1000-year period is used in analyses for decommissioning nuclear materials sites. Decisions that are intended to apply that far into the future rely, in large part, on models that predict radionuclide

transport in the environment. Because such predictions are inherently uncertain, a variety of strategies have been adopted to provide confidence in decisions about nuclear waste disposal:

- Reliance on natural systems to limit transport (e.g., stable rock formations, arid sites);
- Redundant design of engineered components (to prevent catastrophic behavior if one component fails);
- Site-specific characterization of processes potentially affecting transport (to provide understanding of site behavior and to develop and parameterize models);
- Natural analog studies (e.g., of uranium transport);
- Accelerated degradation studies (e.g., of glass waste forms);
- Peer review;
- Monitoring of site behavior and comparison with model predictions;
- Sensitivity and uncertainty analyses of predictive models.

The extent of analyses conducted to support decisions on nuclear waste disposal varies significantly. In the USA, the disposal of HLW at Yucca Mountain, a remote site in southern Nevada, USA, and the disposal of TRU at the Waste Isolation Pilot Plant have been supported by large budgets, detailed site-characterization studies, complex modeling analyses, and extensive site monitoring. In contrast, decommissioning of nuclear materials sites may rely on very limited, site-specific data and relatively simplified modeling analyses. Regardless of the level of effort, however, predictive modeling of radionuclide transport in soils and sediments plays an important role. Assessing the uncertainty inherent in model predictions supports decision-making by increasing the credibility of the modeling analyses, improving the predictability of consequences, avoiding worst-case assumptions, and focusing data collection on variables most valuable in improving predictions and reducing uncertainty.

The sources of uncertainty that potentially have an impact on predictions of radionuclide transport in the environment are many. In a general sense they arise from:

- Measurement or sampling errors in characterizing the system;
- Spatial and temporal variability in the system's properties;
- Disparity among the scale of the system's features, events, and processes and the scales at which sampling and simulation occur;

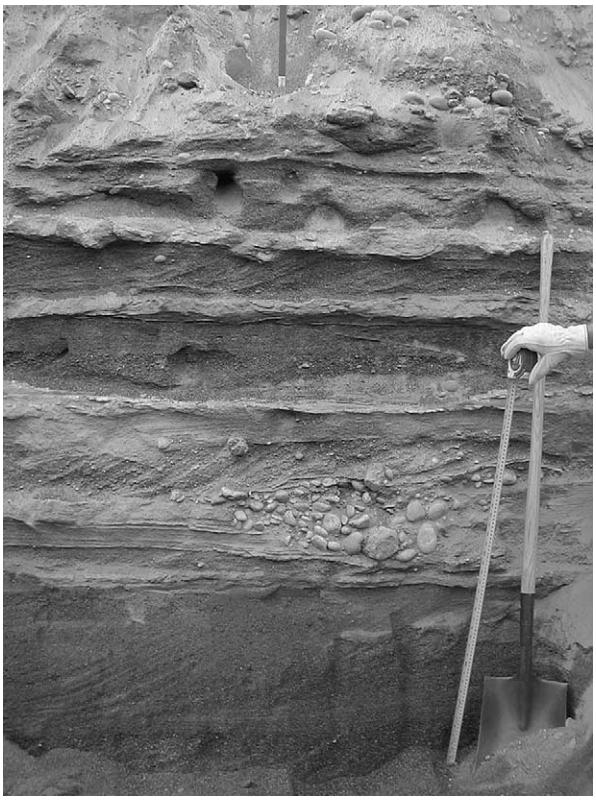


Figure 5 Exposed trench face of layered, heterogeneous sediments in which nuclear wastes are stored at the Hanford Site, Washington state, USA. (Photo courtesy of John Selker, Oregon State University.)

- Randomness in the system's stresses, particularly transient external stresses often occurring over a short period of time (e.g., floods, earthquakes);
- Incomplete knowledge of the system being analyzed (i.e., ignorance of significant features, events, or processes).

The issues of variability and scale are demonstrated in Figure 5, which shows the spatial variability of sediments on a trench face at a location on the Hanford Site (Washington, USA). Fine, silty layers only a few centimeters thick lie between coarse, sand layers with varying amounts of gravel. Parameters describing flow and transport vary over a wide range at this site (by several orders of magnitude in the case of hydraulic conductivity). A small-scale measurement with a sample size of 30 cm or so (such as a neutron-probe measurement of water content) would probably sample over distinctly different materials. In a model, it would not be unlikely that the area of the entire photograph would be represented by a single parameter value. Uncertainty is thus introduced both in the interpretation of measured quantities and in the assignment of parameter values.

Transport of radioactive contaminants through the vadose zone (i.e., the zone between the ground surface and the water table) depends on the magnitude and variability of geochemical and hydrologic parameters. These parameters include particle-size distribution, specific surface area, cation-exchange capacity, chemical retardation factors, water retention, and unsaturated hydraulic conductivity, all of which control fate and transport of chemicals and radionuclides in the vadose zone, and all have significant levels of uncertainty associated with them at a given waste site. For example, understanding how leaks from HLW tanks move through soils presents a formidable challenge. Wastes are often hot, highly saline, and chemically reactive. Waste plume movement is both spatially and temporally affected by geochemical interactions as well as hydrologic controls. Concentrated water flow from water-line leaks and meteorologic sources (winter rains and snowmelt) can cause contaminants to move in preferential flow paths that may accelerate downward movement. On the other hand, horizontal layering of sediments can cause plumes to spread laterally, delaying their arrival at the groundwater table.

In modeling the performance of nuclear waste disposal sites, it is convenient to classify uncertainties into three basic types: scenario uncertainty; model uncertainty; and parameter uncertainty. Scenario uncertainty addresses the possibility that the future state of the system may be significantly different from what is assumed as the initial case. This may be the result of direct human activity (e.g., irrigation or groundwater extraction) or a more natural event (e.g., increased precipitation due to climate change).

Model uncertainty arises when a complex system is conceptualized for the purpose of modeling and when that concept is implemented as a mathematical model for the purpose of prediction. It is frequently the case that multiple conceptual models can be proposed for a single site to explain observed behavior and that site data are insufficient to determine which model best represents reality. An example is the case of flow and transport in the unsaturated zone at the Yucca Mountain site. Understanding of the relative importance of fractures versus the rock matrix in contributing to seepage and radionuclide transport, and the best way to represent the fracture and/or matrix flow and transport processes, has evolved over time in response to improved analyses and additional site data.

Parameter uncertainty addresses our inability to specify precisely the parameters of the models developed to describe flow and transport of radionuclides in the environment. Even if the models are accurate representations of the actual system processes, it is impossible to measure exhaustively the required model parameters due to their spatial andtemporal variability and apparent randomness. Worse yet, the quantities that are often measured are only indirectly related to model parameters and may be on a scale different (either larger or smaller) from the scale of the model parameters. Model calibration (minimizing the difference between observed system behavior and model-predicted behavior by systematically varying parameter values) is an important method to reduce and quantify parameter uncertainty.

Parameter uncertainty is most amenable to characterization using formal probabilistic methods. In soils, this generally means using measured data and/or expert judgment to specify probability distributions for flow and transport characteristics. Because measurement of these characteristics is frequently expensive and time-consuming, more easily measured soil characteristics are often used to infer distributions of the desired parameters. Examples include soil texture measurements, to infer unsaturated hydraulic properties, and soil mineralogy (along with other quantities), to infer chemical effects on radionuclide transport (i.e., adsorption). Model uncertainty is generally quantified by comparing flow and transport results from alternative models, potentially weighting these results by estimates of the relative probability of the models. Scenario uncertainty can be similarly quantified, although in this case it is possible that absolute probabilities of alternative scenarios can be derived (e.g., the probability of a flood occurring in a given time-period).

In relatively extensive studies of uncertainty (such as for Yucca Mountain and the Waste Isolation Pilot Plant), parameter, model, and scenario uncertainties have all been found to contribute significantly to the overall system uncertainty (e.g., uncertainty in the performance of the nuclear waste disposal facilities). In addition, results of the uncertainty analyses have been used to identify model components and parameters that have the greatest impact on system performance. This information has been used to target data collection to provide the greatest reduction in system uncertainty.

Summary

Nuclear waste disposal is a global concern. Discharges to the environment have been significant, with radioactive production and subsequent dispersal of billions of curies. To date these waste inventories have been largely confined to locations near nuclear-production facilities primarily in the USA and FSU. Dispersal from global fallout and from nuclear accidents have produced measurable radiation doses to the global population, but compared with natural

sources have contributed only a small fraction of total annual dose and, as of 2004, do not present a significant health risk to the general public. However, the public needs to be informed of hazards that remain in those areas where soils and waters have been exposed to nuclear waste discharges and remain highly contaminated.

Safe disposal of nuclear waste is a priority among most nations and will continue to be a major expense for those nations involved in past nuclear weapons production or with past and current nuclear-power generation. Treatment, immobilization, and containment are methods that are being proposed to handle nuclear wastes until they decay to harmless levels. Vitrification is an expensive but viable technology for immobilizing high-level radioactive waste, and deep, geologic storage is a scientifically sound method for ensuring long-term public protection. Low-level radioactive waste disposal in shallow landfills, covered by soil covers, is allowed (in 2004) and should be acceptable if longevity of these disposal facilities and their isolation capabilities can be demonstrated. This remains a scientific challenge.

See also: **Isotopes in Soil and Plant Investigations**; **Remediation of Polluted Soils**; **Waste Disposal on Land:** Liquid; Municipal

Further Reading

Bradley DJ (1997) *Behind the Nuclear Curtain: Radioactive Waste Management in the Former Soviet Union*. Columbus, OH: Battelle Press.

Crowley KD and Ahearne JF (2002) Managing the environmental legacy of US nuclear weapons production. *American Scientist* 90(6): 514–523.

Gee GW, Ward AL, and Wittreich CD (2002) *The Hanford Site 1000-Year Cap Design Test*. PNNL-14143. Richland, WA: Pacific Northwest National Laboratory.

Gephart RE (2003) *Hanford. A Conversation about Nuclear Waste and Cleanup*. Columbus, OH: Battelle Press.

Gephart RE and Lundgren RE (1998) *Hanford Tank Cleanup: A Guide to Understanding the Technical Issues*. Columbus, OH: Battelle Press.

Gerber MS (2002) *On the Home Front, the Cold War Legacy of the Hanford Nuclear Site*, 2nd edn. Lincoln, NE: University of Nebraska Press.

Joint Norwegian–Russian Expert Group (1997) *Sources Contributing to Radioactive Contamination of the Techa River and Areas Surrounding the "Mayak" Production Association, Urals, Russia*. Joint Norwegian–Russian Expert Group for Investigation of Radioactive Contamination in the Northern Areas. Osteras, Norway: Gjovik Trykkeri.

League of Women Voters Education Fund (1993) *The Nuclear Waste Primer*. New York: Lyons and Burford.

Stenhouse MJ and Kirko VI (2001) *Defense Nuclear Waste Disposal in Russia: International Perspective*. Dordrecht, the Netherlands: Kluwer Academic Press.

USDOE (2000) *Pioneering Nuclear Waste Disposal (The Waste Isolation Pilot Plant)*. DOE/CAO-00-3124. Carlsbad, New Mexico: US Department of Energy.

USDOE (2002) *Yucca Mountain Science and Engineering Report Review 1*. DOE/RW-0539-1. Las Vegas, NV: US Department of Energy.

Ward AL, Gee GW, and White MD (1997) *A Comprehensive Analysis of Contaminant Transport in the Vadose Zone beneath Tank SX-109*. PNNL-11463. Richland, WA: Pacific Northwest National Laboratory.

NUTRIENT AVAILABILITY

N K Fageria, EMBRAPA, Santo Antônio de Goiás, GO, Brazil
V C Baligar, USDA Agricultural Research Service, Beltsville, MD, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Agricultural production and productivity are directly linked with nutrient availability. For sustained high crop yields, the application of nutrients is required. The term 'nutrient availability' has been used and defined in many ways. Hence, any use of the term needs to be accompanied by an appropriate definition. The Soil Science Society of America has defined available nutrients: (1) the amounts of soil nutrients in chemical forms accessible to plant roots or compounds likely to be convertible to such forms during the growing season, and (2) the contents of legally designated available nutrients in fertilizers determined by specified laboratory procedures which in most states constitute the legal basis for guarantees.

Soil productivity and nutrient availability are interrelated. Soil productivity is the capacity of a soil to produce a certain yield of crops or other plants with a specified system of management. Management of nutrients is an important aspect of maintaining soil productivity. Quantity of nutrient present in the soil is defined in terms of soil fertility, and, generally, soil

analysis is used as a criterion to make fertilizer recommendations for field crops. Nutrient availability is evaluated by observations and tests which are used to predict the response of plants and management of nutrients. The presence of sufficient total quantities of essential nutrients in a soil does not guarantee the availability of these nutrients to growing plants, because other factors such as soil-moisture content, soil temperature, pH, soil physical conditions, presence of toxic elements and/or salts may be limiting. Hence, adequate levels of nutrients alone do not guarantee soil productivity. A productive soil is one that has optimal total environmental conditions for plant growth. Since soil is a continuum, it is a matrix in constant change. It is very difficult under practical conditions to have all crop production factors at an optimal level; hence, most of the crop production factors are usually at a suboptimal level.

In modern agriculture, maximizing and sustaining crop yields are the main objectives. One of the major problems constraining the development of an economically successful agriculture is nutrient deficiency for crop production. As much as 50% of the increase in crop yields worldwide during the twentieth century was due to the adoption of chemical fertilizers. Soil infertility (natural element deficiencies or unavailability) is probably the single most important factor limiting crop yields worldwide.

Essential Nutrients for Plant Growth and Their Functions

The supply and absorption of chemical elements needed for growth and metabolism may be defined as nutrition, and the chemical elements required by an organism are termed 'nutrients.' Sixteen nutrients are essential for plant growth and development. On the basis of quantity required by plants, nutrients are divided into macro and micro groups. The essential nutrients such as carbon (C), hydrogen (H), oxygen (O), nitrogen (N), phosphorus (P), potassium (K), calcium (Ca), magnesium (Mg), and sulfur (S) are known as macronutrients; whereas iron (Fe), manganese (Mn), zinc (Zn), copper (Cu), boron (B), molybdenum (Mo), and chlorine (Cl) are termed micronutrients. Further, the essentiality of silicon (Si), sodium (Na), vanadium (V), and cobalt (Co) has been considered but is not yet proven. Plants obtain C, H, and O from air and water, and the remaining nutrients are absorbed from soil solution as inorganic ions or oxides by growing plant roots. Nutrient absorption by plants is usually referred to as 'ion uptake' or 'ion absorption,' because it is the ionic form in which nutrients are absorbed. Cations and anions may be absorbed independently and may not be absorbed in equal quantities: however, electroneutrality must be maintained in the plant and in the growth medium; therefore, ionic relationship achieves major importance in plant nutrition. The majority of cations in plant tissues are in the inorganic form, predominantly K^+, Ca^{2+}, and Mg^{2+}, and the majority of anions are in the organic form. The organic ions are synthesized within the tissue, while inorganic ions are absorbed from the growth medium. Generally, the monovalent cations are absorbed rapidly, whereas the divalent cations, especially Ca^{2+}, are absorbed more slowly. Similarly, the monovalent anions are generally absorbed more rapidly than polyvalent anions. The so-called physiological acidity or alkalinity of a salt depends upon which ion of the salt, the cation or the anion, is absorbed most rapidly. Thus a salt such as K_2SO_4 would be physiologically acid, since the K would enter the roots more rapidly than the SO_4^{2-}. By the same token, $CaCl_2$ should be physiologically alkaline, since the Ca^{2+} enters slowly and the Cl^- rapidly.

Mineral nutrients can also be absorbed by higher plants when applied as foliar sprays in appropriate concentrations. However, in modern high-yielding cultivars, nutritional requirements (macronutrients) are rarely met with foliar applications. Furthermore, foliar application of macronutrients requires several sprays and can be washed off by rain. The plant needs sufficient leaf area for absorption, and leaf damage by high nutrient concentrations is a serious practical problem. Despite these drawbacks, under some situations foliar application is the most effective method to correct a nutritional disorder. For example, iron deficiency in calcareous soils can be corrected by foliar application of ferrous sulfate solution more efficiently than the soil application of iron sources.

The macro- and micronutrient classification is simply based on the amount required: macronutrients are needed in relatively larger quantities than micronutrients. Each of these essential nutrients performs a specific biophysical or biochemical function within plant cells. Hence all nutrients are equally important for production of higher crop yields. If any one of these nutrients is deficient in the growth medium, plant growth is reduced. The low requirements of plants for micronutrients can be explained by the participation of these elements in enzymatic reactions and as constituents of growth hormones rather than as components of major plant products such as structural and protoplasmic tissue. To obtain higher yields, essential nutrients should be present in adequate concentrations and in available form, as well as in appropriate balance. An imbalance results when one or more nutrients are present in either deficit or excess supply. Table 1 shows forms of absorption and functions of essential plant nutrients.

Table 1 Forms of absorption and functions of essential nutrients in plants

Nutrient	*Forms taken up by plants*	*Functions*
Carbon	CO_2	Basic molecular component of carbohydrates, proteins, lipids, and nucleic acids
Hydrogen	H_2O	Hydrogen plays a central role in plant metabolism. Important in ionic balance and as main reducing agent; plays a key role in energy relations of cells
Oxygen	H_2O, O_2	Oxygen is like carbon in that it occurs in virtually all organic compounds of living organisms
Nitrogen	NH_4^+, NO_3^-	Nitrogen is a component of many important organic compounds ranging from proteins to nucleic acids
Phosphorus	$H_2PO_4^-$, HPO_4^{2-}	Central role in plants in energy transfer and protein metabolism
Potassium	K^+	Helps in osmotic and ionic regulation. Functions as a cofactor or activator for many enzymes of carbohydrate and protein metabolism
Calcium	Ca^{2+}	Calcium is involved in cell division and plays a major role in the maintenance of membrane integrity
Magnesium	Mg^{2+}	Component of chlorophyll and a cofactor for many enzymatic reactions
Sulfur	SO_4^{2-}	Sulfur is like phosphorus in that it is involved in plant cell energetic reactions
Iron	Fe^{2+}, Fe^{3+}	An essential component of many heme and nonheme Fe enzymes and carriers, including the cytochromes (respiratory electron carriers) and the ferredoxins. The latter are involved in key metabolic functions such as N fixation, photosynthesis, and electron transfer
Zinc	Zn^{2+}	Essential component of several dehydrogenases, proteinases, and peptidases, including carbonic anhydrase, alcohol dehydrogenase, glutamic dehydrogenase, and malic dehydrogenase
Manganese	Mn^{2+}	Involved in the O_2-evolving system of photosynthesis; is a component of the enzymes arginase and phosphotransferase
Copper	Cu^{2+}	Constituent of a number of important oxidase enzymes, including cytochrome oxidase, ascorbic acid oxidase, and lactase, and important in photosynthesis, and protein and carbohydrate metabolism
Boron	$B(OH)_3^0$	Activates certain dehydrogenase enzymes, involved in carbohydrate metabolism, synthesis of cell wall components; is essential for cell division and development
Molybdenum	Mo	An essential component of nitrate reductase and N_2-fixation enzymes; is required for normal assimilation of N
Chlorine	Cl^-	Essential for photosynthesis and as an activator of enzymes, involved in splitting water, and functions in osmoregulation of plants growing on saline soils

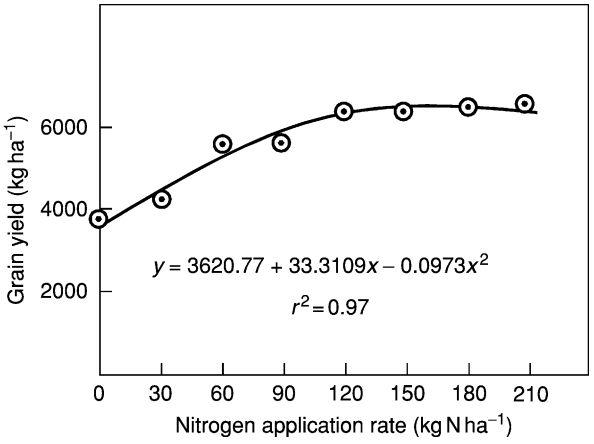


Figure 1 Response of lowland rice to nitrogen fertilization grown on an Inceptisol of central Brazil.

Nutritional disorders (deficiency and/or toxicity) limit crop yields in all types of soils around the world. For example, in Brazilian Oxisols and Inceptisols, deficiency of N, P, K, and Zn limits the yield of almost all field crops. Figure 1 shows the importance of N application in increasing yield of lowland rice grown on a Brazilian Inceptisol. It is clear from Figure 1 that yield of this important field crop increases with increasing levels of N until an adequate level in the soil is achieved. To obtain good yields of crops, nutrient disorders must be corrected. The first step in this direction is to identify the nutritional disorder.

Diagnostic Techniques for Nutrient Availability

Diagnostic techniques for nutrient availability refer to the methods for identifying nutrient deficiencies, toxicities, or imbalances in the soil–plant system. Nutritional deficiency can occur when there is insufficient nutrient in the medium or when it cannot be absorbed and utilized by plants as the result of unfavorable environmental conditions. Table 2 shows soil conditions associated with nutrient deficiencies. Nutritional deficiencies are very common in almost all field crops worldwide. The magnitude varies from crop to crop and region to region. Even some cultivars are more susceptible to nutritional deficiencies than others within a crop species. Four methods assess nutrient availability or sufficiency of mineral nutrients for plant growth: (1) visual symptoms, (2) soil testing, (3) plant analysis, and (4) crop growth response. These four approaches are becoming

Table 2 Soil conditions inducing nutrient deficiencies for crop plants

Nutrient	*Conditions inducing deficiency*
Nitrogen	Excess leaching with heavy rainfall; low organic matter content; burning the crop residue
Phosphorus	Acidic, organic, leached, and calcareous soils; high rate of liming
Potassium	Sandy, organic, leached, and eroded soils; high liming application, intensive cropping system
Calcium	Acidic, alkaline, or sodic soils
Magnesium	Acidic, alkaline, or sodic soils
Sulfur	Low organic matter content of soils; use of N and P fertilizers containing no sulfur; burning the crop residue
Iron	Calcareous soils; soils high in P, Mn, Cu, or Zn; high rate of liming
Zinc	Highly leached acidic soils; calcareous soils; high levels of Ca, Mg, and P
Manganese	High organic matter content; calcareous soils
Copper	Sandy soils; high liming rate in acid soils
Boron	Sandy soils; naturally acidic leached soils; alkaline soils with free lime
Molybdenum	High podzolized soils; well-drained calcareous soils

Table 3 Nutrient-deficiency symptoms in field crops

Nutrient	*Symptoms*
N	Chlorosis starts in old leaves; in cereals tillering is reduced; under field conditions, if deficiency is severe, whole crop appears yellowish and growth is stunted
P	Growth is stunted; purple–orange color of older leaves; new leaves dark green; in cereals tillering is drastically reduced
K	Older leaves may show spots or marginal burn starting from tips; increased susceptibility to diseases, drought, and cold injury
Ca	New leaves become white; growing points die and curl
Mg	Marginal or interveinal chlorosis with pinkish color of older leaves; sometimes leaf-rolling, like drought effect; plants susceptible to winter injury
S	Chlorosis of younger leaves; under severe deficiency whole plant becomes chlorotic and similar to appearance in N deficiency
Zn	Rusting in strip of older leaves with chlorosis in fully matured leaves; leaf size is reduced
Fe	Interveinal chlorosis of younger leaves; under severe deficiency whole leaf becomes first yellow and finally white
Mn	Similar to iron deficiency; at advanced stage necrosis develops instead of white color
Cu	Chlorosis of young leaves, rolling, and dieback
Mo	Mottled pale appearance in young leaves; bleaching and withering of leaves
B	Pale green tips of blades, bronze tints; death of growing points

widely used separately or collectively as nutrient-availability, -deficiency or -sufficiency diagnostic aids. They are extremely helpful, yet are not without limitations.

Visual Symptoms

When the supply of a particular nutrient is at an inadequate level in the soil or when plant roots are not able to absorb required amounts, owing to unfavorable conditions in the rhizosphere, plants show certain growth disorders. These disorders may be expressed in reduced height, reduced tillering in cereals, leaf discoloration, reduced root growth, and reduced growth of newly emerging parts of the plant. Table 3 summarizes the symptoms of nutritional deficiency in crop plants. Using visual symptoms is the cheapest method of diagnosis for nutritional disorders. However, this technique needs a lot of experience on the part of the observer, because deficiency symptoms can be confused with problems caused by drought, insect and disease infestation, herbicide damage, soil salinity, soil acidity and inadequate drainage. Sometimes, a plant may be on the borderline with respect to deficiency and adequacy of a given nutrient. In this situation there are no visual symptoms, but the plant is not producing at its capacity. This condition is frequently called 'hidden hunger.'

Deficiency symptoms normally occur over an area and not on an individual plant. If a symptom is found on a single plant, it may be due to disease or insect injury or a genetic variation. Also, the earlier symptoms are often more useful than late, mature symptoms. Some nutrients are relatively immobile in the plant, while others are more mobile. In general, deficiency symptoms caused by immobile nutrients first appear on the younger or upper leaves. The older leaves don't show any symptoms because immobile nutrients do not move or translocate from older to newer leaves. Immobile nutrients are Ca, Zn, B, Cu, Fe, Mn, and Mo.

In contrast, when there is a deficiency of a mobile nutrient, the symptoms first appear on the older leaves of the plant. This is because the mobile nutrients move out of the older leaves to the younger part of the plant. The mobile nutrients are N, P, K, and Mg. S may behave as a mobile or an immobile nutrient. However, in rice plants S deficiency first appears in younger leaves.

The use of visible symptoms has the advantage of direct field application without the need of costly equipment or laboratory support services, as is the case with soil and plant analysis. A disadvantage is that sometimes it is too late to correct deficiency of a given nutrient, because for some disorders considerable yield loss may have already occurred by the time visible symptoms appear.

Soil Testing

In a broad sense, soil testing is any chemical or physical measurement that is made on a soil. The main objective of soil testing is to measure soil nutrient status and lime requirements in order to make fertilizer and lime recommendations for profitable farming. Soil testing is an important tool in high-yield farming but produces best results only when used in conjunction with other good farming practices:

> There is good evidence that the competent use of soil tests can make a valuable contribution to the more intelligent management of the soil

This statement was made by the US National Soil Test Workgroup in its 1951 report and is still applicable today.

Soil testing involves collecting soil samples, preparation for analysis, chemical or physical analysis, interpretation of analysis results, and finally making fertilizer and lime recommendations for the crops.

The use of soil analysis as a fertilizer recommendation method is based on the existence of a functional relationship between the amount of nutrient extracted from the soil by chemical methods and crop yield. When a soil analysis test shows low levels of a particular nutrient in a given soil, application of that nutrient is expected to increase crop yield. Generally, nutrient analysis is arbitrarily classified as very low, low, adequate, high, and excessive. At very low nutrient levels, relative crop yield is expected to be less than 70% and a larger application of fertilizer for soil-building purposes is required. After the application of the nutrient, growth response is expected to be dramatic and profitable. At the low fertility level, relative yield is expected to be 70–90%. Under these conditions, annual application of fertilizer is necessary to produce maximum response and increase soil fertility. The increased yield justifies the cost of fertilization. When soil analysis testing shows adequate levels, relative crop yield under this situation is expected to be 90–100%. Normal annual applications to produce maximum yields are recommended. In this case more fertilizer may increase yields slightly, but the added yield would not pay back the expense of the additional fertilizer. At high levels of nutrient, there is no increase in yield, but small applications of fertilizer maintain soil level. The recommended amount may be doubled and applied in alternative years. When soil testing shows very high or excessive nutrient levels, yield may be reduced due to toxicity or imbalances of nutrients. In this situation there is no need to apply nutrient until the level drops back into the low range. To get such a nutrient level and yield relationship, it is necessary to conduct fertilizer yield trials in several locations in a given agroecologic region for different crops. Some specific recommendations for soil analysis are summarized here:

1. Soil samples must be representative of the land area in question. The recommendation is to take a minimum of one composite sample per 12–15 ha for lime and fertilizer evaluations. A representative soil sample is composed of 15–20 subsamples from a uniform field with no major variation in slope, drainage, or past fertilizer history. Any of these listed factors, if changed, will have an effect on the number of samples and unit area from which the sample is obtained;
2. Depth of sampling for mobile nutrients such as N should be 60 cm and, for immobile nutrients such as P, K, Ca, and Mg, 15–20 cm. For pasture crops, a sampling depth of 10 cm is normally sufficient to evaluate nutrient status and make liming and fertilizer recommendations;
3. The appropriate extractor must be selected. Three extracting solutions, $0.05\,\mathrm{mol\,l^{-1}}$ HCl + $0.025\,\mathrm{mol\,l^{-1}}$ H_2SO_4 (Mehlich 1), $0.03\,\mathrm{mol\,l^{-1}}$ $NH_4F + 0.025\,\mathrm{mol\,l^{-1}}$ HCl (Bray-P1), and $0.5\,\mathrm{mol\,l^{-1}}$ $NaHCO_3$ at pH 8.5 (Olsen) are the most commonly used extractants for P at the present time and are generally adequate to cover a broad range of soils. Commonly used extractants for K, Ca, and Mg are double acid (Mehlich 1), $1\,\mathrm{mol\,l^{-1}}$ NH_4Ac at pH 7, and NaOAc at pH 4.8. Multielement extracting reagents are replacing the more familiar, single-element extractants. After mixing with an appropriate aliquot of soil, the obtained extract is assayed by an inductively coupled plasma emission spectrometer (ICP-AES). A flow-injection analyzer (FIA) is another multielement analyzer capable of assaying these soil extracts;
4. Optimum soil test values for macro- and micronutrients, which vary from soil to soil, crop to crop, and extraction to extraction, are required. They are usually approximately greater than $10\,\mathrm{mg\,P\,kg^{-1}}$, greater than $50\,\mathrm{mg\,K\,kg^{-1}}$, greater than $600\,\mathrm{mg\,Ca\,kg^{-1}}$, greater than $120\,\mathrm{mg\,Mg\,kg^{-1}}$, and greater than $12\,\mathrm{mg\,S\,kg^{-1}}$ for producing satisfactory results for most soils and crops. For micronutrients, the critical values reported are: Fe, 2.5–5 $\mathrm{mg\,kg^{-1}}$; Mn, 4–8 $\mathrm{mg\,kg^{-1}}$; Zn, 0.8–3 $\mathrm{mg\,kg^{-1}}$; B, 0.1–2 $\mathrm{mg\,kg^{-1}}$, Cu, 0.5–2 $\mathrm{mg\,kg^{-1}}$, and Mo, 0.2–0.5 $\mathrm{mg\,kg^{-1}}$, respectively;
5. pH values should be ascertained; the pH of agricultural soils is in the range of 4–9. It is difficult to define optimum pH values of different plant species. Most food crops grow well in acid soils of pH approximately 6. Lime is called the foundation of crop production or 'workhorse' in acid soils;
6. Fertilizer field trials establish nutrient requirement and are the most important part of the overall

diagnostic process of a nutrient disorder, whereby the full interplay of soil, plant, climate, and management factors can occur.

Plant Analysis

Plant analysis is based on the concept that the concentration of an essential element in a plant or part of a plant indicates the soil's ability to supply that nutrient. This means it is directly related to the quantity in the soil that is available to the plant. For annual crops, the primary objective of plant analysis is to identify nutritional problems or to determine or monitor the nutrient status during the growing season. If deficiency is identified early in the growth stage of a crop, it is possible to correct it during the current season. Otherwise, steps should be taken to correct it in the next crop. Plant analysis can be useful for the prediction of nutrient needs in perennial crops, usually for the year following the time of sampling and analysis. Like soil analysis, plant analysis also involves plant sampling, plant tissue preparation, analysis, and interpretation of analytical results. All these steps are important for a meaningful plant analysis program.

Many factors such as soil, climate, plant species, and their interactions affect absorption of nutrients by growing plants. However, the concentrations of the essential nutrients are maintained within rather narrow limits in plant tissues. Such consistency is thought to arise from the operation of delicate feedback systems, which enable plants to respond in a homeostatic fashion to environmental fluctuations. Table 4 shows average adequate concentrations for essential nutrients in important field crops. No doubt these values vary with soil, climate, crops, and management practice, and it is very difficult or not logical to make generalizations. Generalized values (Table 4) do give the reader some idea about what adequate levels of nutrients are in crop plants.

For the interpretation of plant analysis results, a critical nutrient concentration concept has been developed. This concept is widely used in the interpretation of plant analysis results for diagnosing nutritional disorders. Critical nutrient concentration is usually designated as a single point within the bend of the curve when crop yield is plotted against nutrient concentration where the plant nutrient status shifts from deficient to adequate. The critical nutrient concentration has been defined in several ways: (1) the concentration that is just deficient for maximum growth; (2) the point where growth is 10% less than the maximum; (3) the concentration where plant growth begins to decrease; and (4) the lowest amount of element in the plant accompanying the highest yield. The above definitions are similar, but not identical. It is well known that there is a good relationship between concentration of a nutrient and yield of a given crop. This is shown in Figures 2 and 3, which show a highly significant correlation ($R^2 = 0.94$ and $R^2 = 0.78$) between N in lowland rice plants (*Oryza sativa* L.) and average yields of lowland rice from three field experiments in the central part of Brazil. Similarly, N uptake (N concentration × dry weight) in shoots as well as in grain has a significant quadratic relationship with shoots as well as grain yield in lowland rice (Figures 4 and 5). In addition to the critical nutrient-concentration approach, tissue

Table 4 Average concentration of essential nutrients in dry matter sufficient for adequate growth of field crops[a]

	Concentration	
Nutrient	*g kg^{-1} or mg kg^{-1}*	*μmol g^{-1}*
Carbon	450	37 500
Oxygen	450	28 125
Hydrogen	60	60 000
Nitrogen	14	1 000
Phosphorus	1.9	60
Potassium	9.8	250
Calcium	5.0	125
Magnesium	1.9	80
Sulfur	1.0	30
Iron	112	2
Manganese	55	1
Zinc	20	0.3
Copper	6	0.1
Boron	22	2
Molybdenum	0.10	0.001
Chlorine	106	3

[a]Values of macronutrients are in grams per kilogram and values of micronutrients in milligrams per kilogram.

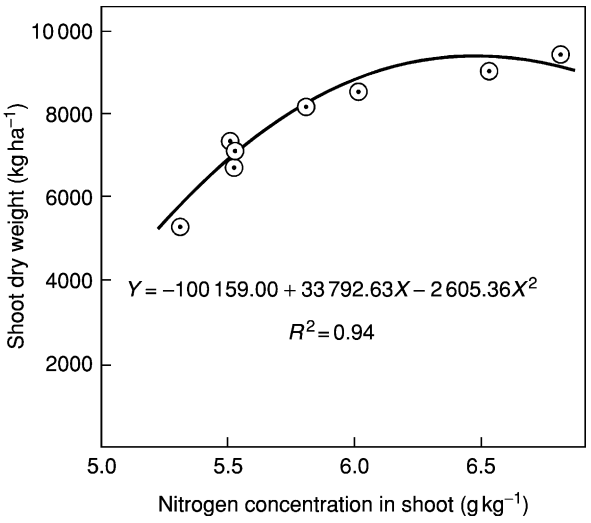


Figure 2 Relationship between nitrogen concentration in shoots and shoot dry weight of lowland rice.

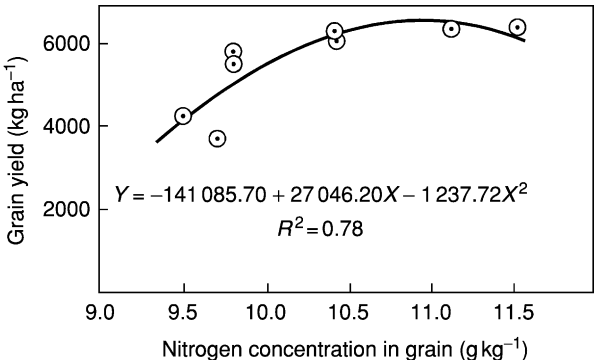


Figure 3 Relationship between nitrogen concentration in grain and grain yield of lowland rice.

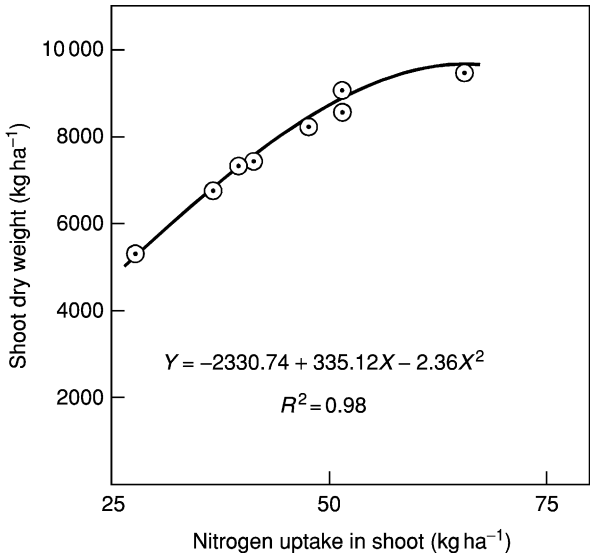


Figure 4 Relationship between nitrogen uptake in shoots and shoot dry weight of lowland rice.

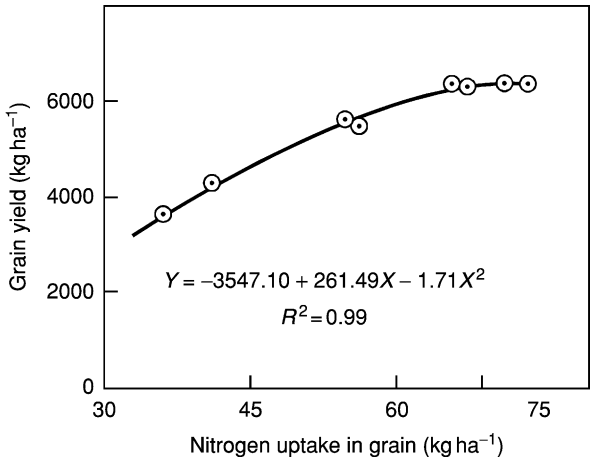


Figure 5 Relationship between nitrogen uptake in grain and grain yield of lowland rice.

Table 5 Approximate nutrient accumulation by important field crops

Crop	Yield (t ha^{-1})	N (kg ha^{-1})	P (kg ha^{-1})	K (kg ha^{-1})	Mg (kg ha^{-1})	S (kg ha^{-1})
Alfalfa (hay)	9	240	28	141	24	25
Cotton (lint)	1	120	20	75	24	20
Peanut	2	170	13	91	12	15
Rice	6	100	22	133	12	10
Oat	4	130	20	133	9	20
Potato	40	175	35	257	24	20
Sweet potato	40	190	33	282	39	–
Sugar beet	45	200	39	249	54	35
Sugar cane	100	130	39	282	48	60
Rapeseed	3	165	31	183	18	65
Barley	5	150	24	125	15	20
Pea	2	125	15	66	9	–
Dry bean	2	125	15	66	9	–
Sunflower	3	120	26	199	33	15
Cassava	40	150	31	291	24	20
Corn	6	120	22	100	24	25
Soybean	3	220	18	141	24	20
Sorghum	4	120	18	83	18	15
Wheat	6	170	33	145	18	30

Reproduced from Potash and Phosphate Institute of Canada (1988) *Potash: Its Need and Use in Modern Agriculture.* Saskatchewan, Canada: Potash and Phosphate Institute of Canada.

nutrient status can also be diagnosed by the diagnosis and recommendation integrated system (DRIS) and composition nutrient diagnosis (CND).

Nutrient concentrations in plant are generally expressed in milligrams per kilogram or grams per kilogram. Sometimes there is confusion between concentration and accumulation. Concentration is content per unit dry weight of plant tissue, whereas accumulation is concentration multiplied by dry matter. Nutrient accumulation is generally expressed in kilograms per hectare. Table 5 shows nutrient accumulation by important field crops. Nutrient accumulation is related to crop yield and crop species. Nitrogen and potassium are accumulated in higher quantities as compared to P, Mg, and S. Uptake and absorption are also very common in expressing plant analysis results. Uptake or absorption normally refers to the process of elemental movement into and within plants and not to quantitative-based terms such as concentration or total accumulation.

Data related to the quantity of nutrients accumulated during a cropping season give an idea of fertility depletion and help in managing soil fertility for the succeeding crops. The best time to determine nutrient accumulation is at flowering or at harvest, when

Table 6 Methods of correcting nutrient deficiency and Al and Mn toxicity

Nutrient/element	*Corrective measures*
Nitrogen	Addition of organic matter to the soil; application of N fertilizers, including legumes in the crop rotation; use of foliar spray of 0.25–0.5% solution of urea
Phosphorus	Adjustment of extreme pH; application of phosphorus fertilizers
Potassium	Application of potassium fertilizers, incorporation of crop residues
Calcium	Liming of acid soils; addition of gypsum or other soluble calcium source where lime is not needed; foliar spray in acute cases with 0.75–1% calcium nitrate solution
Magnesium	Application of dolomitic limestone; foliar application of 2% magnesium sulfate solution
Sulfur	Use of fertilizer salt containing sulfur such as ammonium sulfate and single superphosphate; application of gypsum or elemental sulfur
Zinc	Addition of zinc sulfate to soil; foliar spray of 0.1–0.5% solution of zinc sulfate
Iron	Foliar spray of 2% iron sulfate or 0.02–0.05% solution of iron chelate; use of efficient cultivars
Copper	Soil application of copper source of fertilizer or foliar spray of 0.1–0.2% solution of copper sulfate
Boron	Soil application of boron source or foliar spray of 0.1–0.25% solution of borox
Molybdenum	Liming of acid soils, soil application of sodium or ammonium molybdate; foliar spray of 0.07–0.1% solution of ammonium molybdate
Manganese	Foliar application of 0.1% solution of manganese sulfate
Aluminum/ manganese	Application of lime; use of tolerant species or cultivars

accumulation is expected to be maximum. Grains as well as straw should be analyzed and their weights per unit area determined to calculate total accumulation.

Crop Growth Response

Soil and plant analyses are the common practices for identifying nutritional deficiencies in crop production. The best criterion, however, for diagnosing nutritional deficiencies in annual crops is through evaluating crop responses to applied nutrients. If a given crop responds to an applied nutrient in a given soil, this means that the nutrient is deficient for that crop. Relative decrease in yield in the absence of a nutrient, as compared to an adequate soil fertility level, can give an idea of the magnitude of nutrient deficiency. For example, research conducted on Oxisols of central Brazil provides evidence of which major nutrient is most limiting for upland rice, common bean, corn, soybean, and wheat. The conclusion was that P deficiency was the primary yield-limiting nutrient for annual crop production. Methods for correcting nutrient deficiencies and AL and Mn toxicities are summarized in Table 6.

Summary

Soil is our most precious nonrenewable resource in food production. In the long run we must make sure that the productivity of the soil is maintained or improved. Maintaining adequate level of nutrients is one way to achieve this objective. Historically, the study of soil fertility has focused on managing soil nutrient status to create optimal conditions for plant growth. The amount and availability of soil nutrients are determined by physical, chemical, and biological processes, which are often managed in an attempt to increase or optimize uptake of nutrients and yields. There are 16 essential nutrients for plant growth and their presence in the soil at an adequate level and proportion is fundamental for obtaining higher yields. The important criteria in evaluating nutrient availability or deficiency or sufficiency are visual symptoms, soil testing, plant tissue testing, and crop response to applied nutrients. Soil testing is generally recognized as the best available method for diagnosing soil nutrient limitations before a crop is planted so that correction can be made in that year by appropriate fertilization.

Soil fertility is measured in terms of quantity of available nutrient in a given soil. However, plant nutrition involves transport of nutrients from soil to plant roots, absorption by roots, translocation from absorbing site to various parts of the plants, and finally utilization in the growth and development processes of plants. This means it is a very dynamic process and influenced by soil, climate, and plant factors. A thorough knowledge of these factors is essential to understand properly the subject of mineral nutrition of plants.

Integrated nutrient management systems should be used to improve the overall nutrient availability and use efficiency by crops. Successful application of many practices which are part of the integrated nutrient management can increase the profitability and minimize the environmental impacts of nutrient use, the fundamental goal of supplying adequate amounts of nutrients in appropriate proportions for higher yields.

See also: **Fertigation**; **Fertility**; **Fertilizers and Fertilization**; **Foliar Applications of Nutrients**; **Macronutrients**; **Micronutrients**; **Nutrient Management**

Further Reading

Baligar VC, Fageria NK, and He ZL (2001) Nutrient use efficiency in plants. *Communication in Soil Science and Plant Analysis* 32: 921–950.

Barber SA (1995) *Soil Nutrient Bioavailability: A Mechanistic Approach*, 2nd edn. New York: John Wiley.

Bennett WF (1993) *Nutrient Deficiencies and Toxicities in Crop Plants*. St. Paul, MN: American Phytopathological Society.

Black AM (2000) Bioavailability of major essential nutrients. In: Sumner ME (ed.) *Handbook of Soil Science*, pp. 3–37. Boca Raton, FL: CRC Press.

Fageria NK and Baligar VC (1997) Integrated plant nutrient management for sustainable crop production: an overview. *International Journal of Tropical Agriculture* 15: 1–18.

Fageria NK and Baligar VC (2001) Lowland rice response to nitrogen fertilization. *Communication in Soil Science and Plant Analysis* 32: 1405–1429.

Fageria NK and Baligar VC (2001) Improving nutrient use efficiency of annual crops in Brazilian acid soils for sustainable crop production. *Communication in Soil Science and Plant Analysis* 32: 1303–1319.

Fageria NK and Gheyi HR (1999) *Efficient Crop Production*. Campina Grande, Brazil: Federal University of Paraiba.

Fageria NK, Baligar VC, and Jones CA (1997) *Growth and Mineral Nutrition of Field Crops*, 2nd edn. New York: Marcel Dekker.

Foth HD and Ellis BG (1997) *Soil Fertility*, 2nd edn. Boca Raton, FL: CRC Press.

Grundon NJ (1987) *Hungry Crops: A Guide to Nutrient Deficiencies in Field Crops*. Brisbane, Australia: Department of Primary Industries.

Marschner H (1995) *Mineral Nutrition of Higher Plants*, 2nd edn. New York: Academic Press.

Mengel K, Kirkby EA, Kosegarten H, and Appel T (2001) *Principles of Plant Nutrition*, 5th edn. Dordrecht, The Netherlands: Kluwer Academic Publishers.

Mortvedt JJ (2000) Bioavailability of micronutrients. In: Sumner ME (ed.) *Handbook of Soil Science*, pp. 71–88. Boca Raton, FL: CRC Press.

Tisdale SL, Nelson WL, Beaton JD, and Havlin JL (1993) *Soil Fertility and Fertilizers*, 5th edn. New York: Macmillan Publishers.

NUTRIENT MANAGEMENT

G D Binford, University of Delaware, Newark, DE, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

The management of nutrients in soils is important for both economic and environmental reasons. Plants require certain amounts of each essential nutrient to maintain optimum health. During the production of crops, it is important to have enough of each nutrient present in the soil to optimize productivity and maximize profitability. These nutrients come primarily from native soil minerals, the application of fertilizers, nitrogen fixation, or the application of animal manures. Although these nutrients are essential for optimum plant growth and crop production, it is important from an environmental standpoint that these nutrients are not applied in excess of what is required by the plant. When excessive amounts of nutrients, primarily nitrogen (N) and phosphorus (P), are applied to the soil, these nutrients can leak from the soil and move into nearby water supplies. As surface waters increase in concentrations of N and P, these increases can result in accelerated growth of water plants. An overabundance of plants growing in the water can result in a process known as eutrophication, which can be detrimental to the overall health of the water ecosystem. Excess N, as nitrate, can also leak from soils and get into groundwater supplies, which can be a concern for both human and animal health.

The potential for losses of nutrients is especially great in areas with intensive concentrations of animals, because the animal wastes from these facilities contain significant quantities of nutrients. As a result, proper nutrient management is an especially major concern in areas with large animal production facilities. These animal wastes are usually land applied to provide nutrients for the crops to be grown on the land. These wastes, however, often contain a higher P:N ratio than the ratio needed by the crop. In addition, because N is often needed in relatively high amounts for crops such as corn, when these manures are applied to meet the N demands of the crop, the amount of P applied is much greater than that required by the crop. As a result, after decades of similar applications, P concentrations in soils become quite excessive. Research has shown that, when P concentrations of soils reach a certain level, the risk of P loss from the soil to nearby water supplies greatly increases.

Losses of nutrients are not only confined to operations with intensive livestock operations. In fact, nutrients can be lost from any soil when nutrients are present at levels greater than growing plants can utilize and/or the soil can hold. This can occur on farms, golf courses, lawns, or any other place nutrients are applied to soils. Also, poor application techniques can lead to nutrient losses from soils. For example, nutrients that are applied accidentally to impervious surfaces such as roads or sidewalks will most likely be lost in runoff and could impact surface-water quality.

In recent years, the concept of nutrient-management planning has been promoted as a practice that will help prevent overapplication and improper application of nutrients to the land. The primary goals of nutrient-management planning are to maximize profitability of crop production and to prevent a buildup of excess nutrients on the land, thereby reducing the potential for nutrient losses from land to water.

Nutrient-Management Planning

There is a specific process that is typically followed in the development of a sound nutrient-management plan. It is important that a nutrient-management plan be developed by someone who understands both soil and crop sciences, so that both an environmentally sound and an economically feasible nutrient-management plan can be prepared. The process of developing a nutrient-management plan can be broken down into several steps. These steps include:

1. Development of a map for the operation;
2. Determination of crop or plant nutrient requirements;
3. Inventory of nutrient production on the operation;
4. Determination of nutrient balance on the operation;
5. Emergency action plan;
6. Record keeping.

Step 1: Creating Maps

Detailed maps of the operation are developed and these maps include:

1. Identification of property boundaries;
2. Identification of field boundaries with field identification codes;
3. Location of any areas sensitive to nutrient loading, such as surface waters or wellheads;
4. Direction of flow for any streams on the property;
5. Identification of all soil types where nutrients are applied;
6. If present, location of animal production facilities;
7. Location of nutrient storage (fertilizer and/or manure);
8. Location of roads, neighboring homes, or public facilities.

These detailed maps are important for understanding the layout of the operation; they can be used to guide nutrient applications near areas that are critical to nutrient loading such as surface waters. If necessary, these maps can also be used as a guide to limit applications of animal wastes near public roads, neighboring homes, or areas of public gatherings. The maps with field identification codes are also helpful to the person implementing the plan so that the appropriate nutrient recommendation can be applied to the corresponding area of the field. Maps can be drawn by hand, with computer software, or created from aerial photographs.

Step 2: Nutrients Required by Crops or Plants

Identifying which nutrients are required includes:

1. Determining the crop rotation or plants to be grown on the facility;
2. Determining crop/plant nutrient requirements;
3. Soil testing and/or plant analysis;
4. Nutrient credits for animal waste applications, if applicable;
5. Nitrogen credits for previous legume crops, if applicable;
6. Nitrogen credits for nitrate in irrigation waters, if applicable.

The first step in determining plant nutrient requirements is deciding on what type of crop or plants will be grown and the rotation of crops or plants for the period of the nutrient-management plan. After this decision has been made, a determination of the required nutrients is needed. The primary procedure for making this determination in most situations is by soil testing.

Soil testing is the best method available for determining how much nutrient is needed for optimum production of most crops. For some crops, the use of plant analysis may be the preferred method of determining crop nutrient needs. Making nutrient-application decisions without soil testing or plant analysis can be a risky endeavor because this will probably result in reduced profitability from either under- or overapplying nutrients; in addition, overapplication of nutrients can result in reduced water quality. It is important to follow recommended procedures for taking soil and plant samples.

When making N recommendations for many crops, the determining factor on how much N to apply is based on the expected yield of the crop. The best method for determining expected yield is from previous yield records. It is important to use an expected yield that is realistic, because if an expected yield is used that is much greater than can be achieved realistically then more nutrient will be applied than the crop can utilize. Applying excess nutrient increases the risk of nutrients getting into nearby water supplies and jeopardizing water quality. At the same time, using an expected yield that is too low can limit yields due to nutrient deficiencies, which reduces profitability of crop production.

Once the recommended rate of fertilizer has been determined, it is important to make any adjustments in the recommended rate for nutrient credits from animal wastes, previous legume crops, and N that is present in irrigation water. These credits can be determined for animal wastes by having a nutrient analysis done by a laboratory. Quantification of nitrates in irrigation water can also be determined by a laboratory. Credits for previous legume crops are usually standard values that have been determined by crop-management and soil-fertility research.

Because a nutrient-management plan is typically written to make fertilizer recommendations for all crops throughout the rotation, it is important to estimate the amount of nutrients that will be removed when a crop is harvested. Estimating the amounts of nutrient removed from the soil by a crop will usually be necessary for making appropriate fertilizer recommendations for subsequent crops in the rotation, unless soil samples are taken for each crop. Typically, soil samples for soil testing are taken every 2–3 years depending on the rotation and soil type. In some situations, a crop grower will apply enough nutrients (i.e., nonmobile nutrients) to meet the nutrient requirements of the current crop and the subsequent crop, which reduces application costs. An example of this practice would be a corn–soybean rotation, where a grower may often apply enough P and potassium (K) to the corn crop so that this application not only meets the requirements of the corn crop but also meets the P and K requirements of the following soybean crop.

Step 3: Inventory of Available Nutrients on the Operation

It is only necessary to make an inventory of available nutrients for operations that have animals and are producing animal waste; if no animals are present, then this step will not be needed. If an operation has no animals, such as a golf course or a grain production farm, then there are no nutrients (i.e., animal waste) being generated on the operation, so all nutrient needs on an operation of this type are fulfilled by importing nutrients to the operation. If animals are present on the operation, then it is important to determine quantities of nutrients (i.e., waste) being generated on the operation so that these nutrients can be utilized most appropriately so that potential losses of these nutrients to the environment are minimized and the economic value of these nutrients is maximized.

The first phase in determining nutrient inventory is to determine how much animal waste is being produced on the operation. If there are no records from past animal waste production from the operation, it is necessary to use average animal waste production values for the development of the nutrient-management plan. These average production values have been generated from previous research and can be found through most University/Cooperative Extension programs. To improve the accuracy of determining the amount of animal waste being generated on an operation, records can be kept on the amount of manure that is removed from the operation. These amounts can be determined by weighing each load of animal waste as it leaves the animal production facility or by determining the volume of animal waste as it leaves the facility and then converting this volume to an estimated weight. Actual measurements from the animal operation would be the most accurate and reliable method for determining nutrient inventory on an animal operation, as compared to just using established book (i.e., average) values.

In addition to determining the quantity of animal waste being produced, the nutrient content of the animal waste should be determined. This can be done by sending a representative sample of the animal waste to a laboratory for analysis. It is important to make sure a representative sample is taken and also that the sample is handled properly before sending to the laboratory. The laboratory should be able to provide details on the proper procedure for taking a representative sample of animal waste; procedures vary depending on the type of animal waste being sampled. It is also possible to obtain average book values for the nutrient content of animal wastes; however, laboratory analyses provide a better determination for each specific animal operation than an average book value.

It is important to remember when calculating the amount of available nutrient present on the operation that the availability of N in animal manures will be dependent on how the manure is handled before and during application. For example, significantly less N will probably be available for future crops when manure is applied on the soil surface and not incorporated compared with manure that is applied and

incorporated immediately; this is because of potential volatilization of N from the nonincorporated manure. General guidelines have been developed for determining the amount of N that is potentially available to crops after the manure has been field-applied. It is important to remember that these are just guidelines and that weather conditions have a major impact on the amount of N that is potentially available to subsequent crops. Because of this potential variability in the amount of plant-available N that is present from field-applied manures, it is advisable to use diagnostic tools (e.g., the pre-sidedress soil nitrate test or PSNT, plant tissue analyses, and chlorophyll meters) to monitor the N status of the growing crop when deciding how much additional N to apply as fertilizer to the growing crop. The use of such tools helps to reduce excess applications of N, which reduce profitability and increase potential environmental harm.

Step 4: Nutrient Balance

As in the previous step, balancing of nutrients is only needed if animals are present on the operation. If no animals are present, then this step is also unnecessary. After determining the quantity of nutrients needed by the planned crop rotation through proper soil testing and the quantity of nutrients that are being produced from the animal production facility, the next step is to evaluate the nutrient balance. The main question that has to be answered before the nutrient balance can be calculated is to determine whether the nutrient management plan will be an N- or P-based plan. In some situations, it may be appropriate to have a metals-based plan. In other words, the animal waste is applied based on the metal content of the waste. In the development of the plan, it is necessary to assess each management area (i.e., area from which the soil sample is taken) for its nutrient needs and whether that area will require an N-based or P-based plan.

Historically on many operations, animal waste has been applied at rates to meet the N demands of the crop or at even higher rates in situations where the waste has been considered a disposable by-product of the animal production facility. In recent years, the concept of P-based planning has become popular. This popularity in P-based planning has resulted from areas where animal wastes have been applied at P rates much greater than the amounts of P removed in the harvested portion of the crop. As a result, after decades of this practice, soil-test P levels have increased to levels that are considered a potential threat for off-site movement of P into nearby water supplies.

Before deciding whether a P-based plan should be used, the criteria to be used to define when the risk of P loss warrants the use of P-based planning instead of N-based planning must be determined. Simply defining a soil-test level that is considered too high is one method; however, a method usually referred to as a P site index probably makes more sense and would be more representative of P-loss potential from a soil. There is currently a great deal of research being conducted on the development of an appropriate P site index. It is important that a P site index be adapted and developed for local conditions. A P site index evaluates all aspects of potential P transport from soil to water and not just the current level of P in the soil (i.e., soil test P). A P site index provides a relative risk of P loss from a soil and takes into account such things as soil erosion potential, distance from water, P loading rates, soil-test P, and subsurface drainage.

A farm balance of nutrients can be determined after a decision has been made on which areas will need a P-based plan and which areas will need an N-based plan. The total nutrient requirements of the crops to be grown can be calculated by summing the total nutrient needs of each area that was soil-sampled. The difference between the total supplies of nutrients produced by the animal operation can then be compared with the total nutrient needs of the crops to be grown. If an excess of nutrients (i.e., animal waste) is available on the operation, an alternative plan should be developed for transporting the excess animal waste from the operation. It is also important to consider when the animal waste will be removed from the animal production facility and when it will be applied to the land or removed from the operation so that appropriate storage can be planned.

It is likely that significant future improvements in nutrient farm balances can be achieved through diet modifications of livestock. Currently, many livestock operations are adding P to animal diets at rates greater than needed for optimal animal health. As a result, much of this added P ends up in the manure and has to be land-applied. New technologies in diet formulations and animal genetics also offer opportunities for reducing the P content of animal manures. For example, recent work in Delaware and Maryland, USA, has shown that significant reductions can occur in the P content of poultry manure by utilizing phytase enzyme and low-phytate corn in the feed ration of the birds. In fact, it is possible that future use of these new technologies and with improved feed management the ratio of P:N in animal manures could be much more in balance with the ratio of P:N that is needed by crops. This would allow manures to be applied at rates equal to the amount of nutrient that will be removed by the crop, which would eliminate nutrient buildup in

soils, thereby, reducing the potential for nutrient losses from land to water.

Step 5: Emergency Action Plan

As a part of the overall nutrient-management plan, it is important to have a planned course of action in the event of a nutrient spill during transport or storage. This action plan should be in place for either manure or fertilizer that is collected, stored, transferred, or applied on the operation. This emergency plan should include telephone numbers for reporting a spill and for seeking assistance. The plan should also show anticipated flow paths in the event of a spill or discharge of nutrients. To be most helpful as a part of this action plan, the anticipated flow path should be shown on the overall site map of the operation.

Step 6: Record Keeping

Maintaining detailed records is an important part of having a useful nutrient-management plan. Development of future nutrient-management plans is improved and less time-consuming if detailed records are maintained concerning the operation. Detailed records also provide documentation on procedures used on the operation, which could be helpful should the need arise to prove activities used on the operation. Many of the needed records for determining animal waste production may not be available the first time a nutrient-management plan is written, which will require the use of average book values. Future nutrient-management planning, however, will be made easier and more site-specific if on-site records are available from the operation.

The end result of nutrient-management planning is the same result of any planning process. It does not matter if we are talking about a business plan, a financial plan, or a nutrient-management plan: the more time and effort put into the planning process, the better the results. A well-written and detailed nutrient-management plan should result in an improved efficiency of nutrient use. This improved efficiency should result in an operation that has increased potential profitability and decreased potential damage to the environment.

Summary

Nutrient management planning offers a systematic process for determining amounts of nutrients produced (i.e., animal wastes) and amounts of nutrients needed for optimal plant growth. Any operation that applies nutrients to land should have a nutrient-management plan prepared by an individual with a sound understanding of plant and soil sciences. There are six basic steps to follow when developing a nutrient management plan; however, the details involved with each step will vay depending on the type of operation. It is important to remember that, for normal growth and development, plants must acquire nutrients from soils through root uptake. Therefore, if soil testing indicates a need for nutrients, then nutrients must be added to the soil to support optimal plant growth. Nonetheless, if too much nutrient is applied to land, these excess nutrients can leak from soils and result in water-quality concerns. Therefore, the overall goal of sound nutrient management is to optimize plant health, while minimizing environmental harm.

See also: **Fertility**; **Fertilizers and Fertilization**; **Fertigation**; **Foliar Applications of Nutrients**; **Macronutrients**; **Micronutrients**; **Nutrient Availability**

Further Reading

Beegle DB, Carton OT, and Bailey JS (2000) Nutrient management planning: justification, theory, practice. *Journal of Environmental Quality* 29: 72–79.

Hoeft RG (2002) Developing a Nutrient Management Plan. In: *Illinois Agronomy Handbook*, pp. 132–141. Urbana, IL: University of Illinois Cooperative Extension.

Koelsch R and Shapiro C (1997) *Determining Crop Available Nutrients From Manure*. NebGuide G97-1335A. Lincoln, NE: University of Nebraska–Lincoln.

Leytem AB and Sims JT (1999) *Nutrient Management for Delaware Agriculture – Developing a Comprehensive Nutrient Management Plan*. Fact Sheet NM-02. College of Agriculture, University of Delaware. Newark, DE: University of Delaware Press.

Midwest Farm Service (2002) *Livestock and Poultry Environmental Stewardship (LPES) Curriculum*. Ames, IA: Midwest Farm Service.

Natural Resources Conservation Service (2000) *Comprehensive Nutrient Management Planning Technical Guidance*. National Planning Procedures Handbook, Subpart E, Parts 600.50-600.54; Subpart F, Part 600.75. NRCS, US Department of Agriculture. Washington, DC: US Government Printing Office.

Nutrient Management Planning with Animal Feeding Operations. Handbook for Nutrient Management Planners (1999) Auburn, AL: Alabama Cooperative Extension System/Auburn University Press.

Sims JT and Moore PA Jr (1998) Nutrient management planning: phosphorus or nitrogen-based? In: National Poultry Waste Management Symposium. Auburn, AL: National Poultry Waste Management Symposium Committee.

O

ORGANIC FARMING

C A Francis, University of Nebraska, Lincoln, NE, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

One of the most rapidly growing sectors of the food industry is organic products, in both the USA and Europe. Market share in 2002 was 2% in the USA and up to 5% in other countries. In part this growth is due to increasing concern in the general consumer population about the amounts of chemical pesticides that may be carried over from field production on the surface of fruits and vegetables and eventually find their way into the human food chain. Another concern comes from a recent series of outbreaks of foot and mouth disease, as well as bovine spongiform encephalitis in domestic animals in northern Europe, events that have caused many to begin buying meat from animals produced using organic methods. There are ambitious national goals in Europe to make 10–20% of all agriculture organic and/or ecologic before 2012.

Organic farming is defined as production of crop, animal, and other products without the use of synthetic chemical fertilizers and pesticides, transgenic species, or antibiotics and growth-enhancing steroids, or other chemicals. This is called 'ecological farming' in some northern European countries. It involves a philosophy of improving soil fertility, protecting crop plants and animals through natural methods and system design, maintaining the rural environment as a desirable place to live, and showing concern for the distribution of benefits from agriculture and the food system. In a sense, organic farming is an ancient art that evolved with the first farmers and dominated food production up until approximately 100 years ago. Before that, fossil fuel-derived pesticides or fertilizers were not available, and food was produced essentially by organic methods in most of the world. The landmark book *Farmers of Forty Centuries*, by F.H. King of the US Department of Agriculture, describes the systems used in China, Korea, and Japan, and these would qualify today as organic production systems. What we today call 'organic farming' has its roots in these ancient and highly sustainable systems in Asia.

Organic farming practices have depended on limited scientific research and extensive practical experiences by farmers and proponents of this approach. There are important precedents in Rudolph Steiner's lectures on biodynamic farming in Germany, as well as Sir Albert Howard's extensive observations and experiments in India on composting and improving soil humus. Developments in organic farming over recent decades have included increased attention to building soil fertility, fine-tuning methods of pest management, and emergence of certification organizations that provide a specific label to guarantee the authenticity of food produced by organic methods. Interest in ecologic agriculture in the Nordic region and throughout Europe has been especially intense in recent years. Useful sources that document this history are shown in Table 1 and in the Further Reading section. These books describe the evolution of agricultural practices as well as the growing popularity of organic products over the last century.

Building on a brief history, this article continues with discussions of how organic farmers maintain and build soil fertility, and how they design systems to protect against pest problems, especially weed management. Important to development of organic farming are the contributions of the biodynamic movement and integrated systems research. Although organic crop yields generally are lower per unit area, many people speculate that this is due to less research and development invested in these systems. In fact, the best organic growers often have yields equal to or better than country or regional averages for crops, and most farmers benefit from lower production costs and premium prices. Labeling and marketing are important issues in the commercialization of organic food, and these must include discussion of claims for nutritional benefits from the products of organic farming practices. Finally, there is intense discussion about the industrialization of organic farming and food products, and this will continue as a contentious issue into the future.

Table 1 Key background information on organic farming and gardening, from historical antecedents to current reviews

Key reference	*Description of relevance to organic farming*
King, 1911[a]	Travelogue and descriptions of observations of farming systems by USDA's Dr. King during travels in China, Korea, and Japan in the early 1900s; a primer on classic, sustainable agriculture that has endured for more than 4000 years
Steiner, 1958[b]	Original eight lectures by Rudolph Steiner from Germany, founder of anthroposophy and father of biodynamic agriculture; these provide a source of many unique ideas on management and uniting the spiritual with the biological
Howard, 1943[c]	Detailed instructions on compost management and development of humus in the soil, based on experiences of Alfred Howard in India, plus ideas on integrated farming
Rodale, 1955[d]	Visionary author and publisher who demonstrated organic farming and gardening techniques on his research farm; founder of Rodale Press, he popularized the concept in the USA
Edwards *et al.*, 1990[e]	Comprehensive review of sustainable agriculture in a series of chapters, including Historical Roots written by Richard Harwood; contains many references and updates on organic farming
Lampkin, 1990[f]	Highly prolific writer and scientist who has chronicled the growth of the organic industry and published detailed lists of practices and practical economic guidelines (since 1995) in the UK
Sooby, 2003[g]	Survey results of organic farming research in the US landgrant system, showing that there is proportionately little research compared to conventional farming and market demand
Lotter, 2002[h]	An unusually complete review of the literature on organic farming; focuses on cultural practices as well as conversion process, economics, consumer demand, and certification activities in USA

[a]King FH (1911) *Farmers of Forty Centuries, or Permanent Agriculture in China, Korea, and Japan.* Emmaus, PA: Rodale Press.
[b]Steiner R (1958) *Agriculture: A Course in Eight Lectures.* London, UK: Biodynamic Agriculture Association.
[c]Howard A (1943) *An Agricultural Testament.* London, UK: Oxford University Press.
[d]Rodale JI (1955) *Organic Gardening: How to Grow Healthy Vegetables, Fruits, and Flowers Using Nature's Own Methods.* Garden City, NJ: Hanover House.
[e]Edwards CA, Lal R, Madden P, Miller RH, and House G (eds) (1990) *Sustainable Agricultural Systems.* Ankeney, IA: Soil and Water Conservation Society.
[f]Lampkin N (1990) *Organic Farming.* Ipswich, UK: Farming Press.
[g]Sooby J (2003) *State of the States: Organic Systems Research at Land Grant Institutions, 2001–2003.* Santa Cruz, CA: Organic Farming Research Foundation.
[h]Lotter D (2002) Organic farming. *Journal of Sustainable Agriculture* 21(4): 59–128.

Building Soil Fertility

Organic farmers build and maintain soil fertility through a number of practices that mimic the natural ecosystem, rather than supplying crop needs through application of chemical, fossil fuel-derived commercial fertilizers. One characteristic of organic farmers is their strong belief in the soil as a complex place that is home to a living ecosystem of billions of organisms that work together to provide cycling of nutrients, and not merely a physical medium that holds plants upright and provides a place to store chemical nutrients until they are taken up by plants. The care of this living system and the cycling of its nutrients in field crops, vegetables, pastures, and domestic animals are central to the philosophy of organic farming and a foundation on which farmers build sustainable systems.

Crop diversity, careful soil management, legumes, and animals in the system are key strategies for the design of efficient and sustainable nutrient relationships in each field. These diverse components can enhance the process of nutrient cycling within the system and contribute to minimizing the losses of nutrients. There is no special formula for creating more nutrients within the organic farming system, but rather a logical planning to minimize extraction or loss of nutrients from the field and farm, and to replace those nutrients using internal resources on the farm (see Table 2). The crop sequence and mixture are designed to maximize nitrogen fixation and preservation of organic matter, efficient cycling of nutrients, and retention of nutrients where they are needed on the farm. Some organic farmers maintain that it is impossible to design an efficient system without including ruminant animals, while others increase the proportion of legumes in the system and use this strategy to accomplish the same goals. The bottom line is to incorporate continuously as many nutrients as the amount extracted in harvest and removal from the farm. A goal is to increase organic matter and nutrients to a more healthy level for soil and plants when these factors are deficient, but not to a level that creates a risk for erosion or leaching of soluble nutrients. The concept of balance is paramount, and crop rotation is essential. (*See* **Crop Rotations**.)

Soil organic matter is central to nutrient storage and cycling in the soil, and to providing nutrients to crop and pasture plants at the right time in their growth cycles. Along with the clay mineral fraction of the soil, organic matter provides negatively charged sites for cations (positively charged nutrient ions such as potassium, calcium, and magnesium) to be adsorbed, thus contributing to the cation exchange

Table 2 Nutrient-management practices and processes used by organic farmers to maintain soil fertility adequate for efficient and economical plant growth

Practice or process	*Description and impact on soil quality and crop production*
Legumes in rotation	Nitrogen fixation, contributions to soil organic matter, fibrous root systems contribute to loosening soil and quality
Manure application	Provides nitrogen and other nutrients, increases level of soil organic matter, improves soil tilth and quality
On-farm cycling	Use of primarily green manure crops and animal manure produced by feeding grain or forage on same farm helps balance
Cover crops	Provide cover over the ground surface during much of the year to prevent erosion, trap nutrients, contribute organic matter
Timely seeding	Synchronize crop planting and growth cycle with availability of nutrients from soil reserve
Reduced tillage	Maintains crop residues on top of soil, where they break down more slowly and provide nutrients when needed by later crops
Crop–Ley rotation	Rotation of crops with perennial pastures and grazing animals provides manure, fixed nitrogen, and green biomass to soil
Contour planting	Seeding crops on contours or terraces reduces soil erosion and thus reduces nutrient losses with eroding soil
Crop–animal systems	Integration of a diverse array of animal enterprises provides an internal market for forages and helps cycle nutrients

capacity of the soil. Organic matter is increased by incorporating green manure crops and growing forage crops or pastures, as long as they are fed to ruminants and manure recycled in the field, and these crops are not harvested and sold off the farm. The important organic matter is the active fraction that is being broken down by soil organisms and chemical reactions in the soil solution, thus releasing nutrients available to crops.

Algae are the only soil organisms to capture energy directly from sunlight, while all the others (bacteria, fungi, actinomycetes, and a host of macroorganisms) are dependent on breakdown of organic matter to provide their energy and nutrients, which they release and make available to crop and pasture plants. Soils that are too acid or too basic have lower levels of microbial and macroorganism activity, and thus break down soil organic matter more slowly. Most soil organisms such as earthworms are sensitive to soil acidity, and most have optimum activity at near-neutral soil conditions (pH = 7.0). Organic systems are designed to achieve this balance and build soil organic matter. 'Humus' is a term used for a combination of decomposed organic matter and high-molecular-weight compounds that are generally more resistant to biological decomposition than the active organic matter. Preserving organic matter by preventing erosion, minimizing extraction from the field (such as hay or silage harvest), and increasing the proportion of crop residues or green-manure crops in the sequence can enhance the long-term soil organic matter levels and overall soil quality. Higher levels of soil organic matter also help break the force of rain drops, especially when there is residue on top of the soil surface, and contribute to water storage by acting like a large sponge above and within the total soil mass. Thus soil organic matter contributes nutrients and increases water-holding capacity.

Organic systems depend on legumes as a key to maintaining soil nitrogen and other nutrients at levels adequate for successful crop and pasture-plant growth. The atmosphere is approximately 78% elemental nitrogen, and this can be fixed in plants by symbiotic bacteria, which live on the roots of legumes, and made available to subsequent crops when nitrates are excreted into the soil solution or when the legumes or parts of them are incorporated into the soil as green manure or as crop residues. In conventional farming, nitrogen is supplied primarily by application of chemical forms of nitrate, and higher levels applied in this way actually suppress the nitrogen fixation by legumes. It is possible to fix more than 100–150 $kg\,ha^{-1}$ nitrogen in 1 year from an appropriate legume such as alfalfa or clover. It is also possible to lose 25–150 $kg\,ha^{-1}$ of nitrogen in 1 year if this is applied as a soluble chemical, especially if there are high-rainfall events and low-organic-matter levels in the soil. In organic farming systems, the key is to fix and store adequate nitrogen for crop or pasture needs, while not having an excessive amount from legumes that exceed crop needs. The soluble nitrate ions are susceptible to leaching during rainfall events that exceed the ability of soil to hold this moisture.

Elements such as phosphorus, potassium, sulfur, calcium, magnesium, and others are also important for crop and pasture growth, but are less soluble and thus more tightly stored in the soil profile. Potassium is adequate in most agricultural soils and not easily leached out, and in pastures most is returned to the soil by grazing ruminants. Phosphorus is highly insoluble, and its availability is enhanced by the presence of fungal species called 'mycorrhizae' that infect plant roots and help in absorption of phosphate. Since there are often adequate supplies of phosphorus deep in the soil profile, deep-rooted crops with a strong tap root can 'mine' this element from lower soil strata. These elements and others can be supplied from organic sources, especially straw or feed that is cycled back

to the field as manure from animals, and again the balance in the system is the important goal. A smaller amount of these other elements is needed for plant growth compared with nitrogen, but their management is no less important due to serious effects of deficiencies in some soils. A number of soil-management practices, such as contour planting and reduced tillage, are useful to preserve nutrients in the field, in both organic and conventional production.

Protecting Plants against Pests

Similar to the strategies used for maintaining soil fertility, there are comparable lessons from natural ecosystems that help inform our design of agroecosystems to tolerate debilitating or completely destructive pest populations. Many of these principles relate to cropping diversity in each field and on the farm. Genetic resistance to insects and pathogens are commonly used in both organic and conventional farming systems. Well-adapted and healthy crop, pasture, and domestic animal species with tolerance to prevalent pest problems promote the least-expensive control methods, since this alternative may require no further management costs after the crop is planted or the basic animal enterprise is established. Other practices can be incorporated into the management strategy that will control or minimize losses to pests (see Table 3).

Crop rotations are desirable in most production systems, and these are both prescribed in organic rules and widely used to enhance soil fertility, reduce weed populations, and protect against many insects and plant pathogens. Rotations of unalike crop species will favor different weed species in each season or year and thus break the reproductive cycles of these weed species or aid in their control by mechanical means. Different summer crops such as maize and soybean have different growth habits and timing of competition with weeds, their root systems, and patterns of nutrient uptake, thus favoring a somewhat distinctive weed complex. Rotations of summer with winter crops further exploit this principle, since winter annual weeds that develop with the winter crop can be easily controlled in a subsequent summer crop, and vice versa. Rotating annual crops with perennial crops also provides a means of weed management, since the perennial weeds that develop with perennial crops can often be controlled mechanically in the next annual crop. Lastly, allelopathic species such as winter rye can suppress many weed species that will not be a problem in the succeeding summer annual crop. Cucurbits such as squash grown with maize and beans in the well-known, historical three-sisters crop combination in Central America control many weeds both by shading and by exudates from the leaves and roots. Crop varieties that are rapidly growing early in their cycle can quickly shade weeds and provide a measure of control. These are all methods used by organic farmers. (*See* **Crop Rotations**.)

Spatially diverse cropping patterns such as strip-cropping and intercropping are used to diversify the field planting pattern, confuse some insects that enter from the air, and disrupt their movement once they have entered the field. These patterns in organic agriculture may interrupt or discourage pest reproduction and reduce damage to crops. Pathogens likewise may have more difficulty in dispersion when they occur in a diverse system, compared with their rapid dispersion in monoculture. Diverse agroecosystems may include the added diversity potential of flowering

Table 3 Pest-management practices and processes used by organic farmers to maintain crops, pastures, and domestic animals in a cost-efficient and environmentally sound manner

Practice or process	*Description and impact on pest populations and production*
Crop rotation	Breaks the reproductive cycle of many pest species, provides natural controls through diverse crops in most available niches
Cultivation	Reduces weed populations by mechanical means and reduces their potential for reproduction and spread
Competitive species	Choice of highly competitive crop and pasture species and varieties that will use resources quickly and outgrow weeds
Resistant crops	Genetically resist insects and pathogens that would normally attack susceptible species or crop varieties
Planting dates	Adjustment can allow additional cultivation prior to planting (weeds) or avoidance of problems (insects)
Strip-cropping	Adds spatial diversity to field, habitats for some predators and parasites to aid control of unwanted species
Intercropping	Diversity in the field provides more competition for weeds, also reduces pest outbreaks due to resistance of many individual plants
Allelopathy	Planting certain crops such as winter rye, sorghum, and cucurbits that suppress weeds through chemicals excreted by roots or leaves
Summer–winter crop	Rotation of crops that grow in different seasons disrupts weed reproductive cycles and allows mechanical control
Annual/perennial	Rotation of crops with different life cycles helps control weeds that become more abundants in either species grown continuously
Genetic resistance	Crop or animal resistance or tolerance to pathogens or insect problems is an effective and inexpensive means of control
Floral fencerows	Field borders that are biologically diverse and provide a range of flowering species can harbor predators that reduce crop pests

wild species in fencerows, windbreaks, and field borders, providing habitats for predators and parasites that also help control unwanted pests in the adjacent production fields.

These combinations of genetic resistance, spatial and temporal diversity, and different crops in the field can be combined with adequate nutrient management to produce healthy organic crops. When crops and pasture species are growing rapidly, they often outgrow competition from small weed populations and are more resistant to insect pests and plant pathogens. Likewise, a healthy herd or flock of domestic animals is less susceptible to disease or insect problems. When these animal species in organic systems are allowed free range in pastures and crop residues, they also have fewer problems than the same species raised in confinement. Organic production practices favor this type of outdoor management whenever possible for domestic animals, and appropriate regulations assure the humane treatment of all livestock.

Conversion to Organic Production and Certification

The transition or change from conventional production systems to an organic system is called 'the conversion process,' and this requires 2 years in the UK and 3 years in the USA to complete before a farmer can apply for certification. Most advisors who have worked with farmers say that the most important and often difficult step in the conversion is a psychological change, an embracing of the concept that nonchemical farming is possible, achievable, practical, and economically viable. Armed with this confidence, gained by talking with current organic farmers, reading, attending workshops, or finding motivation from some other source, a farmer is ready to undertake the often considerable changes in farming system design and attendant practices that are required for certification.

Often a number of innovations or major changes are needed in planning the system. It is necessary to plan a conversion to nonchemical methods of providing adequate nutrients for crop and pasture growth. Most often this requires introducing more legumes and green-manure crops, or putting fields to pasture. Farmers often embark upon conversion by planting a forage crop that has no additional nutrient needs or complicated protection against pests. Animal numbers need to be adjusted to fit the carrying capacity of the available area of the farm, since it is highly desirable to minimize the amount of purchased feed or fodder that increases costs. In European countries, there are rules about the amount of nonorganic grains and fodder that can be used for animals while maintaining organic certification. There are also rules in the USA about what is permitted in animal production, and, although these vary somewhat among the private certification groups, there is a basic standard in the new organic legislation that must be followed by all groups. In the overall organic system, there needs to be a balance of areas in crops, in pastures, and in green-manure species to provide a sustainable level of fertility for optimum crop and animal growth, and profitability.

This balance may be achieved by either converting the entire farm at one time, or by converting a fraction of the farm each year over a period of years. The former is desirable because it is possible to take advantage of organic premiums in the marketplace much sooner if all fields are certified after the minimum waiting period. However, it may be too difficult to convert the entire farm at once if this requires planting forage crops in every field and there is not an adequate market for this product. Also, if there is an expected reduction in production until the organic system is established and in balance, it may be too difficult financially to make such a change in all fields at the same time. A more rational approach may be to convert some proportion of the farm each year, perhaps 33% or 25%, so that the process can be completed in 3–4 years. This spreads the risk as well as the reduced income over a longer period of time, and allows for the learning that must take place in making necessary changes.

In addition to changes in crops, cropping sequence, and use of inputs, there are likely to be modifications needed in field equipment, additions of facilities for processing or packaging products, and the development of innovative marketing channels. These changes are not all necessary for every farmer who converts to organic production, and the amount of modification in practices and sales options depends on the number of new enterprises added, the distance to markets and complexity of marketing, and the decisions to add more value to products by some direct sales. If there is a modest change in the balance of crops already grown on the farm, it is possible that no new field equipment will be needed. If there is diversification of current enterprises it is likely that additional equipment for new and specialized uses will be essential. How much processing, packaging, and direct sale will take place on the farm will determine whether additional buildings or equipment are needed for these activities. There is always an option of leasing equipment, time-sharing with neighbors, or custom work on operations that would not justify purchase of equipment for a specialized job. Yet this reduces the opportunity to bring return to a farmer's labor, and most operations are probably

done more economically by the owner or manager of each farm. Each of these steps may require new information, and this can be acquired by reading, from the Internet, or from workshops and classes (offered by Cooperative Extension in the USA), or through a certifying organization. People starting in organic farming may underestimate the need for additional information that will help them make the transition and avoid catastrophes in the field.

Biodynamic Farming, Biological Farming, and Integrated Systems

Several other schools of thought and specific groups have contributed to current philosophy and practices in organic farming. Biodynamic farming grew from the lectures and philosophy of Rudolph Steiner in Germany, founder of the movement called 'anthroposophy.' His teachings embrace a holistic and spiritual interpretation of agricultural systems, and especially the central and special role of humans in those systems. In this movement, the farm is viewed as a self-contained and dynamic organism that can be managed using internal resources and special chemical preparations (called 'preparats') that are essential to the health and vitality of composts, the farm, and its surroundings. The practices reflect a belief in cosmic as well as terrestrial forces on the biological rhythms on the farm. Although most organic farmers do not accept the tenets of Steiner, they do embrace the holistic view of the farm, the rejection of chemical fertilizers and pesticides, and the sound management techniques that were fundamental to the original lectures.

Biological farming includes a range of other movements from the USA, Europe, and India that focus on efficient use of animal manures and compost, green-manure crops, and careful attention to nutrients and soil fertility. Concentration on the humus fraction of soil and its importance in supplying nutrients is central to most systems, and soil health is often related to livestock and human health for those practicing the systems. Such studies have led to more research on the composting of human wastes, a practice long accepted in Asia but not a part of most European or Western Hemisphere fertility strategies in recent years. These systems evolved to include integrated management, a focus that continues today. Integrated farming is currently a strategy that takes into account balance of livestock and crop production, organization of the whole farm into a system that cycles nutrients and provides efficient production of several enterprises, and embraces diversity in those plans. It is easy to see how many of these approaches and strategies for farming overlap and complement each other, although they have different names and are vigorously pursued as different strategies by their supporters.

Marketing and Consumption of Organic Products

The organic food industry is growing rapidly in response to consumer demand and due to the new availability of organic crops and livestock. Countries with the largest areas in organic production include Australia (7.5 million ha), Argentina (3 million ha), Italy (1 million ha), and the USA (900 000 ha) in the year 2000. Those countries with the highest per capita consumption of organic products include Denmark, Switzerland, and Austria. Approximately two-thirds of organic foods marketed are fresh vegetables and fruits, with lower sales in processed and packaged foods. The majority of sales still take place from natural-food stores, although this is changing rapidly as the mainstream supermarkets explore this new and enticing market with their customers. There has been a lack of balance between supply and demand in organic food, for example in the USA there has been more demand than supply, and many farmers are increasing production to meet this demand by the public. In some European countries, such as Germany, there has been adequate production for domestic markets and enough available for organic cereal export to the Nordic region. National planners and farmers in this region are responding by moving toward incentives for producing ecologic grains that can substitute for these imports.

Premiums for organic products are highly variable, ranging from less than 10% where there is limited demand to more than 100% when a product is scarce. A review by Donald Lotter illustrates this spread in great detail. Fresh products in the US market for an average of 60% higher price than comparable conventional foods, while processed organic products may cost 100–300% more. In the European Union (EU) these price spreads are more in the order of 10–50%, depending on the product, country, and nature of subsidies for production. These subsidies for both the conversion process and for continuing to grow organic products have helped to narrow the differences in prices between the two. However, the support programs differ among countries, although the EU is slowly making these more uniform, and are so complicated that a Swedish farmer was heard to

say that the most important advisor in his organic farming business was the one who could help him fill out all the EU forms for subsidies. In this realm, both organic and conventional farmers are in the same situation, with a high dependence on external subsidies to maintain profitability.

Consumer demand for organic products has been studied in a number of countries. It is generally found that more affluent customers, those with higher levels of education and smaller families, are more apt to buy organic food. Consumers appear to be motivated by two major factors, that organic foods are more healthy for them and that the organic systems that produce them are more environmentally safe. A recent consumer survey in Hedmark County in Norway found that people who regularly purchased organic food were similar in most ways to those who purchased conventional food products, although they were more concerned about conditions under which the food was produced, the importance of local production, and animal welfare on the farm. There is a general concern about GMOs (transgenically modified organisms) among those who practice organic farming and those who purchase the products, and these new crops and performance-enhancing products for livestock are not allowed by any of the organic certification organizations, nor in the new national US legislation on organic labels.

There is great public interest and continuing debate about the safety and the health benefits of organic food products compared with those considered conventional. There is no question about the research, and surveys that have shown less pesticide residue (most often zero) on organic than on conventional produce. Thus organic food is a safer product than the general offering of vegetables and fruits in the marketplace. In terms of food value, there are numerous claims by organic proponents of better taste, freshness, and nutritional quality of organic foods. These qualities are difficult to substantiate, although it is obvious when produce comes from a local farmer and is purchased quickly and directly by the consumer that it will be fresher and probably have a better taste than a comparable product that has traveled 1000 km or 1500 miles (average distance that US food travels from production to consumer). Definitive clinical comparisons of nutritive value generally fail to show differences between organic and conventional food, results that are broadcast by the large food companies to support their claims of no differences and despaired of by organic food enthusiasts who are convinced of the nutritive value of their chosen products. This is still a valuable area for further research.

Future Perspectives for Organic Farming and Foods

If consistent growth in markets over a period of nearly two decades is any indication, there is a large potential for the expansion of organic food production and sale. One of the best indications of the growth potential for this sector is the current interest of larger corporations and multinationals in acquiring smaller organic processing and distributing companies. In the International Federation of Organic Agriculture Movements (IFOAM) Congress held in Victoria, BC, Canada, in August of 2002, there was extensive debate about the entrance of large corporations into the production and marketing of organic foods. The concern in traditional organic circles is that large corporations (both large farms and large processors and marketers) may fulfill the legal requirements of no pesticides or chemical fertilizers, thus producing a 'clean' product, but they are not in step with the goals of the majority of smaller organic producers to create local food systems that will promote the local food economy and establish personal relationships between farmers and producers. This industrialization of organic foods is likely to be the center of long debates in the near future, as people weigh the importance of organic certification versus food produced locally that meets more of the social goals that are part of the original organic agriculture movement.

Other questions embedded in the organic farming debate include the value and desirability of subsidies for organic production, whether price premiums will disappear if organic foods are widely grown, and the potentials for organic agriculture to feed a growing world population. Proponents of the completely free market argue that consumers will sort out food values in the marketplace, and thus there is no need for subsidies. Opponents argue that there are additional costs and limited research to support their conversion to organic farming, and financial incentives are still needed for farmers to begin organic farming to meet the obvious demands. If there is widespread adoption of organic farming techniques and there are many more organic products on the market, it is likely that price differentials will be reduced, but it is unlikely that they will disappear completely. On the issue of global food production, the best organic farmers today produce yields equal to those on conventional farms, although average production is 10–15% lower on organic farms. Part of this is due to lack of research and adequate extension recommendations in organic systems, and part is due to a lower intensity of resource use in organic farming

where there is dependence on internal resources and limited import of fossil fuel-derived materials from outside the farm. An organic farming strategy can be profitable if the production costs are lower and if price premiums are available. There is general agreement that the major problem contributing to hunger on a global scale is inadequate distribution of benefits from today's food system, i.e., that poverty is more to blame than lack of production. This will not be solved by producing more, argue some supporters of organic farming; and we can save the environment by reducing applications of pesticides and fertilizers that reduce water quality and quality of life for many rural residents.

The unique model of Cuba illustrates what can be done when fossil fuel supplies are severely limited in agriculture, and an entire country embraces organic farming and gardening as a major strategy for local food production. The majority of food consumed in major cities in Cuba – fruits, vegetables, small animals and their products – is produced within or adjacent to the city limits. Most are marketed directly from farmers to consumers. In 2002 Cuba enjoyed a healthy diet based on this food system, after some difficult years of adjustment or 'conversion' in their entire agricultural system. This is one model embedded in a unique political system, including diversity in land ownership, wide participation in food production, and a reliance on primarily internal resources to produce an adequate diet for the entire population. In conclusion, this is a model that should be studied as a specific case where organic farming is highly successful.

See also: **Crop Rotations**

Further Reading

Altieri MA, Nicholls CI, and Itiel JI (1994) *Biodiversity and Pest Management in Agroecosystems*. Binghampton, NY: Food Product Press/Haworth Press.

Bender J (1994) *Future Harvest: Pesticide-Free Farming*. Lincoln, NE: University of Nebraska Press.

Carson R (1962) *Silent Spring*. New York: Houghton Mifflin.

Daily GC (ed.) (1997) *Nature's Services: Societal Dependence on Natural Ecosystems*. Washington, DC: Island Press.

Edwards CA, Lal R, Madden P, Miller RH, and House G (eds) (1990) *Sustainable Agricultural Systems*. Ankeney, IA: Soil & Water Conservation Society.

Francis CA, Flora CB, and King LD (eds) (1990) *Sustainable Agriculture in Temperate Zones*. New York: John Wiley.

Funes F, Garcia L, Bourque M, Perez N, and Rosset P (eds) (2002) *Sustainable Agriculture and Resistance*. San Francisco, CA: Food First Books.

Howard A (1943) *An Agricultural Testament*. London, UK: Oxford University Press.

King FH (1911) *Farmers of Forty Centuries, or Permanent Agriculture in China, Korea, and Japan*. Emmaus, PA: Rodale Press.

Lampkin N (1990) *Organic Farming*. Ipswich, UK: Farming Press.

Lampkin N and Measures M (eds) (2001) *2001 Farm Management Handbook*, 4th edn. Organic Farming Unit, Welsh Institute of Rural Studies. Aberystwyth, UK: University of Wales.

Lampkin N and Padel S (1996) *Economics of Organic Farming: An International Perspective*. Wallingford, UK: CAB International.

Lotter D (2002) Organic farming. *Journal of Sustainable Agriculture* 21(4): 59–128.

Rodale JI (1995) *Organic Gardening: How to Grow Healthy Vegetables, Fruits, and Flowers Using Nature's Own Methods*. Garden City, NJ: Hanover House.

Rodale R (1983) Breaking new ground: the search for a sustainable agriculture. *The Futurist* 1(1): 15–20.

Sooby J (2003) *State of the States: Organic Systems Research at Land Grant Institutions, 2001–2003*. Santa Cruz, CA: Organic Farming Research Foundation.

Soule JD and Piper JD (1993) *Farming in Nature's Image*. Covelo, CA: Island Press.

Steiner R (1958) *Agriculture: A Course in Eight Lectures*. London, UK: Biodynamic Agriculture Association.

ORGANIC MATTER

Contents
Principles and Processes
Genesis and Formation
Interactions with Metals

Principles and Processes

M Schnitzer, Agriculture and Agri-Food Canada, Ottawa, ON, Canada

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

The presence of organic matter (OM) distinguishes the soil from a mass of rock particles and allows it to become a living system. Soil organic matter (SOM) serves as a soil conditioner, nutrient reservoir, substrate for microbial activity, preserver of the environment, and major determinant for sustaining and increasing agricultural productivity. The OM content of soils ranges from less than 1% in desert soils to close to 100% in organic soils. A typical agricultural soil may contain between 1 and 5% OM in the top 15 cm.

The main emphasis of this article is on the analytical characteristics of SOM fractions and on studies on their chemical structure. Based on the results of extensive researches, a two-dimensional humic acid (HA) model structure was formulated which has been subsequently converted, with the aid of computational chemistry, to a three-dimensional HA model structure.

The term 'SOM,' as used in this article, refers to the sum total of all organic carbon-containing substances in soils. The major components of SOM are humic substances, especially HA; other significant components are carbohydrates, proteinaceous materials, and lipids.

Extraction and Fractionation of Humic Substances

Extraction of Humic Substances from Soils

Because, in soils, humic substances and inorganic soil constituents are closely associated, it is necessary to separate the two before either can be examined in greater detail. The most efficient extractants for humic substances are either dilute base (0.1–0.5 mol l^{-1} NaOH solution) or dilute neutral salt solutions such as aqueous 0.1 mol l^{-1} $Na_4P_2O_7$.

Fractionation of Extracted Humic Substances

Most soil OM chemists partition the humic substances extract into the following three fractions: (1) HA, which is that fraction of the extract that coagulates when the extract is acidified to pH 2; (2) fulvic acid (FA), which is that fraction of the extract that remains in solution when the alkaline extract is acidified, that is, it is soluble in both alkali and acid; and (3) humin, which is that humic fraction that remains with inorganic soil constituents, that is, it is insoluble in both alkali and acid.

A number of objections have been raised to the use of alkaline extractants. The latter may extract silica from minerals, protoplasmic components from organic tissues, and bring about the auto-oxidation of some organic components, and the condensation of amino groups of amino acids with carbonyl groups of reducing sugars to form Maillard-type reaction products. Some of these changes can be minimized by doing the extractions in atmospheres of inert gases, but not all possible changes can be excluded.

In Situ Analysis of SOM

Cross-polarization magic angle spinning ^{13}C nuclear magnetic resonance (CP-MAS ^{13}C NMR) and pyrolysis-field ionization mass spectrometry (Py-FIMS) are used for the *in situ* analysis of OM in soils. These methods identify the major chemical components of SOM without extraction and fractionation and yield valuable information on the main chemical structures of these materials. The major chemical components of SOM that have so far been identified include carbohydrates, phenols, lignin monomers and dimers, saturated and unsaturated fatty acids, alkanes and alkenes, alkyl monodi- and triesters, alkyl benzenes, methyl naphthalenes, methyl phenanthrenes, amino acids, and heterocyclic N-compounds. While the *in situ* analysis of OM in whole soils by the two methods referred to above does away with the extraction of OM from soils and its subsequent fractionation into HA, FA, and humin, it defines instead SOM in terms of its chemical makeup. In contrast to earlier workers, who thought

that HA and FA were chemically pure substances that were well separated by the extraction procedure, we know now from chemical, ^{13}C NMR, and mass spectrometric analyses that the assumptions of workers prior to the early years of the twentieth century are incorrect. HA, FA, as well as humin, are not distinct chemical substances. Each of these fractions consists of hundreds of compounds, which appear to be associated at molecular levels by mechanisms not yet well understood.

If soil chemists want to retain the extraction and fractionation approach, they need to define each fraction not on the basis of color and solubility or insolubility in base or acid, but by its ^{13}C NMR spectrum, its infrared (IR) spectrum, its mass spectrum, its elemental and functional group analyses, and so on.

Analytical Characteristics of HAs and FAs

Elemental Composition and Functional Group Content

The elemental composition and oxygen-containing functional group analysis of a typical HA (extracted from the Ah horizon of a Haploboroll) and a FA (extracted from the Bh horizon of a Spodosol) are shown in Table 1.

A more detailed analysis shows that: (1) the HA contains approximately 10% more C, but 36% less O than the FA; (2) there are quantitatively smaller differences between the two materials in H, N, and S contents; (3) the total acidity and COOH content of the FA are significantly higher than those of the HA; (4) both materials contain per unit weight significant concentrations of phenolic OH, alcoholic OH, and ketonic and quinonoid C=O groups; (5) the HA is richer than the FA in quinonoid C=O groups, while both materials contain relatively few OCH_3 groups; (6) about 74% of the total O in the HA is accounted for in functional groups, but all of the O in the FA is similarly distributed; and (7) the E_4/E_6 ratio (ratio of absorbances at 465 and 665 mm) of the FA is almost twice as high as that of the FA, indicating that the FA has lower molecular weight than the HA.

Table 1 Analytical characteristics of a Haploboroll humic acid and a Spodosol fulvic acid

	HA	*FA*
Element (g kg^{-1})		
C	564	509
H	55	33
N	41	7
S	11	3
O	329	448
Functional groups (cmol kg^{-1})		
Total acidity	660	1240
COOH	450	910
Phenolic OH	210	330
Alcoholic OH	280	360
Quinonoid C=O	250	60
Ketonic C=O	190	250
OCH_3	30	10
E_4/E_6	4.3	7.1

HA, humic acid; FA, fulvic acid.

Reproduced with permission from Schnitzer M (2000) A lifetime perspective on the chemistry of soil organic matter. *Advances in Agronomy* 68: 1–59, © Academic Press, San Diego.

Infrared and Fourier Transform Infrared Spectrophotometry

IR and Fourier transform IR (FTIR) spectra of HAs and FAs show bands at 3400 cm^{-1} (H-bonded OH) 2900 cm^{-1} (aliphatic stretch), 1725 cm^{-1}(C=O of COOH, C=O stretch of ketonic C=O), 1630 cm^{-1} (COO^-, C=O of carbonyl and quinone), 1450 cm^{-1} (aliphatic C–H), 1400 cm^{-1} (COO^-), 1200 cm^{-1} (C–O stretch or OH deformation of COOH), and 1050 cm^{-1} (Si–O of silicates). The bands are usually broad because of extensive overlapping of individual absorbances. IR and FTIR spectra of HAs and FAs reflect the preponderance of oxygen-containing functional groups, i.e., COOH, OH, and C=O in these materials. Some FTIR spectra show the presence of C=C and polysaccharides. While IR and FTIR spectra provide useful information on the functional groups and their participation in metal–humic as well as pesticide–humic interactions, they tell us little about the chemical structure of these materials.

Chemical Structure of SOM

^{13}C NMR Spectroscopy

One of the most important methods currently available for elucidating the chemical structure of organic substances is ^{13}C NMR. For the analysis of SOM and fractions derived from it such as HA, FA, and humin as well for the direct *in situ* analysis of OM in whole soils, CP-MAS or solid-state ^{13}C NMR is especially useful. CP-MAS ^{13}C NMR spectra of a Mollisol HA and a Spodosol FA are shown in Figure 1. Both the HA and FA show distinct resonances in the aliphatic (0–105 ppm), aromatic (106–150 ppm), phenolic (155–169 ppm), and carboxyl (170–190 ppm) regions. The signals in the HA spectrum at 17, 21, 25, 27, and 31 ppm are probably due to alkyl C. The resonance at 17 ppm is characteristic of terminal CH_3 groups and that at 31 ppm of $(CH_2)_n$ in straight paraffinic chains. The peak at 40 ppm could include

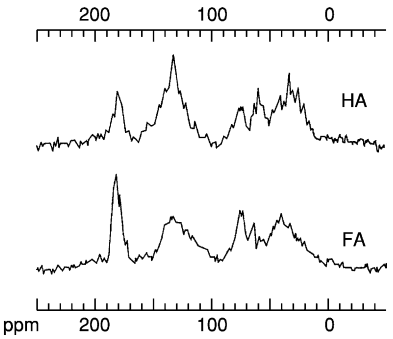


Figure 1 Cross-polarization magic angle spinning ^{13}C nuclear magnetic resonance (CP-MAS ^{13}C NMR) spectra of HA (extracted from the Ah horizon of a Haploboroll and FA (extracted from the Bh horizon of a Spodosol). (Reproduced with permission from Schnitzer M (2000) A lifetime perspective on the chemistry of soil organic matter. *Advances in Agronomy* 68: 1–58, © Academic Press, San Diego.)

Table 2 Distribution of C (%) in a Haploboroll HA and a Spodosol FA as determined by ^{13}C NMR

	% of C	
Chemical shift range (ppm)	*HA*	*FA*
0–40	24.0	15.6
41–60	12.5	12.8
61–105	13.5	19.3
106–150	35.0	30.3
151–170	4.5	3.7
171–190	10.5	18.3
Aliphatic C (0–105 ppm)	50.0	47.7
Aromatic C (106–150 ppm)	35.0	30.3
Phenolic C (151–170 ppm)	4.5	3.7
Aromaticity[a]	44.1	41.6

[a]((Aromatic C + phenolic C)/(aromatic C + phenolic C + aliphatic C)) × 100. Reproduced with permission from Schnitzer M (2000) A lifetime perspective on the chemistry of soil organic matter. *Advances in Agronomy* 68: 1–58, © Academic Press, San Diego.

contributions from both alkyl C and amino acid C. The broad signal at 53 ppm and the sharper one at 59 ppm appear to be due to OCH_3. Amino acid C may also contribute to this region. Carbohydrates in the HA would be expected to produce signals in the 60–65, 70–80, and 90–105 ppm regions; also other types of aliphatic C bonded to O could do so. The aromatic region contains a relatively sharp peak at 130 ppm due to alkyl aromatics. The signal at 155 ppm shows the presence of O- and N-substituted aromatic C (phenolic OH and/or NH_2 bonded to an aromatic C). The broad signal near 180 ppm is due to C in COOH groups, although amide C and C in esters could also contribute to this resonance.

The CP-MAS ^{13}C NMR spectrum of the FA (Figure 1) consists of a number of aliphatic resonances in the 20- to 50-ppm region, followed by signals due to C in OCH_3 groups, and C in amino acids. Signals due to carbohydrate C are present between 50 and 85 ppm. Broad signals between 130 and 133 ppm indicate the presence of C in alkyl aromatics. The strong signal between 170 and 180 ppm shows the presence of C in COOH groups. In general, fewer sharp signals are observed in the spectrum of the FA than that of the HA, possibly because of more H-bonding in the FA.

The CP-MAS ^{13}C NMR data for the HA and FA are summarized in Table 2 in terms of the distribution of C in the different spectral regions. An examination of the data shows similar C distributions in the two humic materials. The HA is slightly more aromatic than the FA, but the FA is richer in COOH groups, which appears to be the main difference between the two substances. Other differences are that the HA is richer in paraffinic C but poorer in carbohydrate C than the FA. However, on the whole, the main structural features, as well as the aromaticity and aliphaticity, are similar so that the HA and FA have similar chemical structures.

As far as the CP-MAS ^{13}C NMR spectrum of humin, that portion of SOM that stays behind after extraction of the soil with dilute alkali or neutral salt solution, is concerned, the spectrum, after extensive leaching, is very similar to that of HA. This indicates that humin is essentially HA-bound strongly to soil minerals.

Electron-Spin Resonance Spectroscopy

Humic substances contain free radicals that can participate in a wide variety of organic–organic and organic–inorganic interactions. The electron-spin resonance (ESR) spectrum of a typical HA consists of a single line devoid of hyperfine splitting. From the magnitudes of the *g*-values (the spectroscopic splitting constant), the prominent free radicals in humic materials appear to be semiquinones or substituted semiquinones. The latter can be produced either by the reduction of quinones or the oxidation of phenols. Except for indicating the presence of phenols, semiquinones, and quinones in these materials, ESR spectroscopy has so far contributed little to our understanding of the structural chemistry of these substances. The main reason for this is that it has been difficult to split the signal.

It is also known that there are two types of free radicals in humic materials: (1) permanent free radicals with long lifetimes, and (2) transient free radicals with relatively short lives (several hours). Transient free radicals in humic substances can be generated in relatively high concentrations by chemical reduction, irradiation, or increases in pH.

Oxidative Degradation

The oxidative degradation of HAs, FAs, humins, and whole soils under a wide variety of experimental conditions produces aliphatic carboxylic, phenolic, and benzene carboxylic acids. Among aliphatic oxidation products are mono-, di-, tri-, and tetracarboxylic acids. Major phenolic acids include aromatic rings substituted by one to three OH groups and by one to five COOH groups, while major benzene carboxylic acids are the di-, tri-, tetra-, penta-, and hexa-forms. From the oxidation products identified and from CP-MAS ^{13}C NMR spectra of humic substances, it appears that the aromatic rings are cross-linked by paraffinic chains. Oxidation converts CH_3 and CH_2 groups closest to the aromatic rings to COOH groups, whereas CH_2 groups in alkyl chains are oxidized to mono- and dicarboxylic acids and/or to CO_2. The formation of significant concentrations of CO_2 from the oxidation of side chains may explain the low oxidation yields of aliphatic acids compared with phenolic and benzene carboxylic acids.

The model structure shown in Figure 2 has an aromaticity of 50% if functional groups are excluded. Of special interest are the following: (1) isolated aromatic rings are important structural components of all humic substances; (2) aliphatic chains link aromatic rings to form an alkyl aromatic network; and (3) the structure contains voids of various dimensions that can trap organic and inorganic soil constituents. These characteristics are typical of soil humic substances.

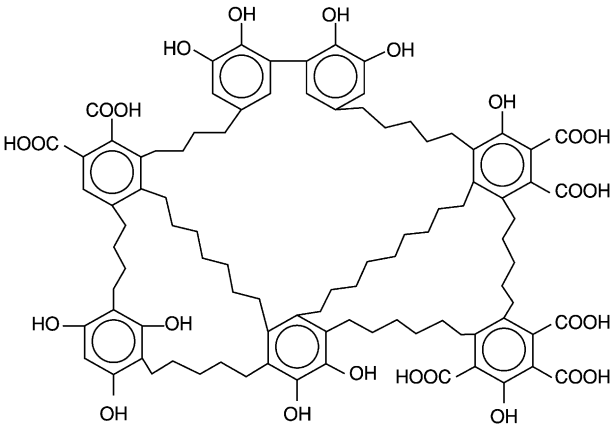


Figure 2 Chemical structure of humic substances based on oxidation products. (Reproduced with permission from Schnitzer M (2000) A lifetime perspective on the chemistry of soil organic matter. *Advances in Agronomy* 68: 1–58, © Academic Press, San Diego.)

Reductive Degradation

Relatively mild reduction of humic substances with Na amalgam produces phenols and phenolic acids. But the harsher Zn-dust distillation and Zn-dust fusion yield methyl-substituted naphthalene, anthracene, phenanthrene, pyrene, and perylene. The methyl groups on the polycyclics are probably the remains of longer alkyl chains linking the polycyclics in HA, FA, and humin structures.

Pyrolysis-Field Ionization Mass Spectrometry of HAs, FAs, and Humins

Compounds identified in the Py-FI mass spectra of a whole soil, an HA, an FA, and a humin are summarized in Table 3. The most abundant compounds identified in all three fractions are carbohydrates, phenols, lignin monomers, lignin dimers, *n*-fatty acids, *n*-alkyl diesters, and *n*-alkyl benzenes. Minor components include *n*-alkyl monoesters, methylnaphthalenes, methylphenanthrenes, and N-compounds. HA tends to be enriched in *n*-fatty acids and the whole soil and the humin in *n*-alkyl diesters and *n*-alkyl benzenes.

Curie-Point Pyrolysis–Gas Chromatography/Mass Spectrometry of HAs and FAs

Curie-point pyrolysis–gas chromatography/mass spectrometry (Cp Py-GC/MS) of HAs and FAs indicates the presence of relatively high concentrations of aromatic rings substituted by alkyl chains. Of special significance is the identification of a series of C_1–C_{22} *n*-alkyl benzenes. In addition, ethylmethyl benzene, methylpropyl benzene, methylheptyl benzene, methyloctyl benzene, and methylundecyl benzene are also detected. Other compounds identified

Table 3 Compounds identified by Py-GC/MS in the initial Armadale soil and in HA, FA, and humin fraction isolated from it

Compound identified	*Soil*	*HA*	*FA*	*Humin*
Carbohydrates	++[a]	++	++	++
Lingin monomers	++	++	++	++
Lingin dimers	++	++	++	++
n-Fatty acids	+	+++	+	++
n-Alkyl monoesters	+	+	+	+
n-Alkyl diesters	++	+	+	++
n-Alkyl benzenes	+		+	++
Methylnaphthalenes	+			+
Methylphenanthrenes	+			+
N compounds	+	+		+
n-Alkanes	+			

[a]+, weak (relative intensity <20%); ++, intense (relative intensity 20–60%); +++, very intense (relative intensity >60%).
Reproduced with permission from Schnitzer M and Schulten H-R (1995) Analysis of organic matter in soil extracts and whole soils by pyrolysis-mass spectrometry. *Advances in Agronomy* 55: 168–217, © Academic Press, San Diego.

are trimethyl- and tetramethyl benzenes, alkyl naphthalenes, and alkyl phenanthrenes. The alkyl substitution on the naphthalenes ranges from one to five methyls, whereas that on the phenanthrene rings ranges from one to four methyls.

Other Characteristics

From surface pressure, surface tension, and viscosity measurements at different pHs and varying concentrations of HA and neutral electrolytes, it appears that, under conditions normally prevailing in most agricultural soils, HA behaves like a flexible, linear polyelectrolyte. Additional support for this view comes from transmission electron micrographs of HA which show that in dilute aqueous solution HA forms flat, elongated, multibranched filaments, 20–100 nm in width. With increasing pH or HA concentration, a finely woven network of elongated fibers is formed which then coalesces into a sheet-like structure, perforated by voids of varying dimensions. The overall impression is that of a relatively flexible and open structure.

Model Structures of HA

Two- and Three-Dimensional Model Structures of HA

The two-dimensional HA model structure of HA shown in Figure 3 is based on long-term chemical, ^{13}C NMR spectroscopic, degradative, electron-microscopic, X-ray, Py-FIMS, and Cp Py–GC/MS studies. In this structure, *n*-alkyl aromatics play a significant role. Oxygen is present in the form of COOH, phenolic and alcoholic OH, ester, ethers and ketones, whereas nitrogen occurs in nitriles and hetercyclic structures. The carbon skeleton contains voids of various dimensions which can trap and bind other organic and inorganic soil constituents as well as water. The molecular formula of the HA is $C_{308}H_{328}C_{90}N_5$, its molecular weight is 5540 Da, and its elemental analysis is 66.8% C, 6.0% H, 26.0% O, and 1.3% N. It is assumed that carbohydrates and proteinaceous materials are adsorbed on external HA surfaces and in internal voids and that hydrogen bonds play an important role in their

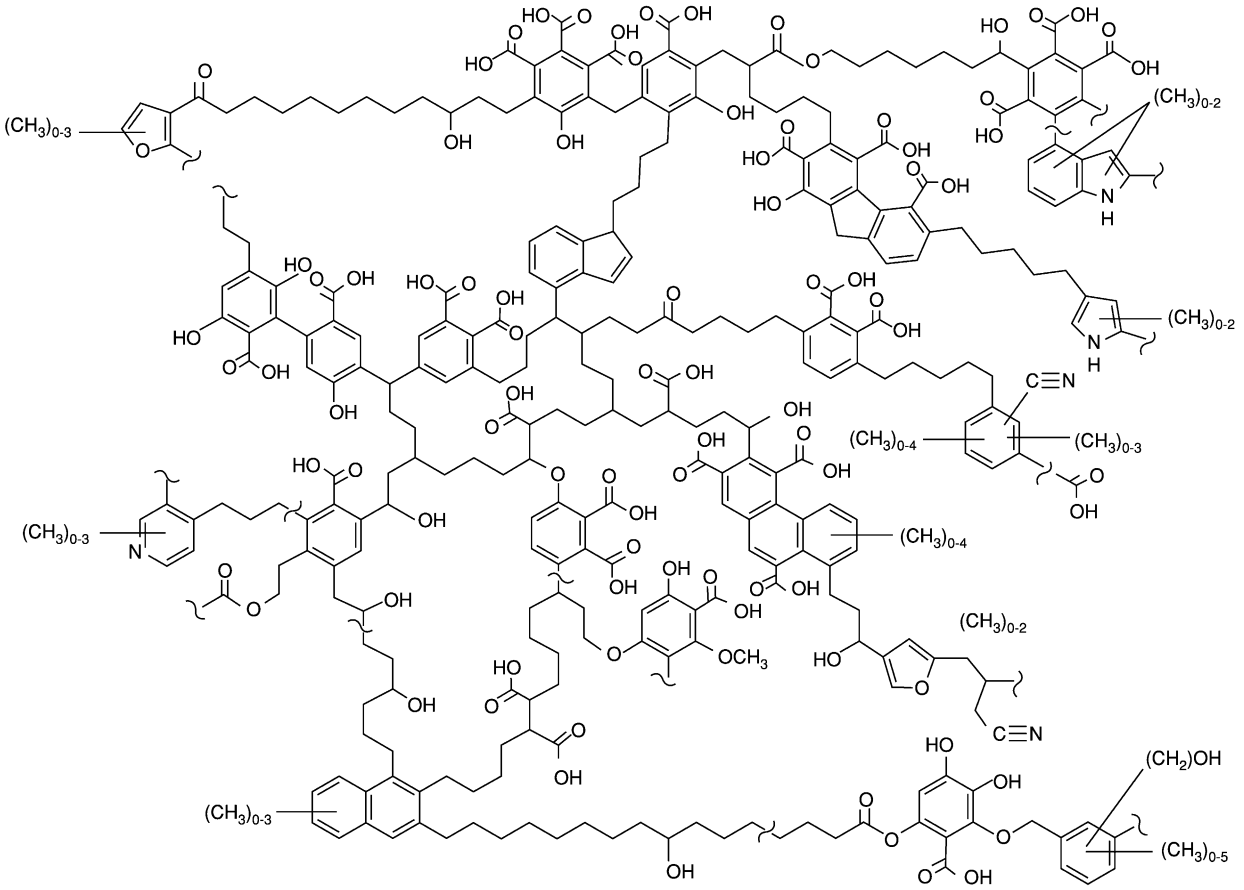


Figure 3 Two-dimensional HA model structure. (Reproduced with permission from Schulten H-R and Schnitzer M (1993) A state of the art structural concept for humic substances. *Naturwissenschaften* 80: 29–30, Fig. 1, © Springer-Verlag, Heidelberg.)

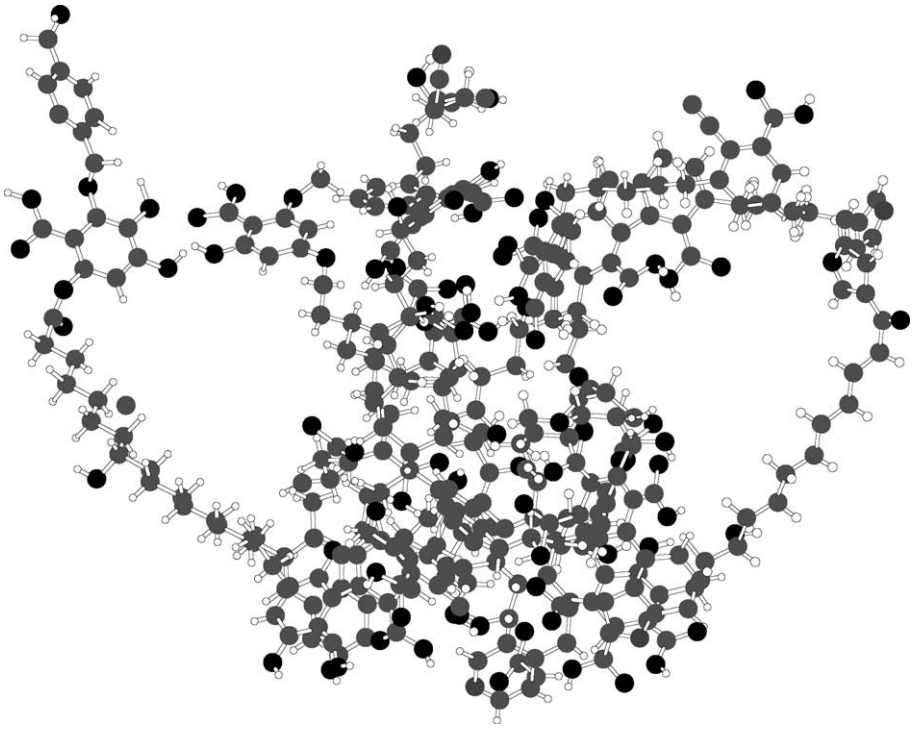


Figure 4 Three-dimensional humic acid model structure. Element colors are: carbon (light gray), hydrogen (white), oxygen (black), and nitrogen (dark gray). (Reproduced from Schulten H-R and Schnitzer M (1997) Chemical model structures for soil organic matter and whole soils. *Soil Science* 162: 115–130, © Lippincott, Williams & Wilkins, Baltimore.)

immobilization. Aside from carbohydrates and proteinaceous materials, the voids can also trap and bind lipids and biocides, as well as inorganics such as clay minerals and hydrous oxides.

The conversion, with the aid of HyperChem software, of the two-dimensional HA structure (**Figure 3**) to a three-dimensional HA structure is shown in **Figure 4**. The three-dimensional HA model structure has a molecular weight of 5547 and an elemental analysis of 66.7% C, 6.1% H, 26.0% O, and 1.3% N.

It is likely that applications of computational chemistry to the development of SOM structural models will increase in the future and lead to a better understanding of the spatial arrangements of the molecular constituents of SOM. A more comprehensive knowledge of the chemistry and reactions of SOM will certainly be beneficial to a sustainable agriculture and help to protect the environment.

Reactions of SOM with Metals and Organics

Formation of Water-Soluble Metal–Organic Complexes

Reactions in water near pH 7 between di- and trivalent metal ions and HAs and FAs are likely to proceed via one or more of the mechanism shown in **Figure 5**, taking a divalent metal ion M^{2+} as an example. According to eqn [1] in **Figure 5**, one COOH group on the HA or FA reacts with one metal ion to form an organic salt or monodentate complex. Equation [2] in **Figure 5** describes a reaction in which one COOH and one adjacent OH group react simultaneously with the metal ion to form a bidentate complex or chelate. According to eqn [3] in **Figure 5**, two adjacent COOH groups interact simultaneously with the metal ion to form a bidentate chelate. Equation [4] shows the metal ion M^{2+} linked to the HA or FA not only by electrostatic bonding, but also through a water molecule (water bridge) in its primary hydration shell to a C=O group of the ligand. The latter type of interaction is especially important when the cation has a high solvation energy and so retains its primary hydration shell. Equations [2] and [3] in **Figure 5** describe the formation of strong, inner-sphere complexes, whereas eqn [4] refers to the formation of a weaker, outer-sphere complex. While a number of workers have published stability constants of water-soluble metal–HA and metal–FA complexes, problems have been encountered with the analysis and interpretation of the data. Probably the most serious obstacle to progress in this area is our lack of knowledge of the chemical structures of HAs and FAs, the ligands.

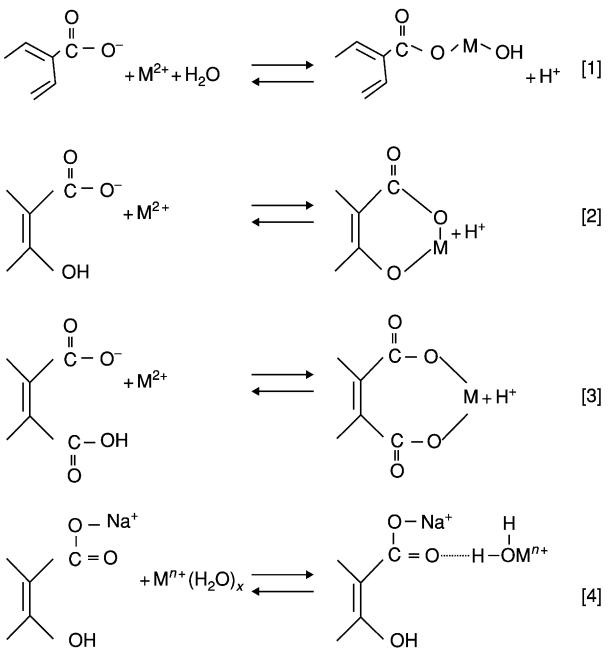


Figure 5 Major metal–HA and –FA reaction mechanisms. (Reproduced with permission from Schnitzer M (1986) Binding of humic substances by soil colloids. In: Huang PM and Schnitzer M (eds) *Interactions of Soil Minerals with Natural Organics and Microbes*, pp. 77–101. Special Publication No. 17. Madison, WI: Soil Science Society of America.)

Other Types of Metal SOM Interactions

The formation of mixed-ligand complexes of the type M^{2+}–FA – secondary ligand (Y) is known to occur in soils. Secondary ligands (Y) can be citrate tartrate, salicylate, phosphate, bicarbonate, sulfate, aspartate, glycinate, etc. In neutral to weakly acid solutions, these complexes are more stable than simple complexes. The formation of mixed ligand complexes prevents the precipitation of metal ions by hydrolysis at elevated pHs.

Due to their ability to complex di-, tri-, and tetravalent metal ions, dilute aqueous FA solutions and aqueous HA solutions at pH >6.5 can attack and degrade minerals to form water-soluble and water-insoluble metal complexes. Thus, the weathering of minerals in soils and sediments is often enhanced by the action of naturally occurring organic substances, especially FA. Because of its abundance in soils, its solubility in water, and its ability to complex with metal ions and interact with silica, the latter may increase the concentrations of these soil constituents in aqueous solutions to levels that far exceed their normal solubilities. In this manner, aqueous FA solutions may not only bring about the dissolution and degradation of existing minerals, but also lead to the synthesis of new minerals by permitting the complexed and dissolved metals and silica to form new combinations. Conversely, active surfaces of inorganic soil constituents may catalyze either the degradation or synthesis of HAs and FAs.

The extent of adsorption of HA and FA on mineral surfaces depends on the characteristics of the surface, the pH of the system, and its water content. For example, sepiolite, which has a channel-like surface, readily adsorbs FA. In untreated sepiolite, these channels are filled by bound and/or zeolitic water, which can be displaced by undissociated FA in aqueous solution at pH 3.

The interlayer adsorption of FA by expanding clay minerals is pH-dependent, being highest at low pH but no longer occurring at pH >5. The FA cannot be displaced from the clay interlayers by leaching with $1\,\mathrm{mol\,l^{-1}}$ NaCl; an inflection is observed in the adsorption–pH curve near the pH corresponding to the p*K* of the acid species of FA so that the adsorption can be classified as a ligand-exchange reaction. In this type of reaction, the anion (FA) is thought to penetrate the coordination shell of the dominant cation in the clay and displace water coordinated to the dominant cation in the clay interlayer.

Nitrogen in Soils and SOM

Almost 95% of total soil N is associated with SOM. This N occurs in the forms of amino acids, peptides, proteins, amino sugars, heterocyclic N compounds, and ammonia. The amino acid composition of soils and SOM from widely differing origins is remarkably similar and resembles that of bacteria. This appears to indicate that microbes play a major role in the synthesis of amino acids, peptides, and proteins in soils. The most prominent amino sugars in soils and SOM are D-glucosamine and D-galactosamine. Present in smaller amounts are muramic acid, D-mannosamine, *N*-acetylglucosamine, and D-fucosamine.

By Py-FIMS and Cp Py-GC/MS, more than 100 heterocyclic N compounds have been identified in soils and SOM. These N compounds include nonsubstituted and substituted pyrroles, pyrrolidines, imidazoles, pyrazoles, pyridines, pyrazines, nitriles, indoles, quinolines, isoquinolines, benzothiazoles, and pyrimidines. Low-mass N-compounds detected are hydrocyanic acid, dinitrogen, dinitrogen monooxide, isocyanomethane, acetamide, and hydrazoic acid. N-derivatives of benzene detected include benzamine, benzonitrile, and isocyanomethylbenzene. According to the latest estimates, the distribution of total N in soils and SOM is as follows: proteinaceous materials (amino acids, peptides, proteins), 40%; amino sugars, 5–6%; heterocyclic N (including purines and pyrimidines), 35%; and NH_3-N, 19%. Thus, proteinaceous materials and N heterocyclics are the major N components. While there are indications that the

proteinaceous materials and amino sugar are strongly adsorbed on the HA and SOM surfaces and in the voids, the heterocyclic N components appear to be chemically bonded to HA and SOM components; that is, they are integral structural components of HAs and SOM.

Phosphorus and Sulfur in Soils and SOM

According to current estimates, about 75% of the total P in soils occurs in organically bound forms. Principal organic P forms include inositol phosphates, phospholipids, nucleic acids, phosphoproteins, and metabolic phosphates. Other organic P compounds identified by ^{31}P NMR in soils and SOM are alkyl phosphonic esters, alkyl phosphonic acids, choline phosphate, and orthophosphate mono- and diesters.

It is estimated that more than 90% of the total soil S in most noncalcareous soils occurs in organic forms. These include phenolic sulfates, sulfated polysaccharides, choline sulfate, and sulfated lipids. Also present are S-containing amino acids (cystine and methionine), cystathionine, djenkolic acid, taurine, biotin, and thiamine. In poorly drained soils, the decomposition of organic S compounds produces volatile compounds such as carbon disulfide, carbonyl sulfide, methyl mercaptan, diethyl sulfide, dimethyl sulfide, and dimethyl disulfide.

Interactions of SOM with Pesticides and Herbicides

The persistence, degradation, bioavailability, leachability, and volatility of pesticides and herbicides in soils are directly related to the concentration of SOM. The pesticides and herbicides may sorb on SOM and be retained by van der Waals forces, hydrophobic bonding, hydrogen bonding, charge transfer, ion exchange, and ligand exchange. Pesticides and herbicides may also be adsorbed unchanged on external SOM surfaces or in internal voids. SOM can promote the nonbiological degradation of pesticides and herbicides in addition to chemically binding residues arising from their biological and chemical degradation. These processes may play important roles in the detoxification, protection, and preservation of the environment.

List of Technical Nomenclature

Ah	humus-enriched A horizon
Bh	humus-enriched B horizon
^{13}C	carbon-13
C=O	carbonyl group
COOH	carboxyl group
CO_2	carbon dioxide
CP-MAS ^{13}C NMR	cross-polarization magic angle spinning carbon-13 nuclear magnetic resonance spectroscopy
CP Py-GC/MS	Curie-point pyrolysis–gas chromatography/mass spectrometry
cmol kg^{-1}	centimoles per kilogram
E_4/E_6	ratio of absorbances at 465 nm and 665 nm
ESR	electron-spin resonance spectroscopy
FA	fulvic acid
FTIR	Fourier-transform infrared spectrophotometry
g kg^{-1}	grams per kilogram
HA	humic acid
N	nitrogen
NaOH	sodium hydroxide
$Na_4P_2O_7$	sodium pyrophosphate
OCH_3	methoxyl group
OH	hydroxyl group
OM	organic matter
P	phosphorus
Py-FIMS	pyrolysis-field ionization mass spectrometry
ppm	parts per million
S	sulfur
SOM	soil organic matter
TII	total ion intensity
Zn	zinc

See also: **Carbon Cycle in Soils:** Dynamics and Management; **Carbon Emissions and Sequestration**; **Cover Crops**; **Crop Rotations**; **Manure Management**; **Mulches**; **Organic Farming**

Further Reading

Chen Y (1998) Electron microscopy of soil structure and soil components. In: Huang PM, Senesi N, and Buffle J (eds) *Structure and Surface Reactions of Soil Particles*, pp. 155–182. Chichester, UK: John Wiley.

Schnitzer M (1995) Organic–inorganic interactions in soils and their effect on soil quality. In: Huang PM, Berthelin J, Bollag J-M, McGill WB, and Page AL (eds) *Environmental Impacts of Soil Component Interactions*, pp. 3–19. Boca Raton, FL: Lewis Publishers.

Schnitzer M (2000) A lifetime perspective on the chemistry of soil organic matter. *Advances in Agronomy* 68: 1–58.
Schnitzer M (2001) The *in situ* analysis of organic matter in soils. *Canadian Journal of Soil Science* 81: 249–254.
Schnitzer M and Schulten H-R (1995) Analysis of organic matter in soil extracts and whole soils by pyrolysis-mass spectrometry. *Advances in Agronomy* 55: 168–217.
Schulten H-R and Schnitzer M (1997) Chemical model structures for soil organic matter and whole soils. *Soil Science* 162: 115–130.
Schulten H-R and Schnitzer M (1998) The chemistry of soil nitrogen: a review. *Biology and Fertility of Soils* 26: 1–15.
Schulten H-R, Leinweber P, and Schnitzer M (1998) Analytical pyrolysis and computer modelling of humic and soil particles. In: Huang PM, Senesi N, and Buffle J (eds) *Structure and Surface Reactions of Soil Particles*, pp. 281–324. Chichester, UK: John Wiley.
Stevenson FJ (1994) *Humus Chemistry*. New York: John Wiley.
Wilson MA (1987) *NMR Techniques and Applications in Geochemistry and Soil Chemistry*. Oxford, UK: Pergamon Press.

Genesis and Formation

K M Haider, University Halle, Halle, Germany
G Guggenberger, University of Bayreuth, Bayreuth, Germany

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

The primary sources of soil organic matter (SOM: all organic C-containing compounds in the soil) are dead plant materials in the form of leaves, straw, twigs, roots, and other plant litter materials. The global biomass of higher plants amounts to approximately 500–700 Gt C (1 Gt = 10^9 t) and contributes approximately 100 Gt C in litter annually in the soil, either on the surface or deposited belowground as rhizodeposition. About the same amount of carbon is annually released from soil as CO_2. The dead plant material consists of lignocelluloses with an average composition of 15–60% cellulose, 10–30% hemicelluloses, 5–30% lignin, and 2–15% protein. Minor components are phenols, sugars, amino acids, and peptides, as well as numerous secondary metabolites. Most of the compounds are used as nutrient and energy sources for microbial growth.

SOM is a natural product resulting from microbial activity in the inorganic and/or organic soil environment. The amount and accumulation of SOM are controlled by the composition and amounts of the plant residues, by climatic conditions, and soil texture. Other important factors are microbial activity, soil redox conditions, and other soil chemical and physical properties.

Soils may contain several tons of SOM per hectare. Most of it can be only slowly degraded and metabolized by soil organisms as a consequence of both its physical and chemical stabilization. Plant debris, with structural relationship to its origin, may become resistant to microbial degradation by interaction with minerals, whereas its completely humified products are produced by random condensation of refractory plant and microbial products. Their diversity and lack of regular polymeric structures do not favor efficient enzymatic degradation and energy production. Therefore, chemical composition and structure of the humified residues and their physical protection by spatial distribution are equally important for accumulation of soil carbon, which amounts to approximately 1400–1600 Gt C on a global scale.

Combining the concepts of chemical stability and physical protection, it is assumed that there are two categories of natural organic matter: labile (readily degrading) and resistant (slowly degrading). Furthermore, there are three categories of microenvironments: free unprotected particles, particles occluded or entrapped in an inorganic or organic microenvironment, and particles attached or adsorbed to surfaces by physical or chemical mechanisms. This gives rise to six categories of SOM:

1. Free particulate, intrinsically labile;
2. Free particulate, intrinsically resistant;
3. Occluded particulate, potentially labile but currently protected as a result of inaccessibility;
4. Occluded particulate, intrinsically resistant and further protected by its inaccessibility;
5. Adsorbed, potentially labile, but currently protected by adsorption;
6. Adsorbed, intrinsically resistant and further protected by adsorption.

Soil Fabric and its Impact on SOM Stabilization

Transformation and Stabilization Processes of Free and Occluded Particulate Organic Matter

The location of organic matter within the soil structural units has been demonstrated to control SOM dynamics by physical or chemical mechanisms. Crushing of macroaggregates (greater than 250 μm) provides plant-derived material from roots or leaves colonized by microorganisms, which decompose

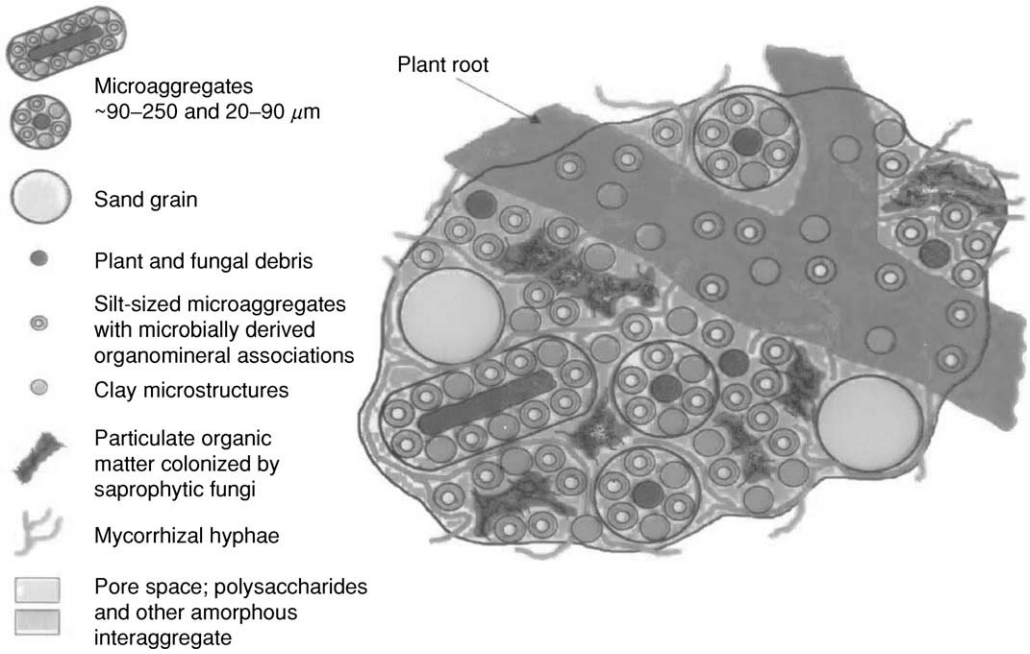


Figure 1 Mineral and organic components within a macroaggregate. (Adapted from Jastrow JD and Miller RM (1998) Soil aggregate stabilization and carbon sequestration: feedbacks through organomineral associations. In: Lal R, Kimble JM, Follett RF, and Stewart BA (eds) *Soil Processes and the Carbon Cycle*, pp. 207–223. Boca Raton, FL: CRC Press.)

them and form humified material, microaggregates (less than 250 μm), and sand grains (Figure 1).

Microscopic and electron-microscopic evaluation gives direct access to the origin and the degree of decomposition of SOM within the different fractions. Mostly large, undecomposed root and plant fragments exist in the 500–2000-μm aggregate fractions; whereas, in the 10–100-μm aggregate fractions, more decomposed materials have been observed.

Proportions of free particulate organic matter vary widely and are affected by land use, vegetation and soil type, climate, and other factors. Its accumulation is favored in cold and dry climates and in continuously vegetated soils with a large input of plant litter, e.g., as in grass or forest vegetation areas. There, free particulate SOM can account for 15–40% of the total SOM in surface horizons, whereas in tilled soils it contributes less than 10% of the total SOM and reflects differences in cropping sequences and tillage. Free particulate SOM also includes organic particles which adhere loosely to surfaces of organomineral complexes. This type of particulate SOM can be recovered by minimal dispersion (e.g., density fractionation) of soil samples in which aggregates remain intact. It consists either of partly decomposed litter particles from recently deposited crop residues or of older, uncomplexed SOM previously occluded in aggregates but released by their destruction. These particles tend to be readily depleted when soils previously under a permanent native vegetation are brought into cultivation.

Plant-derived materials in soil are rapidly colonized by microorganisms. Microorganims and their mucilages are attached to clay particles and form microoaggregates which slow their biodegradation. With time the materials in macroaggregates become more fragile and mainly humified material is adsorbed on mineral surfaces in microaggregates, where it is hardly available to microorganisms and stabilized for a long time (Figure 2).

Many studies on different soil types have shown that macroaggregates contain younger C than microaggregates. Examples from investigations in Oxisols, Alfisols, and Inceptisols, where C3 vegetation was replaced by C4 vegetation, are shown in Table 1. The photosynthetic switch and the corresponding differences in $\delta\,^{13}$C-values from C4 to C3 vegetation are frequently used to calculate average turnover times of aggregate-associated organic matter. The average turnover time for old C4-derived C is approximately 400 years for microaggregates compared with an average turnover time of approximately 140 years for macroaggregates.

Stabilization of SOM by Sequestration in Microporous Structures or on Hydrophobic Surfaces

Physical factors confer biological stability to organic matter in soils through the constraints they place on the opportunities for reactions between substrates and enzymes or decomposer organisms. Microbial

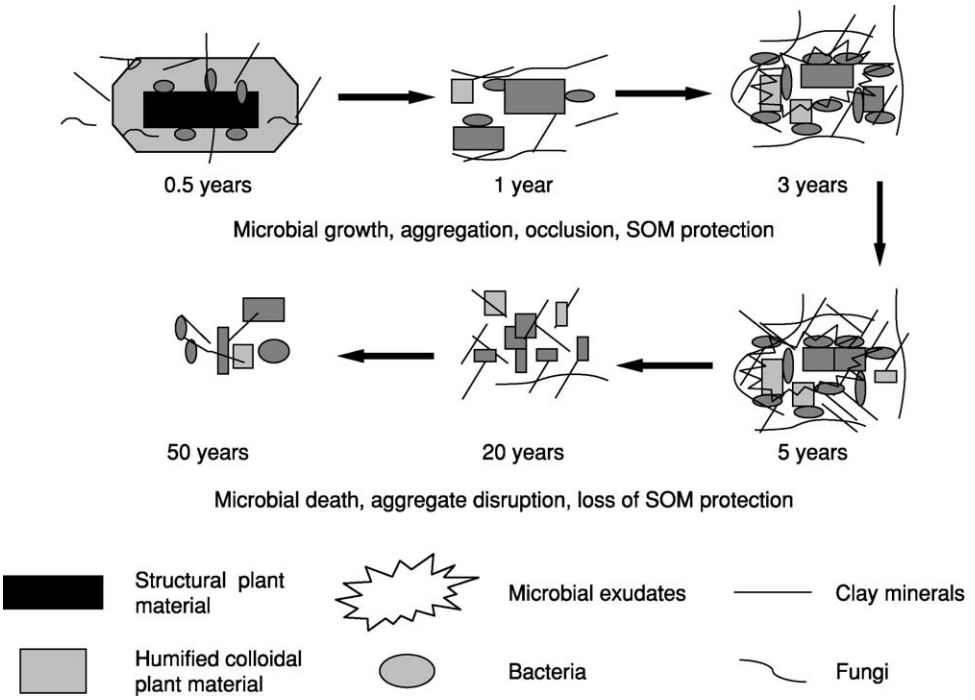


Figure 2 Schematic presentation of the fate of soil organic matter (SOM) and soil aggregates. (Source: Balesdent J, Chenu C, and Balabane M (2000) Relationship of soil organic matter dynamics to physical protection and tillage. *Soil and Tillage Research* 53: 215–230.)

Table 1 Percentage of young C (C4-derived C) in aggregates

Soil type (soil taxonomy)	*Site*	*Experiment*	*Aggregate size (μm)*	*C4-C (%)*
Oxisol[a]	Australia	C4 grass for 35 years	200–2000	49
		on former C3 rainforest	<200	41
		C4 grass for 83 years	200–2000	75
		on former C3 rainforest	<200	67
Alfisol[b]	France	Corn (C4) for 6 years	200–6000	15
		on former C3 arable crops	<200	6
		Corn (C4) for 23 years	200–6000	46
		on former C3 arable crops	<200	38
Inceptisol[b]	Canada (Quebec)	Corn (C4) for 15 years	2000–6000	21
		on former (C3) hay system	1000–2000	18
			500–1000	11
			200–500	8
			50–200	3
			<50	1

[a]Aggregates obtained by hand-grinding.
[b]Water-stable aggregates of slaked soil.
Adapted from Guggenberger G and Haider K (2002) Effect of mineral colloids on biochemical cycling of C, N, P and S in soil. In: Huang PM, Bollag J-M, and Senesi N (eds) *Interactions Between Soil Particles and Microorganisms and Their Impact on the Terrestrial Environment.* IUPAC Series on Analytical and Physical Chemistry of Environmental Systems, vol. 8, pp. 267–322. Chichester: John Wiley, with permission.

access to substrates when located in small pores or sorbed on solid surfaces may be limited, because they are inaccessible to microorganisms and their enzymes, and are thereby protected from rapid decomposition. Carbohydrates, lignin-degradation products, and nitrogenous substances can interact with soil colloids by sorption, polymerization, or entrapment in voids of inorganic or organic soil constituents.

Xenobiotics such as pesticides or polyaromatic hydrocarbons can be sorbed in the microporous structure of SOM or minerals. Similarly to xenobiotics, also natural SOM can be sorbed in micropores. Since such sorption sites are limited, the

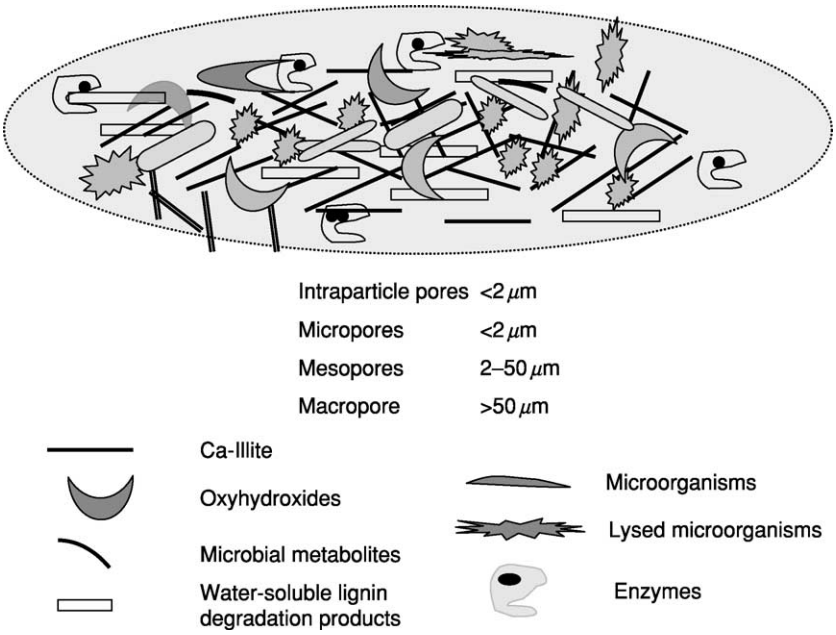


Figure 3 Conceptual model of stabilization of soil organic matter within nanopores.

mineralization of phenols or phenolic acids, of amino acids, and of glucose or other carbohydrates can be more strongly retarded at low concentrations (10 ppm or less) than at higher concentrations. It can be presumed that the observed continuous decline in bioavailability (and extractability) is associated with a slow diffusion of the molecules to more remote sites. Intraparticle and intraorganic diffusion into nanopores is believed to account for these slow phases of sorption and desorption of nonionic molecules. Dissolved organic matter (DOM) can diffuse into nanopores of primary particles or sorb on active surfaces. Within pores less than 8 nm, DOM is inaccessible to exoenzymes. A conceptual model of the stabilization of sorbed SOM within nanopores is shown in Figure 3.

Chemical Information About Organic Matter in Aggregates and Organomineral Complexes

Microscopic evidence suggests that SOM in the coarse sand fraction is mainly composed of little-decomposed plant residues, whereas that associated with fine sand consists of quite well-altered organic debris and fine particles of roots. This is corroborated by chemical analyses. Significantly lower C-to-N (C/N) ratios in fine fractions than in the coarser ones indicate a more pronounced microbial alteration in the fine-size separates (Figure 4a). This is also obvious from the pattern of monosaccharides released upon acid hydrolysis in the fractions. Individual sugar monomers released from SOM upon acid hydrolysis culminate in sand and clay separates and have their minimum in the silt-size fraction. There is a shift from primarily plant-derived monosaccharides in the coarse fractions to a dominating microbial contribution in the fine fractions (Figure 4b). Ratios of (galactose + mannose) to (arabinose + xylose) and of (rhamnose + fucose) to (arabinose + xylose) increase from sand to silt to clay, indicating primarily plant-derived polysaccharides in the sand-size separates and larger fractions of microbial polysaccharides in the clay. Amino acids, in particular diaminopimelic acid, which is confined to the peptidoglycans of cell wall of the prokaryotes, are recovered in elevated yields from clay. Because most of the bacterial biomass has been found in the vicinity of clay, high concentrations of bacterial-derived compounds (diaminopimelic acid, muramic acid) in the clay-size separates may be caused by a high production of microbial-derived substances.

As revealed by CuO-oxidation products, particle-size fractions of soils differ in the contribution of lignin to total organic carbon, but also in the degree of lignin alteration. Highest lignin contents obtained by the CuO-oxidation method are measured in the coarse fractions and decrease continuously with decreasing particle size (Figure 4c). A higher degree of oxidative alteration of lignin in finer-size separates compared with coarser fractions suggests that lignin in finer fractions had undergone more intensive microbial degradation.

Cross-polarization/magic angle spinning (CP/MAS) ^{13}C nuclear magnetic resonance (NMR) spectroscopy

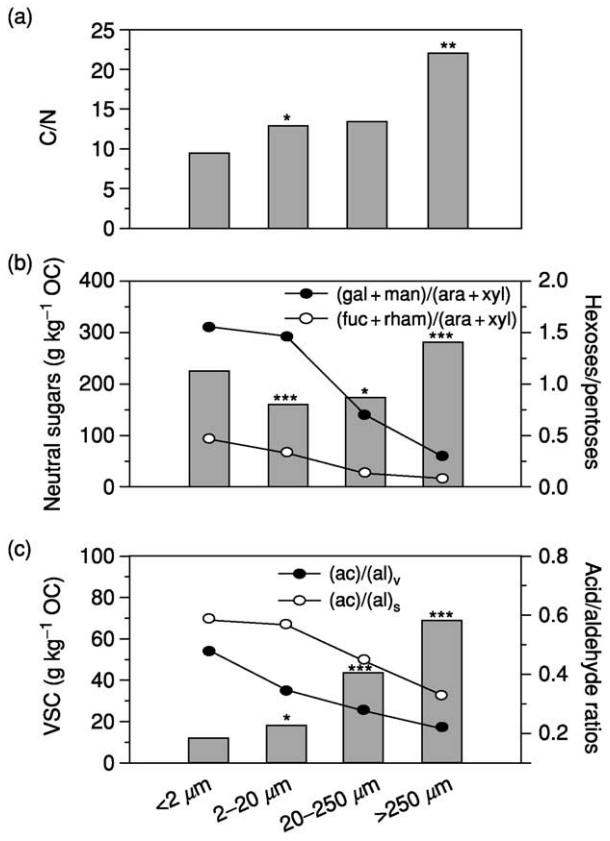


Figure 4 Carbon-to-nitrogen ratio (C/N) in particle-size fractions of native topsoils (0–10 cm) (a); the concentration and the pattern of monosaccharides released upon acid hydrolysis (the pentoses arabinose (ara) and xylose (xyl) are considered primarily plant-derived, while the desoxysugars rhamnose (rham) and fucose (fuc) are considered synthesized mostly by microorganisms, the hexoses mannose (man) and galactose (gal) occur in both types of organisms) (b); and the concentrations of lignin-related phenols and phenolic acid/aldehyde ratios (filled circles, acid/aldehyde ratio vanillyl; open circles, acid/aldehyde ratio syringyl) after CuO oxidation of SOM in different particle sizes in native prairie soil (c). OC, organic carbon. *, **, *** Significant at the 0.05, 0.01 and 0.001 probability. (Adapted from Amelung W, Flach KW, and Zech W (1999) Lignin in particle-size fractions of native grassland as influenced by climate. *Soil Science Society of America Journal* 63: 1222–1228.)

is, at present, probably the most favored tool for detecting individual forms of C. Although this method is problematic with respect to quantification, it provides important insights into the chemical structure of SOM associated with different primary-particle fractions (Figure 5).

In general, O-alkyl resonances, mainly derived from carbohydrates or from ether bonds in degraded lignin predominate across the entire textural range, but they are minimized in the silt fraction. The strongest O-alkyl signals are always observed in the sand fraction, and spectra of the organic matter associated with sand strongly resemble those of fresh plant litter. Aromatic C, and in particular the aromatic C with other C substituents, is maximal in silt but is lowest in the clay fraction(s). Alkyl C steadily increases with decreasing particle size and represents the major or the second major C species in the clay(s). Long-chain aliphatic material derived from natural waxes is a major component of this alkyl C. Furthermore, microorganisms also synthesize long-chain aliphatic compounds.

Using the natural ^{13}C-labeling technique where there is a shift in the photosynthetic pathway from C3-vegetation to C4-vegetation or vice versa (see also Table 1), it was shown that particle-size fractions represent SOM pools with different turnover times. Organic compounds in sand-size fractions are turned over rather rapidly within several years, whereas organic matter associated with silt and clay particles show markedly higher turnover times, thus being involved in mid- and long-term dynamics.

Stabilization of SOM by Humification

As well as stabilization by spatial separation of organic matter particles with structural relationships to plant tissue material at different stages of microbial decomposition, the chemical stabilization of the completely humified products of the particles together with microbial compounds makes an important contribution to SOM. Accumulation of refractory humic compounds can be caused by preferential sorption of refractory fractions of organic matter in the dissolved stage (DOM). It results from leaching of dead plant residues, particularly those of the litter layer, with water or by solubilization of already humified compounds in the soil solution. Only small proportions of DOM can be identified in the form of organic acids, sugars, amino acids, and phenols. But most DOM is complex humified organic matter. The carbohydrate fraction of DOM contains a higher proportion of hexose and deoxysugars, which are typical for a microbial origin.

Adsorption of DOM on mineral phases results in a drastic decrease in its microbial degradability. Laboratory studies show that soil samples can adsorb DOM rapidly and there is evidence of a fast and strong adsorption of DOM to Al and Fe oxides and hydroxides. Sorption of the hydrophobic DOM fraction is strongly favored over the binding of the hydrophilic fractions in soils and on hydrous oxides. Furthermore, ^{13}C NMR spectroscopy shows that the sorption of DOM to soil material leads to a preferential depletion of carbonyl and aromatic C, whereas alkyl C accumulates in the soluble portion. Also a preferential sorption of N compounds to oxides and clay minerals has been reported. Especially in the

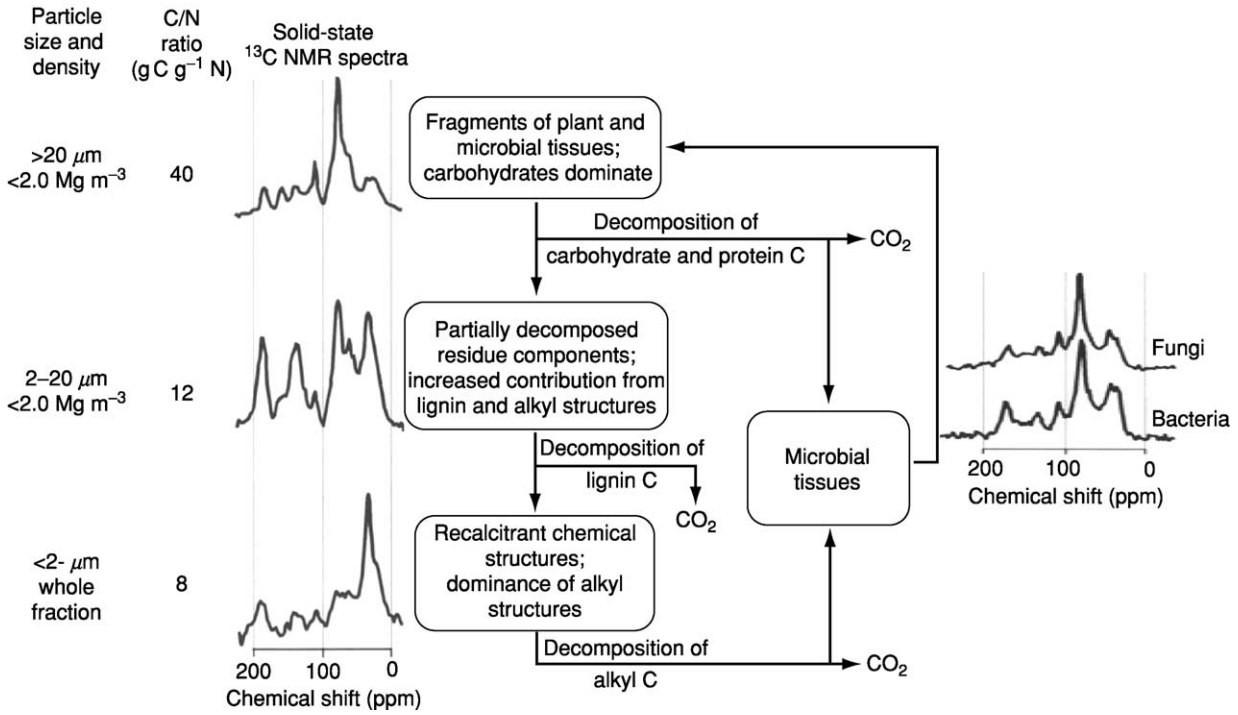


Figure 5 Cross-polarization/magic angle spinning ^{13}C-NMR spectra of the organic matter associated with different particle-size fractions. (Adapted from Baldock JA and Skjemstad JO (2000) Role of the soil matrix and minerals in protecting natural materials against biological attack. *Organic Geochemistry* 31: 697–710, with permission.)

deeper layers of the forest floor and in the A-horizon, low rates of DOM decomposition are found, which qualifies DOM as a source for the formation of stable humic substances. Sorption of DOM can create hydrophobic surfaces which limit microbial access by the reduction of wettability.

Biochemical Stabilization Mechanisms of Main Components from Plant Residues in SOM

Polysaccharide-related materials With respect to quantity, polysaccharides account for 5–25% or more of soil organic carbon. Although most plant polysaccharides are readily decomposed, several microbially synthesized polysaccharides are relatively resistant against decomposition. Based on model experiments, mechanisms for their stabilization have been proposed, which may be caused by a close association with mineral colloids, an interaction with metal ions through uronic acid and mannose units, or by their interaction with tannins. Polysaccharides that are stabilized in such ways may represent a stable fraction of SOM. A further protection against rapid degradation arises from the special role of polysaccharides as 'bridges' between mineral particles, being directly involved in the formation of stable aggregates. There are indications that even more stable products of microbial resynthesis are also involved in the stabilization of SOM. One substance of importance is 'glomalin,' a glycoprotein with N-linked oligosaccharides, which is produced by hyphae of arbuscular mycorrhizal fungi and which is abundant and stable in soil.

Lignin-related materials Lignin, which is the second most abundant component of plant residues in terrestrial ecosystems, is relatively recalcitrant by itself, and degradation rates of lignocellulosic materials are negatively correlated to their lignin content or to their lignin-to-N ratio. Lignin as an aromatic biomolecule is degraded at a much slower rate than cellulosic and noncellulosic polysaccharides, and proteins. This recalcitrance is caused by the presence of nonhydrolyzable C–O–C and C–C bonds between the phenylpropanoid units, and the structural heterogeneity of stable ether and C–C bonds, which can only be cleaved by oxidative mechanisms through ligninolytic enzymes. These enzymes are directly or indirectly involved in the generation of lignin radicals by removing electrons from phenolic units. The resulting unstable radicals subsequently undergo a variety of spontaneous rearrangement reactions. During these reactions, water and oxygen molecules, and thereby additional hydroxyl, carboxyl, and C=O functions are introduced, which increase the oxygen content in the remnant lignin. Another peculiarity of

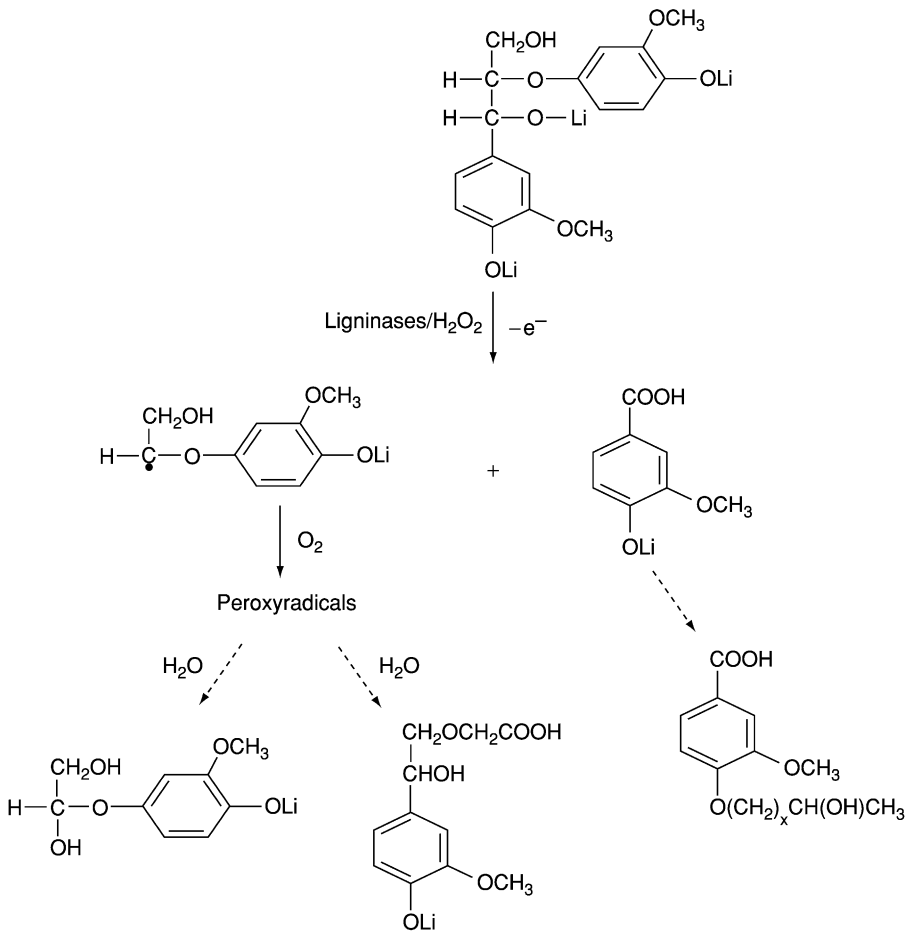


Figure 6 Structural features of a microbially attacked lignin. Li, residues of lignin; •, radical (unpaired electron). (Adapted from Haider K (1992) Problems related to the humification processes in soils of temperate climates. In: Stotzky G and Bollag J-M (eds) *Soil Biochemistry*, vol. 7, pp. 55–94. New York: Marcel Dekker.)

lignin biodegradation, compared with other plant compounds, is the cometabolic character of the degradation process, by which it does not provide a source of energy or carbon for ligninolytic organisms. Thus, for an effective lignin degradation, a readily metabolizable cosubstrate such as carbohydrates is required.

Preservation of lignin constituents in humic substances has attracted attention for many years. Similarities and differences in structure and chemical reactivity between lignin and humic compounds have been reviewed recently in terms of existing humification hypotheses. ^{13}C-NMR spectroscopy of microbially degraded lignin in soils, its oxidative CuO-splitting or its reductive HI-splitting products, pyrolysis in combination with previous chemolysis by tetramethylammonium hydroxide (TMAH) indicate that intact or only demethylated lignin does not appear to be stabilized in the long term in soils or in any of its fractions. However, the latter method reveals that phenolic structures characteristic for altered lignin structural units contribute significantly to the stable SOM pool in soils.

Based on investigations with ligninolytic organisms, enzymatic lignin degradation is a result of random peroxidative splitting of the lignin macromolecule in the α-β and other bonds of the side chain and in the aromatic rings, and by demethylation of the methoxyl groups (Figure 6). Thereby, the compact and hydrophobic structure of the lignin molecule becomes dispersed and, by introduction of additional carboxyl and hydroxyl groups, more hydrophilic. Additional hydroxyl and carboxyl groups in vicinal positions are apt to form complexes with metal ions in solution or on surfaces of clays or metal oxides.

Impact and Function of the Soil Microflora in Relation to SOM Stability

In order to understand the role of microorganisms in the generation and turnover of refractory SOM, a better knowledge of population structure and the activity of the microbial communities in the soil

habitat is necessary. Soils generally contain several tons of organic matter per hectare, but soil microorganisms can degrade most of this SOM pool only very slowly.

Models to describe the turnover of SOM assume the presence of easily and less-easily degradable pools in soil. These pools are, however, not clearly separated, but exist in a continuous sequence of overlapping residence times. Approaches to characterize SOM pools with different turnover rates distinguish between active, slow, and passive pools. The active or easily degradable pools with mean residence times of 0.1–4 years consist of free or only loosely adsorbed particulate organic matter (e.g., the light fraction, the microbial biomass). The slow-degrading SOM pools with mean residence times between 25 and 50 years consist of physically stabilized organic matter, i.e., occluded or sequestered in small aggregates. The passive or stable SOM pool contains the chemically stabilized humic materials, sorbed to fine silt and/or to clay, and has mean residence times of approximately 400–3000 years. On average, approximately 5–10% of the total C in SOM is involved in the active pools, and approximately 20–40% in the slow pools. To maintain a relatively constant nutrient level and release by mineralization, the pools have to be renewed continuously by fresh plant residues. The balance between decay and renewal processes controls nutrient availability and is sensitive to management practices. The slow pools are also affected by cultivation, as physical disturbance (caused, e.g., by ploughing) destroys macroaggregates and larger microaggregates and thus releases physically protected SOM. The passive pools contain the most stable SOM, which is protected by chemical and physical mechanisms. A strict distinction between physical and chemical effects of soil minerals on the stabilization of SOM can often not be made. The passive pool represents 50–70% of the total C in SOM and is seldom affected by management practices.

Whereas the compounds in the active and also some of the ingredients in the slow pools are used by microorganisms for metabolic and energy-producing processes, the great SOM portion of humic compounds in the passive pool is not readily available to microorganisms as a source of energy and products for microbial biosynthesis. On the other hand, small amendments of readily available substrates such as glucose or fructose, but even more in the form of cellulose, wheat straw, root exudates, or green manure, cause 2–5 times more C to be released from soil as CO_2 than was added with the substrate. This effect is termed the 'priming effect.' It arises either immediately or very shortly after the addition of available substrates and is caused by microbial activity, since similar effects have never been observed under sterile conditions. In the case of a 'positive priming effect,' the extra mineralization originates at least partly from the labile humus fractions or by an enhanced turnover of the native soil microbial biomass. In the case of a passive soil fraction, an increased induced enzyme production leading to a cometabolic decomposition has been found.

As for lignin, the completely humified products consist of randomly condensed intermediate compounds. In the case of a passive humus fraction, they represent refractory plant and microbial residues. Their diversity and lack of regular polymeric structures and the large number of enzymatic steps necessary to release CO_2 do not allow efficient enzymatic degradation and energy production. The slow decomposition of the passive SOM, however, becomes accelerated by small inputs of available substrates. Even when these inputs are relatively small compared with the large size of the passive fraction in most soils, they 'trigger' the synthesis of enzymes for the cometabolic degradation of a portion of the otherwise passive SOM fraction. This priming effect only lasts for a relatively short period after addition and disappears after metabolization of the added substrates. It can be hypothesized that this effect may limit the ultimate accumulation of passive SOM.

See also: **Carbon Cycle in Soils:** Dynamics and Management; **Carbon Emissions and Sequestration**; **Factors of Soil Formation:** Biota; **Organic Matter:** Principles and Processes; **Pedology:** Basic Principles

Further Reading

Amelung W, Flach KW, and Zech W (1999) Lignin in particle-size fractions of native grassland as influenced by climate. *Soil Science Society of America Journal* 63: 1222–1228.

Baldock JA and Skjemstad JO (2000) Role of the soil matrix and minerals in protecting natural materials against biological attack. *Organic Geochemistry* 31: 697–710.

Balesdent J, Chenu C, and Balabane M (2000) Relationship of soil organic matter dynamics to physical protection and tillage. *Soil and Tillage Research* 53: 215–230.

Batjes N (1996) Total carbon and nitrogen in the soils of the world. *European Journal of Soil Science* 47: 151–163.

Christensen BT (2001) Physical fractionation of soil and structural and functional complexity in organic matter turnover. *European Journal of Soil Science* 52: 345–353.

de Jonge H, de Jonge LW, and Mittelmeijer-Hazelegger MC (2000) The microporous structure of organic and mineral soil materials. *Soil Science* 165: 99–108.

Desjardins T, Andreux F, Volkoff B, and Cerri CC (1994) Organic carbon and ^{13}C contents in soils and soil size-

fractions, and their changes due to deforestation and pasture installation in eastern Amazonia. *Geoderma* 61: 103–108.
Fontaine S, Mariotti A, and Abbadie L (2003) The priming effect of organic matter: a question of microbial competition? *Soil Biology and Biochemistry* 35: 837–843.
Guggenberger G and Haider K (2002) Effect of mineral colloids on biochemical cycling of C, N, P and S in soil. In: Huang PM, Bollag J-M, and Senesi N (eds) *Interactions Between Soil Particles and Microorganisms and Their Impact on the Terrestrial Environment*. IUPAC Series on Analytical and Physical Chemistry of Environmental Systems, vol. 8, pp. 267–322.
Haider K (1992) Problems related to the humification processes in soils of temperate climates. In: Stotzky G and Bollag J-M (eds) *Soil Biochemistry,* vol. 7, pp. 55–94. New York: Marcel Dekker.
Jastrow JD and Miller RM (1998) Soil aggregate stabilization and carbon sequestration: feedbacks through organomineral associations. In: Lal R, Kimble JM, Follett RF, and Stewart BA (eds) *Soil Processes and the Carbon Cycle*, pp. 207–223. Boca Raton, FL: CRC Press.
Kalbitz K, Solinger S, Park J-H, Michalzik K, and Matzner E (2000) Controls on the dynamics of dissolved organic matter in soils: a review. *Soil Science* 165: 277–304.
Kögel-Knabner I (1993) Biodegradation and humification processes in forest soils. In: Bollag J-M and Stotzky G (eds) *Soil Biochemistry,* vol. 8, pp. 101–137. New York: Marcel Dekker.
Ladd JN, Foster RC, Nannipieri P, and Oades JM (1995) Soil structure and biological activity. In: Stotzky G and Bollag J-M (eds) *Soil Biochemistry,* vol. 9, pp. 23–78. New York: Marcel Dekker.
Shevchenko SM and Bailey GW (1996) Life after death: lignin–humic relationship reexamined. *Critical Reviews of Environmental Science and Technology* 26: 95–153.
Zech W, Guggenberger G, Haumaier L *et al.* (1996) Organic matter dynamics in forest soils of temperate and tropical ecosystems. In: Piccolo A (ed.) *Humic Substances in Terrestrial Ecosystems*, pp. 101–170. Amsterdam, the Netherlands: Elsevier.
Zepp RG and Sonntag CH (eds) (1995) *The Role of Nonliving Organic Matter in the Earth's Carbon Cycle*. Chichester, UK: John Wiley.

Interactions with Metals

N Senesi and E Loffredo, Università di Bari, Bari, Italy

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Metal ions in soil are rarely found in the free state, but mostly occur in interactions with other soil organic, inorganic, and biological components, in the liquid and solid phases. The distribution and speciation of metal ions in soil, especially in trace concentrations, are extremely complex and are governed by a variety of reactions, among which complexation with soil organic matter (SOM) plays a key role.

Thus, understanding the molecular, mechanistic, and quantitative aspects of metal–SOM complexation is extremely important in the prediction of processes in which metals are involved in soil. These include: pedogenesis, mobility and transport, fixation and accumulation, chemical and biological reactivity, bioavailability, and toxicity to higher plants and microorganisms. The topic is also of considerable practical interest because of the continuous and increasing release of various metals in soil by numerous modern agricultural practices.

Molecular and Mechanistic Aspects

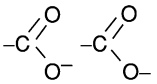
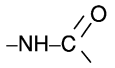
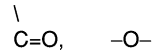
Metal Reactivity, SOM Binding Sites, and Complexation Stoichiometry

Metal cations may be subdivided into three classes with respect to their reactivity with organic ligands: (1) 'hard' cations such as alkaline-earth metals Ca^{2+} and Mg^{2+} that are mainly involved in electrostatic interactions and may form rather weak, outer-sphere complexes only with hard oxygen ligands; (2) 'soft' cations such as Cd^{2+}, Pb^{2+}, and Hg^{2+} that possess a strong affinity for and tend to form covalent bonds with intermediate (N) and soft (S) ligands; and (3) 'borderline' cations, including most transition metals such as Fe^{3+}, Cu^{2+}, Zn^{2+}, and Mn^{2+} that have a character intermediate between hard and soft metals, and possess appreciable affinity for both hard and soft ligands.

The most important metal-binding functional groups present in SOM, classified according to their affinity for hard, borderline, and soft metals, are listed in Table 1. For hard cations the order of donor atom affinity is: O > N > S, whereas a reverse order is observed for soft metals. For bidentate ligand sites, affinity for a given hard metal decreases with the overall hardness of the donor atoms, in the order: (O, O) > (O, N) > (N, N) > (N, S); whereas this order is reversed for soft metals. In general, the competitive reactions for a given ligand essentially involve hard and borderline metals for O sites, and borderline and soft metals for N and S sites, with competition between hard and soft metals being weak. However, the type, source, and concentration of SOM can affect metal-binding affinity, and the relative affinities are often dependent on the method used to measure metal bonding and pH.

Humic substances (HS), i.e., humic acids (HA) and fulvic acids (FA) are universally recognized to

Table 1 Most important metal organic ligands of soil organic matter classified according to their preference for hard, borderline, and soft metals

Ligands preferred by hard cations (hard bases)	*Ligands preferred by borderline cations*	*Ligands preferred by soft cations (soft bases)*
 (Ester)	$-NH_2$, $=NH$, $\equiv N$	R^-
(Carboxylate)	(Primary, secondary, tertiary amino groups)	(Alchil anion)
–OH (Alcoholic and phenolic)		$-SH$, $-S^-$
	(Amide)	(Sulfhydryl, sulfide)
		–S–S–, –S–
(Carbonyl) (Ether)		(Disulfide) (Thioether)

represent the most numerous (between 70 and 80%) and most metal-reactive SOM components. This is because they contain a large number of binding sites per molecule and thus behave as natural 'multiligand' systems.

Metal ion–HS interactions may yield mononuclear complexes, with the central group being either the HS macromolecule or the metal ion, and polynuclear complexes of various stoichiometries. Accumulated evidence shows that aromatic carboxyls and phenolic groups play a prominent role in the 1:1 binding of metal ions by HA and FA, by forming chelates that involve two COOH groups in a phthalic-type site (**Figure 1a**) and both COOH and phenolic groups in a salicylate-type site (**Figure 1b**). Nonaromatic carboxyl and hydroxyl sites, e.g., pyruvic and glycolic acid sites (**Figure 1c, d**), may also be involved in metal ion binding by HA and FA. The most stable complexes are believed to be formed with the more strongly acidic COOH groups, whereas the less stable complexes involve weakly acidic COOH and phenolic OH groups. Other possible combinations involve two phenolic OH, quinone, NH_2, and sulfhydryl groups, and conjugated ketonic structures (**Figure 1e**). Metal ions may also coordinate with ligands belonging to two (or more) HS molecules, forming 2:1 complexes (**Figure 1f, g**) and/or chelates (**Figure 1h, i**). Eventually, a chain structure (**Figure 1j**) may be produced that may result in the aggregation and precipitation as the chain grows at high metal-to-HS ratios.

Two main types of complexes may be formed between metal ions and HS, which are: (1) inner-sphere complexes that result in the formation of bonds with some covalent character between the ligand atom(s) and the metal ion, both completely or partially dehydrated; and (2) outer-sphere complexes that result in the electrostatic attraction between the ligand(s) and the metal ion that remains completely hydrated. For simplicity, **Figure 1a–j** shows formation of inner-sphere complexes, but they may represent outer-sphere HS complexes if the cation remains solvated (e.g., **Figure 1k**).

A number of molecular properties of HS and physical and chemical characteristics of the surrounding medium govern metal–HS binding processes. These include: polyfunctionality, polyelectrolyte character, hydrophilicity/hydrophobicity and extent of ionization of major acidic groups of HS, electronic and steric environment of the HS ligand site, and pH, ionic strength, and metal concentration of the medium.

Experimental Evidence

Spectroscopic techniques such as infrared (IR), ultraviolet-visible (UV-VIS), fluorescence, electron-spin or electron paramagnetic resonance (ESR or EPR), X-ray absorption (XAS), Mössbauer, and nuclear magnetic resonance (NMR) spectroscopies have provided a number of important molecular and mechanistic insights of the chemical and physical nature of HS binding sites and metal–HS interactions.

IR, UV-VIS, and fluorescence spectroscopies Fourier transform IR (FT IR) and diffuse reflectance FT IR (DRIFT) spectroscopies, often in conjunction with selective group blocking techniques and/or with

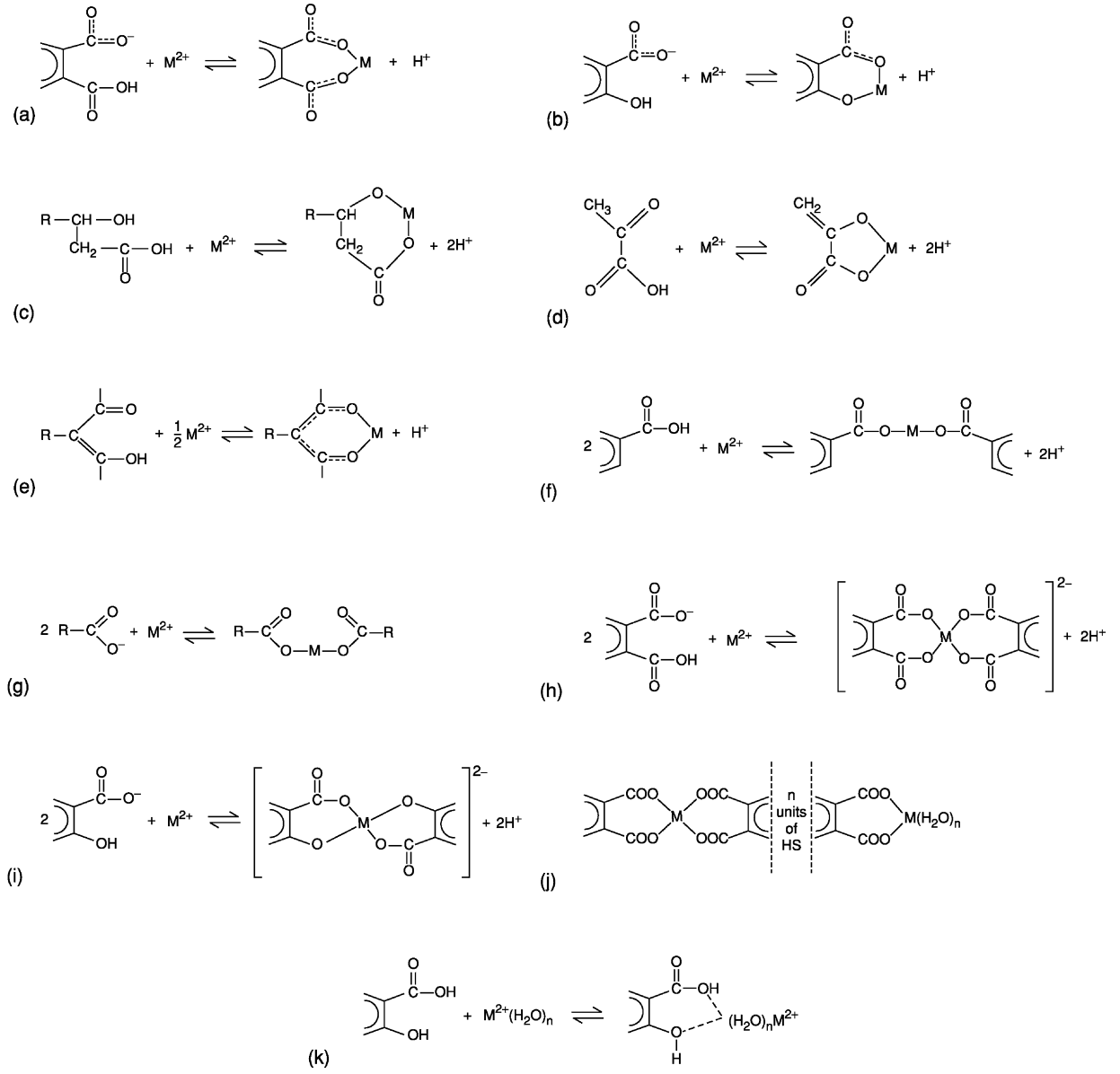


Figure 1 The most common binding sites for metal ions in humic substances and stoichiometries of complexation: (a) a phthalic-type binding site; (b) a salicylate-type binding site; (c) a pyruvic acid binding site; (d) a glycolic binding site; (e) a conjugated ketonic binding site; (f) a 2:1 aromatic carboxyl complex; (g) a 2:1 aliphatic carboxyl complex; (h) a 2:1 phtalic-type chelate; (i) a 2:1 salicylate-type chelate; (j) a chain structure; (k) a salicylate-type outer-sphere complex.

chemical derivatization methods, have provided useful information about the nature and reactivity of HS functional groups that interact with metal ions.

The prominent role played by COOH groups of HA and FA in the complexation of several metal ions, including Cu^{2+}, Mn^{2+}, Zn^{2+}, Pb^{2+}, Co^{2+}, Ni^{2+}, Ca^{2+}, Mg^{2+}, Sr^{2+}, Fe^{3+}, and Al^{3+}, has been demonstrated by the strong reduction or disappearance of the typical IR absorption bands of COOH groups at approximately 1710 and 1200 cm^{-1}, and the simultaneous reinforcement or appearance of the typical IR bands of COO^- groups near 1600 and 1380 cm^{-1}, upon reaction of HA and FA with the metal (Figure 2). The extent of COO^- to COOH conversion varies in the decreasing order: $Fe^{3+} > Cu^{2+} > Al^{3+} > Ca^{2+} > Mg^{2+}$. The increase in COO^- vibration frequency for a number of metal–HA and –FA complexes suggests that high-covalent bonds are preferentially formed at low levels of metal ion, whereas bonding becomes increasingly ionic as the system is enriched with the metal. The large separation measured between the frequencies at approximately 1600 and 1380 cm^{-1} for the COO^- group complexed by Cu^{2+}, Fe^{3+}, Co^{2+}, and Zn^{2+}, with respect to that of

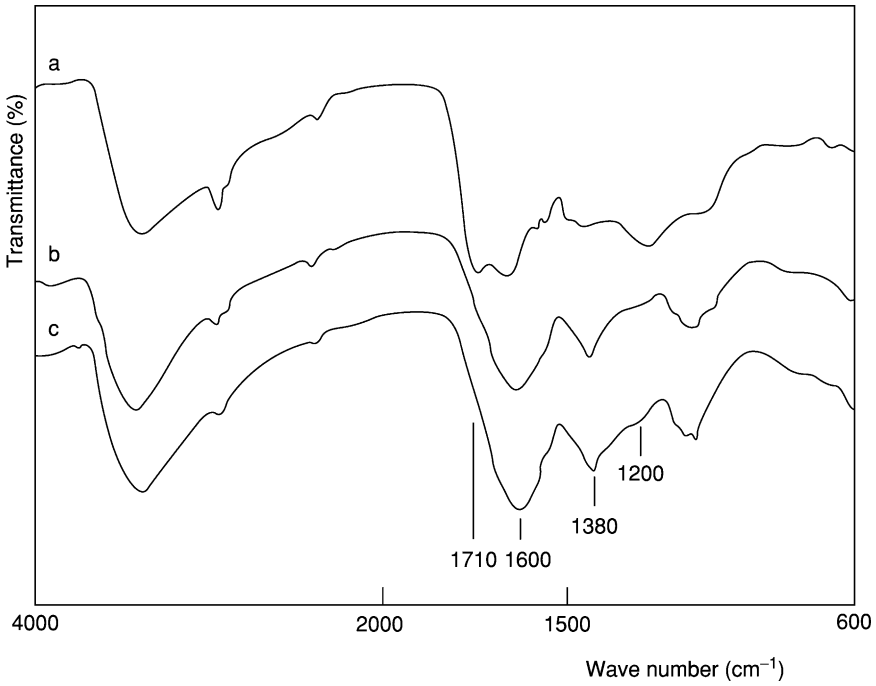


Figure 2 Infrared spectra of a soil humic acid (HA) (a) and its complexes with Cu^{2+} (Cu^{2+}:HA = 0.04) (b), and Cu^{2+} and Fe^{3+} ((Cu^{2+} + Fe^{3+}):HA = 0.08) (c). (Adapted from Senesi N, Sposito G, and Martin GP (1986) Copper (II) and iron (III) complexation by soil humic acids: an IR and ESR study. *Science Total Environment* 55: 351–362, with permission.)

the uncomplexed COO^- ion, suggests the preferential formation of unidentate metal complexes.

The net IR shift of the OH band, from 3500–3400 cm^{-1} in HA and FA to 3300–3200 cm^{-1} in their complexes with Zn^{2+}, Fe^{3+}, Cu^{2+}, Pb^{2+}, Co^{2+}, and Mn^{2+} is ascribed to the involvement of phenolic and/or alcoholic OH groups in metal binding. The extent of the OH stretching shift toward lower frequencies follows the order: Fe > Cu > Co > Mn. Modification of the typical band at approximately 1070 cm^{-1} ascribed to OH groups of polysaccharide components of FA suggests that these structures are also involved in metal complexation. The shift to a lower frequency measured for the band at 1610 cm^{-1} in Cu^{2+}–FA complexes is attributed to the involvement of conjugated ketones in Cu^{2+} binding. IR evidence is also provided for metal binding sites involving amide N (and possibly amide C=O) and sulfonic groups by N- and S-rich HS. Finally, the very sharp IR bands observed in the region 1130–1080 cm^{-1} and at between 890 and 697 cm^{-1} for HA and FA complexes with Cu^{2+}, Zn^{2+}, Al^{3+}, or Fe^{3+} are attributed to the metal–O vibration of bound hydroxylated and/or hydrated metal ions.

The stoichiometry of metal–HS complexes has been investigated by UV-VIS spectroscopy based on the Job method. For example, the ions Cu^{2+}, Fe^{3+}, and Al^{3+} form 1:1 complexes with FA at pH 3; whereas, at pH 5, Cu^{2+} and Fe^{3+} form 2:1 molar complexes with the FA, and Al^{3+}–FA complex composition remains at 1:1.

The capacity of paramagnetic transition metal ions such as Cu^{2+}, Fe^{3+}, Fe^{2+}, Mn^{2+}, Cr^{3+}, and VO^{2+} to quench efficiently the natural fluorescence of some HS ligands is attributed to their capacity to form strong, inner-sphere complexes with HS, involving strongly acidic carboxyl groups in bidentate salicylic- or phthalic-type chelation sites. In contrast, the low fluorescence-quenching ability exhibited by cations such as Mn^{2+}, Co^{2+}, Ni^{2+}, Pb^{2+}, and Al^{3+} suggests the formation of weak, outer-sphere complexes with HS. The quenching effect generally increases with increasing pH (Figure 3). Fluorescence-quenching studies also suggest the involvement of aromatic amino acidic groups such as tyrosine and phenylamine moieties, and citrate and malonate moieties in the complexation of Cu^{2+} by FA.

ESR, XAS, Mössbauer, and NMR spectroscopies The ESR technique applied to HA and FA complexes with paramagnetic transition metals of major importance in soil, including Fe, Cu, Mn, and V, has provided useful information about oxidation states of metals bound, symmetry and type of coordination sites in HS, binding mechanisms of metals to HS, and identity of ligand atoms and groups involved in metal complexing.

The representative ESR spectrum of a soil HA shown in Figure 4a exhibits a number of patterns

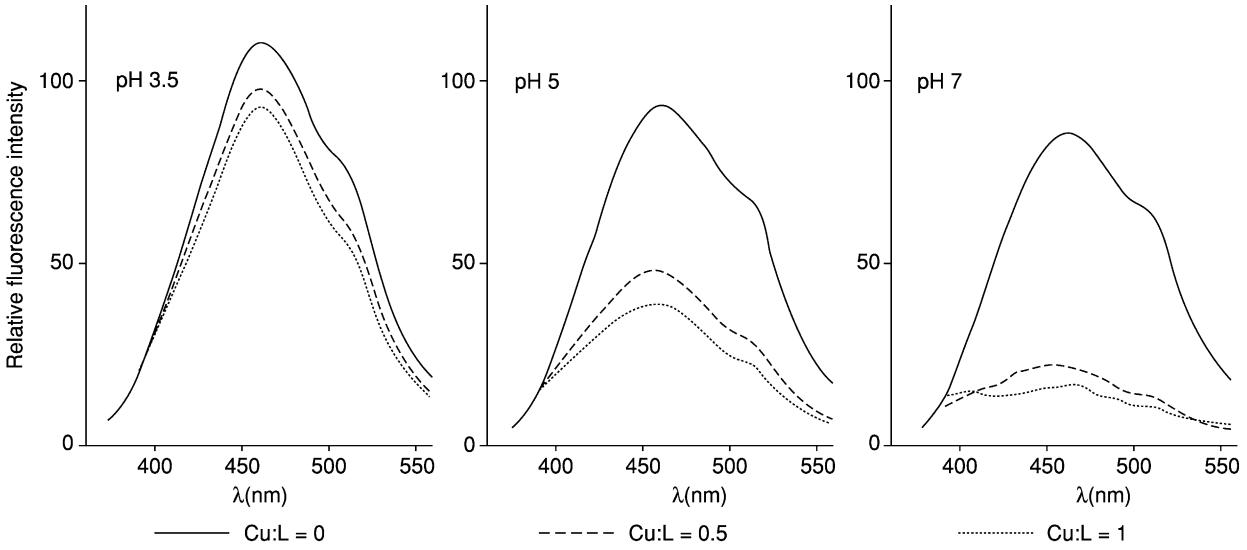


Figure 3 Fluorescence emission spectra ($\lambda_{exc} = 355$ nm) of a soil fulvic acid (FA) and its complexes with Cu^{2+} obtained at three pHs and two different Cu^{2+}-to-FA ligand (L) ratios. (Reproduced with permission from Bartoli F, Hatira A, Andre JC, and Portal JM (1987) *Soil Biology Biochemistry* 19: 355–362.)

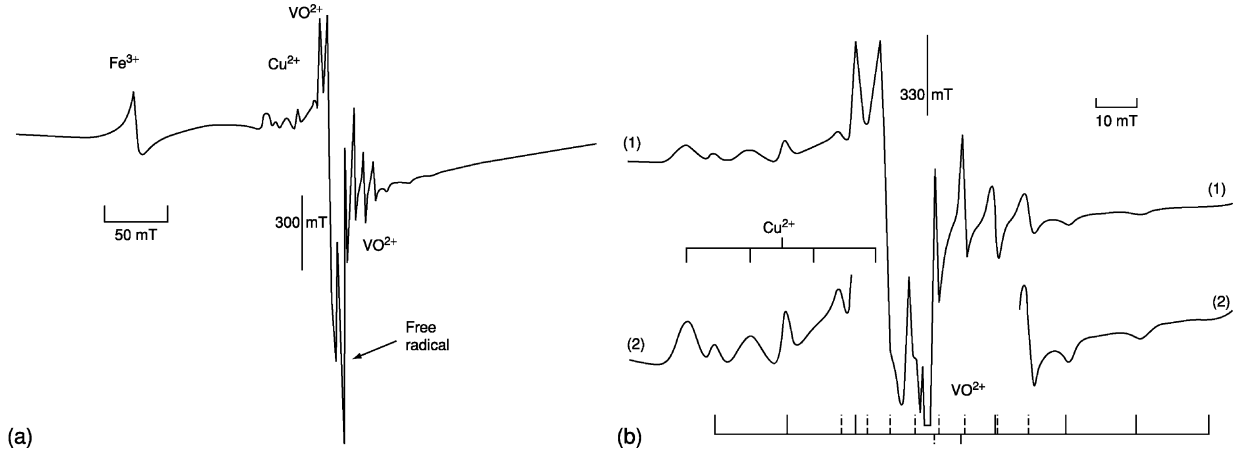


Figure 4 Electron-spin resonance (ESR) spectra at 77 K of a loam soil HA recorded on scan ranges of 800 mT (a) and 200 mT (b) (1, high-field region; 2, higher gain). (Reproduced with permission from Senesi N (1990) Application of electron-spin resonance (ESR) in soil chemistry. In: Stewart BA (ed.) Propriétés fluorescentes et colloidales d'une solution organique de podzol au cours du processus de complexation par le cuivre. *Advances in Soil Science*, vol. 14, pp. 77–130. New York: Springer-Verlag.)

that can be resolved, analyzed, and attributed to various metal–HS complexes. The asymmetrical, isotropic resonance line appearing at low magnetic field is consistent with high-spin Fe^{3+} ions in tetrahedral or octahedral sites of rhombic symmetry bound to O functional groups, possibly carboxylic and/or phenolic groups of HA. This form of Fe exhibits considerable resistance to proton and metal exchange, and to chemical reduction, thus suggesting that Fe^{3+} is strongly bound and protected in inner-sphere complexes in HS. The broad signal often exhibited by soil HS at higher field is preferentially ascribed to Fe^{3+} ions in slightly distorted octahedral sites. Iron in such sites is easily reduced by chemical agents and easily extracted by complexing agents, thus suggesting a weak binding of Fe^{3+} on external surfaces of HS.

The richly structured pattern exhibited at higher field (Figure 4a), which is better analyzed using the enlarged spectrum of the region (Figure 4b), can be resolved into two distinct, overlapping rigid-limit spectra of the 'axial' type. One comprises the typical anisotropic pattern ascribed to Cu^{2+} ions held in inner-sphere complexes with ligands arranged in a square planar (distorted octahedral) coordination site (tetragonal symmetry), involving either only O functional groups (carboxyls, phenolic hydroxyls,

carbonyls, and often water molecules) of HS, or both O and N ligands, or even only N atoms (i.e., a tetraporphyrin site). The other pattern is consistent with a VO^{2+} ion rigidly bound in HS as an inner-sphere complex in a square planar coordination site to O ligands, mostly phenolate, or possibly water molecules and, more rarely, N ligands.

ESR spectral parameters also provide evidence of a high covalent bond for Cu^{2+} and VO^{2+} complexed in HS, which renders Cu^{2+} and VO^{2+} ions difficult to remove from these sites, even with acid leaching. Covalent bonding is favored by low metal loadings in HS, where complexation to N-containing groups is preferred to O-containing ligands, whereas at high loadings the metal is bound largely to O-containing ligands and has a higher degree of mobility.

The ESR spectra of some soil HA and FA feature a well-resolved isotropic pattern consistent with high-spin, hexahydrated Mn^{2+} bound by electrostatic forces as outer-sphere complexes in a distorted octahedral environment to six O atoms of negatively charged carboxylate and phenolate groups. However, at pH > 8, or temperature $T > 50°C$, Mn^{2+} can enter inner-sphere, multiligand complexation sites. Thus, the type and stability of Mn^{2+}–HS complexes and, in turn, their ease of exchangeability and bioavailability in natural systems are strongly dependent on pH and temperature.

ESR evidence has been provided that HS possess a high residual binding capacity toward added Cu^{2+}, Fe^{3+}, VO^{2+}, and Mn^{2+} ions that can form complexes of various stability against various physical and chemical treatments. Most of the added metal is bound to surface sites in relatively labile and exchangeable forms.

In general, ESR analysis shows that high pH values that generate a greater availability of negatively charged O ligands favor inner-sphere complexation for metals, whereas, at lower pH, metals are preferentially retained as hydrated ions. Inner-sphere coordination is also preferred when competing water ligands are removed by dehydration, thus forcing the metal to enter into direct bonding with HS ligands. ESR data confirm that the more electronegative the metal ion, the stronger the metal bound to HS, the higher the degree of bond covalency. The ESR-related, electron-nuclear double resonance (ENDOR) and electron spin-echo envelope modulation (ESEEM) spectroscopies have the potential to overcome some limitations and yield more information than the classic ESR approach about the chemical environment of paramagnetic metal ions.

The XAS techniques, X-ray absorption near-edge structure (XANES) and extended X-ray-absorption fine-structure (EXAFS) spectroscopies, have yielded additional details about the coordination of the central ion to HS ligands and values of bond lengths and distances of metal-ligand also for nonparamagnetic metal ions that cannot be revealed by ESR. Analysis of XANES spectra, radial structure function (RSF) plots derived from EXAFS spectra, and corresponding simulations for Cu^{2+} bound to HS at various pHs (Figure 5) indicate the presence of a tetragonally distorted octahedral binding site for Cu^{2+} in HS, involving four equatorial O (or N) atoms at a mean distance of 1.92–1.97 Å and two axial O atoms at a mean

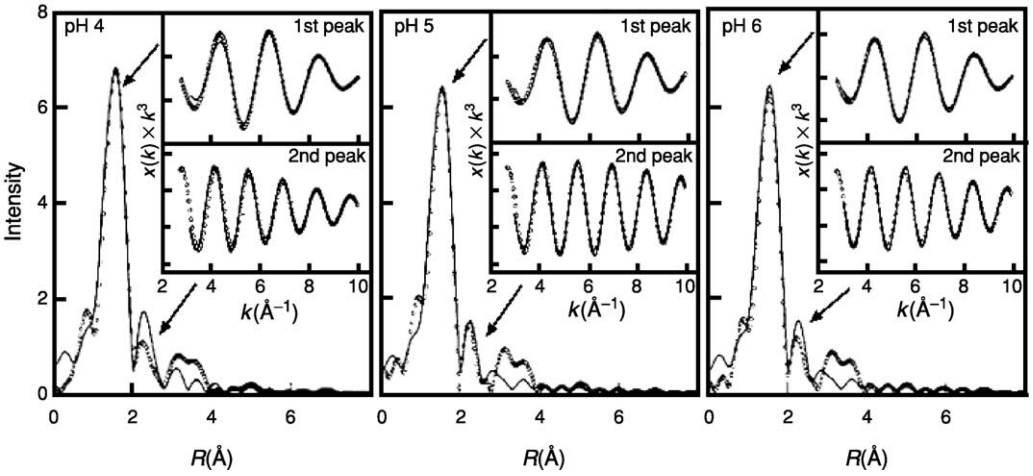


Figure 5 Experimental radial structure function for Cu^{2+}–humic substance complexes at pH 4, 5, and 6 (dotted line) and simulations (solid line) for an adjusted model of the coordination site derived from bond network analysis. The inset shows plots of experimental (dotted line) and fitted (solid line) inverse Fourier-transformed scattering curves for the first atomic shell (Cu–O) and second atomic shell (Cu–C). *R*, distance (in Å) of the ligand atom from Cu^{2+}. (Reproduced with permission from Xia K, Bleam WF, and Helmke PA (1997) Studies of the nature of Cu^{2+} and Pb^{2+} binding sites in soil humic substances using X-ray absorption spectroscopy. *Geochimica Cosmochimica Acta* 61: 2211–2221.)

distance of 2.02 Å in the first coordination shell, and four C atoms at a mean distance of 3.13 Å in the second shell. Similar XANES and EXAFS studies of Pb^{2+}–HS complexes indicate that Pb^{2+} is bound to four O atoms in the first shell at mean distances, decreasing from 2.46 to 2.32 Å, with increasing pH from 4 to 6, and two C atoms at a mean distance of 3.26 Å in the second shell. These results suggest inner-sphere bonding of Pb^{2+} to two O-containing, possibly salicylate- and catechol-type functional groups of HS. Complexes of Ni^{2+}, Co^{2+}, and Zn^{2+} with soil HS at pH 4 are interpreted as inner-sphere octahedral coordination, with six O atoms in the first shell for Ni^{2+} and Co^{2+}, and with four O and two S atoms for Zn^{2+}. The bond distances for Ni—O and Co—O are 2.10 Å and 2.04 Å, respectively, and for Zn—O and Zn—S are 2.13 Å and 2.33 Å, respectively. There is one C atom in the second shell for Co, at a distance of 2.87 Å, and there are two for Ni and Zn, at a distance of 2.94 Å and 3.29 Å, respectively. At very high Zn^{2+} concentration, however, most of the Zn (80–90%) is found to be bound as an outer-sphere complex. XANES and EXAFS spectra of Fe^{3+} and Mn^{2+} complexes with HA suggest that both ions are held in HA by six nearest-neighbor O or N atoms at a distance of 2.01 ± 0.05 Å and 2.20 ± 0.04 Å, respectively, and that Fe^{3+} occupies more than one binding site. Detailed EXAFS studies of Hg^{2+} complexation by soil HA suggest a two-coordinate binding environment for Hg^{2+}, with one O atom and one S atom in the first coordination shell at distances of 2.02 and 2.38 Å, respectively, and one C atom and a second S atom in the second shell at 2.78 and 2.93 Å, respectively. These results suggest thiol, disulfide/disulfane, COOH, and phenol OH groups of HA as ligands for Hg complexation.

Although Mössbauer spectroscopy is highly specific for the isotope ^{57}Fe, this technique is able to distinguish unambiguously between Fe^{3+} and Fe^{2+} ions that occur mostly in high-spin forms in combination with HS. Mössbauer parameters calculated from the representative Mössbauer spectrum of a soil HA shown in Figure 6a suggest the presence of three sites for Fe^{3+} ions in the HA, two with octahedral and one with tetrahedral coordination. Chemical reduction with hydrazine of the HA sample results in a Mössbauer spectrum in which one of the two doublets assigned to Fe^{3+} ions in octahedral sites disappears (3a in Figure 6a), and a new doublet appears that is typical of Fe^{2+} ions in the same type of sites (3b in Figure 6b). Mössbauer studies of several Fe^{2+}–HA complexes suggest the presence of Fe^{2+} ions either as a completely hydrated form in outer-sphere complexes or as a partially hydrated form bound to COO^- and phenolic and alcoholic OH- and N-containing groups of HA in elongated octahedral arrangements in inner-sphere, chelate-type complexes. Mössbauer evidence of partial oxidation of Fe^{2+} to Fe^{3+} is obtained on exposure to air of some Fe^{2+}–HA complexes.

The slight decrease observed in the line widths of the water NMR signal in a water solution of soil FA and Mn^{2+} suggests the formation of outer-sphere complexes of hexahydrated Mn^{2+} ion with FA in a slightly distorted octahedral environment. Differently, the great decrease measured in solutions of FA with Fe^{3+} or Cu^{2+} suggests the formation of inner-sphere coordination of these ions with FA. The involvement of O- and N-containing groups is suggested by ^{113}Cd-NMR in Cd complexation by soil FA in acidic conditions. Modern, two-dimensional NMR techniques such as total correlation (TOCSY), nuclear Overhauser effect (NOESY), heteronuclear Overhauser effect (HOES), and inverse-detection heteronuclear multiple quantum coherence (HMQC) spectroscopies are very promising in the studies of metal–HS interactions.

Quantitative Aspects

Metal-Binding Capacity of HS

In general, the binding or complexation capacity (CC) of a multiligand system such as HS is defined as the maximum capacity of the system for binding metal ions. To a good approximation, the CC of HS may be considered as the weighted average of the complexation capacities of the individual ligands $(CC)_i$ in the system, according to:

$$CC = \frac{\sum (CC)_i [\text{weight}]_i}{\sum [\text{weight}]_i} \qquad [1]$$

where $[\text{weight}]_i$ is a weighting factor related to the relative abundance of the ligand in the multiligand system. Given the relative abundance of acidic functional group ligands, primarily COOH, and the low concentrations of other O-, N-, and S-containing ligands in HS, the maximum binding capacity of HS approximates the total acidic functional group content.

In practice, the measured complexation capacity of HS is dependent on experimental conditions, being a function of various factors that include pH, ionic strength, HS concentration and properties, and nature of the metal ion. The effect of pH on metal–HS complexation results from both changes in the extent of ionization of COOH groups and hydrolysis reactions of the metal ion involving the formation of oxyhydroxides. In general, the complexation capacity

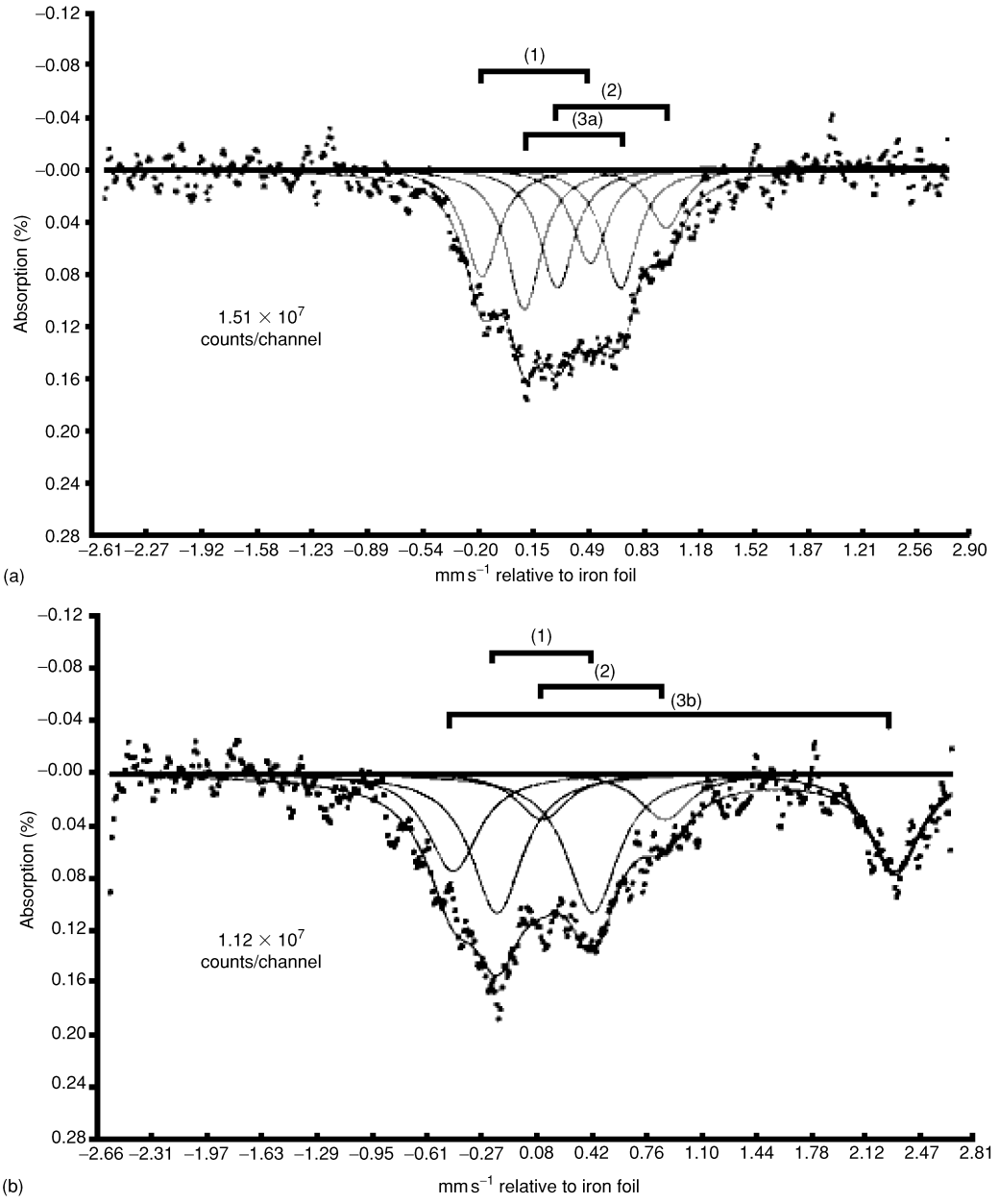


Figure 6 Mössbauer spectra of a soil humic acid (HA) (a) and the same HA after reduction with hydrazine (b). Doublets (1), (2), (3) refer to three different sites for Fe in HA. Doublets (3a) and (3b) refer, respectively, to Fe^{3+} and Fe^{2+} ions in similar octahedral sites. (Reproduced with permission from Senesi N, Griffith SM, Schnitzer M, and Townsend MG (1977) Binding of Fe^{3+} by humic materials. *Geochimica Cosmochimica Acta* 41: 969–976.)

of HS is reported to increase with increasing pH from acidity up to approximately neutrality and increasing HS concentrations, to decrease at high ionic strength, and to vary with the nature of the metal ion. For any given pH and ionic strength, trivalent cations are bound in greater amounts than divalent ones, of which those forming strong coordination complexes (e.g., Cu^{2+}) are bound to a greater extent and at stronger binding sites than those forming weak complexes (e.g., Mn^{2+}).

Stability Constants and Related Parameters of Metal–HS Complexes

The stability constant is probably the most important quantitative parameter for the characterization of a metal–ligand complex in that it provides a numerical

index of the affinity of the metal cation for the ligand. Several different theoretical and experimental approaches have been attempted for the determination of stability constants of metal–HS complexes.

The overall reaction of a single metal ion (M) and a single ligand (L) to form the complex M_mL_n:

$$mM + nL \rightleftarrows M_mL_n \qquad [2]$$

can be described by:

$$K = \frac{\{M_mL_n\}}{\{M\}^m\{L\}^n} = \frac{[M_mL_n]}{[M]^m[L]^n} \times \frac{\gamma_{M_mL_n}}{(\gamma_M)^m(\gamma_L)^n} = K_c \times \Gamma \qquad [3]$$

where braces and square brackets denote activities and concentrations, respectively, γ-values are activity coefficients, K is the true 'overall thermodynamic stability (equilibrium) constant,' K_c is the 'concentration quotient,' and Γ is the activity coefficient ratio. For convenience, the charges on the metal, ligand, and complex are not shown explicitly, even if their values are extremely important.

Most experimental investigations of metal–HS complexation reactions yield concentrations, rather than activities, of reactants and products. Thus, thermodynamic constants cannot be measured directly but must be indirectly obtained by calculation or by extrapolation methods from the concentration quotients whose values are not constant but depend, as well as those of the activity coefficient ratios, on the ionic concentration (ionic strength) of the solution. Basic electrostatic considerations (Debye–Hückel theory) indicate that Γ equals 1, and thus K_c values equal K values, at zero ionic strength; and Γ increases, and K_c tends to decrease, with increasing ionic strength. The concentration of the complex M_mL_n is thus expected to decrease upon addition of a background electrolyte to the solution of metal and ligand.

Competition from side reactions, especially the acid–base chemistry of both the metal ion and the ligand, i.e., hydrolysis of the metal ion to produce hydroxy complexes and protonation of the ligand, is another factor that affects the extent of complexation of M by L, which must be duly considered. In general, protonation of ligands occurs at low pH, and hydrolysis of metals at high pH. Thus, the most favorable condition for measuring the complexation of M by L is at intermediate pH values.

Metal–ligand complexation reactions are usually investigated at constant pH and using experimental techniques that are able to distinguish the metal–ligand complex from the uncomplexed metal and all forms of the ligand that are not bound to the metal ion. Thus, a 'conditional stability constant,' or more precisely a 'conditional concentration quotient,' K_c^*, can be written:

$$K_c^* = \frac{[M_mL_n]}{\frac{[M]^m}{(\alpha_M)^m} \times \frac{[L]^n}{(\alpha_L)^n}} = (\alpha_M)^m(\alpha_L)^n \frac{[M_mL_n]}{[M]^m[L]^n}$$
$$= (\alpha_M)^m \times (\alpha_L)^n \times K_c \qquad [4]$$

where α_M and α_L are the side-reaction coefficients of the metal ion and the ligand, respectively. Eqn [4] describes the formation and stability of the complex relative to all other forms of the metal and ligand that may exist under the actual experimental conditions. Further, eqn [4] shows that the conditional concentration quotient, K_c^*, includes a pH-invariant but ionic strength-dependent term, K_c, and pH-dependent terms, α_M and α_L.

Assuming for simplicity a 1:1 stoichiometry for all ML_i complexes, the complexation reaction of a single metal ion M with a multiligand system such as HS containing a number of ligands L_i can be described by an 'average stability constant,' or, more appropriately, an 'average concentration quotient':

$$\bar{K}_c = \frac{\sum[ML_i]}{[M]\sum[L_i]} \qquad [5]$$

where $\sum[ML_i]$ and $\sum[L_i]$ are, respectively, the sum of the concentrations of all 1:1 complexes and all uncomplexed ligands in the system. In practice, the metal ion will distribute itself among the various ligands on the basis of its stability constant with each ligand and of the relative concentration of that ligand. The metal ion will thus tend to complex preferentially with ligands having the highest stability constant and/or highest concentration.

Similar to single-ligand systems, an 'average conditional concentration quotient,' also called 'stability function,' $\bar{K}_c^*$, that takes pH effects into account, can be defined for multiligand systems such as HS. This quantity is a measure of the overall binding of a single metal ion to a multiligand system. However, it cannot be calculated rigorously for HS, because there is not sufficient information available to enable the full description of side reactions and the calculation of side-reaction coefficients of HS for which individual components are unknown. Thus the function $\bar{K}_c^*$ is often calculated directly from experimental data as:

$$\bar{K}_c^* = \frac{C_M - [M]}{[M](C_L - C_M + [M])} \qquad [6]$$

where C_M and C_L are the total stoichiometric concentrations of the metal and ligand in the system examined and [M] is the concentration of the free metal ion. The value of $(C_M - [M])$ corresponds to the sum

of the concentrations of all complexes formed between M and the multiligand system, and is calculated by neglecting the presence of inorganic metal complexes. The quantity $(C_L - C_M + [M])$ represents the sum of the concentrations of all binding sites that are not associated with M, and is calculated assuming an average 1:1 metal-to-ligand stoichiometry for the total of binding sites. Stability functions can also be defined for complexes of stoichiometries other than 1:1.

Mean $\bar{K}_c^*$ values are ultimately functions of ionic strength, pH, and degree of saturation of the multiligand system with the metal ion. At a given pH and ionic strength, $\bar{K}_c^*$ will decrease steadily as the total metal-to-ligand ratio (C_M:C_L) increases. At low metal-to-ligand ratios, the occurrence of preferential reactions of stronger ligands will determine the functional nature of $\bar{K}_c^*$. Thus, it is not possible to obtain true stability constants for metal–HS complexation, and reported values are not actually constant, because the functional dependence is a fundamental feature of systems such as HS, which cannot be eliminated by any experimental method unless total fractionation be achieved into pure ligands that can be studied separately.

In conclusion, the ill-defined nature of multiligand and polyelectrolytic HS systems and of the types, concentrations, and strengths of their several nonidentical binding sites prevent metal–HS complexation reactions to be described in rigorous mathematical terms. However, the stability function concept provides an important overview and useful insights into the nature of metal binding by HS, and may also provide useful indications for the most appropriate experimental methods with which to investigate metal–HS interactions.

Modeling and Models

The large body of field and laboratory data accumulated on the binding of metal ions and protons by HS has determined the development of chemical speciation models that could encapsulate the ample knowledge and information acquired, and allow applications to 'real-world' situations. The prevalent effort of any modeling approach is the quantitative description of the relative concentrations and strengths of the many nonidentical binding sites contained in HS. The various models applied to describe metal–HS complexation can be classified as 'discrete' ligand models and 'continuous' multiligand models.

In the discrete ligand approach, only a few ligands or binding sites are required to fit the experimental data. Since early modeling work, the competition for HS binding sites between metal ions and protons has been recognized, and metal binding has been estimated from displacements of pH titration curves. Successively modified and refined discrete models have incorporated site heterogeneity effects, i.e., variability in the intrinsic binding strengths among sites, and electrostatic effects related to the net negative charge of HS. Recently, a simplified and improved model (model VI), also based on the discrete-site/electrostatic formulation but moving toward a more distributional approach in describing binding sites for metals, has been developed and applied with success to describe published data sets on a range of metals.

The continuous multiligand distribution models are based on the assumption that a large number of heterogeneous ligands with a range of binding affinities are involved in metal binding with formation of complexes characterized by a continuous (e.g., Gaussian) distribution of the stability constants. The initial, noncompetitive Gaussian distribution models that do not account for proton–metal competition have been extended in the competitive Gaussian distribution models that characterize HS binding-site heterogeneity by using a continuous distribution of equilibrium constants, together with electrostatic submodels, and by assuming monodentate binding of metals to proton-binding sites. The most recent of these models is the NICA (nonideal competitive adsorption)–Donnan model that fits very well extensive experimental data sets obtained for a wide range of metal ion concentrations. The incorporation of multidentate sites into the NICA-Donnan model is under consideration.

A conceptually simple and calculation-convenient model is the 'quasiparticle' model, which includes previous concepts and models, and describes mathematically an aqueous system containing HS by replacing it with a set of hypothetical, average noninteracting molecules whose behavior mimics that of the actual HS system.

Since the primary purpose of any metal–HS binding model is to perform chemical speciation calculations for field situations, validation of the model on field data is required. Models formulated have been applied with variable success. More recent models reproduce trends quite well but lack of knowledge still exists on some key factors, and empirical adjustment of these factors is necessary to fit the data. Thus, more extensive field-testing is required to establish the usefulness or otherwise of any model of metal–HS binding.

Experimental Methods

A variety of separation and nonseparation methods have been used to speciate metal ions in the presence of HS, to assess the complexing capacity of HS, and to calculate conditional stability constants or quotients

of metal–HS complexes. Most commonly used separation approaches include: (1) proton-release titration; (2) cation exchange with synthetic resins; (3) gel filtration or permeation chromatography (GPC); (4) high-performance cation exchange chromatography (HP-CEC); (5) conventional dialysis and equilibrium dialysis–ligand exchange (EDLE); (6) ultrafiltration (UF); and (7) capillary electrophoresis (CE).

Separation approaches generally present the great advantage of allowing measurement of nearly any metal ion by sensitive techniques such as atomic absorption and plasma spectrometries. However, dialysis and chromatographic methods suffer from the disadvantages of disturbance of complexation equilibria and adsorption of interacting species on the membrane or chromatographic material, whereas equilibrium dialysis and cation exchange present uncertainties in the definition of chemical forms actually measured by the method. An important advantage of the UF methods is that they do not take such a long time to carry out as dialysis and do not disturb the complexation equilibria.

Most important nonseparation methods include: (1) ion-selective electrode (ISE) potentiometry; (2) voltammetric techniques such as anodic-stripping voltammetry (ASV); and (3) various spectroscopic techniques, including fluorescence, ESR, and NMR spectroscopies. Some nonseparation methods such as ISE and ASV, although being the most sensitive and advanced, also suffer from the problems of adsorption and shifts in equilibria. Further, these methods are applicable adequately only to a limited typology of metal ions, e.g., Cu^{2+}, Cd^{2+}, Pb^{2+}, and Zn^{2+}.

The fluorescence method presents a number of advantages over other methods. Titration curves of HS fluorescence quenching versus concentration of added metal quencher have allowed direct measurement of the complexing capacity of HS ligands and the stability constants of HS–metal complexes. The method is relatively rapid, since no separation is required between bound and free metal ion, thus errors associated with the separation step in most speciation methods are avoided. Neither supporting electrolyte, nor buffer, nor adsorbing material are required to be added to the sample; it is even more sensitive than ASV and ISE potentiometry and is sensitive enough for application to unmodified, natural organic ligands without preconcentration. The major disadvantage of fluorescence spectroscopy is, however, its efficiency only with strongly binding, paramagnetic metal ions such as Cu^{2+}. This limitation can be overcome by the use of fluorescent probes, such as in the lanthanide ion probe fluorescence spectroscopy (LIPS), where HS samples are titrated by Eu^{3+} ions, and the titration plot of the ratio of the intensities of two emission lines of Eu^{3+} is used to estimate the amount of bound and free species of the probe ion.

The ESR approach allows the determination of the free metal ion concentration and therefore of its degree of complexation with the ligand. ESR spectroscopy has been applied to obtain stability constants and stoichiometries in aqueous solution of VO^{2+} and Mn^{2+} complexes with FA. On these bases, the VO^{2+}–FA complexes are modeled as bis(phthalato)(salicylato)-oxovanadium (IV) and mono(salicylato)-oxovanadium (IV). The K_c values measured for Mn^{2+}–FA complexes by ESR are in excellent agreement with K_c values determined by an ion-exchange method. Major advantages of ESR spectroscopy are its high sensitivity and the minimal or no-sample pretreatment required, whereas its intrinsic limitation is applicability only to paramagnetic metal ions such as Cu, Fe, Mn, and V.

Few NMR experiments have been conducted to relate structure of HS to their metal-binding capacity. The positive correlation found between aromaticity of several HA determined by solution-state ^{13}C-NMR and stability constants of their complexes with Cd^{2+} and Zn^{2+} ions suggests the involvement of phenolic groups in the complexation of these metals. The higher content of carboxylate groups in a soil FA than in a soil HA measured by distortion enhancement by polarization transfer (DEPT)-^{13}C-NMR and quaternary-C only (QUAT) subspectra is consistent with the higher binding constant measured for Eu^{3+} with FA than with HA. Dipolar dephasing cross-polarization magic angle spinning (CPMAS) ^{13}C-NMR has shown that the stability constant of a Cd^{2+}–sediment HA is greater than that of Cd^{2+}–soil HA, probably due to the higher content and stronger binding of amino acidic N groups, in addition to carboxylic groups, of the sediment HA. Further, great metal stability constants of Zn^{2+}–HA and Cd^{2+}–HA complexes are found to be associated with a higher degree of substituted aromatic C groups, possibly due to chelation effects.

List of Technical Nomenclature

α_L	Side-reaction coefficient of the ligand
α_M	Side-reaction coefficient of the metal ion
Γ	Activity coefficient ratio
Å	Ångstrom (10^{-8} cm)
K	Thermodynamic stability (equilibrium) constant
$\bar{K}$	Mean stability constant or mean concentration quotient

K_c	Concentration quotient
K_c^*	Conditional stability constant or conditional concentration quotient
$\bar{K}_c^*$	Mean conditional concentration quotient or stability function
mT	Millitesla

See also: **Chemical Equilibria**; **Colloid-Facilitated Sorption and Transport**; **Electron-Spin Resonance Spectroscopy**; **Fluorescence Spectroscopy**; **Fourier Transform Infrared Spectroscopy**; **Heavy Metals**; **Organic Matter:** Principles and Processes; **Oxidation–Reduction of Contaminants**; **Sorption:** Metals; **Surface Complexation Modeling**

Further Reading

Buffle J (1988) *Complexation Reactions in Aquatic Systems: An Analytical Approach*. Chichester, UK: Ellis Harwood.

Clapp CE, Hayes MHB, Senesi N, Bloom PR, and Jardine PM (eds) (2001) *Humic Substances and Chemical Contaminants*. Madison, WI: SSSA Publications.

MacCarthy P and Perdue EM (1991) Complexation of metal ions by humic substances: fundamental considerations. In: Bolt GH, De Boodt MF, Hayes MHB, and McBride MB (eds) *Interactions at the Soil Colloid–Soil Solution Interface*, pp. 469–492. NATO ASI Series, vol. 190. Dordrecht, The Netherlands: Kluwer.

Senesi N (1990) Molecular and quantitative aspects of the chemistry of fulvic acid and its interaction with metal ions and organic chemicals, part II. The fluorescence spectroscopy approach. *Analytica Chimica Acta* 232: 77–106.

Senesi N (1990) Application of electron-spin resonance (ESR) spectroscopy in soil chemistry. In: Stewart BA (ed.) *Advances in Soil Science,* vol. 14, pp. 77–130. New York: Springer-Verlag.

Senesi N (1992) Metal–humic substance complexes in the environment. Molecular and mechanistic aspects by multiple spectroscopic approach. In: Adriano DC (ed.) *Biogeochemistry of Trace Metals*, pp. 429–496. Boca Raton, FL: Lewis.

Sposito G (1986) Sorption of trace metals by humic materials in soils and natural waters. *CRC Critical Reviews in Environmental Control* 16: 193–229.

Stevenson FJ (1976) Binding of metal ions by humic acids. In: Nriagu JO (ed.) *Environmental Biogeochemistry. Metals Transfer and Ecological Mass Balances,* vol. 2, pp. 519–540. Ann Arbor, MI: Ann Arbor Science.

Stevenson FJ (1994) *Humus Chemistry. Genesis, Composition, Reactions*, 2nd edn. New York: Wiley Interscience.

Stevenson FJ and Fitch A (1986) Chemistry of complexation of metal ions with soil solution organics. In: Huang PM and Schnitzer M (eds) *Interactions of Soil Minerals with Natural Organic Microbes*, pp. 2–58. Madison, WI: SSSA Publications.

ORGANIC RESIDUES, DECOMPOSITION

A J Franzluebbers, USDA Agricultural Research Service, Watkinsville, GA, USA

Published by Elsevier Ltd.

Introduction

Organic residues are carbon-containing compounds of biological origin. Decomposition is the breakdown of these complex organic materials into simpler components. Decomposition of organic residues in soil is an important ecological function whereby heterotrophic organisms consume various components, resulting in the physical and biochemical breakdown of organic materials and transformation and cycling of constituent elements. Biochemical composition, environmental factors, and diversity of organisms (*See* **Biodiversity**) play major roles in the dynamics and fate of organic residues in soil (*See* **Humification**; **Organic Matter:** Principles and Processes; Genesis and Formation). Decomposition is an important soil process that affects nutrient availability on a local scale (*See* **Nutrient Availability**) and environmental quality on a global scale by controlling carbon transformations (*See* **Carbon Cycle in Soils:** Dynamics and Management; **Carbon Emissions and Sequestration**), greenhouse gas emissions (*See* **Greenhouse Gas Emissions**), and nutrient fluxes (*See* **Nitrogen in Soils:** Cycle). The most common methods of determining decomposition are by mass loss of residues or by complete mineralization of organic constituents into end products of inorganic carbon (CO_2, CH_4), nitrogen (NH_4, NO_3), or phosphorus (PO_4).

Types of Organic Residues

Organic residues that enter the soil ecosystem are derived primarily from: (1) autotrophic organisms, including a wide diversity of plants, algae, and lichens; (2) animal tissue and manures; and (3) various organic additions from human activities, including pesticides,

oil products, and synthetic compounds (*See* **Pollutants:** Biodegradation). The presence of heterotrophic organisms in soil, such as fauna and microorganisms, contributes directly to the type of organic residues through synthesis of unique compounds in their bodies, as well as indirectly through transformations of organic residues during decay and subsequent humus synthesis (*See* **Bacteria:** Soil; **Fauna**; **Fungi**; **Nematodes**).

Plant residues are, by far, the most dominant input of organic residues to the soil ecosystem. The quantity of residues produced varies with environmental conditions, but typically ranges from 200 to 2000 $g \cdot m^{-2}$ per year. The allocation of total plant biomass between roots and shoots depends upon a number of factors, including annual or perennial vegetation, soil fertility, climate, and various other soil physical and chemical characteristics. Generally, annual plants tend to have shoot to root ratios of more than 1, while perennial plants tend to have lower shoot to root ratios. Rooting depth of plants depends upon factors similar to those affecting shoot to root ratio. In general, about 50% of total root biomass is found in the surface 20 cm of soil. Accurate characterization of root inputs is difficult because of the intimate contact between roots and the soil-decomposer community, which acts quickly on substrates that are unprotected in the moist soil environment. The portion of root biomass that decomposes rapidly and does not persist in soil is often called 'rhizodeposition.' Rhizodeposits are typically composed of sloughed root cap cells, organic acids actively secreted by growing roots, lysates of root tissues, and high molecular weight root mucilage.

Plant residues are composed of: (1) highly labile intracellular compounds such as protein, starch, fructan, and chlorophyll and other pigments; (2) moderately resistant structural compounds such as cellulose and hemicellulose; and (3) resistant structural compounds such as lignin, polyphenols, lipids, and cutin. If isolated from the whole plant, individual components would decompose at different rates due to their susceptibility to microbial utilization (Figure 1). The rate of decomposition of whole-plant materials varies in response to the relative proportion of different intracellular and structural compounds, as outlined in the following.

For many field crop residues, C concentration is relatively constant at 41% of dry matter (Figure 2a). Protein concentration (estimated as 6.25 times the N concentration), however, varies greatly among residues by species, plant part, age, and management. The amount of C and N mineralized from crop residues varying in N concentration during incubation in soil is highly related to the C to N ratio of residues (Figure 2b).

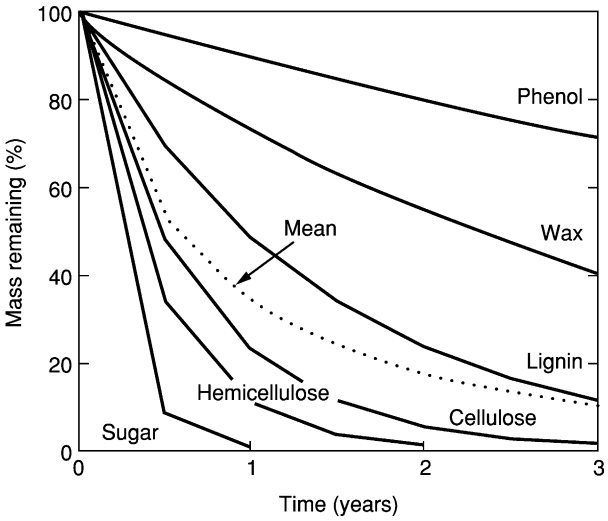


Figure 1 Theoretical decomposition of various components of plant cells, expressed as mass remaining during a 3-year period. Phenol, wax, and lignin groups decompose slower than the average rate of decomposition of plant material, and cellulose, hemicellulose, and sugar decompose faster. Mean proportion of the whole plant as phenol is 5%, as wax is 5%, as lignin is 40%, as cellulose is 20%, as hemicellulose is 15%, and as sugar is 15%. These proportions will vary depending upon plant species, age, fertility, and environmental conditions. Decomposition of organic residues typically follow first-order kinetics ($y_t = e^{-k \cdot t}$, where y = mass remaining at time t and k is a nonlinear rate constant). (Source: Minderman G (1968) Addition, decomposition and accumulation of organic matter in forests. *Journal of Ecology* 56: 355–362.)

Leguminous herbaceous and woody plant species are often applied to soil as mulch or incorporated as green manure to improve soil fertility, especially in the humid tropics where decomposition conditions are ideal. Decomposition of these materials is dependent not only upon N concentration (positive), but also on the concentration of polyphenolics (negative), which are reactive compounds that can form stable polymers with various N compounds and slow decomposition (Figure 3).

The amount of lignin in plant tissue has a negative impact on the rate of decomposition, although some tissues may have a high protein concentration (Figure 4). Lignin is a high molecular weight polymer composed of phenyl propane units in cell walls, giving plants rigidity and resistance to microbial attack. Aside from polysaccharides, lignin is one of the most abundant biopolymers in nature. Lignin is relatively resistant to microbial decomposition, being limited primarily to a group of white-rot fungi that can completely decompose it (*See* **Fungi**). In addition, a diverse community of soil organisms is able to work in concert to decompose parts of the lignin structure slowly (*See* **Biodiversity**).

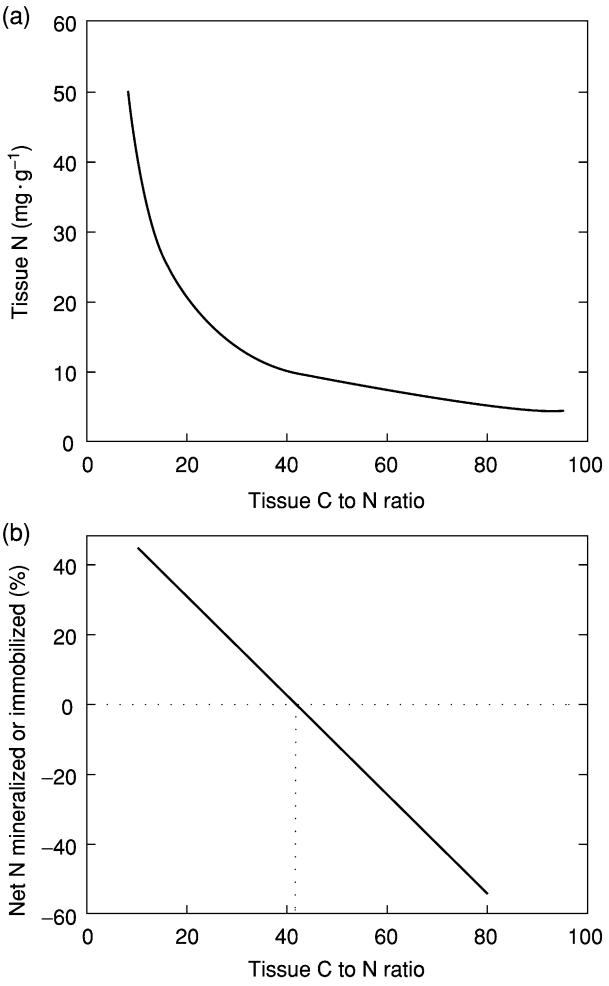


Figure 2 Relationship of C to N ratio of various crop residues with (a) N concentration and (b) net N mineralization during aerobic incubation. The close relationship between C to N ratio and N concentration (a) indicates uniform C concentration (410 mg·g^{-1}) and varying N concentration. Under defined incubation conditions (i.e., optimum temperature and moisture for 20 ± 10 weeks), the percentage of N mineralized from plant tissues varying in C to N ratio (b) is positive when the C to N ratio is less than 42 (i.e., net mineralization of N) and is negative when the C to N ratio is more than 42 (i.e., net immobilization of N). (Source: Vigil MF and Kissel DE (1991) Equations for estimating the amount of nitrogen mineralized from crop residues. *Soil Science Society of America Journal* 55: 757–761.)

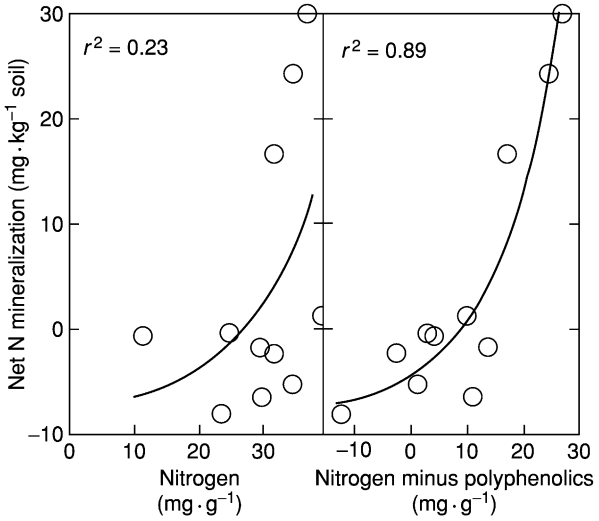


Figure 3 Relationship of N mineralization from different legume green manures (a) with N concentration and (b) with N concentration minus polyphenolic concentration. Although N concentration alone is typically a good predictor of potential N mineralization, condensation of N with polyphenolics can inhibit mineralization. Therefore, information on polyphenolic content of residues is needed to obtain a better prediction of N mineralization. (Source: Palm CA and Sanchez PA (1991) Nitrogen release from the leaves of some tropical legumes as affected by their lignin and polyphenolic contents. *Soil Biology and Biochemistry* 23: 83–88.)

Environmental Factors

Water directly affects decomposition through its controls on activity and transport of soil microorganisms, solubilization of organic constituents, oxygen supply, and soil pH (*See* **Microbial Processes:** Environmental Factors). An optimum water content for decomposition is within a range that is not too high to limit oxygen availability and that is not too low to limit substrate availability and mobility of soil microorganisms. The ideal soil moisture condition for decomposition is between −10 and −50 kPa (or 30–60% water-filled pore space). Fluctuating moisture conditions in soil are common, but they do not necessarily have a negative impact on the ability of microorganisms to respond quickly to available substrates upon rewetting (Figure 5).

Temperature controls the biochemical activity of intra- and extracellular enzyme activities, as well as metabolic activity of most soil organisms responsible for the biodegradation of organic residues. For every 10°C rise in temperature, decomposition generally increases one- to two-fold.

Oxygen is a requirement for the dominant heterotrophic soil organisms responsible for decomposition. Oxygen serves as the most efficient electron acceptor during microbial respiration. When oxygen becomes limiting in soil, decomposition proceeds at a much slower rate. Alternative electron acceptors, including NO_3^-, Fe^{3+}, Mn^{4+}, SO_4^{2-}, and CO_2, can be utilized by anaerobic bacteria such as denitrifiers, sulfate reducers, and methanogens.

Soil pH affects decomposition of organic materials because of its direct effect on the activity of heterotrophic microorganisms. As soil pH decreases to less than 4, microbial activity is severely retarded and accumulation of organic matter can occur, as

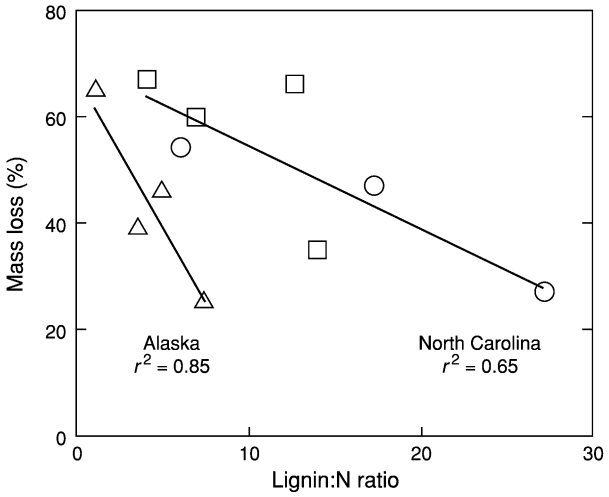


Figure 4 Decomposition of surface-placed residues as affected by their lignin-to-N ratio. Lignin is a polymer that is difficult for most soil microorganisms to decompose. Therefore, high lignin content reduces decomposition rate. At very low lignin content, decomposition of plant material would be similar whether the climate would be cold (Alaska) or warm (North Carolina). However, with increasing lignin content, decomposition is greatly reduced in a cold climate (Alaska) relative to a warm climate (North Carolina). Data from North Carolina are mass loss at the end of 1 year (circles: Blair JM (1988) Nitrogen, sulfur, and phosphorus dynamics in decomposing deciduous litter in the Southern Appalachians. *Soil Biology and Biochemistry* 20: 693–701.) and C loss at the end of 44 weeks (squares: Buchanan M and King LD (1993) Carbon and phosphorus losses from decomposing crop residues in no-till and conventional till agroecosystems. *Agronomy Journal* 85: 631–638.). Data from Alaska are mass loss at the end of 1 year. (Koenig RT and Cochran VL (1994) Decomposition and nitrogen mineralization from legume and non-legume crop residues in a subarctic agricultural soil. *Biology and Fertility of Soils* 17: 269–275.)

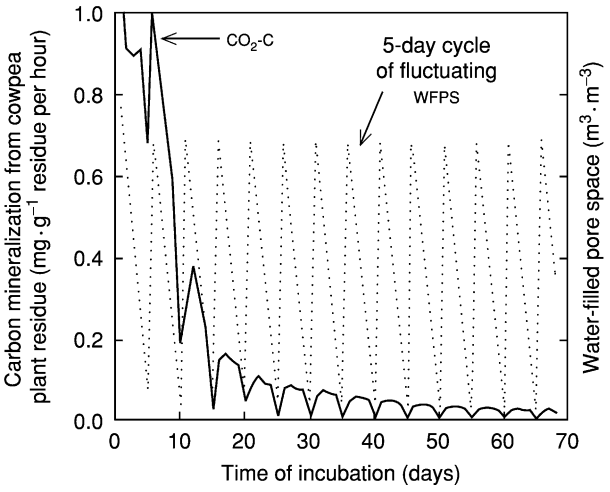


Figure 5 Carbon loss rate of decomposing leaf tissue during successive cycles of drying and wetting. Lowest points of each cycle have a water potential of −10 MPa. Extreme drying reduces decomposition, but decomposition resumes very rapidly upon rewetting. WFPS, water-filled pore space. (Source: Franzluebbers K, Weaver RW, Juo ASR, and Franzleubbers AJ (1994) Carbon and nitrogen mineralization from cowpea plants part decomposing in moist and in repeatedly dried and wetted soil. *Soil Biology and Biochemistry* 26: 1379–1387.)

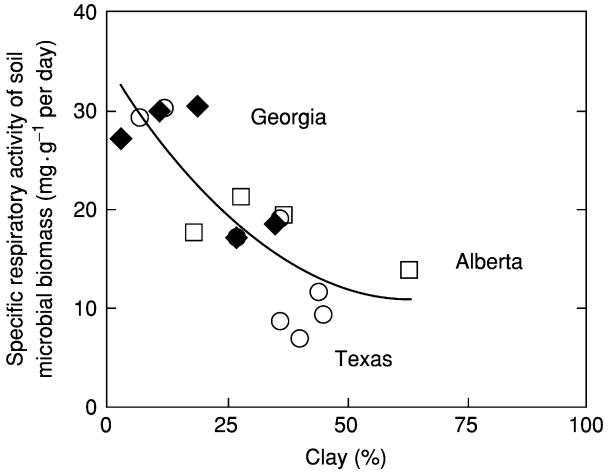


Figure 6 Relationship of microbial-specific mineralization with soil clay content (AJ Franzluebbers, unpublished data). Soils with fine texture have higher clay content and typically more aggregated soil that can protect a portion of the soil organic matter within aggregates. The higher specific respiratory activity of microbial biomass in coarse-textured soil than in fine-textured soil suggests a lower level of protection of the microbial community against predation by other microorganisms and by soil fauna.

especially evidenced by the thick litter layer that often develops under forests on acidic soils (e.g., mor).

Organic materials can become isolated, and even protected from decomposition to some degree, once they become incorporated into water-stable soil aggregates (*See* **Aggregation:** Microbial Aspects). Protection within these aggregates may be due to the physical barrier that limits soil faunal activity and, at times, due to low oxygen concentration that limits soil microbial activity (*See* **Structure**). Soil texture (*See* **Texture**) offers a similar physical limitation on organic matter decomposition, where the pores in coarse-textured soils are freely accessed by a wide variety of micro-, meso-, and macrofauna. Pores in fine-textured soils are often more diverse, with micropores that limit faunal activity and protect organic matter from decomposition typically not present in coarse-textured soil (Figure 6).

Soil management can have large impacts on residue decomposition by altering the physical environment and controlling the supply of substrates. Whether organic residues are left on the soil surface or incorporated into soil (*See* **Cultivation and Tillage; Conservation Tillage**) affects the moisture and temperature regimes they occupy. Generally, incorporated residues decompose faster owing to more consistent moisture

and temperature conditions in the soil than on the soil. The type of cropping can alter the timing, quantity, and quality of residues added to soil (*See* **Crop Rotations**). Residues that become available to soil organisms for decomposition during dry or cold periods will decompose at a different rate and possibly undergo different transformations because of different participating organisms than residues that become available during hot or wet periods of the year. Whether crops are perennial or annual can influence: (1) the level of soil disturbance, (2) the type, timing, and quantity of substrates supplied to the soil, and (3) the placement of residues.

Organism Interactions

Various types of organisms feed upon specific components of organic materials in soil. This complexity of interactions among organisms can be illustrated as a food web (Figure 7), where bacteria and fungi are principal decomposers at many points within the web and progressively larger organisms tend to feed on more specific components of organic material within the food web (*See* **Food–Web Interactions**). The soil microbial biomass plays a pivotal role in the decomposition of organic inputs from plant and animal residues and the breakdown and transformation of organic matter to and from slow and passive pools (Figure 8).

The presence of soil fauna (*See* **Fauna**) often facilitates decomposition by: (1) comminuting plant residues, which exposes a greater surface area to attack by soil microorganisms; (2) transporting organic residues to new locations in the soil, which facilitates decomposition, interaction with soil nutrients, and isolation from certain environmental conditions; (3) inoculating partially digested organic residues with specific bacteria and enzymes (*See* **Enzymes in Soils**); and (4) altering physical characteristics of soil by creating burrows, fecal pellets, and distribution of soil particles, which influence water (*See* **Water, Properties**), air (*See* **Aeration**), nutrient (*See* **Nutrient Availability**), and energy (*See* **Energy Balance**) retention and transport in the soil.

Nutrient Cycling

Nitrogen is a key element that is held within organic residues primarily as proteins and amino acids and released to the soil environment as NH_4^+ through the process of mineralization (*See* **Nitrogen in Soils:** Cycle). Decomposition of organic residues is

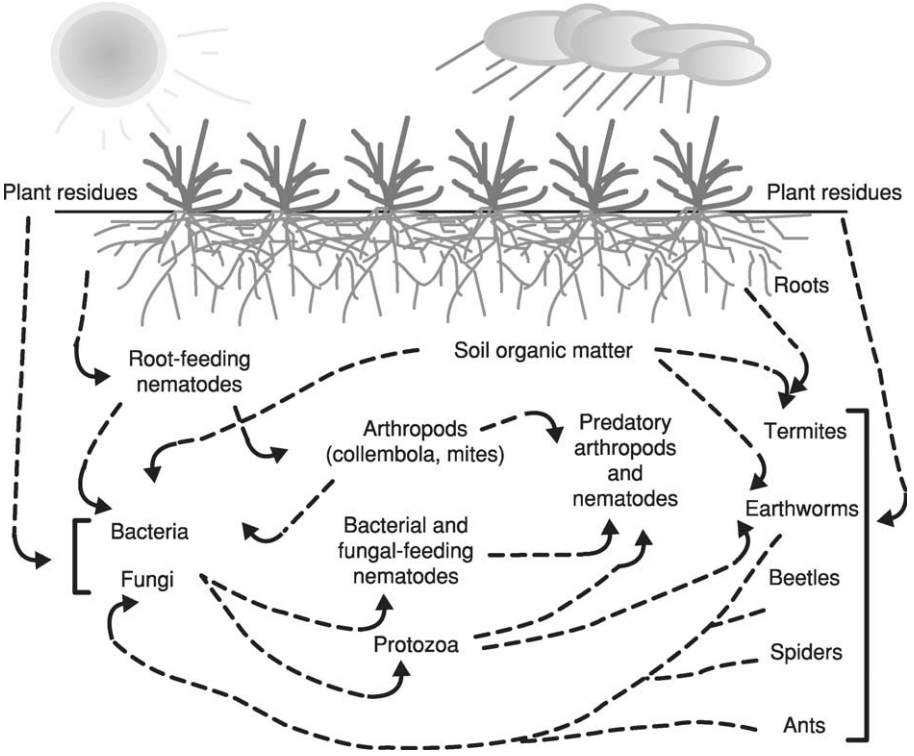


Figure 7 A soil food web. Plant residues are fed upon by various organisms residing in soil. In turn, the growth and remains of these organisms are fed upon by other heterotrophic organisms in soil until nearly all energy embedded within the organic molecules has been exhausted.

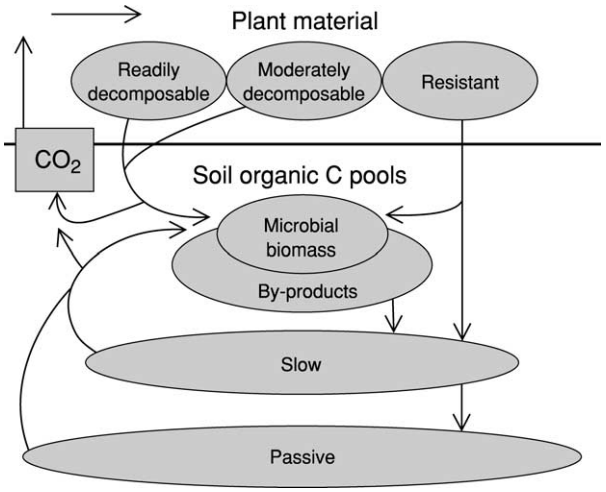


Figure 8 Pools of carbon in soil and the central role of microbial biomass in transforming organic residues into soil organic matter and releasing carbon dioxide into the atmosphere.

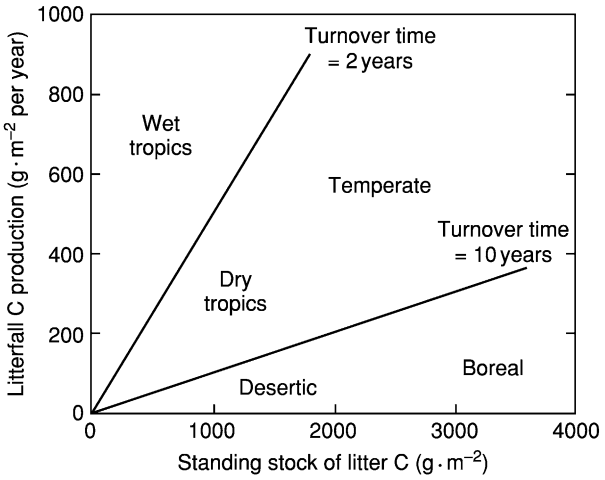


Figure 9 Decomposition estimates based on annual litterfall and standing stock of litter at the forest floor under different environmental conditions in the world. (Source: Olson JS (1963) Energy storage and the balance of producers and decomposers in ecological systems. *Ecology* 44: 322–331.) With high stock of litter at the forest floor and low litterfall production, slow decomposition occurs due to limited moisture or cold temperature. With low stock of litter and high litterfall production, rapid decomposition occurs due to plentiful moisture and warm temperature.

necessary to release organically bound N into inorganic N, which can then be taken up by plants.

Phosphorus is bound organically in plants primarily as deoxyribonucleic acid (DNA) and the energy-transfer molecules adenosine diphosphate (ADP) and adenosine triphosphate (ATP). Mineralization of P from organic residues can supply not only a portion of plant requirements, but also the rapidly growing microbial population that has a high demand for P to meet its ATP production requirements (*See* **Phosphorus in Soils:** Overview).

Carbon is the primary element in organic materials, representing 35–45% of the dry weight of most plant materials, 40–50% of the dry weight of animal manures, and 45–55% of the dry weight of soil humus. In general, C undergoes complete cycling from atmosphere to plants to soil organic matter to atmosphere (*See* **Carbon Cycle in Soils:** Dynamics and Management; **Greenhouse Gas Emissions**). The residence time of C in the terrestrial sphere can be relatively short in warm–moist environments and longer in cold–dry environments (Figure 9), but globally the quantity of C fixed within plant biomass each year is believed to be balanced by the quantity of C released to the atmosphere through decomposition.

Decomposition is an important process that returns organically bound nutrients back to nature for utilization and cycling. On a global scale, this vital ecosystem function balances the autotrophic production of organic matter through photosynthesis. An imbalance in these opposing processes would lead to either: (a) accumulation of organic residues, with a consequent reduction in plant-available nutrients, which could ultimately reduce net primary productivity without external nutrient inputs; or (b) depletion of soil organic matter reserves, which could facilitate greenhouse gas emissions and global warming, thereby depleting inherent soil fertility and reducing net primary productivity. Decomposition is a vital ecosystem function that sustains life on Earth.

Future Advances

Decomposition of organic residues has been historically and extensively studied, and yet much remains to be learned about the decomposition mechanisms and fate of residues in different environments. Of particular deficiency in our current knowledge are the dynamics and interactions with soil of plant root production and decomposition. More detailed chemical composition and fate of different organic residues are needed. Analytical techniques have relied in the past upon acid digestion procedures that are not particularly well suited for analyzing specific molecular classes of compounds. Future analytical techniques will probably rely more on a combination of spectroscopic, chemolytic, and thermolytic analyses to describe both initial chemical properties and the temporal changes in properties during decomposition.

List of Technical Nomenclature

%	Percent
ADP	Adenosine diphosphate

ATP	Adenosine triphosphate
C	Carbon
C	Celsius
d	Day
DNA	Deoxyribonucleic acid
g	Gram
kg	Kilogram
kPa	Kilopascal
m	Meter
mg	Milligram
Mpa	Megapascal
N	Nitrogen
P	Phosphorus
yr	Year

See also: **Aeration**; **Aggregation:** Microbial Aspects; **Bacteria:** Soil; **Biodiversity**; **Carbon Cycle in Soils:** Dynamics and Management; **Carbon Emissions and Sequestration**; **Conservation Tillage**; **Crop Rotations**; **Cultivation and Tillage**; **Energy Balance**; **Enzymes in Soils**; **Fauna**; **Food–Web Interactions**; **Fungi**; **Greenhouse Gas Emissions**; **Humification**; **Microbial Processes:** Environmental Factors; **Nematodes**; **Nitrogen in Soils:** Cycle; **Nutrient Availability**; **Organic Matter:** Principles and Processes; Genesis and Formation; **Phosphorus in Soils:** Overview; **Pollutants:** Biodegradation; **Structure**; **Texture**; **Water, Properties**

Further Reading

Alexander M (1994) *Biodegradation and Bioremediation*. San Diego: Academic Press.

Coleman DC and Crossley DA Jr (1996) *Fundamentals of Soil Ecology*. San Diego: Academic Press.

Kögel-Knabner I (2002) The macromolecular organic composition of plant and microbial residues as inputs to soil organic matter. *Soil Biology and Biochemistry* 34: 139–162.

Paul EA and Clark FE (1996) *Soil Microbiology and Biochemistry*. San Diego: Academic Press.

ORGANIC SOILS

D L Mokma, Michigan State University, East Lansing, MI, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Organic soils consist of organic materials derived primarily from plants, in contrast to mineral materials, from which most soils of the world formed. These soils form where production of organic matter exceeds decomposition, usually under conditions of water saturation: as a result, most organic soils are wetlands. Organic soil materials have more than 12–18% organic C depending on the clay content of the mineral fraction (Figure 1). Organic soils are commonly called peatlands. Several terms are commonly used to separate different organic soils. Highly acid organic soils are referred to as bogs. Less acid organic soils are called fens if they lack trees and are dominated by grasses and/or sedges, or swamps if they are at least partially forested. Marsh refers to wet mineral soils that are dominated by grasses and/or sedges.

Formation

The main soil-forming process active in organic soils is accumulation of organic materials. Paludization is the accumulation of organic materials under anaerobic conditions where preservation allowed a net gain

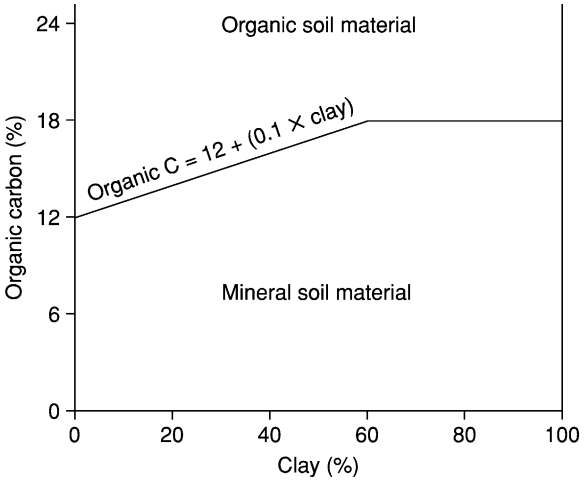


Figure 1 Organic carbon requirements for organic soil materials based on clay content.

through time. Little transformation of parent materials has occurred because of the anaerobic environment in which they form. Decomposition of organic materials is controlled by interrelated factors, primarily moisture content, temperature, acidity, microbial activity, composition of the deposit, and time. Organic soil formation began when microbial activity was promoted by entry of air into the organic deposit. The processes are physical, chemical, and biological ripening. Physical ripening involves a decrease in volume and depends on the nature of organic matter, content of mineral material, and depth of the water table. Chemical ripening is the decomposition of organic materials. Biological ripening involves the reduction in particle size and mixing of organic materials by living animals. The earth can be divided in two organic soil-forming zones, temperate zone and tropical zone. In temperate zones the median rate of organic soil formation is about 0.6 mm $year^{-1}$ and the carbon contribution is about 21.4 g m^{-2} $year^{-1}$. In tropical zones the rate of formation is about 2.2 mm $year^{-1}$ and the carbon contribution is about 128 g m^{-2} $year^{-1}$.

Parent Materials

Accumulated organic materials are primarily composed of plant remains. Small amounts of animal materials are included. The initial deposit of organic materials is considered geogenic rather than pedogenic. The deposit is the parent material. The dominant pedogenic process is the alteration of recognizable organic forms of leaves, stems, and roots, to unrecognizable organic materials. The characteristics of organic soils depend primarily on the nature of the vegetation that was deposited in the water and the degree of decomposition. Vegetation growing on organic soils varies with local climatic conditions, groundwater, and topography. Vegetation may differ in different portions of a single area. Changes in vegetation may also occur as the organic materials accumulate.

Topography

Organic soils may form in closed depressions, coastal marshes, on slopes where there is seepage, or blanket a dissected landscape. The common feature is water that may come from several sources. Climate controls where organic soils form on a regional scale; however, topography determines where they form on a specific landscape. Organic soils form on parts of landscapes where there is concentration of run-on, retention of precipitation, or discharge of groundwater. These conditions most often occur in topographic depressions or very flat areas.

Organisms

Organic soils in their natural environment can exhibit large numbers of contrasting flora. The flora includes those submerged and floating on water, growing along the shoreline and on the bog surface. The sequence of flora on the bog surface is usually meadow, heath, shrub, and finally forest association. Chemical composition of the organic and inorganic fractions, water level, and pH greatly affect plant associations. As a result, a single area of organic soil may have several plant associations at any one time. Differences in vegetative cover change organic soil characteristics that influence utilization of the soil. Generally, the first stage in organic soil formation is plankton growth in water and chara, hornwort and others in deep water and pond weeds and water lilies in shallow water (**Figure 2**). After some accumulation of organic materials, conditions become favorable for growth of sedges, reeds, and grasses (**Figure 3**). Later the area is covered with heath. Finally, succession leads to shrubs and forests (**Figure 4**).

Plant communities with high primary production usually have high decomposition rates, therefore, organic matter accumulation is negligible. Communities with low or modest primary production frequently have slow decomposition rates, therefore, organic matter accumulation is significant. The anaerobic environment of organic soils preserves pollen, therefore they contain a record of past climates and flora.

Climate

Organic soils can form in virtually any climatic region; however, water must be available in the region for the organic materials to accumulate. Accumulation of organic materials is favored by cold and wet climates because they inhibit decomposition of the materials. In the high latitudes of Asia, Europe, and North America average precipitation exceeds annual potential evaporation. In the northern part of these latitudes organic soils are not common because short growing seasons inhibit the production of organic materials. In lower latitudes organic soils are primarily limited to coastal plains that are nearly level with high annual precipitation and no dry season or to environments with high water tables maintained by tidal waters or coastal marshes. In arid, semiarid, and arctic areas, organic soil formation is very slow or negligible due to climate.

Time

Most organic soils have formed since the end of the last ice age. In high latitudes they are concentrated in regions that were covered by glaciers and formed

Figure 2 First stage of organic soil development: water lilies.

Figure 3 Sedges and grasses on organic soil.

following retreat of glaciers. In lower latitudes they were also impacted by glaciation. Large quantities of water were incorporated in glacial ice. Melting of glaciers caused sea level to rise at a rate that was more rapid than that of initial vegetation colonizing tidal regions. When the sea level rise slowed, marsh vegetation could become established and organic materials began to accumulate, producing the parent materials for organic soils. About 3.6 m of organic materials accumulated between 4400 BP and 1914 near Belle Glade, FL. In this low-latitude region the average rate of accumulation was about 1.4 mm year^{-1}. In Michigan, a more northern latitude, the average rate was 0.55 mm year^{-1}. In the slightly glaciated region of southwest Wisconsin, the average rate of accumulation was about 0.68 mm year^{-1}. In Iceland it was 0.39 mm year^{-1}.

Distribution

Organic soils comprise about 2% of the Earth's ice-free land surface. This includes 0.8% organic soils with permafrost. Organic soils can form in virtually any climate, even in arid areas, as long as water is available. They are most prevalent in the cool, humid boreal forest areas of northern Asia, Europe, and

Figure 4 Leather leaf on acidic organic soil.

North America. Every province of Canada and almost every state in the USA has organic soil. More than 95% of the total peat reserves of the world are located in the temperate zone of the northern hemisphere. In these areas precipitation usually exceeds evapotranspiration and summers are relatively cool. At lower latitudes organic soils occur locally on humid coastal plains, for example, South-East Asia, Indonesia, and east North America. These coastal areas are inundated by tidal waters.

In many central European countries organic soils have been extensively used and no longer qualify as organic soils. Nearly all organic soils in the Netherlands, Germany, and Switzerland have been lost. In more northern European countries the organic soils are relatively virgin. About 75% of Europe's remaining organic soil resources are located in Finland, Norway, and Sweden. Because of the loss of organic soils resulting from intensive utilization, it is difficult to obtain an exact amount of organic soil resources remaining.

Utilization

Most organic soils occur in wetland environments, thus they provide wetland functions of protecting surface and ground water quantity and quality, preserving biological diversity, reducing surface water runoff, assisting in flood protection, and providing erosion and sedimentation control. Early humans avoided organic soils because they were low, wet, and harbored pests. Malaria was a major problem in some areas. Some people believed evil spirits resided in these areas. Individuals seeking refuge from slavery or the law frequently used them. About the turn of the eighteenth century, interest in organic soils (Figure 5) was stimulated by the need for fuel. Wood,

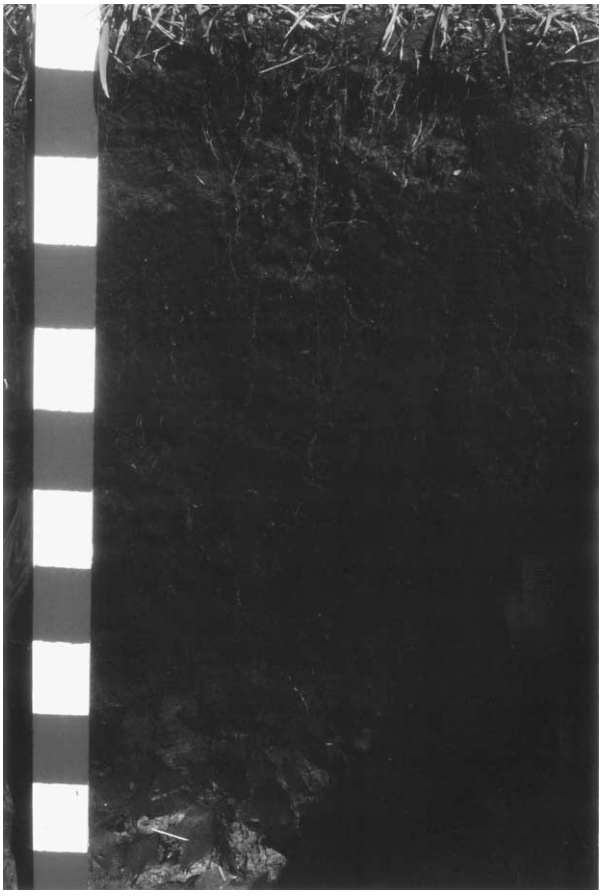


Figure 5 Profile of organic soil in Finland.

Figure 6 Use of organic soil for fuel.

the common fuel of that day was in short supply. Peat is still used for fuel (Figure 6) and for generation of electricity in some areas.

Utilization of organic soils is affected by several of their properties: soil reaction, cation exchange capacity, bulk density, water retention, water content, hydraulic conductivity, shear strength, and shrinkage. A profile description of a cultivated organic soil is given in Table 1. Selected physical and chemical properties of the same pedon are given in Table 2. The pH of organic deposits is largely a function of the nature of the water that flowed into the organic soil area and the vegetation. Organic soils in landscapes dominated by acid, sandy Spodosols are typically very acid and have a pH of less than 4.5. Also, organic soils that receive water only from precipitation, no groundwater flow, tend to be acidic. Coastal organic soils that received sulfur from seawater are usually very acid and may be acid sulfate soils. Organic soils in landscapes that receive subsurface flow that contains significant amounts of bases and/or bicarbonate are usually less acidic. Because organic soils are composed of primarily decomposed vegetative materials, composition of vegetation exerts a strong influence on soil reaction. Acid-loving plants, such as sphagnum moss (Figure 7), produce more acidic materials than base-loving plants. Very acid organic materials have low levels of microbial activity compared to less acidic materials. Less decomposed organic soil materials are more acidic than more decomposed materials. The availability of phosphorus and many micronutrients is dependent on pH. Acid organic soils are usually not limed above pH 5.5 in order to reduce the likelihood of causing micronutrient deficiencies in crops.

Table 1 Profile description of a cultivated organic soil

Oap	0–23 cm. Black (N 2/0), black (10YR 2/1) rubbed; about 5% fibers, 1% rubbed; weak fine granular structure; friable; few fine roots; estimated mineral content 20%; abrupt smooth boundary
Oa1	23–48 cm. Black (N 2/1), very dark brown (10YR 2/2) rubbed; about 20% fibers, 4% rubbed; weak coarse blocky structure; friable; few fine roots; estimated mineral content 25%; clear smooth boundary
Oa2	48–76 cm. Very dark brown (10YR 2/2) broken face and rubbed; about 40% fibers, 10% rubbed; weak thick platy structure; friable; very few fine roots; estimated mineral content 15%; clear smooth boundary
Oa3	76–115 cm. Very dark brown (10YR 2/2) broken face and rubbed; about 25% fibers, 4% rubbed; weak thick platy structure; friable; estimated mineral content 15%; clear smooth boundary
Oa4	115–190 cm. Dark brown (7.5YR 3/2), very dark brown (10YR 2/2) rubbed; about 35% fibers, 7% rubbed; weak thick platy structure; friable; estimated mineral content 12%; clear smooth boundary

The cation exchange capacity of organic soils results primarily from carboxyl radicals. Other functional groups that contribute are phenolic, enolic, and quinonic hydroxyls and heterocyclic nitrogen structures. Organic soils have high cation exchange capacities compared to mineral soil. The cation exchange capacity is pH-dependent and ranges from 10 to 20 mmol per 100 g at pH 3.5–4.0 to 100–200 mmoL per 100 g at pH 7.0.

One of the most outstanding characteristics of organic soils is their generally low bulk density. Bulk densities range from as little as 0.02 to 0.30 g cm^{-3}. Some studies found that highly decomposed materials

Table 2 Selected properties of a cultivated organic soil

Horizon	Depth (cm)	Bulk density ($g\ cm^{-3}$)	Water content (%)	$CaCl_2$ pH	Organic C (%)	Fiber		Extract color[a]
						Unrubbed (%)	Rubbed (%)	
Oap	0–23	0.30	255	6.2	45.0	–	5	7.5YR 3/2
Oa1	23–48	0.17	498	6.2	43.8	24	3	7.5YR 3/2
Oa2	48–76	0.12	699	5.8	39.0	20	3	7.5YR 4/2
Oa3	76–115	0.13	667	5.7	32.0	38	2	7.5YR 4/2
Oa4	115–190	0.12	705	5.6	49.5	40	2	7.5YR 4/3

[a]Color of pyrophosphate extract.

Figure 7 Sphagnum moss.

have a greater bulk density than less decomposed materials; however, a recent study found no difference in bulk density with difference in degree of decomposition. Bulk density of organic soils increases with increasing mineral content.

Organic soils are wet not only because they exist in saturated conditions, but also because they have a great affinity to take up and hold water. Water retention is related to degree of decomposition. Large pores in slightly decomposed organic soil materials drain easily at low suction, whereas small pores in more decomposed materials retain water at low suctions. Water contents of organic soils are related to degree of decomposition. Fibric soil materials retain from 8.5× to more than 30× on a weight basis, hemic soil materials hold between 4.5 and 8.5×, and sapric soil materials usually retain less than 4.5×. Water contents ranged from 100% (by volume) for undecomposed, near-surface sphagnum moss peat to less than 80% for deeper, more decomposed peat. Hydraulic conductivity of organic soils is also related to degree of decomposition of the organic materials because of differences in pore size distribution. Saturated hydraulic conductivity of sapric material ranges from 0.5 to 15 $cm\ h^{-1}$, that of hemic materials ranges from 1.5 to 15 $cm\ h^{-1}$, and that of fibric materials is greater than 15 $cm\ h^{-1}$. Weakly or undecomposed peats had rather rapid hydraulic conductivities, whereas more decomposed peats had slow conductivities. After drainage, the hydraulic conductivity of an organic soil decreases, due in part to greater decomposition and in part to the decrease of the buoyant effect of the groundwater.

Shear strength is a measure of a soil's internal resistance to stress. In mineral soils, shear strength decreases as water content approaches saturation. In organic soils, this relationship is not as clear. Draining a fibrous peat reduces the water content but increases the shear strength because of the 'fibrous interlock.' However, drainage also increases decomposition that destroys the 'fibrous interlock.' As decomposition increases, organic soils behave more like mineral soils with respect to shear strength.

Most organic soils undergo considerable shrinkage upon drying. Well-decomposed organic soil materials may shrink 70% or more depending on the mineral content of the material. The degree to which the shrinkage is irreversible is partially related to degree of decomposition. Organic soils are vast reservoirs of carbon. Carbon is stored during formation of these soils. During warmer periods small amounts of organic materials accumulate, whereas during cool, moist periods considerable amounts accumulate. Excavation of organic materials for various purposes

(fuel, manufacture of peat products for agriculture and horticulture, medical applications) accelerates decomposition. Drainage leads to increases in decomposition of organic materials and emission of carbon dioxide. Methane and carbon dioxide stored in organic soils are released by drainage. Release of carbon dioxide as a result of these activities increases the concentration of greenhouse gases.

One of the early more intensive uses of organic soils was grazing, especially during the summer when water tables were naturally lower. Low bearing capacity and strength limited the usefulness of these soils. Artificial drainage not only lowered the water table but increased decomposition, thereby increasing bearing capacity and strength. Initially drainage was accomplished with open ditches and later with tile systems. Ditches have been constructed in some forests to increase productivity (Figure 8). Tile systems were more costly and required more intensive use of the land. Agronomic crops in high-latitude areas suffered frost damage in the spring after planting (Figure 9) and fall before maturity, thus the change from low-value crops fed to animals to high-value crops that tolerate frost or have a shorter growing season and are consumed by humans. Rice, carrots, onions, celery, potatoes, lettuce, mint, sugarcane, and

Figure 8 Ditch installed in forest to lower water table and increase production.

Figure 9 Frost damage of crop on organic soil.

cranberries are grown extensively on organic soils (Figures 10 and 11). Peat is also mined for use as a soil amendment in agriculture and horticulture. The low bulk density of organic soils makes them ideal for grass sod production (Figure 12) where large volumes of soil are moved long distances.

Utilization of organic soils requires careful management. Wise management of water is essential. In their natural state, organic soils are too wet. When they are drained, they become aerobic and organic materials decompose, causing rapid subsidence (Figure 13).

The most rapid subsidence occurs immediately after drainage. Subsidence in Hunts, England was 3.4 cm $year^{-1}$, Quebec, Canada 2.1 cm $year^{-1}$, Ontario, Canada 3.3 cm $year^{-1}$, Minnesota 5.1 cm $year^{-1}$, Indiana 3.0 cm $year^{-1}$, Florida 3.0 cm $year^{-1}$, and California 2.3 cm $year^{-1}$. An area in Michigan had to be retiled 28 years after it was established as a result of subsidence (about 3 cm $year^{-1}$). Organic soils have limited life expectancies. Rate of subsidence, degree of decomposition, and depth of organic materials determine the life of organic soils. Careful attention to wind erosion control and water table

Figure 10 Carrots growing on organic soil.

Figure 11 Organic soil used for cranberry production.

Figure 12 Low bulk density makes organic soil ideal for sod production.

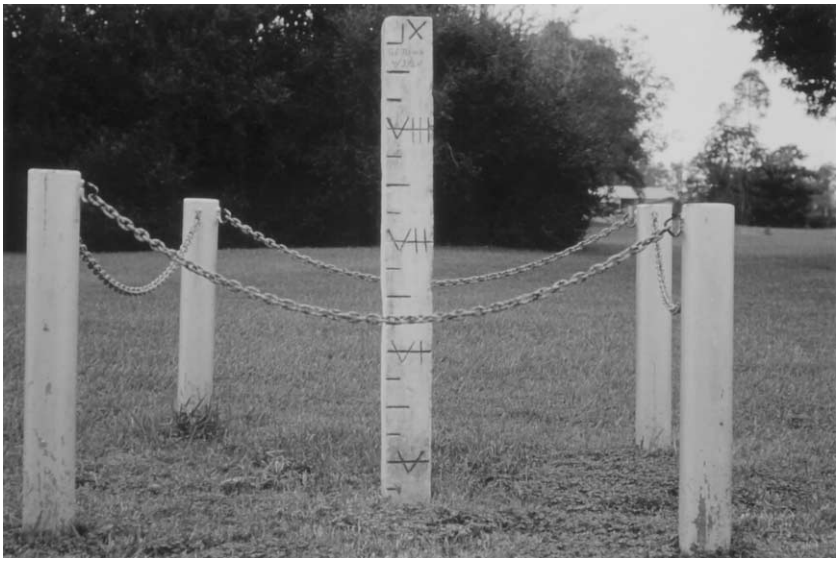


Figure 13 In 1924 this 9-feet (3-m) post was placed on the underlying bedrock with the top of the post at the surface of the organic soil at Belle Glade, FL. This photograph, taken in 1972, shows that subsidence has been about 2.5 cm year^{-1}.

height can pay dividends in conservation of organic soils. To minimize subsidence water tables should be maintained as high as crop and field conditions will permit. Raising the water tables to at or near the soil surface (Figure 14) during the noncrop period can be effective in reducing subsidence and wind erosion. After drainage, organic soils are very susceptible to wind erosion because of the low bulk density. Relatively low wind velocities cause movement of the light-weight materials if unprotected. When wind erosion is very severe, organic soil material in the air reduces visibility, thereby causing highways to be closed.

Fire hazard increases after drainage of organic soils. These fires are difficult to control and may burn underground for several months. Some fires were accidental, caused by sparks from rail locomotives or farm tractors. Other fires were deliberately started to reduce excess acidity and/or increase nutrient levels. Some pasture and old hay lands were burned to destroy unwanted plant residues and brush. As a result of fires, undesirable materials such as sand, gravel, clay, and marl were exposed. These fields usually had marked variability in pH. Some burned areas had greater than pH 8 because of marl or other calcareous materials. Marl in the plow layer

Figure 14 Raising water tables to at or near the organic soil surface will reduce subsidence and wind erosion.

Figure 15 Low bearing capacity of organic soils causes cracks in roads that require patching.

causes P and micronutrients to be unavailable for plant growth. In some Michigan fields available soil P levels ranged from 2 to 300 kg ha^{-1}. Levels below 10 kg ha^{-1} occurred in plow layers that contained marl, whereas the high levels occurred in plow layers that lacked marl. Fertilizer application in a field had been based on soil test analyses of the problem area.

The major engineering problem with the use of organic soils is compressibility. Compressibility is related to nature of the organic materials, degree of decomposition, water content, structure, and density. Compression and subsidence of organic soil materials continue for many years after loading. Construction of structures and roads on organic soils is usually difficult. Structures are usually constructed on piles that were driven into the underlying mineral soil to support foundations. Floating road designs have been used on large areas of organic soils. These roads frequently crack because of the low bearing capacity (Figure 15). If the area is not large, the organic materials can be removed and replaced with sand and gravel. Ultimately, structures are left standing above the ground level. Drainage accelerates decomposition and shrinkage of the organic materials, therefore subsidence of driveways, sidewalks, and/or lawns leaves structures perched above soil surfaces

Figure 16 Subsidence of lawns and sidewalks in Michigan necessitated addition of steps built on pilings to enter the apartment building.

Figure 17 Subsidence of lawn and driveway in New Orleans, Louisiana, requires capping of steps and decreases the accessibility of carport.

(Figures 16 and 17). Garages and carports may stand above the driveway and be inaccessible to cars. Many of these garages and carports have been converted into another room, since both the house and garage or carport were built on piles. Acidity of organic soils causes corrosion of concrete and metal materials. Organic soils have high porosity and high adsorptive capacity, therefore they are useful for treatment of wastewater. They have the potential to remove heavy metals and hydrocarbon. Mined peat is used for treatment of septic tank effluent before dispersing into soil.

All these activities have one thing in common – they bring stored carbon into atmospheric circulation.

Classification

Organic soils are saturated and have 40 cm or more of organic soil materials according to Food and Agriculture Organization of the UN–Unesco Soil Map of the World, Revised Legend and Soil Taxonomy. If the organic materials consist of at least 75% moss or have a bulk density of less than 0.1 g cm^{-3}, the thickness must be at least 60 cm. The World Reference

Table 3 Criteria used in the classification of organic soils

Food and Agriculture Organization of the UN Unesco	*Soil Map of the World, Revised Legend*
Soil units	Presence of permafrost
	Presence of a sulfuric horizon or sulfidic materials
	Degree of saturation
	Degree of decomposition of organic materials
World Reference Base	
Soil units	Presence of permafrost
	Presence of a sulfidic or sulfuric horizon
	Presence of a salic horizon (highly soluble salts)
	Presence of a folic horizon (well-aerated, organic materials)
	Degree of decomposition of organic materials
Soil Taxonomy	
Suborder	Degree of saturation
	Degree of decomposition of organic materials
Great group	Soil temperature regime
	Special materials (sphagnum fibers, sulfuric horizon, sulfidic materials, humilluvic materials)
	Soil moisture regime
Subgroup	Presence of water
	Presence of lithic contact (bedrock)
	Thickness of organic materials
	Presence of mineral materials within organic materials
	Presence of sphagnum
	Presence of limnic layers
	Degree of decomposition of organic materials
	Electrical conductivity

Base for soil resources broadened the requirement by decreasing the minimum depth to 10 cm. The criteria used in the classification of organic soils differ in the three soil classification systems (Table 3). Organic soil materials that are not saturated must contain at least 20% organic C.

Further Reading

Aandahl AR, Buol SW, Hill DE, and Bailey HH (eds) (1974) *Histosols: Their Characteristics, Classification and Use*. Soil Science Society of America special publication 6. Madison, WI: Soil Science Society of America.

Everett KR (1983) Histosols. In: Wilding LP, Smeck NE, and Hall GF (eds) *Pedogenesis and Soil Taxonomy. II. The Soil Orders*, pp. 1–53. Amsterdam: Elsevier.

FAO (1990) FAO-UNESCO *Soil Map of the World, Revised Legend*. World Soil Resources Report 60. Rome: Food and Agriculture Organization.

FAO-ISRIC-ISSS (1994) *World Reference Base for Soil Resources*. World Soil Resources Report 84. Rome: Food and Agriculture Organization.

Farnham RS and Finney HR (1965) Classification and properties of organic soils. *Advances in Agronomy* 7: 115–162.

Gore AJP (ed.) (1983) *Mires: Swamp, Bog, Fen and Moor*. Amsterdam: Elsevier.

Heinselman ML (1970) Landscape evolution, peatland types and the environment in the Lake Agassiz Peatlands Natural Area, Minnesota. *Ecology Monograph* 40: 235–261.

Lucas RE (1982) *Organic Soils (Histosols). Formation, Distribution, Physical and Chemical Properties and Management for Crop Production*. MI State University Agricultural Experimental State Research Report 435. East Lansing, MI: Michigan State University.

Moore PD and Bellamy DJ (1974) *Peatlands*. New York: Springer Verlag.

Rabenhorst MC and Swanson D (2000) Histosols. In: Sumner ME (ed.) *Handbook of Soil Science*, pp. E-183–E-209. Boca Raton, FL: CRC Press.

Soil Survey Staff (1999) *Soil Taxonomy: A Basic System of Soil Classification for Making and Interpreting Soil Surveys*. USDA Agricultural Handbook 436. Washington, DC: US Government Printing Office.

OVERLAND FLOW

T S Steenhuis, L Agnew, P Gérard-Marchant, and M T Walter, Cornell University, Ithaca, NY, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

For many years, it was believed that the occurrence of surface runoff is primarily controlled by the infiltration characteristics of the ground. Specifically, runoff is generated whenever rainfall or irrigation water is applied to the ground at a higher rate than the soil's infiltration capacity. Robert E. Horton was responsible for some of the early foundational work behind this concept and hence the process is ubiquitously referred to as 'Hortonian runoff.' The synonymous term 'infiltration-excess overland flow' is also widely used, again emphasizing that surface runoff occurs when precipitation exceeds the infiltration capacity of the soil. This process is very important in many areas of the country: arid areas (where significant soil crusting and/or surface sealing occurs during rain events), in irrigated fields, in urban areas, and more generally during storms with very high rainfall intensities.

However, the Hortonian runoff concept does not meaningfully explain storm runoff in many regions where the infiltration capacity of the ground is typically much greater than the average rainfall intensities. Many researchers have found that hydraulic conductivities published in soil surveys, usually based on disturbed soil samples, underestimate, by a factor of 10 or more, the conductivity in the field for vegetated soils which contain preferential flow paths in the form of worm channels and root passages. Consequently, only the most intense summer storms can produce runoff under the Hortonian runoff scenario. This is in direct contrast to observations of stream response to almost all rain, even small storms, during the set periods of the year, which means that there must be some other mechanism generating runoff.

In many regions, runoff is most commonly generated on relatively small portions of the landscape that are susceptible to becoming completely saturated. Once the soils in these areas saturate to the surface, any additional rainfall (irrespective of intensity) becomes overland flow. This process is termed 'saturation-excess overland flow.' Therefore, the propensity of an area to produce runoff is largely independent of rainfall intensity. Instead, the total rainfall amount and landscape factors such as soil depth (i.e., available water storage capacity), upland watershed area, and local topography are the important factors determining whether or not a particular area in a watershed will generate runoff. Moreover, as rainfall continues, the saturated area grows in extent, increasing the area generating runoff (hence the term 'variable source area,' VSA). This is in contrast to Hortonian or infiltration-excess runoff generation, which depends on soil type (infiltration rate versus rain intensity) and is independent of position in the landscape, i.e., with the runoff generating area is always the same. Examples of regions in the USA where VSA hydrology is significant include the Northeast and the Pacific Northwest, as well as forested mountain areas. It also occurs throughout the world in many similar areas.

Saturation-Excess Overland Flow

Saturation-excess overland flow comes from two distinguishable sources. In one case, rain falling on already saturated soil has no option but to run off – this case is termed 'direct precipitation on saturated areas' (DPSA). The other source, termed 'return flow,' occurs if the rate of interflow entering a saturated area from upslope exceeds the capacity for interflow to leave the area by flowing downhill through the soil. The excess interflow, thus, 'returns' to the surface as runoff, hence the term. Whereas DPSA runoff only occurs during and just after a rainfall event, return flow seepage can continue as long as an interflow excess exists.

Areas prone to saturation either have a high groundwater table or hard pan (fragipan) at shallow depth. Interflow results in the formation of saturated areas at the bottom of slopes, usually in concave areas, or it quickly resurfaces in seeps and ditches. Some hydrological processes observed in upstate New York, primarily the Catskill Mountains region, can further serve to elucidate mechanisms resulting in saturation. Soils in the Catskills are generally permeable (relative to rainfall intensity) and underlain by a shallow, low-permeable, restrictive layer, typically bedrock or fragipan. Rainwater easily permeates the soil and, by-and-large, runs laterally as interflow on top of the restrictive layer downslope. Many researchers have observed evidence of the accumulation of interflow water at the bottom of slopes in the Catskills during wet periods in the form of increased moisture content at hill bottoms relative to the steep parts of hills. Some common locations where saturation occurs are areas where the soil above the restricting layer is shallow, in places where the downhill topographic slope decreases such as the toe-slope of a hill, or in topographically

converging areas (Figure 1). All three incidences are locations in the landscape where the Darcy flow capacity, or interflow capacity, is reduced either by a decrease in hydraulic transmissivity or hydraulic gradient. When interflow capacity is sufficiently restricted, the soil will saturate. During periods of enhanced rainfall, interflow will be higher and often expand the extent of saturation around saturation-prone areas; conversely, dry periods will decrease interflow and extent of saturation. This is illustrated for wet and dry seasons in Figure 2.

VSA hydrology is an extension of the saturation-excess concept, recognizing that the extent of saturated areas in a watershed will expand and contract, i.e., vary temporally. The variation in the extent of saturated areas has been studied over a range of temporal scales, ranging from storm duration, hours and days, to seasons (Figure 2). Although the original, underlying concept for VSA hydrology was developed by the US Forest Service, the term 'variable source area' is often attributed to Hewlett and Hibbert. As stated above, VSA hydrology is applicable to many landscapes throughout humid regions that have undulating topography.

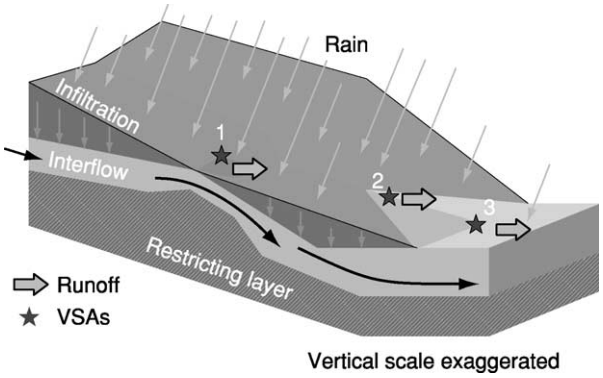


Figure 1 Incidence of saturation-excess hydrology: 1, shallow soil; 2, convergence area; and 3, decreasing downhill slope. VSA, variable source area.

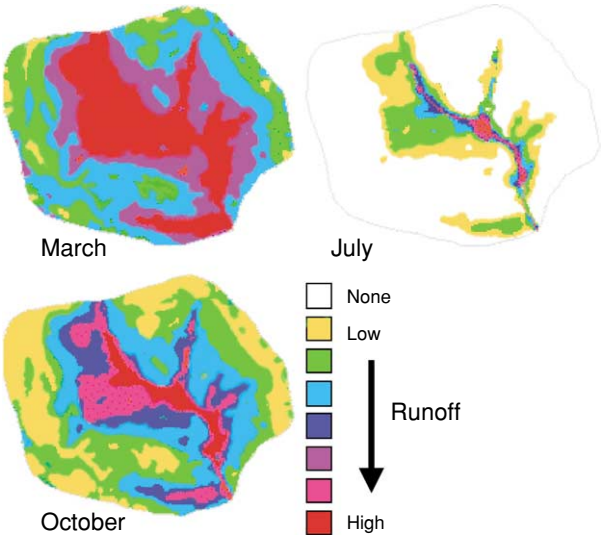


Figure 2 Seasonal changes of saturated areas (variable source areas) in a watershed in New York state. Color denotes areas more prone to saturation.

The VSA hydrology concept, thus, goes beyond conceptualizing the water–soil interface at or below the soil surface and making the simple assumption that water either runs off or infiltrates. The VSA hydrology concept includes the soil reservoir in a spatial landscape context to describe locations where the soil reservoir becomes saturated and begins to generate runoff. The soil reservoir methodology utilizes the bedrock, impermeable soil layers (relative to above layers), and/or the depth to the water table as the underlying hydrologic boundary instead of the soil surface. Thus, both hydrologic and soil water (i.e., porous media) concepts are combined to evaluate potential runoff areas in the landscape.

Calculation Method

TOPMODEL is used most often to predict the saturated-excess overland flow. It is based on calculating a groundwater table height that, when it intercepts the soil surface, will result in runoff. In TOPMODEL, the assumption is made that the watershed is underlain by a groundwater table. This is not realistic for many mountainous areas. However, since most of the saturated areas occur near streams, this limitation is not severe as a groundwater table is present in these areas. How the water gets to the groundwater is often inconsequential. An alternative method to find the saturated areas is the Soil Moisture Routing (SMR) model, a physically based, fully distributed, geographic information system (GIS) integrated code specifically designed for soils underlain by a hardpan. This model solves the water balance for cells in a grid that covers a watershed. What makes it different from most other models is that the spatial location of the grid cells is preserved in the landscape and water is routed by interflow from one cell to the next, resulting in the simulation of saturated areas within the landscape. This model, despite its simplicity, predicts the saturated areas in the landscape amazingly well.

TOPMODEL and SMR are admittedly more complicated than the popular Soil Conservation Service 'Curve Number' method (SCS-CN), which was developed in the 1950s and 1960s by the USDA Soil Conservation Service (currently the Natural Resource Conservation Service). The traditional SCS-CN

method is a rainfall-runoff model originally developed for predicting flood-flow volumes from ungauged watersheds for hydraulic engineering design. Despite the limited scope of intended application and several identified theoretical problems with the traditional SCS-CN method, it is ubiquitously used in water quality models to generate the fraction of rainfall that runs off. The original model was justified by Victor Mockus, to whom the traditional SCS-CN method is largely attributed, "on grounds that it produces rainfall-runoff curves of a type found on natural watersheds." Subsequently, hydrologists have shown that the basis of the method can be described in ways that are nominally consistent with both infiltration-excess theory and saturation-excess or VSA hydrology.

Figure 3 Manure deposition in exactly the wrong place from a hydrologically sensitive-area standpoint (courtesy of Glenn Warner).

Water Quality Impacts of Variable Source Area Hydrology

VSA hydrology has been shown to have significant consequences on surfacewater quality, having direct impacts on specific areas such as animal waste management, drinking water and human health, nutrient and pesticide management, pollution assessment and prevention, and watershed management. It is important that the development and implementation of the next generation of land-management practices for improving or preserving surface water quality be consistent with our most current hydrological understanding. Interestingly, many current practices aimed at reducing contaminant transport are based on Hortonian concepts. However, the effectiveness of these practices is diminished where Hortonian processes are not governing runoff generation.

Wherever VSA hydrology is a dominant process, there will be regions within a watershed that are more susceptible to producing runoff and delivering it to surfacewater bodies than other regions. These areas can be considered hydrologically sensitive areas (HSA). Recognizing the existence of HSAs allows watershed-scale water quality efforts to be focused on those areas where HSAs coincide with land uses that potentially contribute pollutants. The intersection of an HSA and a pollutant loading area is referred to as the 'critical management zone.' The most obvious best-management practice for this area would be to limit or prohibit potentially polluting activity from this region. A contrasting approach is finding methods and means of eliminating the hydrological sensitivity of the critical zones, although recent research suggests attempts to remove hydrologic sensitivity through drainage practices may simply reroute pollutants from overland flow to subsurface flow, with little reduction in concentration. In either case, the HSAs warrant primary attention when trying to preserve or improve water quality.

One example of where potential management practices derived from a recognition that VSA hydrology is in direct conflict with currently mandated management practices is whether dairy operators in the New York city watersheds should spread manure on steep slopes or in flat areas. The current dogma is to avoid spreading on steep slopes and to maximize spreading in flat areas. Under Hortonian flow, steep areas might arguably produce the most rapidly moving runoff and, therefore, the greatest potential for erosion and transport of manure. However, VSA hydrology is the dominant process in these watersheds and the steep slopes infiltrate essentially all rainwater and drain very rapidly, resulting in almost no substantial runoff. Conversely, the flat areas, especially at the base of hillslopes, are especially prone to saturation and, thus, prone to runoff generation. Figure 3 shows manure deposition in, from a HSA standpoint, exactly the wrong place. A manure-spreading policy that is more consistent with the recognized hydrology would promote spreading high in the watershed and minimize spreading on low, flat areas susceptible to runoff generation.

Further Reading

Beven KJ (2001) *Rainfall-Runoff Modelling: The Primer*. Chichester, UK: Wiley.

Dingman SL (2002) *Physical Hydrology*, 2nd edn. Englewood Cliffs, NJ: Prentice Hall.

Dunne T (1970) *Runoff Production in Humid Areas*. US Department of Agriculture Publication ARS-41-160. Washington, DC: USDA.

Dunne T and Leopold LB (1978) *Water in Environmental Planning*. New York: W.H. Freeman.

Frankenberger JR, Brooks ES, Walter MT, Walter MF, and Steenhuis TS (1999) A GIS-based variable source area model. *Hydrological Proceedings* 13: 804–822.

Hewlett JD and Hibbert AR (1967) Factors affecting the response of small watersheds to precipitation in humid regions. In: Sopper WE and Lull HW (eds) *Forest Hydrology,* pp. 275–290. Oxford, UK: Pergamon Press.

Horton RE (1933) The role of infiltration in the hydrologic cycle. *Transactions of the American Geophysical Union* 14: 446–460.

Kirkby MJ (1978) *Hillslope Hydrology*. New York: Wiley.

Kuo WL, Steenhuis TS, McCulloch CE, Mohler CL, Weinstein DA, DeGloria SD, and Swaney DP (1999) Effect of grid size on runoff and soil moisture for a variable-source-area hydrology model. *Water Resources Research* 35: 3419–3428.

OXIDATION–REDUCTION OF CONTAMINANTS

C J Matocha, University of Kentucky, Lexington, KY, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

In the soil environment, inorganic and organic contaminants can be transformed by reduction and oxidation (redox) processes that can occur chemically, photochemically, or biologically. Historically, microbial oxidation of organic matter in soils and sediments ranks the sequence of oxidants as molecular oxygen (O_2) reduction, nitrate (NO_3^-) reduction, manganese dioxide (MnO_2) reduction, iron hydroxide ($Fe(OH)_3$) reduction, sulfate (SO_4^{2-}) reduction, and methane (CH_4) production. This reduction sequence is based on the oxidant yielding the most free energy change per mole of organic carbon oxidized and proceeds until all oxidants or all oxidizable organic matter are consumed in well-defined depth zones. It is well accepted that, in certain field situations, this thermodynamic reduction sequence is an oversimplification and does not coincide with the kinetic reduction sequence, because it provides no information about reaction rates and other potentially competing abiotic pathways.

In abiotic reaction pathways, the role of catalysis is important. Naturally occurring solid Fe and Mn oxides have been shown abiotically to catalyze transformations of organic and inorganic pollutants, which themselves could serve as electron donors and acceptors. Most reactions involving these metal oxides are heterogeneous, with complex, multistep processes such as adsorption, precipitation, electron transfer, and diffusion. The focus here is the role of reactive soil Fe and Mn species in the redox transformations of inorganic and organic contaminants.

Redox Status of Reactive Fe and Mn Species

Selected half-reactions and standard reduction potentials of important, solid-phase Mn(III,IV) and Fe(III) oxides in soils are compared below with other important redox half-reactions:

$$\frac{1}{2}\,\delta\text{-}MnO_x(s) + 2H^+ + e^- \rightarrow \frac{1}{2}\,Mn^{2+} + H_2O \qquad E_h^0 = 1.29\,V \qquad [1]$$

$$Fe(OH)_3(s) + 3H^+ + e^- \rightarrow Fe^{2+} + 3H_2O \qquad E_h^0 = 1.057\,V \qquad [2]$$

$$\frac{1}{4}\,O_2(g) + H^+ + e^- \rightarrow \frac{1}{2}\,H_2O \qquad E_h^0 = 1.229\,V \qquad [3]$$

$$\frac{1}{8}\,SO_4^{2-} + \frac{5}{4}\,H^+ + e^- \rightarrow \frac{1}{8}\,H_2S + \frac{1}{2}\,H_2O \qquad E_h^0 = 0.303\,V \qquad [4]$$

Birnessite (δ-$MnO_x(s)$) is the most commonly identified Mn(III,IV) oxide in soils, controlling solubility of plant-available Mn(II). Ferrihydrite, an amorphous Fe hydroxide ($Fe(OH)_3(s)$), is depicted as controlling Fe(II) solubility. Half-reactions are useful to represent the stoichiometry and thermodynamics of redox reactions. The standard state potential (E_h^0) in volts fixes the activity of reactants, products, and H^+ at unity by definition. The Nernst equation can be written for each half-reaction and E_h–pH diagrams constructed under more realistic environmental conditions ((Mn^{2+}), (Fe^{2+}), and $(SO_4^{2-}) = 10^{-5}\,mol\,l^{-1}$ and partial pressures of O_2 and H_2S of 0.02 and 0.0001 MPa):

1. δ-$MnO_x(s)/Mn^{2+}$: $E_h = 1.378 - 0.118\,pH$
2. $Fe(OH)_3(s)/Fe^{2+}$: $E_h = 1.352 - 0.177\,pH$
3. $O_2(g)/H_2O$: $E_h = 1.219 - 0.059\,pH$
4. SO_4^{2-}/H_2S: $E_h = 0.362 - 0.7375\,pH$

When summing two half-reactions, it is assumed that they are in equilibrium with one another (coupled). As long as the calculated E_h for the oxidant half-reaction is higher than that of the reductant half-reaction, the forward reaction is favorable. The free energy change, ΔG, of a redox reaction, can be assessed by knowledge of the standard reduction potentials of the half-reactions involved. E_h is directly related to ΔG of the system by:

$$\Delta G = -nFE_h \qquad [5]$$

where n is the number of electrons involved in the redox reaction and F is the Faraday constant ($F = 96\,500\,C\,mol^{-1}$). Redox reactions where $E_h > 0$ would give a negative ΔG and be thermodynamically favorable. Over a wide range of pH values, including those often encountered in soil environments (pH 4–8), oxidant strength from a thermodynamic point of view decreases in the order $O_2 > \delta$-$MnO_x > Fe(OH)_3 > SO_4^{2-}$ (Figure 1). Several lines of field evidence have shown that relying solely on thermodynamics to predict oxidant consumption can be misleading, and kinetic and catalytic factors need to be considered.

Sulfate reduction preceding Fe(III) reduction in lake pore-water sediments has been reported (Table 1). Overlap of these reduction sequences was ascribed to a wide range of Fe(III) oxide stabilities, giving rise to Fe(III) reduction rates that vary considerably and to the type of oxidizable organic C. The most unstable, poorly crystalline Fe oxides such as ferrihydrite are reduced preferentially to well-crystallized phases such as goethite (α-FeOOH). This situation is further complicated in soils containing Fe-rich phyllosilicate minerals that can undergo reduction and oxidation cycles.

Table 1 Pore-water chemistry with depth of Lake Greifen. Note the occurrence of sulfate reduction precedes that of Fe(III) reduction

Depth (cm)	*SO_4^{2-} (μM)*	*[H_2S] (μM)*	*Fe(II) (μM)*
7	170	0	0
1	145	0	0
−1	67	3.0	0
−2	10	10.0	0
−5	9	9.0	12.0
−10	5	4.0	100
−25	10	2.0	300
−30	10	2.0	350

Adapted from Wersin P, Hohener P, Giovanoli R, and Stumm W (1991) Early diagenetic influences on iron transformations in a freshwater lake sediment. *Chemical Geology* 90: 233–252.

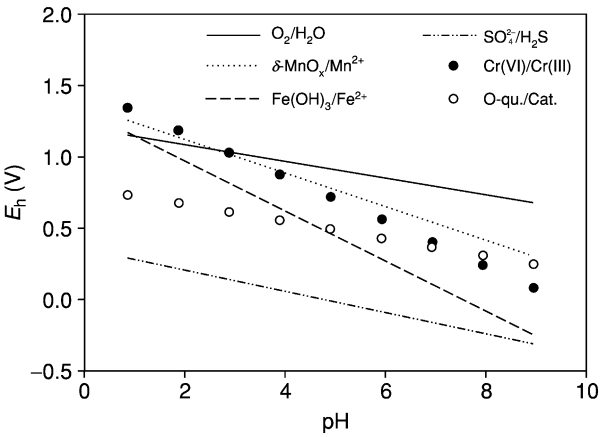


Figure 1 Relationship of redox potential (E_h) to pH for several important redox couples in soils compared with representative inorganic and organic (*ortho*-quinone (O-qu.)/catechol (cat.) couple) contaminants. Dissolved Mn, Fe, S, Cr, and organic species were assigned an activity of $10^{-5}\,mol\,l^{-1}$, and dissolved oxygen and sulfide were given values of 0.02 and 0.0001 MPa.

Reduced products from the primary redox reactions can participate in abiotic secondary reactions, affecting oxidant availability and organic matter degradation pathways. Consumption of dissolved O_2 is dependent on microbial activity and availability of suitable electron donors such as dissolved Fe(II) and Mn(II). Dissolved Fe(II) and Mn(II) can diffuse upwards from anaerobic environments into oxic zones. The thermodynamic driving force for capture of downward-diffusing O_2, which is in ample supply from the atmosphere, by dissolved Fe(II) and Mn(II) is favorable over a wide soil pH range. Yet at pH >6, Fe(II) oxygenation is rapid, while Mn(II) oxygenation occurs only after years and coexistence of dissolved Mn(II) and O_2 is not uncommon. Surface catalysis by hydrous oxides and microorganisms increases Mn(II) oxygenation rates, thus, only heterogeneous oxygenation of Mn(II) is included in reactive transport models because of the slow kinetics of homogeneous Mn(II) oxygenation. Depending on the environmental setting, as much as 50% of the dissolved O_2 can be consumed by these rate-limited secondary reactions and the remainder by microbial respiration.

Kinetics and Mechanism

The goal of collecting fundamental kinetic data is to derive an elementary rate expression with a rate constant and to propose a mechanism. The usefulness of the kinetic approach is that it expresses the rate as

changes in reactant and product concentration over time and empirically determines the reaction orders. Consider the following simple forward reaction:

$$A \xrightarrow{k_1} P \quad \text{rate} = -\mathrm{d}[A]\mathrm{d}t^{-1} = \mathrm{d}[\mathrm{P}]\mathrm{d}t^{-1} = k_1[A]^a \qquad [6]$$

where a is the power dependency or reaction order and k_1 is the rate constant, assuming the reaction stoichiometry is known and constant over the time period investigated. The value of k_1 should depend only on temperature, thus it should be specified, and be time-invariant. The rate expression is defined once the value of a is determined. If unity, then the reaction is first-order in A and first-order overall. Fractional-order rate expressions are not uncommon, particularly in multistep processes such as heterogeneous redox reactions. The rate, $-\mathrm{d}[A]\,\mathrm{d}t^{-1}$, is usually given in terms of moles per liter per second, and the units of k_1 for a first-order reaction are per second, per minute, or per hour. This information is used in reactive transport models that attempt to predict spatial and temporal distributions of redox-active species. It is important to realize that the empirical rate expressions may not always obey the overall reaction stoichiometry. Identification of the rate-limiting step, represented by the rate expression, is important because it represents the slowest step, thus any manipulation of it can either speed up (for instance adding a catalyst) or slow down the overall reaction.

For redox reactions to occur, the oxidant and reductant must encounter one another in order to allow the reaction to proceed. This is the case for both homogeneous and heterogeneous reactions. The general mechanism for heterogeneous redox reactions includes the following steps: (1) precursor-surface complex formation between oxidant and reductant, (2) electron transfer, and (3) product release. Both electronic and steric effects will influence the rate and extent of precursor complex formation, especially when considering sorption behavior at a surface. The precursor complex can involve a direct, inner-sphere bond between oxidant and reductant or an outer-sphere bond, where no bond breaking or making occurs and the transition state is unaffected. The rate-limiting step for outer-sphere electron transfer mechanisms is the electron-transfer step, while either precursor complexation or electron transfer can be rate-limiting in inner-sphere mechanisms. Bridging ligands with filled sigma (σ) and pi (π) orbitals such as oxo surface groups (O^{2-}, OH^-, and H_2O) on minerals can facilitate electron transfer for inner sphere mechanisms because the lone pairs of electrons function as 'bridges.' If the overlap of molecular orbitals for oxidant and reductant is perfect, then inner-sphere electron transfer mechanisms do not require a bridging ligand. Interfacial characteristics can also determine electron transfer rates, an additional consideration for heterogeneous reactions.

Rates of most redox reactions increase with temperature, and rate constants (k_{obs}) can be described by the Arrhenius equation to determine activation energies (E_a):

$$k_{obs} = A_f \exp^{-E_a/RT} \qquad [7]$$

where A_f is the frequency factor, R is the gas constant, and T is absolute temperature. Low E_a values ($<42\,\mathrm{kJ\,mol^{-1}}$) indicate that the rate-limiting step is a diffusion-controlled process, while high E_a values represent surface chemical control of the rate-limiting step. Conducting kinetic experiments over a temperature range of 4–40°C (277–313 K) can simulate conditions in aquifers and soil environments.

Thermodynamic considerations are an important first step in the appraisal of whether a given inorganic or organic contaminant will undergo redox reactions in a soil environment. Delineation of redox zones by analysis of soil solution and pore water for major redox-active species is also helpful in predicting the fate of the contaminant, particularly if rate expressions have been determined already for the specified redox couple.

Inorganic Contaminants

Manganese as an Oxidant

The importance of Mn(III,IV) (hydr)oxides environmentally lies in their ability to adsorb and in some cases oxidize certain trace metals. Chromium (Cr) is a well-known contaminant, and typical oxidation states are Cr(III) and Cr(VI). Hexavalent Cr is mobile in soils because it is an oxyanion and repelled by negatively charged soil particles. In contrast, Cr^{3+} is more strongly adsorbed and forms relatively insoluble oxide and hydroxide compounds. Chromium(III) and Cr(VI) speciation in solution are functions of pH, but will not be accounted for here. The half-reaction and standard reduction potential for the Cr(VI)-Cr(III) couple can be written as:

$$CrO_4^{2-} + 8H^+ + 3e^- \rightarrow Cr^{3+} + 4H_2O \qquad E_h^0 = 1.51\,\mathrm{V} \qquad [8]$$

where $E_h = 1.51–0.1573\,\mathrm{pH}$.

Oxidation of Cr(III) to Cr(VI) by O_2 is more energetically favorable than by Mn(III,IV) (hydr)-oxides or Fe oxides (Figure 1). However, rates of

Cr(III) oxidation by O_2 are slow over a wide pH range. Soil Mn oxides and laboratory-prepared Mn oxides can directly oxidize Cr(III) at faster rates than O_2 or Fe oxides, according to the following overall reaction:

$$Cr^{3+} + 1.5MnO_2 + H_2O \rightarrow CrO_4^{2-} + 1.5Mn(II) + 2H^+ \quad [9]$$

Typical rate equations and E_a values (more than $42\,kJ\,mol^{-1}$) indicate that the slowest step is either precursor-surface complex formation or electron transfer. Temporal variability in Cr(III) oxidation rates is observed as initial Cr(III) concentration and pH increases and is attributed to surface precipitation of $Cr(OH)_3(s)$ that inhibits the reaction. In some cases, formation of an intermediate Mn(III) oxyhydroxide phase is reported to retard the reaction also. These findings are significant because Cr(VI) is more mobile and toxic than Cr(III).

The coupling of Mn reduction with Cr(III) oxidation to Cr(VI) can be explained by frontier molecular orbital theory. Oxides and hydroxides of Mn(III) and Mn(IV) are kinetically better oxidants than O_2, because the lowest unoccupied molecular orbitals (LUMO) for Mn(III) and Mn(IV) are the e_g orbitals, which are of sigma (σ^*) symmetry and available to accept electrons from σ- and π-donor contaminant metals (Figure 2). For Mn(IV), with a $d^3(t_{2g}^3\ e_g^0)$ electronic configuration, both e_g orbitals are vacant and a concerted two-electron transfer can occur. In contrast, the LUMO for O_2 is of π^* symmetry (π_{2px}^{1*} π_{2py}^{1*}) and, being singly occupied, can only accept one electron at a time. Oxo groups present on the Mn oxide surface serve as bridging ligands, because they have filled orbitals of π and σ symmetry to make the Cr(III)-to-Mn(IV) electron transfer facile.

Other examples of Mn oxide-catalyzed oxidations of trace metal contaminants include oxidation of Co(II) to Co(III), As(III) to As(V), and Se(IV) to Se(VI). In all of these reactions, there is a rate dependence on reductant and oxidant concentration. The rates in oxygenated systems are negligible without the presence of the solid-phase Mn oxide.

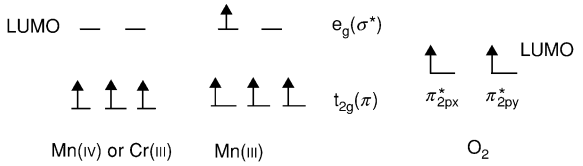


Figure 2 Lowest unoccupied molecular orbitals (LUMO) for Mn(III) and Mn(IV) are shown as e_g orbitals with σ^* symmetry. For O_2, the LUMO are antibonding π^* molecular orbitals resulting from overlap of 2px and 2py atomic orbitals.

Fe(II) as a Reductant

Rate studies of oxidation of Fe(II) in solution at pH > 5 shows a second-order dependence on OH^- concentration and first-order dependence on O_2 concentration:

$$-d[Fe(II)]/dt = k[Fe(II)][OH^-]^2[O_2] \quad [10]$$

because dissolved OH^- donates electron density to Fe(II) through the σ and π systems and enhances the metal's reducing power. Thus, O_2 reacts predominantly with the hydrolyzed species $FeOH^+(aq)$ and $Fe(OH)_2$ (aq) as the pH increases. Fe(II) coordinated with σ- and π-donating ligands, with oxygen as the ligating atom, such as oxo groups in minerals and dissolved organic matter, lowers the aqueous Fe(III)–Fe(II) reduction potential, making coordinated Fe(II) a better reductant than dissolved Fe(II). Naturally occurring, solid-phase Fe(II) minerals such as siderite ($FeCO_3(s)$), pyrite ($FeS_2(s)$), and sulfate green rust ($[Fe_4^{II}\ Fe_2^{III}\ (OH)_{12}]\ [SO_4 \bullet 3H_2O]$) are stronger reductants than aqueous Fe(II) also.

During reductive dissolution of Fe(III) (oxy)hydroxides or Fe-bearing phyllosilicates in in soils and sediments by either microorganisms or abiotic reductants, Fe(II) is released to solution. After initial release, one of the processes that dissolved Fe(II) can undergo is adsorption back on to the surface. Adsorbed Fe(II) is significant in anaerobic soil environments for two reasons. First, in pure mineral systems, dissolved Fe(II) resulting from microbial reduction of goethite has been shown to adsorb to remaining goethite surface sites and limit the extent of reduction. In Fe(III) oxide-rich subsoils with mixed mineralogy, dissolved biogenic Fe(II) adsorbs strongly to kaolinite to limit bioreduction. The second reason is that Fe(II) adsorbed as an inner-sphere surface complex is an effective reductant and can catalyze oxygenation. The conceptual model is shown in Figure 3, where >M—O represents a surface metal (Fe(III)) oxide or phyllosilicate mineral functional group. The oxide or hydroxide group behaves as a bridging ligand and 'pumps' electron density from adsorbed Fe(II) to the O_2. Surface OH groups on metal oxides and phyllosilicate minerals can catalyze electron transfer from Fe(II) to O_2 analogous to

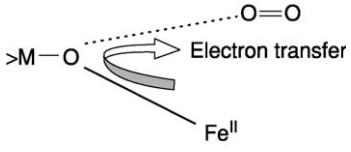


Figure 3 Conceptual model of adsorbed Fe(II) as an inner-sphere surface complex that enhances oxygenation.

dissolved OH^-. Therefore, adsorbed Fe(II) bound to surface OH groups in an inner-sphere surface complex is a better reductant than dissolved Fe(II). A kinetic estimate for the redox potential of adsorbed Fe(II) on goethite is $E_h^0 = 0.36$ V, similar in magnitude to $FeOH^+$ ($E_h^0 = 0.34$ V). Characterization of the reducing power of adsorbed Fe(II) on different mineral surfaces remains a challenge. Adsorbed Fe(II) species as reductants are important to consider in any environmental setting because, at the onset of anaerobic conditions, solid Fe(III) oxides are the most abundant oxidants, in concentrations approximately 60-fold greater than solid Mn oxides.

Investigators have recently shown that adsorbed Fe(II) coordinated to surface hydroxyl functional groups on metal oxides can catalyze the reduction of Cr(VI), nitrite (NO_2^-), uranium(U(VI)), and technetium(Tc(VI)). Reduction of toxic Cr(VI) by reduced Fe(II) species is beneficial because the solubility of Cr is lowered by formation of sparingly soluble Cr(III) precipitates. This multistep process is rapid in the presence of dissolved Fe(II), structural Fe(II), and adsorbed Fe(II). Rate expressions are typically second-order:

$$-\mathrm{d}[\mathrm{Cr(VI)}]/\mathrm{d}t = k_{\mathrm{obs}}[\mathrm{Fe(II)}]_x[\mathrm{Cr(VI)}] \qquad [11]$$

where k_{obs} is the second-order rate coefficient and $[Fe(II)]_x$ represents dissolved, adsorbed, or structural Fe(II). The values of k_{obs} for dissolved, adsorbed, and structural Fe(II) species can be compared qualitatively. Hydrous oxide surfaces such as goethite adsorb Fe(II) in an inner-sphere complex that is very reactive toward reduction of dissolved Cr(VI), with k_{obs} values ranging from 6.6 to 72 $l\,mol^{-1}\,s^{-1}$ (Table 2). Dissolved Fe(II) and solid-phase Fe(II) minerals such as green rust are also effective reductants of Cr(VI) (Table 2) and minimize contamination of soil and water supplies.

Nitrite reduction to N_2O by Fe(II) adsorbed to lepidocrocite (γ-FeOOH) is rapid, and the maximum rate of 10 $\mu mol\,h^{-1}$ occurs at pH 8.5. The postulated mechanism Fe(II) adsorbed as a bidentate inner-sphere complex with surface oxo groups provides an ideal environment for NO_2^- adsorption and subsequent N–N double-bond formation, similar to the function of the Fe metal center in nitrite reductase enzyme. No reduction of NO_2^- occurs in the absence of lepidocrocite or in the presence of dissolved Fe(II). The N_2O produced may take part in global warming and atmospheric ozone destruction, because it is a more potent greenhouse gas than CO_2 and represents an irreversible loss of N from soil.

Table 2 Chromate (Cr(VI)) reduction rates based on observed second-order rate coefficients (k_{obs}) by naturally occurring reactive Fe(II) species

Fe(II) species	k_{obs} *($l\,mol^{-1}\,s^{-1}$)*
Dissolved Fe(II)	3.7[a]
Solid Fe(II) (green rust)	6.3[a]
Adsorbed Fe(II)	
Goethite	6.6–72[a]
Aluminum oxide	3.5–7.5[a]

[a]Source: Buerge IJ and Hug SJ (1999) Influence of mineral surfaces on chromium(VI) reduction by iron(II). *Environmental Science and Technology* 33: 4285–4291.

Organic Contaminants

Background

For reduction and oxidation processes to occur for organic compounds, the same general steps of precursor complexation, electron transfer, and product release are involved. Relevant structural properties of the organic compound that have to be considered are identity and location of functional groups, acidity and basicity of those groups, and overall molecular size and shape of the molecule. These properties in turn affect the water solubility of the organic compound and hence its ability to encounter the relevant oxidant or reductant in the soil environment. Active sites on organic compounds are the functional groups. Electron-donating functional groups that are able to form surface complexes, particularly chelate complexes, are effective reductants of oxidized Fe and Mn species. In general, removal of electrons from organic compounds can occur by two different pathways. The heterolytic pathway involves the transfer of two electrons, while homolytic pathways describe one electron transfer and are often referred to as free-radical pathways. Organic free radicals are usually reactive and unstable.

Oxidation and Reduction in Soil Organic Matter

Soil organic matter (SOM) consists of a wide range of compounds derived from algal, plant, and microbial material with different structures and consequently different properties. Primary functional groups in SOM include O- and N-containing groups that are electron-donating and undergo processes to form humic substances (HS), a subclass of SOM consisting of high-molecular-weight biopolymers that react profoundly with inorganic and organic contaminants. The exact structure of HS is unknown, and the humification process is not completely understood, yet it is a way of sequestering C. Characteristics of the precursor organic carbon and the transformations it has undergone are important factors influencing the

chemical properties of HS. Humification of natural SOM can also tie up organic contaminants which themselves often contain electron-donating functional groups through nucleophilic addition reactions. Implicit in the SOM-decomposition model is the paramount role of microorganisms, because they can derive energy from the reduced C source. Faster SOM degradation rates have been shown to result in greater overlap of reduction sequences with depth, indicating the likelihood of abiotic SOM oxidation.

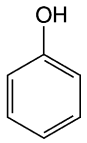
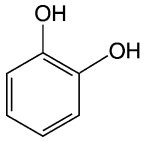
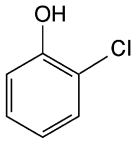
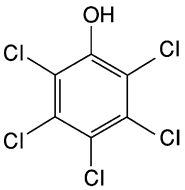
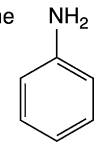
Manganese and Iron as Oxidants

Direct reaction of organic compounds with O_2 is exceedingly slow, especially for organic pollutants. Solid-phase Mn(III,IV) (hydr)oxides can function as important electron acceptors of SOM, chlorinated phenols, and anilines. Catechol (1,2-dihydroxybenzene) is a suitable analogue of SOM, because semiquinones and quinones have been identified in soil HS. Catechol is known for its oxidation and reduction properties and represents a chelating and polymerizing polyphenol.

Naturally occurring birnessite (δ-$MnO_{1.7}$) catalyzes oxidation of catechol more rapidly than O_2 and FeOOH based on measurements of Mn(II) release (Table 3). Efficient dissolution of birnessite is probably aided by the ability of catechol to form a surface chelate during the precursor-surface complex step with metal center surface sites. In addition, electrons can be transferred immediately along the bonding axis as an inner-sphere process from the filled catechol p orbitals of σ symmetry to the vacant LUMO of σ^* symmetry of the Mn(III,IV) metal centers after precursor-surface complex formation. This proper symmetry overlap probably explains why Mn oxides are kinetically better oxidants than O_2 (see Figure 2) and Fe(III) oxides, because the LUMO for Fe(III) is singly occupied (d^5, t_{2g}^3 e_g^2).

Oxidation of substituted phenols by metal ion oxidants is complex, because product distribution and reaction rates are highly dependent on the nature of the oxidant, reactant concentrations, and reaction conditions. Several competing reactions may occur that accompany phenol oxidation and reductive dissolution of Mn(III,IV) (hydr)oxides, such as oxidative coupling of organic radicals or phenolic dimers which form complex mixtures of polymeric oxidation products. In fact, the rapid production of dissolved Mn(II) is accompanied by formation of polymeric reaction products via oxidative coupling reactions from the catechol precursor that resembles soil humic substances (Table 4). Spectroscopic indicators of humification are E_4:E_6 ratios, the ratio of absorbance at 465 nm to that at 665 nm. Lower E_4:E_6 ratios (approx. <5) correspond to higher organic radical

Table 3 Initial oxidation rates of soil organic matter (as catechol), chlorophenols, and aniline by naturally occurring Mn and Fe oxides based on Mn(II) and Fe(II) release

Organic compound	*MnO_x d[Mn(II)] dt^{-1} ($\mu mol\ l^{-1}s^{-1}$)*	*Fe oxides d[Fe(II)] dt^{-1} ($\mu mol\ l^{-1}s^{-1}$)*
Phenol 	0.005[a]	BD[b]
Catechol 	2.1[c]	0.0009[b]
Chlorophenols 	0.003[a]	
	0.001[a]	
Aniline 	2.2[d]	

BD, below detection.
[a]Source: Ulrich HJ and Stone AT (1989) Oxidation of chlorophenols adsorbed to manganese oxide surfaces. *Environmental Science and Technology* 23: 421–428.
[b]Source: Lakind JS and Stone AT (1989) Reductive dissolution of goethite by phenolic reductants. *Geochimica et Cosmochimica Acta* 60: 3169–3175.
[c]Source: Matocha CJ, Sparks DL, Amonette JE, and Kukkadapu RK (2001) Kinetics and mechanism of birnessite reduction by catechol. *Soil Science Society of America Journal* 65: 58–66.
[d]Source: Laha S and Luthy RG (1990) Oxidation of aniline and other primary aromatic amines by manganese dioxide. *Environmental Science and Technology* 24: 363–373.

Table 4 Spectroscopic indicators of humification based on E_4:E_6 ratios for products of the birnessite–catechol mixture compared with extracted humic acid from a typical agricultural soil in Kentucky (Woolper silt loam)

Sample	E_4/E_6
Birnessite–catechol (pH 4)	20.3
Birnessite–catechol (pH 6.5)	3.1
Woolper humic acid	3.17

contents and a greater degree of humification and stabilization of soil humus. The E_4:E_6 ratios of filtrates for the birnessite–catechol reaction mixture at high pH are comparable with humic acid extracted in a typical Kentucky soil. Direct abiotic coupling of oxidizable organic carbon (C) with solid-phase Mn contributes to sequestration of C through humification processes.

Chlorophenols are well recognized as being highly toxic and represent an organic pollutant class that can be abiotically degraded by solid-phase Mn oxides. Their water solubility and reactivity decrease as the number of chloro substituents added to the phenol increase (Table 3). Detoxification of chlorinated phenols by Mn(III,IV) (hydr)oxides occurs by oxidation, dechlorination, and polymerization processes that can result in less-toxic products that can be biodegraded more readily in soils. During the oxidative coupling (polymerization) reactions, xenobiotic substances such as chlorinated phenols form insoluble precipitates that can be removed from water by sedimentation or filtration. Anilines, another organic pollutant class found in industrial byproducts and components in pesticides, can also be oxidized rapidly in the presence of birnessite (Table 3) and undergo further oxidative coupling reactions to form a variety of polymeric products. Although not as abundant, Mn oxides are more reactive catalysts than Fe oxides for a wide variety of organic pollutant redox reactions.

Adsorbed Fe(II) as a Reductant

In anaerobic soils and aquifers, Fe(II) species adsorbed on solid Fe(III) oxide sites are important reductants of oxidized organic contaminants such as halogenated solvents and nitroaromatic compounds. A common theme in past research in this area has been the greater reduction rates with Fe(II) adsorbed as an inner-sphere surface complex when compared with the presence of solid-phase or dissolved Fe(II) alone. Surface-bound Fe(II) is not only more reactive but also may be continuously regenerated by the activity of Fe-reducing microorganisms and subsequent adsorption from solution to form new

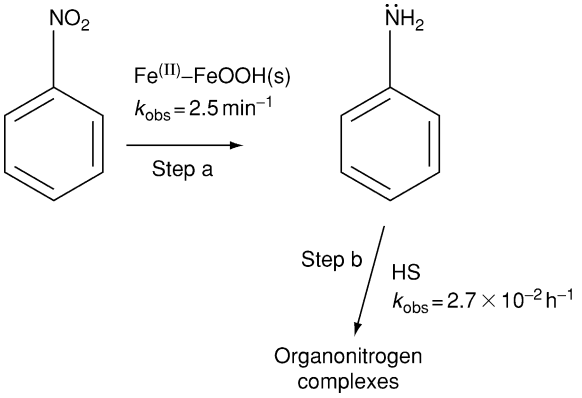


Figure 4 A two-step remediation strategy for nitrobenzene.

reactive Fe(II) species. The surface density of adsorbed reactive Fe(II) can be manipulated by varying pH values due to the pH-dependent sorption behavior. Thus, reduction rates typically increase with pH because of the greater surface density of reactive adsorbed-Fe(II).

For nitroaromatic compounds that are prevalent contaminants in the soil at munitions facilities, a two-step remediation strategy has been proposed. The first step involves the complete reduction of the nitroaromatic group to the corresponding aromatic amine compounds, followed by nucleophilic addition reactions such as oxidative coupling with natural SOM to form tight covalent bonds (bound residues) that pose minimal environmental threat. Both reduction and covalent binding reactions generally are scaled with pH. A generalized scheme is shown in Figure 4 for nitrobenzene. Adsorbed Fe(II)-goethite species reduce nitrobenzene at rapid rates (step a) to the reduced aniline product. Aniline is then reacted with natural SOM groups and incorporated into humic–organonitrogen complexes. Step b can be catalyzed by addition of Mn oxides to minimize environmental contamination of nitrobenzenes and anilines. Structural analyses of the products of these reactions have been made recently by state-of-the-art ^{15}N- and ^{13}C-labeled nuclear magnetic resonance techniques.

See also: **Metal Oxides**; **Redox Potential**; **Redox Reactions, Kinetics**

Further Reading

Buerge IJ and Hug SJ (1999) Influence of mineral surfaces on chromium (VI) reduction by iron(II). *Environmental Science and Technology* 33: 4285–4291.

Fendorf S (1995) Surface reactions of chromium in soils and waters. *Geoderma* 67: 55–71.

Fendorf S, Wielinga BW, and Hansel CM (2000) Chromium transformations in natural environments: the role of

biological and abiological processes in chromium(VI) reduction. *International Geology Review* 42: 691–701.

Laha S and Luthy RG (1990) Oxidation of aniline and other primary aromatic amines by manganese dioxide. *Environmental Science and Technology* 24: 363–373.

Lakind JS and Stone AT (1989) Reductive dissolution of goethite by phenolic reductants. *Geochimica et Cosmochimica Acta* 60: 3169–3175.

Luther GW III (1990) Frontier molecular orbital theory in geochemical processes. In: Stumm W (ed.) *Aquatic Chemical Kinetics: Reaction Rates of Processes in Natural water*, pp. 173–198. New York: Wiley Interscience.

Luther GW III, Kostka JE, Church TM, Sulzberger B, and Stumm W (1992) Seasonal iron cycling in the salt marsh sedimentary environment – the importance of ligand complexes with Fe(II) and Fe(III) in the dissolution of Fe(III) minerals and pyrite, respectively. *Marine Chemistry* 40: 81–103.

Matocha CJ, Sparks DL, Amonette JE, and Kukkadapu RK (2001) Kinetics and mechanism of birnessite reduction by catechol. *Soil Science Society of America Journal* 65: 58–66.

McBride MB (1994) *Environmental Chemistry of Soils*. New York: Oxford University Press.

Postma D and Jakobsen R (1996) Redox zonation: equilibrium constraints on the Fe(III)/SO_4-reduction interface. *Geochimica et Cosmochimica Acta* 60: 3169–3175.

Schwarzenbach RP, Gschwend PM, and Imboden DM (1993) *Environmental Organic Chemistry*. New York: Wiley Interscience.

Sparks DL (1995) *Environmental Soil Chemistry*. San Diego, CA: Academic Press.

Stumm W (1992) *Chemistry of the Solid–Water Interface*. New York: Wiley Interscience.

Ulrich HJ and Stone AT (1989) Oxidation of chlorophenols adsorbed to manganese oxide surfaces. *Environmental Science and Technology* 23: 421–428.

Wehrli B (1990) Redox reactions of metal ions at mineral surfaces. In: Stumm W (ed.) *Aquatic Chemical Kinetics*, pp. 311–336. New York: John Wiley.

Wersin P, Hohener P, Giovanoli R, and Stumm W (1991) Early diagenetic influences on iron transformations in a freshwater lake sediment. *Chemical Geology* 90: 233–252.

P

PADDY SOILS

C Witt, Potash & Phosphate Institute/Potash & Phosphate Institute of Canada and International Potash Institute, Singapore
S M Haefele, International Rice Research Institute, Metro Manila, Philippines

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Wetland rice or paddy soils provide the staple diet for nearly half the world's population. Worldwide, the total harvested rice area covers approximately 150 million hectares, and approximately 90% of this rice area is situated in Asia, with a similar share in rice production. The flooded soil environment is one of the specific characteristics associated with rice cultivation of the traditional rice-growing areas in the river deltas of China, Southeast Asia, and the Indian subcontinent. The term 'paddy soil' does not describe a specific soil type but rather denotes soils in lowland rice production systems with a prolonged period under submergence. 'Water regime' is used to classify the environment of paddy soils into broad ecosystems, distinguishing irrigated and rainfed rice grown predominantly in the lowlands, upland rice grown in the uplands under aerated soil conditions, and deep-water rice in flood-prone river deltas and coastal areas.

Cropping Environments in Asia's Lowlands

With an increasing demand for rice from a growing population, rice-cropping and paddy-soil management in Asia has undergone substantial intensification since the first modern high-yielding varieties with shorter growth periods, reduced height, and better fertilizer response were released in the early 1960s. Rice production tripled from 1948 to 1990 because of a larger physical land area cropped to rice, an increasing number of crops grown per year, and higher yields as a result of Green Revolution technologies such as improved germ plasm, irrigation infrastructure, water and pest management, and greater use of inorganic fertilizer. The intensification was accompanied by changes in chemical, biological, and physical processes in the soil–floodwater system. The duration of soil submergence and tillage operations increased, especially in the expanding double and triple rice-cropping systems. In such systems, periods of aerobic soil conditions are shorter and are mostly restricted to dry fallows. The duration of the dry fallow between two rice crops varies from a few weeks up to 2–3 months in the typical annual double rice-cropping system in Asia.

In the mid-1990s, the harvested area in Asia's rice ecosystems included 73 million hectares of irrigated rice, 46 million hectares of rainfed lowlands, 9 million hectares of upland rice, and approximately 4 million hectares of flood-prone rice ecosystems. While the overall physical rice area will probably not change much in the future, it is forecast that total rice production in Asia will have to increase by approximately 25% from 530 million tonnes in 1999 to 662 million tonnes in 2020 to meet the increasing rice demand in Asia (Table 1). Given the expansion of rice area with irrigation at the expense of upland rice area and the small potential for increasing rice yields in the uplands, yields in irrigated rice would have to increase from $5.3\,t\,ha^{-1}$ in 1999 to $6.7\,t\,ha^{-1}$ in 2020. The expected changes in rice demand, environment and socioeconomic conditions, and their effects on productivity and management of paddy soils in Asia's intensive rice systems are depicted in Figure 1. Most of the expected production increase will have to come from grain yield increases in the lowland rice ecosystems, and technologies developed in recent years provide substantial opportunities for farmers to increase productivity and profitability as rice demand increases. With sufficient investment, rice production can undergo an intensification that is environmentally sound and sustains the soil resource base. However, in some areas, crop management may change in response to environmental concerns, resource availability or low profitability, with possible adverse effects on rice yields and soils.

Table 1 Yield, area, and production of irrigated and nonirrigated rice for Asia in 1999, and required changes in yield based on projected changes in area and production requirements until 2020 (Dawe D, IRRI, personal communication.)

	Grain yield			*Area*			*Production*		
	t ha^{-1}			*10^6 ha*			*10^6 t*		
System	*1999*	*2020*	*Δ%*	*1999*	*2020*	*Δ%*	*1999*	*2020*	*Δ%*
Irrigated	5.36	6.72	+25%	76.0	78.4	+3.2%	406	527	+30%
Nonirrigated	2.08	2.39	+14%	59.5	56.5	−5.0%	124	135	+9%
Total	3.92	4.90	+25%	135.2	135.2	0.0%	530	662	+25%

Projected changes in area and production are adapted from Rosegrant MW, Sombilla MA, and Perez N (1995) *Global Food Projections to 2020: Implications for Investment. Food, Agriculture and the Environment.* Discussion Paper No. 5. Washington, DC: International Food Policy Research Institute.

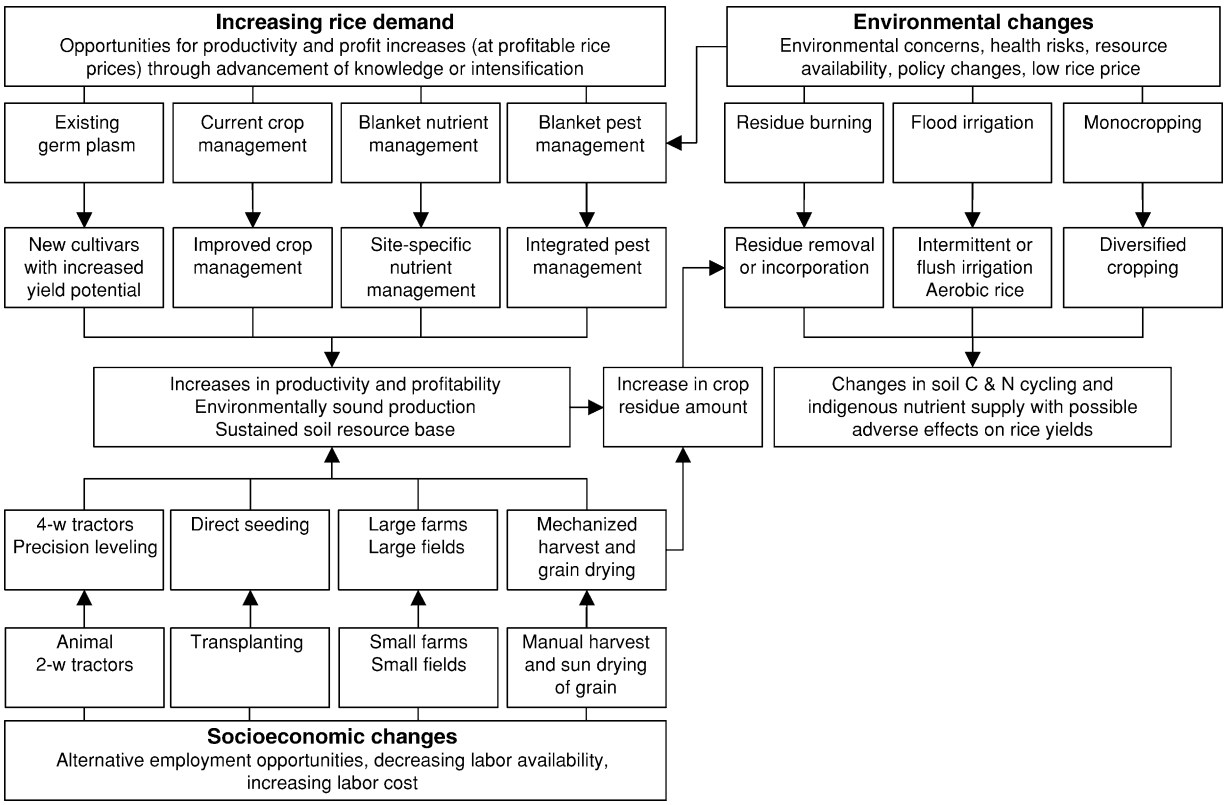


Figure 1 Expected changes in rice demand, environment, and socioeconomic conditions affecting the productivity and management of paddy soils in intensive rice systems in Asia. Some of these changes have already occurred or are in progress in certain areas. w, wheel.

Distribution and Important Characteristics of Paddy Soils

Paddy soils constitute an important group within the wetland soils, where water saturation dominates soil development and the types of plant and animal communities. Thus, paddy soils are defined as wetland soils with standing water in bunded and leveled fields used for cultivation of rice (bunds are small dikes around the field to keep the water from running off). Paddy soils may support an upland, nonrice crop in the dry season when soils are aerated, but insufficient drainage capacities often make rice the only possible crop choice in the lowlands of the humid tropics. Free surface water may occur naturally or rainfall, runoff, or irrigation water may be retained by field bunds and/or compacted subsoil layers.

According to the Food and Agriculture Organization of the UN (FAO) World Reference Base for Soil Resources, most paddy soils are designated as Anthrosols (having a puddled surface layer and a plow pan) or as Gleysols, Fluvisols, Planosols, Plinthosols, and Histosols. Smaller areas of paddy soils fall within the gleyic soil units of Arenosols, Andosols, Cambisols,

Solonetz, Solonchaks, Luvisols, Lixisols, Acrisols, and Alisols. Although Vertisols, Nitisols, and Ferralsols have no gleyic soil units, these soils may be artificially flooded and used for rice cultivation. The US Department of Agriculture Soil Taxonomy does not recognize wetland soils at the level of soil orders, but classifies soils with aquic conditions at the suborder level and soils with hydromorphism at the subgroup level. Most paddy soils would be assigned to the aquic suborders of Andosols, Oxisols, Vertisols, Ultisols, Mollisols, Alfisols, Inceptisols, and Entisols.

Natural or artificial flooding of paddy soils has marked effects on electrochemical, chemical, and microbial processes, which in turn affect soil fertility and thereby crop growth in a dynamic manner. The main electrochemical changes include a decrease in redox potential and changes in soil and solution pH, microbially mediated processes that are mainly controlled by organic matter and reducible Fe contents of the soil. With decreasing redox potential, oxidized forms of soil redox systems serve as electron acceptors in microbial respiration. The general sequence of reduction is O_2–NO_3^-–Mn(III, IV)–Fe(III)–SO_4^{2-}–CO_2–H^+. Some of these redox systems are overlapping (e.g., O_2, NO_3^-, and Mn(III, IV)), whereas others are not (e.g., O_2 and Fe(III); SO_4^{2-}, and CO_2). Reduction of CO_2 causes the emission of methane from flooded paddy soils in freshwater systems. The electrochemical changes induce a pH increase in most acid soils, whereas the pH decrease upon flooding of calcareous and sodic soils is due to the accumulation of carbon dioxide. Two distinctly different topsoil layers develop: a thin aerobic surface layer of a few millimeters depth caused by O_2 diffusion through the aerobic floodwater layer (and around the rice root surface due to O_2 transport through the rice plant's aerenchyma), and the anaerobic soil layer below, with a variable depth. Most roots are concentrated in the first 20 or 30 cm below the soil surface. The presence of aerobic and anaerobic soil compartments causes considerable soil N losses through mineralization–nitrification–denitrification reactions (Figure 2). In most soils, flooding increases the plant availability of P, K, Ca, Mg, Na, Mn, Fe, and Si at least temporarily. Reduction of Fe and Mn oxides releases sorbed and occluded elements such as P, Zn, Cu, B, and Mo. Increasing pH reduces plant availability of Zn and Cu, whilst the solubility of Mo increases. Reduction of SO_4^{2-} reduces plant-available S.

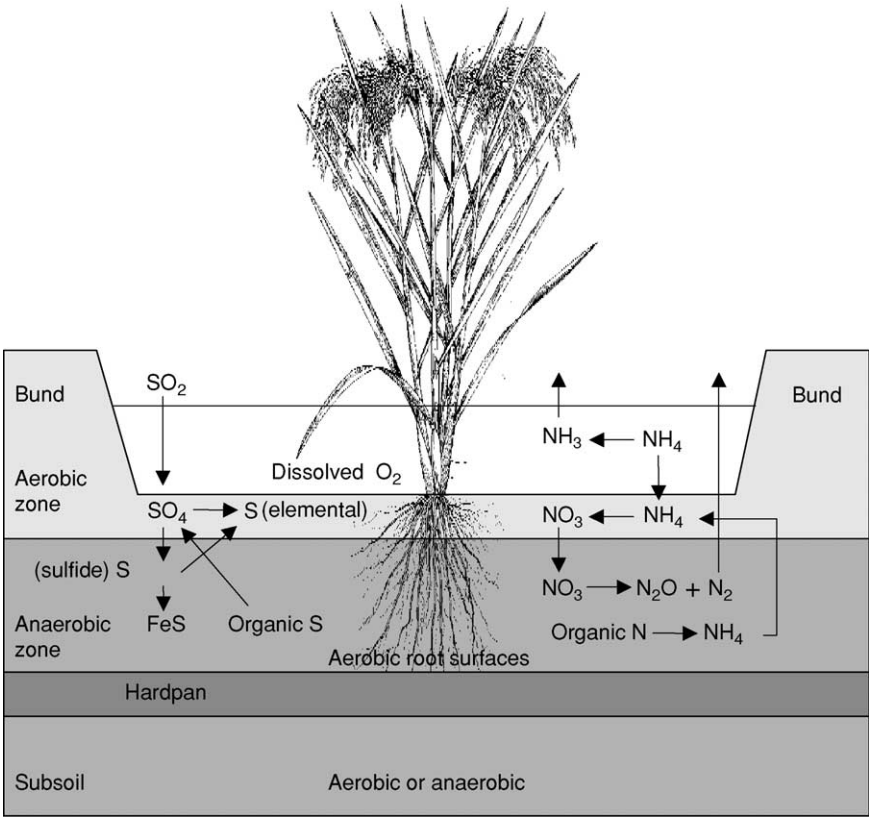


Figure 2 Aerobic and anaerobic soil compartments in a flooded paddy soil, and related sulfur and nitrogen dynamics. (Adapted from Greenland DJ (1997) *The Sustainability of Rice Farming*, pp. 1–273. Wallingford, UK: CABI Publishing/Manila, Philippines: International Rice Research Institute, with permission.)

A widespread and traditional soil-preparation practice in lowland production systems is puddling. Puddling refers to the practice of mixing surface soil with water to make it soft for transplanting, the traditional crop establishment method for rice. Other short-term puddling effects are the destruction of soil aggregates, irrigation-water savings due to reduced percolation, altered soil bulk density, increased soil water-holding capacity, decreased hydraulic conductivity, an acceleration of the electrochemical processes described above, and good weed control. A long-term effect of puddling is the development of a hardpan (plow pan, traffic pan) in the subsoil below the puddled layer, often associated with the accumulation of Fe, Mn, and Si. The puddled surface layer is often characterized by a coarser texture caused by clay mineral decomposition and/or transportation processes. Hardpan formation may take from 3 to 200 years, depending on soil type, climate, hydrology, and puddling frequency.

Paddy Soil Fertility and Nutrient Management

The nutrient-supplying capacity of paddy soils in the traditional rice-growing environments is largely determined by their parent material, the alluvial deposits in the river deltas and flood plains. Natural nutrient inputs are derived from sedimentation, irrigation and rainwater, organic residues or manure, biological N_2 fixation, and carbon assimilation by floodwater flora and fauna. Efforts to maintain soil fertility have historically centered around maintaining soil organic matter (SOM) through the addition of green manures, animal waste, and crop residues, but their significance has decreased continuously in the last decades.

The total annual C input in a double rice-cropping system may currently range from 4000 to 7000 kg C ha^{-1}, including inputs of 1000–2000 kg C ha^{-1} from the photosynthetic and heterotrophic biomass in rice fields. Despite high temperatures in the tropical soil floodwater environment and rapid initial breakdown of organic inputs, average SOM concentrations in paddy soils are higher than in aerated soils, mostly because of high organic matter inputs and a slower decomposition rate of humus. Crop intensification with prolonged soil flooding changes the nature of SOM toward more aromatic compounds (chemical structure with six C-atoms arranged in a ring) because of changes in the breakdown of organic matter under increasingly anaerobic conditions. These changes appear to have limited effects on yield stability under rice monocropping, except in systems with intensive triple cropping and nearly continuous soil submergence.

A recent analysis of yield trends in long-term experiments with intensive rice-based cropping systems in tropical and subtropical regions of Asia has suggested that yields in most experiments have remained stable with recommended fertilizer N, P, and K addition. There is also evidence that organic matter application does not improve grain yield trends in rice-based systems vis-à-vis a balanced mineral fertilizer application. However, the application of organic matter, unless purchased, is often profitable if used as a nutrient source complementary to inorganic fertilizer. The typical irrigated lowland rice soil maintains an average carbon concentration of approximately 1–2% and a nitrogen concentration of approximately 0.1–0.2%, but considerably lower concentrations are found in most rainfed lowland environments.

A unique contribution to the indigenous N supply of paddy soils derives from biological N_2 fixation (BNF), providing 15–50 kg N ha^{-1} per cropping cycle from blue-green algae in the floodwater and heterotrophic bacteria in the rice rhizosphere. The mineralization of algae and bacteria provides a constant supply of ammonium (NH_4^+), the dominating inorganic N form available to plants under anaerobic soil conditions. Because of efficient plant uptake of mineralized N, larger amounts of extractable NH_4^+ are found only after soil flooding during the initial decomposition of residues or if fertilizer N is applied (**Figure 3**).

The major N loss mechanism during the cropping season is ammonia (NH_3) volatilization from the flooded soil, particularly when NH_4^+ concentrations in the floodwater rise after N fertilizer application. Nitrogen is also leached, denitrified, and/or immobilized by soil microbial biomass and SOM. Approximately 70% of the fertilizer N applied by farmers is currently not recovered by crop N uptake, but N recovery efficiencies of applied fertilizers can reach 50% or more in farmers' fields with adequate real-time N management strategies. Considerable amounts of soil nitrate only accumulate in aerated soil during fallow periods or during upland cropping cycles in rice-based rotations. Almost all nitrate is lost after soil flooding because of leaching and denitrification. Including upland crops in rice-based systems affects mineralization, plant availability, and loss of inorganic N. Soil aeration also accelerates the turnover of SOM and changes its composition toward a less aromatic nature, with possible adverse effects on the amount of soil indigenous N supplied to rice. However, few long-term data are available from experiments that directly compare crop rotations.

Other anticipated future changes in crop management include mechanization of soil preparation and

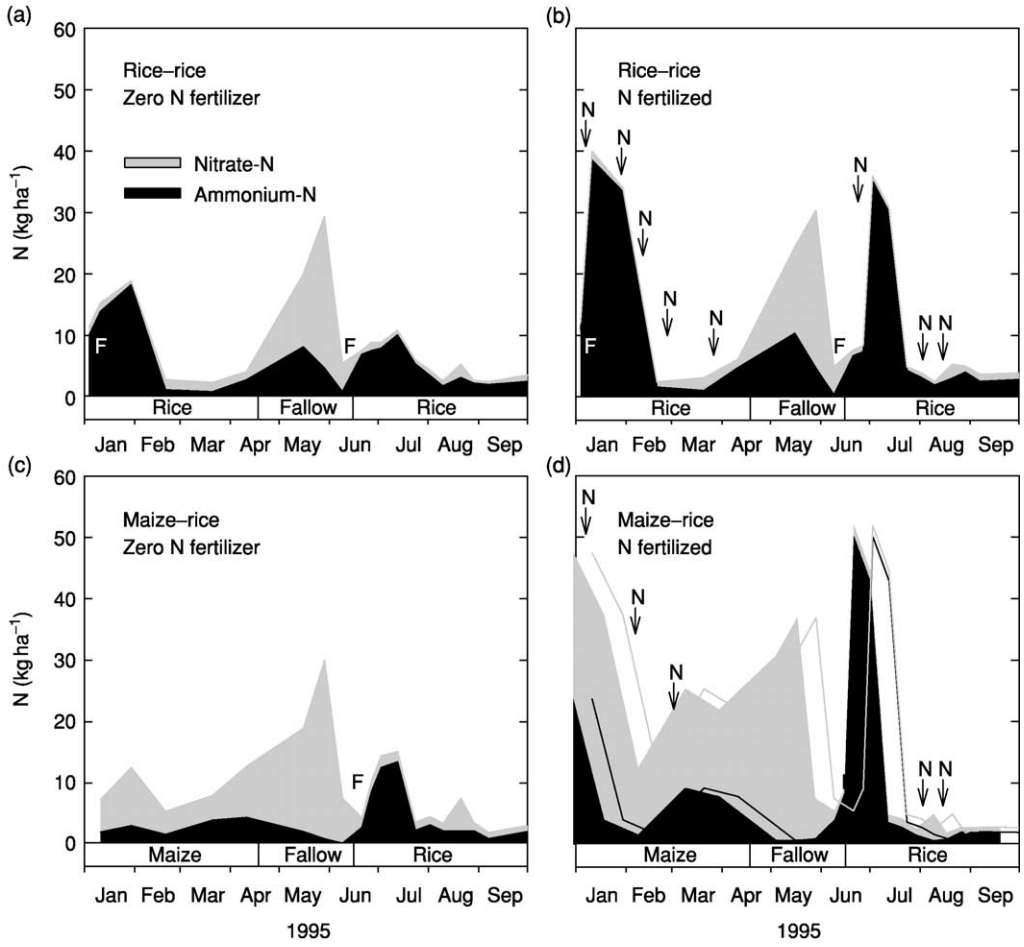


Figure 3 Nitrate (NO_3^-) and ammonium (NH_4^+) dynamics in a rice–rice and a maize–rice cropping experiment with and without fertilizer N application, IRRI experimental farm, 1995. Fertilizer N was applied in several split applications with a total of 190 kg N ha^{-1} in the dry season (Jan to mid-Apr) and 100 kg N ha^{-1} in the wet season (mid-Jun to early Oct). F, flooding; N, fertilizer N split application.

harvest operations, changes in water management toward increasing soil aeration, and larger amounts of crop residues remaining in the field for farmers to deal with (Figure 1). These changes will affect soil C and N cycling in paddy soils, and research currently aims to develop a portfolio of technology and management options to mitigate potential adverse effects.

Rice requires an adequate supply of nutrients from various sources to produce high and sustainable yields. A rice crop with a yield of 6 t ha^{-1}, for example, accumulates approximately 100 kg N, 18 kg P, 100 kg K, 0.3 kg Zn, 11 kg S, 480 kg Si, 21 kg Mg, 24 kg Ca, 3 kg Fe, and less than 0.1 kg Cu and B. Without fertilizer, the supply of indigenous N limits yield in most paddy soils. Other nutrients that have to be supplied with fertilizer in intensive rice systems include P and K, and in some areas also S and/or Zn. Required fertilizer rates are determined by the deficit between plant nutrient requirement and indigenous nutrient supply. This is particularly relevant for N, because residual effects of fertilizer N application are small and long-term trends show little change in indigenous N supply with time in farmers' fields. For nutrients other than N, fertilizer application is often recommended even in the absence of a yield response (maintenance rates) to avoid depletion of soil nutrient particularly at elevated yield levels at which nutrient removal with grain and straw is high. However, a comparison of current fertilizer rates used by farmers in seven major irrigated rice areas in Asia has shown that fertilizer use is often not in congruence with the deficit between yield level and indigenous nutrient supply (Table 2). On average, farmers use 90–140 kg N ha^{-1}, 10–25 kg P ha^{-1}, and 10–55 kg K ha^{-1} to achieve grain yields

Table 2 Average seasonal indigenous N, P, and K supply, fertilizer use, and grain yield in intensive rice systems of South and Southeast Asia. Ranges refer to interquartile ranges (25th, 75th percentiles) of on-farm data

	Indigenous nutrient supply (kg ha^{-1})			*Fertilizer (kg ha^{-1})*			
Site	*N*	*P*	*K*	*N*	*P*	*K*	*Yield (t ha^{-1})*
JI	60–79	18–24	97–141	147–192	14–23	41–75	5.3–6.7
HA	49–68	12–19	61–102	90–118	13–30	50–80	5.5–6.7
AD	43–59	13–20	69–109	93–137	17–27	22–54	5.3–6.6
TH	30–44	14–19	47–65	83–124	9–25	27–45	4.6–6.2
SU	36–74	10–14	65–91	86–148	0–17	0–0	3.6–5.4
MA	43–65	10–18	65–100	75–142	9–20	8–31	3.9–5.3
OM	27–49	11–18	55–94	96–123	13–25	9–37	3.4–5.5
Mean	41–65	12–19	61–102	92–141	10–25	11–56	4.4–6.2

JI, Jinhua, Zhejiang (China); HA, Hanoi, Red River Delta (Vietnam); AD, Aduthurai, Old Cauvery Delta, Tamil Nadu (India); TH, Thanjavur, New Cauvery Delta, Tamil Nadu (India); SU, Sukamandi, West Java (Indonesia); MA, Maligaya, Central Luzon (Philippines); OM, Omon, Mekong Delta (Vietnam).
Source: Dobermann A, Witt C, and Dawe D (eds) (2004) *Increasing Productivity of Intensive Rice Systems through Site-Specific Nutrient Management*. Enfield, NH: Science Publishers/Los Baños, Philippines: International Rice Research Institute (IRRI).

of 4.4–6.2 t ha^{-1} in the typical annual double or triple rice-cropping systems of South and Southeast Asia. This is approximately 50–75% of the climate-adjusted, genetic yield potential of 8.3 t ha^{-1} that has been estimated for existing germ plasm across sites, seasons, and years. In general, yields and fertilizer rates are higher in the typical dry season, with greater solar radiation, than in the typical wet season, with greater cloud cover. The indigenous supply of N, P, and K in irrigated rice in Asia was on average 41–65 kg N ha^{-1}, 12–19 kg P ha^{-1}, and 61–102 kg K ha^{-1}, but varied widely with site, year, season, farm, and field. In rainfed lowlands, yields and fertilizer doses are generally lower, decreasing from favorable to unfavorable lowlands with increasing production risk.

Blanket fertilizer recommendations prevail in most of Asia's lowland rice areas without considering amount and spatial distribution of indigenous nutrient supplies. There is also concern that fertilizer P and K rates are not optimally adjusted to long-term needs. A typical input–output balance for P and K in farmers' fields is given in Table 3. On-farm measurements in irrigated rice have confirmed for a wide range of soils and growing environments that farmers' current fertilizer P and K management causes slightly positive P balances, while fertilizer K application is insufficient to balance K removal with grain and straw. Modern site-specific nutrient management (SSNM) strategies have been developed in recent years to manage nutrients more efficiently and supply fertilizers as required by the crop considering indigenous nutrient supplies. Estimates of indigenous nutrient supplies need to be verified if general crop management of paddy soils changes (Figure 1). Changes in fallow period and residue management or system diversification are likely to introduce changes in indigenous nutrient supplies that are sufficiently large to require fertilizer adjustments.

Table 3 Typical input–output balance for P and K during one cropping cycle in intensive irrigated rice systems in South and Southeast Asia

Parameters for P and K balance	*Unit*	*P*	*K*
Input with inorganic fertilizer	kg ha^{-1}	10–25	10–55
Input with irrigation water	kg ha^{-1}	0–1	10–30
Input with rainwater	kg ha^{-1}	0–1	0–10
Losses with percolation	kg ha^{-1}	0–1	5–15
Removal with grain and straw	kg ha^{-1}	10–15	50–65
Nutrient balance	kg ha^{-1}	0 to +11	−35 to +15

Source: Dobermann A, Witt C, and Dawe D (eds) (2004) *Increasing Productivity of Intensive Rice Systems through Site-Specific Nutrient Management*. Enfield, NH: Science Publishers/Los Baños, Philippines: International Rice Research Institute (IRRI).

The deficiency of Zn is the most widespread micronutrient disorder in rice, not only on soils with a traditionally low supply of Zn (e.g., neutral and calcareous soils containing a large amount of bicarbonate), but increasingly also on other soils with intensive rice cropping. Sulfur deficiency occurs mainly in upland soils containing allophane (an aluminosilicate with primarily short-range structural order which occurs as exceedingly small spherical particles, especially in soils formed from volcanic ash), soils low in organic matter, highly weathered soils containing large amounts of Fe oxides, and sandy soils with high percolation. Lowland rice areas with common S deficiencies are located in Bangladesh, China, India, Myanmar, Pakistan, Sri Lanka, and Thailand. Other nutrient disorders are only of local importance.

Lowland Rice Systems and Declining Water Availability

Today, nearly 90% of the fresh water diverted for human use in Asia goes to agriculture and, of this, more than 50% is used to irrigate rice. With the growing demand for water for nonagricultural uses (domestic, municipal, industrial, and environmental), the proportion available for agriculture is projected to decline from 90% to 73% in developing countries in the near future. Indicators of a looming water crisis include falling groundwater tables in regions where mainly tube wells are used for irrigation and severe water shortages downstream of some major rivers in Asia (e.g., Yellow River, Indus, and Ganges). Figure 4 gives the current distribution of irrigated and rainfed lowland rice and major hot spots of groundwater depletion and reduced river water flows. The productivity of intensive rice systems depends strongly on the availability of sufficient irrigation water, and production is likely to decrease with water shortage unless suitable mitigation options are developed. By 2020, physical water scarcity is projected for wet-season irrigated rice in North China (2.5 million hectares), Pakistan (2.1 million hectares), and north and central India (8.4 million hectares), as well as on 2 million hectares of dry-season rice in central India.

Water productivity at the field level can be increased by: (1) increasing yield per unit cumulative evapotranspiration through the use of high-yielding, short-duration varieties and optimal, site-specific agronomic practices; (2) more effectively using rainwater by using dry-seeding rice technologies to synchronize the start of the cropping season with the start of the rainy season; and (3) reducing unproductive water losses (seepage, percolation, evaporation) through adequate soil and water management. Examples of the latter include saturated soil culture (daily flooding up to approximately 0.01 m water depth) or alternate wetting-and-drying regimes (flooding every 2–7 days up to 0.02–0.05 m water depth). These measures frequently cause a reduction in yield and profitability, although they often result in an increase in water productivity.

A more radical approach is to grow rice under permanently aerobic soil conditions by keeping the soil water content in the root zone always at or below field capacity, a technology called 'aerobic rice.' In contrast to 'upland rice,' this system is intended to be high-yielding, but more research is required to evaluate long-term effects on SOM and nutrient cycling. Such systems may gain local importance where water is limiting and rice prices are high, but yields may not reach those achieved in well-managed flooded systems. The adoption of such water-saving irrigation technologies is expected to cause a shift from mainly anaerobic soils to partly or even completely aerobic soils. Changes in the soil-moisture regime of paddy soils will have profound effects on SOM turnover, nutrient dynamics, carbon sequestration, soil productivity, weed ecology, and greenhouse gas emissions in irrigated systems. Water-saving technologies must therefore develop into integrated natural resource management approaches to maintain the productivity, profitability, and sustainability of the evolving systems.

Water productivity in rainfed lowland systems can be increased through more efficient and effective use of all inputs. A general increase in productivity will also reduce the pressure for more intensive management and area expansion of irrigated lands. Direct-seeding techniques offer further opportunities to achieve higher (water) productivity by timely crop establishment. Avoiding drought stabilizes yields, and a second, upland crop can be grown after rice using the residual soil-water content. Such crop intensification must be accompanied by balanced fertilization and adequate residue management on the often poor soils in rainfed systems to maintain or build up SOM and thereby the indigenous nutrient supply.

Sustainability and Environmental Issues

In general, rice cultivation is sustainable even where rice is continuously grown in monocropping at high production levels. The quality of the soil resource base can be maintained through adequate management practices, including proper crop residue and water management, tillage operations, soil aeration during fallows, and balanced fertilization. The sustainability of rice cropping may be at risk, however, where management-induced changes strongly influence soil characteristics and the soil environment. This may include prolonged soil submergence in intensive systems with three rice crops per year as well as prolonged soil aeration because of crop diversification (Figure 1). In the following, we focus on selected issues that are of general importance for sustainable paddy soil management and the environmentally sound production of rice.

When crops are grown with irrigation, the sustainability of agriculture is potentially threatened by several water-related processes such as salinization and alkalization. In the semiarid tropics, irrigation may cause salinization of soils by rising saline groundwater tables. Alkalization can occur when irrigation water with a positive residual alkalinity is used. However, ample irrigation can leach salts from the topsoil through percolation if sufficient drainage is given.

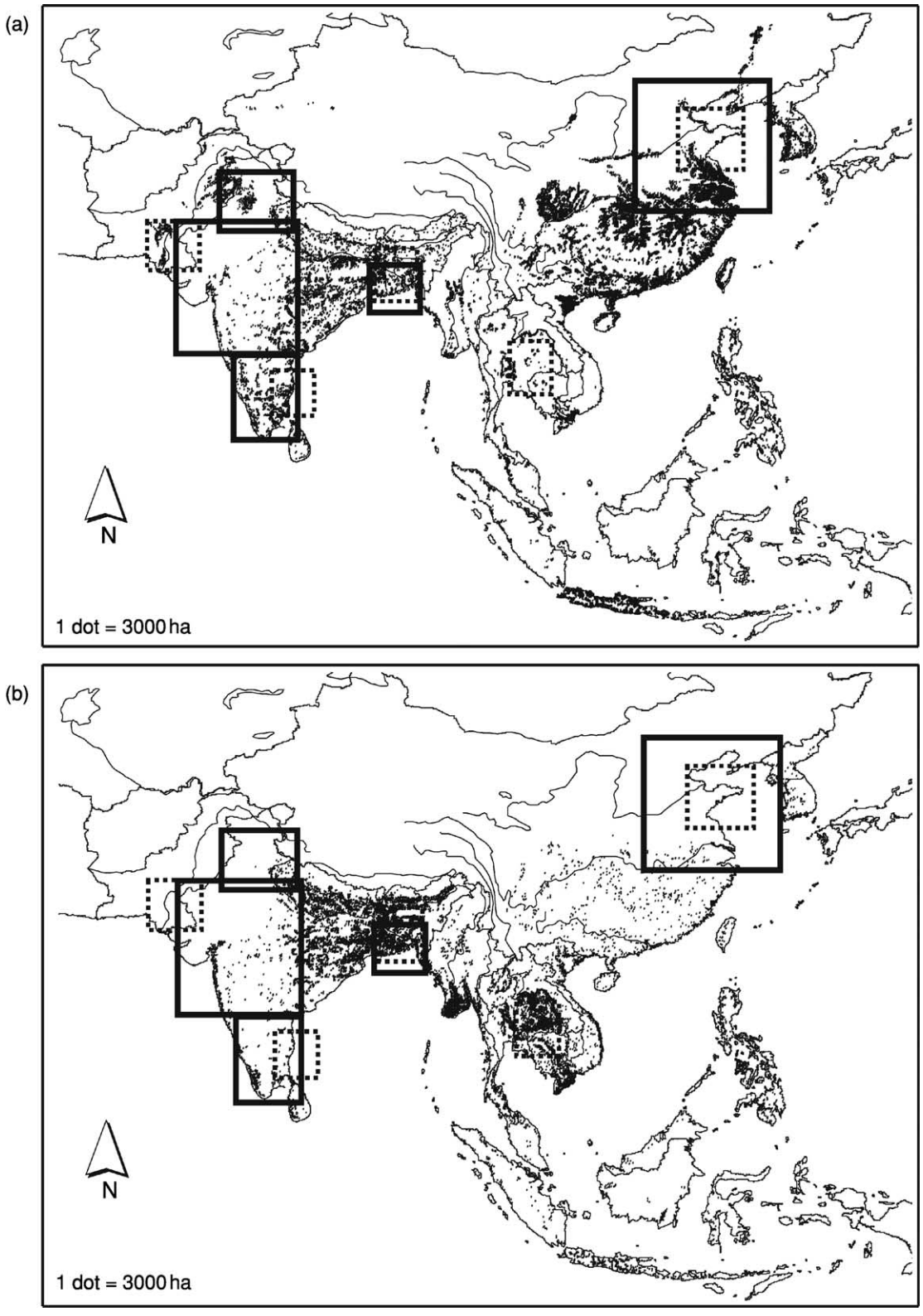


Figure 4 Current distribution of irrigated (a) and rainfed lowland rice (b) in Asia (Data source: Huke RE and Huke EH (1997) *Rice Area by Type of Culture: South, Southeast, and East Asia*. Manila, Philippines: International Rice Reasearch Institute (IRRI), and major hot spots of groundwater depletion (squares with solid lines) and reduced river water flows (squares with broken lines).

This makes irrigated rice a reclamation crop for such problem soils. A very different water-related problem is accidental flooding caused by inadequate irrigation control in some parts of tropical Asia, taking valuable land out of agricultural production; for example, this occurs in approximately 6 million hectares of arable land in India annually.

Other problem soils in the lowlands include Thionic Fluvisols or acid sulfate soils, which are widespread in Thailand, Vietnam, and Indonesia. Potential acid sulfate soils strongly acidify upon drainage, whereas actual acid sulfate soils are characterized by pH values of approximately 4. Rice cropping under submergence offers one of the few sustainable uses of these soils. Management strategies include: (1) complete soil oxidation through drainage and leaching of acidity from the crop root zone, which offers a permanent solution but may cause a contamination of surface water affecting crops, soils and aquatic organisms in surrounding regions; and (2) the limitation of soil oxidation by maintaining a permanent, high groundwater table, which requires sufficient irrigation water.

Environmental concern is increasing because the intensification of rice cropping in the past has led to an excessive use of fertilizer N and pesticides in many intensive rice-growing areas in Asia. This has been recognized, and much progress has been made in recent years in developing and disseminating adequate management strategies such as real-time N management and integrated pest management. Current changes in today's cropping environment include a shift from irrigated transplanted rice to direct seeding and/or the use of water-saving technologies. This causes greater weed pressure and may lead to an increased use of herbicides, particularly at sites with further intensification and mechanization because of labor shortages. Research efforts aim to integrate better water, nutrient, weed, and pest management approaches to enhance direct environmental and economic benefits for farmers.

Paddy soil environments have also caused global environmental concerns because methane (CH_4) is second in importance to CO_2 as a greenhouse gas and flooded paddy soils contribute approximately 10% of global CH_4 emissions. Although the CH_4 production of paddy soils is much smaller than originally assumed, it is still an important contribution. Methods to reduce this share include water-saving technologies as described above, improved crop-residue management, and direct seeding instead of transplanting. Care has to be taken to avoid significant increases in N_2O emissions, another greenhouse gas, which is 10 times more 'effective' than CH_4. Soil-moisture regimes keeping the soil redox potential between -100 and $+200$ mV could minimize both CH_4 and N_2O emissions.

Summary

Paddy soils are defined through their production environment, the typically irrigated and rainfed lowland rice-cropping systems. Cropping intensification since the 1960s has greatly increased rice yield and production, and further intensification is needed to achieve food security at the national and household level. Productivity increases at the farm level could be mainly achieved in two ways. One option is the continuing optimization of the existing rice-based cropping systems with two to three annual rice crops and short fallow periods. There is strong evidence that rice double-cropping in Asia can provide high, sustainable yields despite an increase in the duration of soil submergence and a shift from organic to inorganic nutrient sources since the 1960s. Sufficient inputs of external nutrients with fertilizer are necessary to match the insufficient supply of nutrients from indigenous sources, but inefficient fertilizer use, nutrient imbalances, and excessive pesticide use have become widespread and must be overcome through more site- and season-specific crop, soil, and water management to manage all inputs efficiently. The second option for farmers is to turn to crop diversification, adding more profitable, nonrice crops to the rice-based cropping system or to grow rice under aerobic soil conditions. Conditions imposed in the nonrice season will influence important characteristics of paddy soils and remove typical paddy soil characteristics temporarily or permanently. A key issue is whether the high productivity of current rice-cropping systems can be sustained in such diversified systems.

List of Technical Nomenclature

Indigenous N, P, and K supply ($kg\,ha^{-1}$)	The cumulative amount of a nutrient from all indigenous sources circulating through the soil solution surrounding the entire root system during one crop cycle. Indigenous nutrient supply is estimated as aboveground plant nutrient in plots where the element in question is the only limiting factor
Paddy	(1) Wetland rice; (2) bunded and leveled field used for cultivation of rice. The original meaning of paddy (Malay *padis*) is 'threshed, unhulled rice'

See also: **Anaerobic Soils**; **Redox Potential**; **Redox Reactions, Kinetics**

Further Reading

Dawe D, Dobermann A, Moya PF *et al.* (2000) How widespread are yield declines in long-term rice experiments in Asia? *Field Crops Research* 66: 175–193.

Dawe D, Dobermann A, Ladha JK *et al.* (2003) Does organic matter improve the sustainability and profitability of intensive rice systems? *Field Crops Research* 83: 191–213.

Dobermann A and Fairhurst T (2000) *Rice: Nutrient Disorders and Nutrient Management*. Singapore: Potash & Phosphate Institute (PPI), Potash & Phosphate Institute of Canada (PPIC), and Los Baños, Philippines: International Rice Research Institute (IRRI).

Dobermann A, Witt C, and Dawe D (eds) (2004) *Increasing Productivity of Intensive Rice Systems through Site-Specific Nutrient Management*. Enfield, NH: Science Publishers/Los Baños, Philippines: International Rice Research Institute (IRRI).

Greenland DJ (1997) *The Sustainability of Rice Farming*. Manila, Philippines: International Rice Research Institute (IRRI)/Wallingford, UK: CABI Publishing.

Huke RE and Huke EH (1997) *Rice Area by Type of Culture: South, Southeast, and East Asia*. A revised and updated database. Manila, Philippines: International Rice Research Institute (IRRI).

IRRI (1978) *Soils and Rice*. Los Baños, Philippines: International Rice Research Institute.

Kawaguchi K and Kyuma K (1977) *Paddy Soils in Tropical Asia: Their Material Nature and Fertility.* Honolulu, Hawaii: University Press of Hawaii.

Kirk GJD (2004) *The Biogeochemistry of Submerged Soils*. Chichester, UK: John Wiley.

Kirk GJD and Olk DC (eds) (2000) *Carbon and Nitrogen Dynamics in Flooded Soils*. Los Baños, Philippines: International Rice Research Institute (IRRI).

Kyuma K (2004) *Paddy Soil Science*. Kyoto, Japan: Kyoto University Press/Melbourne, Australia: Trans Pacific Press.

Ladha JK, Hill JE, Duxbury JM, Gupta RK, and Buresh RJ (eds) (2003) *Improving the Productivity and Sustainability of Rice–Wheat Systems: Issues and Impacts*. Proceedings of an International Symposium cosponsored by ASA, IRRI, and CIMMYT at Charlotte, NC, USA, 22 October 2001. ASA Special Publication No. 65. Madison, WI: American Society of Agronomy.

Maclean JL, Dawe D, Hardy B, and Hettel GP (eds) (2002) *Rice Almanac*. Los Baños, Philippines: IRRI Bouaké, Côte d'Ivoire: West Africa Rice Development Association (WARDA), Cali, Colombia: International Center for Tropical Agriculture (CIAT), Rome, Italy: FAO.

Moormann FR and van Breemen N (1978) *Rice: Soil, Water, Land*. Los Baños, Laguna, Philippines: International Rice Research Institute (IRRI).

Rosegrant MW, Sombilla MA, and Perez N (1995) *Global Food Projections to 2020: Implications for Investment. Food, Agriculture and the Environment*. Discussion Paper No. 5. Washington, DC: International Food Policy Research Institute.

Tuong TP and Bouman BAM (2003) Rice production in water-scarce environments. In: Kijne JW, Barker R, and Molden D (eds) *Water Productivity in Agriculture: Limits and Opportunities for Improvement. The Comprehensive Assessment of Water Management in Agriculture Series*, vol. 1, pp. 53–67. Wallingford, UK: CABI Publishing.

Wassmann R, Lantin RS, and Neue HU (eds) (2000) Methane emissions from major rice ecosystems in Asia. Reprinted from *Nutrient Cycling in Agroecosystems* (2000) vol. 58(1–3). Los Baños, Philippines: International Rice Research Institute (IRRI)/Dordrecht, the Netherlands: Kluwer Academic Publishers.

Parent Material *See* **Pedology:** Basic Principles; **Factors of Soil Formation:** Parent Material

PEDOLOGY

Contents
Basic Principles
Dynamic

Basic Principles

M J Singer, University of California–Davis, Davis, CA, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Pedology has two meanings according to the *American Heritage Dictionary of the English Language*. The first is the study of the physical and mental development and characteristics of children, and the second is the scientific study of soils, including their origins, characteristics, and uses. The Greek word 'pedon' means ground and 'logy' means study. The International Union of Soil Scientists (IUSS) and many of the member scientists within the organization interpret 'pedology' as encompassing all scientific study of the soil, and hence all soil scientists are pedologists. In this encyclopedia the narrower definition of 'pedology' that characterizes it as the study of the formation, properties, classification, and management of soil is used. Others define pedology as "the earth science that quantifies the factors and processes of soil formation including the quality, extent, distribution, spatial variability and interpretation of soils from microscopic to megascopic scales."

Pedology is the study of soils in their natural setting (Figure 1) and has been equated with the study of soil genesis and soil classification. The people who study soils in these ways are pedologists. Pedology includes pedogenesis, the study of the origin and formation of soil; soil morphology, which is the characterization of the soil as a natural body; the distribution of soil on landscapes, which is soil survey; the organization of soil information into natural and technical classification systems; and the interpretation of soil properties for use and management of the soil resource. Pedology is "...an integrative and extrapolative science." Pedologists integrate understanding of landscapes, vegetation patterns, climate, and human activity into knowledge about soils, their distribution, mode of formation, and usefulness. Pedologists also extrapolate information from one scale to the next, to build a coherent model of soils from microscopic observations to the landscape scale. Integration and extrapolation are necessary because soils form a highly variable continuum over the Earth's surface.

History

Pedology is a relative newcomer to natural science, having emerged as a discipline in the nineteenth century. Although the agricultural value of soil and the importance of the chemical properties of soil were known well before the emergence of pedology, the concept that soil could be studied as a natural body was a nineteenth-century idea. A German scientist, A. Fallou, introduced the term 'pedology' in 1862. Most soil scientists credit the Russian soil scientist V.V. Dokuchaev as the person who made pedology a scientific discipline. He did so by recognizing that the five soil-forming factors (parent material, climate, biota, topography, and time) interact to produce the morphological properties of the three-dimensional soil individual, the pedon (see Figure 2, below). Hilgard is regarded as the first North American-based scientist to recognize pedology as an independent discipline. Hans Jenny is recognized as the nineteenth-century US soil scientist who took the

Figure 1 Pedologists examining a soil during the 2002 post-meeting field tour of the International Union of Soil Scientists, in Thailand.

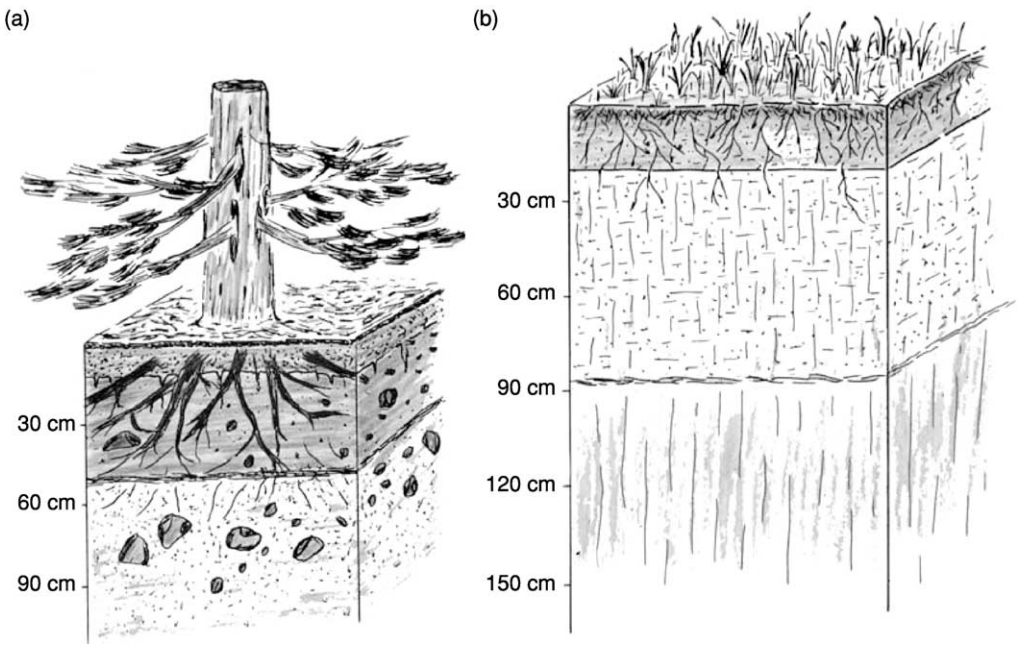


Figure 2 Soils forming under different climatic conditions with various types of vegetation develop different characteristics that are recognized by pedologists: (a) forest soil; (b) grassland soil.

concept of five soil-forming factors and gave it substance in his 1941 book *Factors of Soil Formation*. He further elaborated on his original text in his 1980 text *The Soil Resource*.

From these early beginnings, which were mostly centered on the mapping of soils and their description, emerged a science that built upon the early foundation and that today is a vibrant mix of individuals who study the soil as a natural body on landscapes, who are continuing to develop concepts about the development of soils through complex biogeochemical pathways, and who are looking for new ways to classify the soil to enhance and protect this resource. As scientists reach out to the moon and beyond, the definition of soil and the role of pedology are expanding.

Overview

Soil Classification

Why do we classify soils? The frequently cited reason is to utilize the large body of information about soil. Soil classification systems are grouped into natural and technical systems. Natural systems are those that classify what is seen on the landscape, while technical systems classify soils based on perceived strengths and weaknesses for specific uses. The entries in this encyclopedia focus on the natural classification systems from several countries. These represent diverse methods for categorizing the many kinds of soils that are recognized by pedologists. Although the language of natural soil classification appears foreign and esoteric to the casual reader, soil classification organizes soils with similar properties into useful groups. Most taxonomic systems are hierarchical, with the broadest groupings at the highest levels and the most narrow, most specific groupings at the lowest levels. The informed user can thus learn something about the properties of each group by understanding the construction of the system and the meaning of the terminology.

Classification systems reflect the interests and understanding of the soil scientists who create the systems. Hence the US system of soil classification Soil Taxonomy differs in structure from other systems, because it was developed from its own historical base. The US system has evolved over time, from a system based on perceived processes of soil formation and qualitative definitions that separated broad groupings of soils, to the present system that is based on quantitative definitions. Quantification is the path followed by all sciences and, as the discipline of pedology has matured, the various facets of the science have become more quantitative.

The IUSS has adopted the World Reference Base for Soil Resources (WRB) as the official system for correlating various national soil classification systems. It has also been developed to help users view broad divisions of soils and to facilitate understanding of soils throughout the world. Although each existing soil classification system is different, each is

based on carefully defined and quantified criteria that allow the system to be useful to the widest range of individuals.

Technical systems have also evolved throughout the world, with the intent of creating groups of soils with similar capabilities or limitations. These systems are not limited to agriculture; they exist for most of the common uses of soil. For example, systems exist for rating the usefulness of soils for on-site waste disposal, for construction of recreation areas, and as a source of material for road-fill. Because soil underlies and supports much of our civilization, technical systems are critically important guides to the proper selection of soils for different uses.

Soil Genesis

Soil genesis or pedogenesis is the study of the origin and formation of soil. Pedologists conceptualize soil as a natural, three-dimensional body on the Earth's surface (**Figure 2**). This three-dimensional body, the pedon, is large enough to include all of the properties of the soil individual. Typically the smallest pedon is 1 m^2 and 2 m deep. Pedons form within the framework of the five soil-forming factors. 'Parent material' is the term pedologists use to describe the starting material for soil genesis. The parent material may be consolidated (such as rock) or unconsolidated (such as wind-blown silt or sand). The amount of consolidation influences the rate at which water can enter the parent material and therefore the rate of soil formation.

Equally important are the constituent minerals that comprise the parent material. It is the size, crystalline nature, and chemistry of the minerals that determine the rates at which the inherited (or primary) minerals are changed to new minerals or are dissolved and removed from the parent material, thus creating soil. The mineralogy also determines the fertility of the parent material and soil and hence the biological activity on the parent material and soil. This oversimplifies the complexity of pedogenesis, which consists of many interacting processes, which proceed simultaneously at multiple rates.

A special case of how parent material greatly influences soil properties is the formation of soils on parent material rich in clay minerals. Such parent materials form soils that are rich in clay-size minerals and clay minerals. These soils, called 'shrink–swell soils,' tend to change volume greatly as they wet and dry, which makes use of the soils for agriculture and construction difficult.

Topography describes the landscape on which the soil is found. It typically includes the slope angle, slope shape (concave, convex, simple, or complex), and slope direction or aspect. Topography is complex in how it affects pedogenic processes. For example, in the northern hemisphere, a south-facing slope intercepts more solar energy than does a north-facing slope. In dry climates, the south-facing slope will be more water-limited than the north, and the south-facing slope is likely to have less organic matter accumulation, lower biological activity, less organic acid production, and slower mineral weathering than on a north-facing slope. In the southern hemisphere and in moist cold climates, the results are opposite.

The biotic factor includes animal and plant organisms of all sizes and their direct and indirect effects on pedogenic processes. Vegetation is part of the biotic factor of soil formation. Vegetation, under which soils form, mediates the cycle of carbon and nutrients into and through soil. The most obvious effect of plants is the input of carbon into and on to the soil through roots and litter fall. Plant roots and microorganisms respire large amounts of carbon dioxide that decreases soil pH, increases weathering, and has other important effects on soils. Plants also alter the chemistry of precipitation as it falls through the canopy and flows down the stems of plants. Animals mix the soil, consume and modify carbon, and create pores through which flow water and air. All of these processes and more are part of the biotic factor of pedogenesis.

Precipitation and annual temperature are two climatic factors that pedologists consider when studying soil genesis. Others are the timing of the precipitation (e.g., winter versus summer precipitation), rainfall versus snowfall as the dominant precipitation type, precipitation intensity as it affects soil destruction through erosion, and differences between winter and summer temperatures. The climatic factor is usually focused on mean annual temperature and precipitation, because these are the 'macro' climatic factors, but microclimate also influences the many soil-forming processes considered within the context of soil genesis.

Microclimate is the 'on-the-ground' climate that influences small areas of soil, compared with the regional climate. A topographic low that accumulates water and enhances weathering and translocation of weathering products creates a pedologically distinct 'climate' compared with upslope positions. Cold-air drainage into a valley can result in a distinctly different weathering regime than would be characterized by the general climate.

These factors describe the environment and conditions under which a soil forms and determine the properties of the 'product' of these processes. Time is the final natural soil-forming factor. Pedologists have shown that the net rate of soil formation is slow. The time required for horizons to form and for a soil to develop its distinct properties may be

thousands of years. Many processes contribute to the differentiation of the parent material into a vertically differentiated soil and they all act at different rates. Thus the question 'How old is this soil?' has no easy answer.

Soils record the history of the Earth's surface and sometimes soils that were once at the Earth's surface are buried. Once buried, they cease to form and they may represent climate and vegetation different from the modern climate and vegetation. These soils are paleosols and they are valuable records of natural history.

The influence of humans on soil development has become a growing area of interest to pedologists. This topic was not of major interest until the end of the twentieth century, when pedologists recognized the profound affect that humans were having on soils. Humans create new areas of soil, reclaim areas that have been destroyed through mining and road building, and destroy soils by modifying them for agriculture or urban uses. In particular, large areas of soil in the USA have been deep-plowed or ripped to remove horizons that are limiting to roots or water penetration. When this is done, the soil, as a natural body, is destroyed and the pedological clock is returned to the starting point.

Observing the vertical distribution of properties made visible by digging a hole or trench or removing vegetation on a vertical road excavation is typically the process used to study soil. Once the properties have been described and the horizons designated, samples are collected by depth increments or by horizon for laboratory analyses. Many chemical, physical, and biological analyses are performed on the samples. Interpretation of the analyses is used to reconstruct the pedogenic history from the time the parent material stabilized and pedogenesis began to the present. This method of studying soil has been referred to as 'static pedology.' Another method of study involves collecting and analyzing the soil solution from within horizons and at horizon boundaries. The data originating from these analyses are interpreted to indicate the types of processes extant in the soil profile. This method has been called 'dynamic pedology.'

Soil Morphology

Soils form a more or less continuous blanket over the terrestrial surface of the Earth and, in order to distinguish individuals from within this continuum, soil morphology is described in the field. Conceptually the pedon is the soil individual that is described (**Figure 2**), but in reality what is described is the two-dimensional slice through the pedon that soil scientists call the soil profile. These descriptions begin in the natural environment where the soil is observed, hence the morphological properties are those that can be measured or discerned by a trained observer. The field descriptions are supplemented and enhanced by laboratory analyses and micromorphological observations using thin sections and light and electron microscopes.

The major properties that are described in the field are the thickness, color, texture, and structure of individual layers, called 'soil horizons.' In addition, properties such as the physical behavior at different moisture contents (consistence), pH, boundaries between horizons, and presence of fragments larger than 2 mm in diameter are described. These properties characterize the soil individual, and they are interpreted to predict how soils behave when used for agriculture, recreation, urban development, and waste disposal.

Most pedologists study soils that are well drained. That is, they are well aerated and oxygen-rich for most of the year. The processes that form these soils and their morphology are better understood than the processes and morphology of soils formed under conditions where water remains and oxygen is in short supply for long periods of time. These hydric soils are described using the same properties as well-drained soils, although the soils themselves may look and act quite differently. Another group of soils found under wet, oxygen-poor conditions are the organic soils. Although the vast majority of soils are formed on mineral parent materials in which aluminosilicate minerals are the starting point for soil development, a small but important minority is formed from plant material that is slow to decay because the cold or wet conditions are unsuitable for microbial decomposition.

Organic soils or Histosols consist primarily of dead plant material in various stages of decay. The horizon nomenclature and morphology of these soils are different because of the different processes acting on the parent material. For example, rather than A, B, and C horizons that are standard for mineral soils, the organic soils consist of O horizons. An O horizon is one in which organic material is a major constituent. Each O horizon is given a subscript to denote its organic character. In the US Soil Taxonomy, for example, the subscripts a, e, and i are used to denote strongly decomposed, moderately decomposed, and mostly undecomposed plant material. The terminology used to describe soil morphology has sufficient breadth to encompass all soils. The breadth is accompanied by carefully crafted definitions that allow the specific characteristics of the soil profile to be described.

Modern pedology recognizes that the type of vegetation influences soil morphology, but that morphological properties are not purely a result of the vegetation under which the soil forms. Hence the familiar terms 'forest soil' or 'grassland soil' are more correctly 'soil formed under forest' or 'soil formed under grass' (Figure 2).

Soil Mapping

Pedology developed from the disciplines of agriculture and geology. Both require an understanding of the relationship between the soil individual and the landscape on which the individual is found. Creating soil maps that represent this relationship has been a more or less continuous process in the USA since the end of the nineteenth century. The process accelerated in the twentieth century as aerial photography replaced plane-table mapping and as new concepts of soil formation were learned and applied.

Soil mapping is not the sole provenance of the USA: the developing as well as the developed countries of the world make resource assessments on which development can be based. The Food and Agriculture Organization of the United Nations has led an international effort to create a soil map of the world to help nations and individuals understand the distribution of soil. Maps such as the world soil map are at a low level of resolution, and more detailed maps are required for resource management, but all scales of map are useful in displaying the distribution of soils.

Map scale determines the smallest area that can be easily identified on the map. The scale is selected based on the intended use of the map, the cost, and the finances available. Carefully designed transects across the landscape with many observations at close intervals is the mapping technique used to make very detailed maps. Such soil maps are made for construction projects or very-high-value crops. Fewer observations are made as map scales become less detailed. Soil classification follows the map scale with more general levels in the taxonomy used as map units on generalized maps and more detailed levels in the taxonomy used on more detailed maps.

Summary

Pedology, according to one broadly accepted definition, is the discipline that studies soil as a part of nature. Pedologists study the origin and formation of soil, the morphology of soil, the distribution of soil on landscapes, and the management of soil. The pedology entries in this encyclopedia bring to life these topics and explain the role of pedology in the study of the most important material of the Earth's surface, soil.

List of Technical Nomenclature

Aluminosilicates	Minerals with aluminum and silicon as major constituents. Examples include the clay minerals
Dynamic pedology	The study of soil formation processes by sampling and analyzing the soil solution in different horizons or at horizon boundaries
Histosols	Organic soils that have organic soil materials in more than half of the upper 80 cm, or that are of any thickness if overlying rock or fragmental materials that have interstices are filled with organic soil materials (an order in the US system of Soil Taxonomy)
Horizon	A layer of soil or soil material approximately parallel to the land surface and differing from adjacent genetically related layers in physical, chemical, and biological properties or characteristics such as color, structure, texture, consistency, kinds and number of organisms present, degree of acidity or alkalinity, etc.
Hydric soils	Soils that are wet for long enough to produce periodically anaerobic conditions, thereby influencing the growth of plants
Micromorphology	The study of soil morphology by microscopic (light optical and, less frequently, by submicroscopic) methods, often using thin-section techniques
Natural soil classification	The systematic arrangement of soils into groups or categories on the basis of their characteristics. Broad groupings are made on the basis of general characteristics and subdivisions are made on the basis of more detailed differences in specific properties
Paleosols	A soil that formed on a landscape in the past with distinctive morphological features resulting from a soil-forming environment that no longer exists at the site. A paleosol (or component horizon) may be classed as relict if it has persisted in a land-surface position without major alteration of morphology by processes of the prevailing pedogenic environment. An exhumed paleosol is one that was formerly buried and has been reexposed by erosion of the covering mantle
Parent material	The unconsolidated and more or less chemically weathered mineral or organic matter from which the solum of soils is developed by pedogenic processes

Pedogenesis	The processes that produce soils from parent materials
Pedologists	The individuals who study soils
Pedon	A three-dimensional body of soil with lateral dimensions large enough to permit the study of horizon shapes and relations. Its area ranges from 1 to 10 m^2
Soil-forming factors	Those external factors that control the kinds and rates of processes that form soils: parent material, climate, biota, topography and time and the five soil-forming factors
Soil taxonomy	The US system of soil classification
Static pedology	A method for studying soil formation by sampling the soil solid phase and measuring properties of the solids
Technical soil classification	A soil classification system that rates soils based on desired use of the soil

Acknowledgment

Many of the definitions within this List of Technical Nomenclature were reproduced with permission from the Soil Science Society of America Glossary of Soils Science Terms.

See also: **Classification of Soils**; **Factors of Soil Formation:** Biota; Climate; Human Impacts; Parent Material; Time; **Pedology:** Dynamic

Further Reading

Birkeland PW (1999) *Soils and Geomorphology*, 3rd edn. New York: Oxford University Press.

Buol SW, Hole FD, and McCracken RJ (1980) *Soil Genesis and Classification*, 2nd edn. Ames, IA: Iowa State University Press.

Food and Agriculture Organization of the United Nations (2001) *Lecture Notes on the Major Soils of the World*. World Soil Resources Report 94. Rome, Italy: FAO–UNESCO.

Jenny H (1941) *Factors of Soil Formation: A System of Quantitative Pedology*. New York: McGraw-Hill.

Jenny H (1961) E.W. Hilgard and the birth of modern soil science. In: Berkeley, CA: Farallon Publications/ Pisa, Italy: Agrochimica, Istituto di Chimica Agraria dell' Universita. *Agrochimica*.

Jenny H (1980) *The Soil Resource, Origin and Behavior*. New York: Springer-Verlag.

Tandarich JP, Darmody RG, Follmer LR, and Johnson DL (2002) Historical development of soil and weathering profile concepts from Europe to the United States of America. *Soil Science Society of America Journal* 66: 335–346.

Pickett J (ed.) (2000) *The American Heritage Dictionary of the English Language*, 4th edn. Boston, MA: Houghton Mifflin.

Wilding L (2000) Pedology. In: Sumner ML (ed.) *Handbook of Soil Science*, pp. E1–E4. Boca Raton, FL: CRC Press.

Yaalon DH (1997) History of soil science in context: international perspective. *Advances in Geoecology* 29: 1–15.

Dynamic

F C Ugolini, Università degli Studi di Firenze, Firenze, Italy

© 2005, Elsevier Ltd. All Rights Reserved.

The expression 'dynamic pedology' was first introduced in the open literature in 1978. Later, in 1985, the term was adopted and the approach used to examine the genesis of a Gray Luvisol in the boreal forest of Canada. Dynamic pedology singles out pedological investigations conducted with the use of soil solution obtained in the field through lysimetry; it does not include experimental pedology where solution is obtained by leaching or displacement from soil columns, or laboratory experiments conducted to simulate weathering. Dynamic pedology is distinguished from traditional pedology or 'static pedology' in which only the solid phase is investigated. Paraphrasing a biological expression 'dynamic pedology' is equivalent to an *in vivo* versus an *in vitro* study of the soil.

The collection, analysis, and interpretation of the soil solution allow evaluation of the contemporaneity of the pedogenic processes and hence the dynamic aspect of the soil. In contrast, the analyses of the solid phase tend to average the sum total of all the processes that have occurred since the parent material was laid down. Furthermore, the proton donors considered responsible for the formation of soils can be identified in soil solution. If pedogenesis is defined as a combination of weathering and transportation processes occurring in the presence or absence of living and dead biota, proton donors initiate weathering reactions and participate in transport of ions released during weathering and biological activity. Questions inquiring about the nature and the actuality of the processes cannot be answered by traditional pedology. This is particularly crucial when processes are at their limits as the process of podzolization in the High Arctic or in those environments where soil formation is dictated by catastrophic events. In terrestrial ecosystem studies, the watershed approach has been useful for providing a balance of the ionic

inputs and outputs; however, it has not provided an insight to soil processes. In these studies the soil is considered as a black box placed between the incoming precipitation and the stream discharge. Dynamic pedology, by examining the soil solution exiting the genetic horizons, could, on the other hand, explain the changes in ionic composition caused by the processes occurring at the horizon level. The dynamic pedology approach is suitable to examine perturbations such as substitution of species, effects of fire, addition of fertilizers, pesticides, and pollutants. During the early 1970s, when acid rain became a large issue in Western Europe and North America, soil solution collected via lysimeters became the best way to monitor soil acidification and pollutants' movement in soils.

Early Stages of Soil Development on Recent Volcanic Ejecta

An example of soil development comes from a study conducted on an Entisol (Figure 1) developed on a pyroclastic flow emplaced during the 1980s eruption of Mount Saint Helens (USA). Soil solution collected between 1984 and 1986 shows that vapors of strong acids (H_2SO_4, HCl, HNO_3, and HF) produced by volcanic emanations during the eruption reacted with the airborne tephra, resulting in the formation of soluble salts that were promptly leached once the tephra was deposited on the ground. In addition, acid precipitation (pH 3.6–5.2) from gaseous emissions produced during the dome-building stage after the eruption has contributed to the introduction in the soil of proton donors, mostly H_2SO_4 and HCl. These strong mineral acids have been neutralized within the first 5 cm of soil (Figure 2). Calcium, Mg^{2+}, Na^+, and K^+ associated with SO_4^{2-}, Cl^-, and NO_3^- dominate the soil solution together with Si. Colonization by nitrogen-fixing lupine (*Lupinus lepidus*) has introduced a new regimen in the weathering and soil genesis. Soils under lupine plants show a pH higher than the near-barren sites. The increase in pH is due to the neutralizing effect of the acid deposition by the lupine foliage, which exchanges base cations with the H^+ of the rain. In the barren areas, the consumption of protons occurs in the first 5 cm; in the presence of lupines, a redistribution of the proton load occupies the entire rooting zone. Hydrogen ions are released by the roots in exchange for cations that buffer the acidity in the foliage. As a result of lupine invasion, the soil solution under these plants contains more dissolved organic matter (DOM), H_2CO_3/HCO_3^-, NH_4^+, and Cl^-.

In summary, in the proximity of an active volcano, soil development on pyroclastic material 4 years old must be interpreted as a function of those events that have occurred prior to, during, and after the deposition of volcanoclastic material. On the barren sites, strong mineral acids are the major proton donors, but they have been neutralized in the upper part of the soil (Figure 2). Under this weathering

Figure 1 Soil profile (Entisol) developed on Mount Saint Helens pyroclastic flow. The surface of the flow is visible in the background. The measuring tape is 1.0 m, with 10-cm intervals.

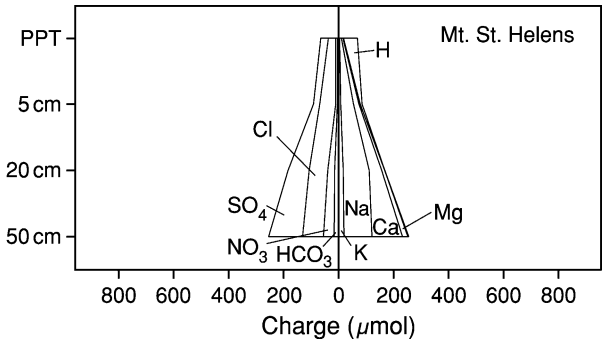


Figure 2 Charge balance for the precipitation (PPT) and for the soil solution of the Entisol, Mount Saint Helens. Concentrations of the different charged species as they exit the depths marked on the left. (Reproduced from Ugolini FC and Dahlgren RA (2002) Soil development in volcanic ash. *Global Environmental Research* 6: 73.)

regimen, there is a loss of Si, Ca, Mg, Na, and K; Fe and Al are not mobilized. With the colonization of nitrogen-fixing plants, the acidity of the incoming precipitation is neutralized by the foliage of the lupine rather than by the soil. Here, there is a shift in proton donors compared with the barren sites. Under the lupines, DOM and H_2CO_3/HCO_3^- are the major proton donors. The concentration of cations, anions, and Si under the lupines is higher than in the barren sites.

Migration of Particles in the Process of Podzolization

The unique morphology of Spodosols, supported by analytical data, has been attributed to the movement of DOM and metal cations in solution or suspension. Migration of clay particles has not been considered of importance in this process, although it has been detected in the early evolutionary stages of Spodosols. Lysimeters emplaced in well-developed Spodosols formed under Pacific silver fir (*Abies amabilis* (Dougl.) ex J. Forbes) forest, in the Washington Cascades (USA), have shown the existence of two distinct chemical compartments: the upper one starting from the top of the canopy and ending at the boundaries of the Bhs/Bs horizons, and the lower one below this boundary that reaches into the parent material. Weathering and ion transport in the upper compartment are governed by DOM that acidifies the soil (pH 3.8) and prevents the dissociation of the H_2CO_3. In the lower compartment, the pH is higher (approx. 5) than in the higher compartment, owing to the liberation of OH^- during the ligand exchange reactions between the DOM and the surface of amorphous materials. This compartment is dominated by HCO_3^- (Figure 3). Fulvic acid (FA) has been detected in the upper compartment, with 70% reduction across the Bhs/Bs. Electrodialysis, adsorption, molecular sieving, and other techniques indicate that FA is transported in solution in particulate form. However, direct evidence of their migration in soil is lacking. Lysimeters placed at the boundaries of the E, Bhs, and Bs of a Spodosol collect soil solution and suspended material, including humic substances. Upon filtration through a 0.1-μm Nucleopore filter, the suspended material is collected, lyophilized, and observed on a scanning electron microscope (SEM) equipped with energy-dispersive X-ray analysis (EDAX). Through the SEM it is possible to observe the presence of filaments 2–12 μm long extending from oblong bodies 2–3 μm wide. These bodies contain minor quantities of Al, Fe, P, S, and Si (Figure 4). Leachates collected below the Bs, on the other hand, show almost exclusively the presence of mineral particles ranging in size approximately from 2.0 to 22 μm.

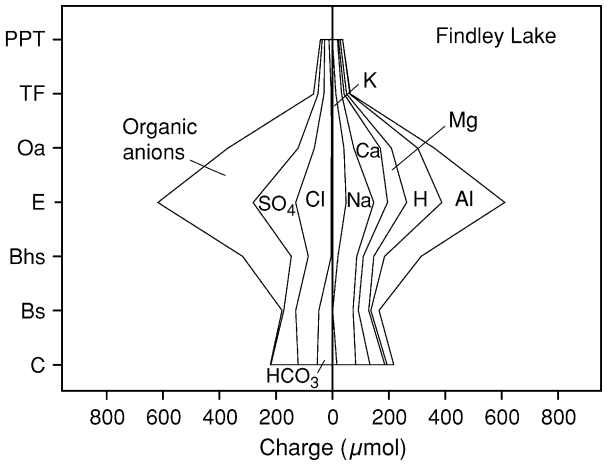


Figure 3 Charge balance for the soil solution of the Spodosol, Washington Cascades. Concentrations of the charged species for precipitation (PPT), canopy throughfall (TF), and the soil horizon solutions. The anion deficit was assumed to be due to organic anions. (Reproduced from Ugolini FC and Dahlgren RA (2002) Soil development in volcanic ash. *Global Environmental Research* 6: 73.)

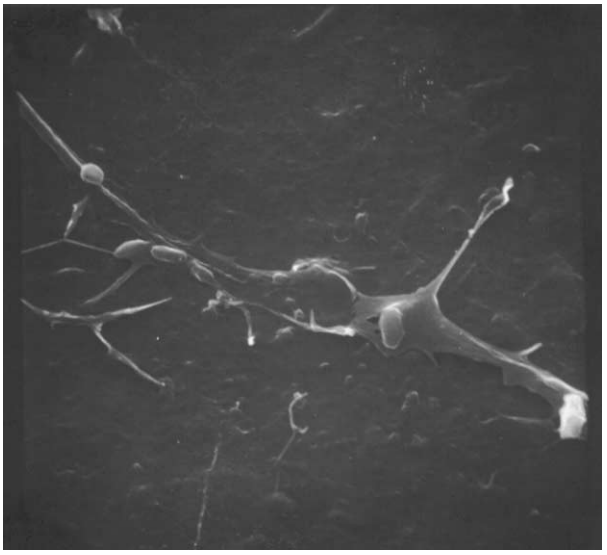


Figure 4 Scanning electron micrograph of an organic particle collected on 0.1-μm Nucleopore filter from soil solution obtained below the Oa horizon of a Spodosol, Washington Cascades. The filaments are 5–12 μm long, the oblong bodies are 2–3 μm wide (×5000).

EDAX spectra of these particles show that they are made of Si, Al with some Fe, Ca, and K, reflecting the composition of silicate minerals. This study indicates that the suspended material in the soil solution reflects the trend of the solution itself. The reduction of DOM across the Bhs/Bs is also verified by the presence of organic particles in the upper profile (Figure 4) and their absence below the Bhs/Bs (Figure 5). This study provides direct evidence for the mechanism of podzolization and the migration of particulate matter.

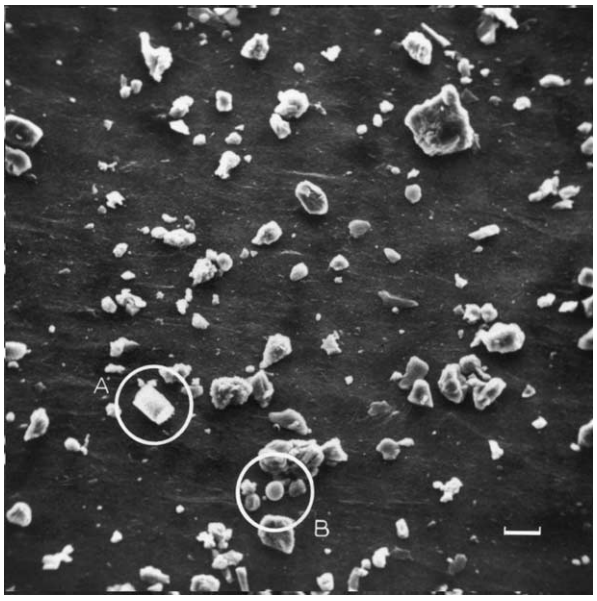


Figure 5 Scanning electron micrograph of silicate particles collected on 0.1-μm Nucleopore filter from soil solution obtained below the Bs horizon of a Spodosol, Washington Cascades. Circle A encloses a diatom, circle B, a spherical siliceous particle. Scale bar 10 μm ($\times$450). (Reproduced with permission from Ugolini FC, Dawson HJ, and Zachara J (1977) Direct evidence of particle migration in the soil solution of a Podzol. *Science* 198: 604.)

Evidence of Catastrophic Events and Horizon Inversion

Traditionally, soil formation is considered to be a relatively slow process, representing a continuum in time and space. In the absence of adequate techniques, short–term or catastrophic events in soil development have not been appreciated. However, improvement in soil-solution collection and analytical procedures has made the investigation of isolated events possible. At Walker Lake, Alaska (67° 09′ N and 154° 22′ W), the use of lysimeters has provided new evidence of threshold-controlled episodes associated with rapid solution-transport and particle migration. These events result in the formation of horizon inversion. At Walker Lake, in the boreal forest, the Spodosols have formed in the boreal forest on glacial drift deposited *c.* 13 000 BP. The horizon sequence shows a perplexing inversion of the Bhs below the Bs (Figure 6). From 25–28 July 1981, an intense storm followed by a prolonged leaching episode caused a radical departure from the average trend. The sequential change in solution composition shows a distinct flashing effect associated with normal leaching episodes. The flashing effect is a well-known phenomenon associated with antecedent moisture conditions, and it is typical of semiarid and subhumid climates. In the absence of complete

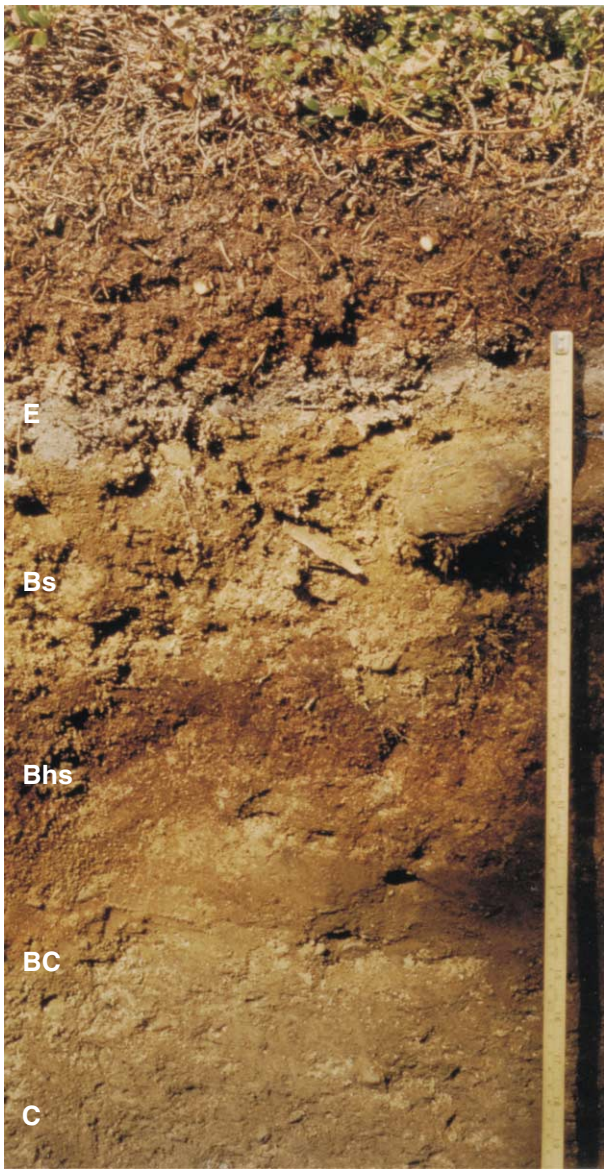


Figure 6 Boreal forest Spodosol showing an inverted sequence: a Bs over a Bhs; Walker Lake, Alaska. The yardstick is 50 cm long.

weather information for this remote site, sodium can be used as an indicator of antecedent moisture conditions. High initial concentrations of this ion indicate that a long period of drought has occurred. The high-intensity, low-frequency rainstorm delivered over 150% of the average total precipitation anticipated for the entire month. The pedogenic response was a progressive and dramatic surge in subsurface translocation processes (Figure 7). The successive stages in these processes, using iron as an indicator parameter, are shown in Figure 7. The first stage displays a normal leaching trend: depletion of Fe from the E horizon and accumulation into the Bs. After continuous leaching, phase 2 was initiated. In

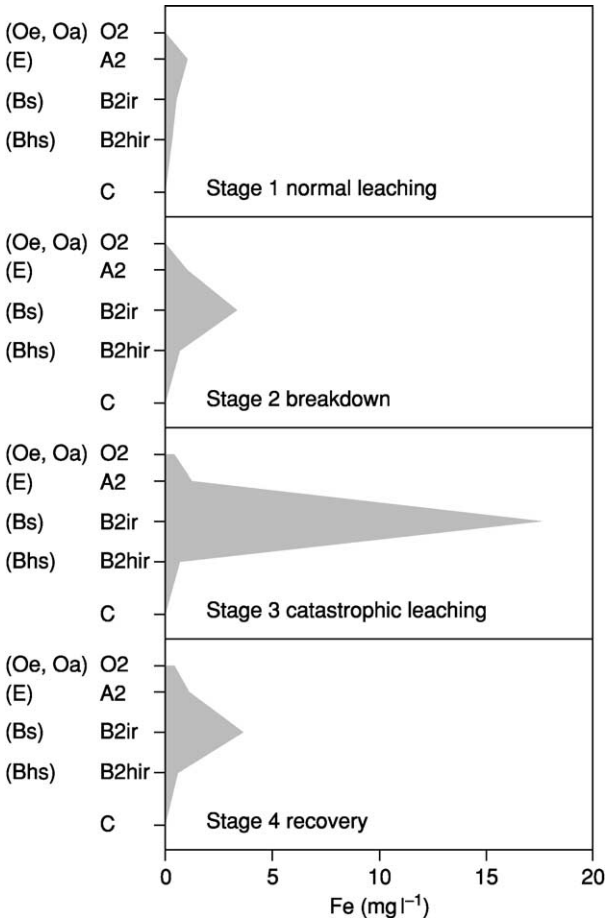


Figure 7 Dissolved iron (milligrams per liter) used as an indicator to separate the stages of a catastrophic leaching episode in a boreal forest Spodosol, Walker Lake, Alaska. (Adapted from Stoner MG and Ugolini FC (1988) Arctic pedogenesis. 2. Threshold-controlled subsurface leaching episodes. *Soil Science* 145: 49 © 1988, Lippincott Williams & Wilkins, with permission.)

this stage there is a depletion of metal cations and DOM from the Bs and translocation into the Bhs. Since no abnormal concentration of metal cations and DOM appears to leave the O and E horizon leachates, this surge indicates that the source is the Bs itself. Evidently, this horizon had been undergoing a degradation of the microstructure with the liberation of metal–organic (metal-organo) particles that had become suspended in the soil solution. Stage 3 shows a maximum liberation of particulate matter that leaves the Bs and is deposited in the Bhs. Filtered and unfiltered solutions show that 80% of Fe and Al is in particle form. This influx of suspended material tends to build up the Bhs horizon below the Bs. Stage 4 represents a recovery. This monitored response to a major rainstorm helps to explain the observed horizon inversion – the Bs over the Bhs – a feature observed in other Spodosols of the Arctic. Storms of less intensity are able to moisten only the upper part of the profile. During these episodes, the penetrating wetting front leaves mobile metal-organo complexes stranded in the upper Bs. These complexes are only temporarily detained, because their metal-to-C ratio is too low to induce their precipitation. This example serves to demonstrate the usefulness of lysimetry for capturing the impact of individual rainstorms. Further, it appears that in subhumid climates soil development occurs through episodic pulses corresponding to rainstorms of moderate magnitude. This trend is punctuated by catastrophic events that are responsible for the unusual migration of Fe, Al, and organic matter in particulate form. The catastrophic events are responsible for the observed B-horizon inversion. A similar horizon inversion has been observed for a Podzol located in the Massif Central (France). However, this phenomenon is explained by two separate pedogenic phases involving truncation and burial.

Effects of Species Substitution

Dynamic pedology is particularly useful when applied to study the impacts of replacement of species either naturally or anthropically induced. In northern Idaho, disturbed areas such as clearcuttings in the Grand Fir Mosaic Ecosystem are poorly revegetated by the native firs. Bracken (*Pteridium aquilinum* (L.) Kuhn) rapidly occupies the disturbed sites preventing the regeneration of the Grand Fir (*Abies grandis* (Dougl. ex D. Don) Lindley). Solutions collected under the bracken show a lower pH (minimum 4.8), a higher Al concentration (maximum 1.6 mg l^{-1}), and higher DOM than the undisturbed forest during 1994 and 1995.

Another study on the effects of species substitution has been conducted in northeastern Japan, in the Miyagi prefecture, to examine, through soil solution, the effects of Japanese oak trees (*Quercus serrata* Thumb.) on Alic Melanudands (nonallophanic Andisols). The oaks were established 50 years ago on a site originally occupied by Japanese pampas grass (*Miscanthus sinensis* Anders). The soil-solution composition collected under the Japanese oaks has been compared with that of the pampas grass growing in a plot next to the oaks (Figure 8). The acid precipitation (pH 4.8) caused by the presence of HNO_3 and H_2SO_4, of volcanic origin, and paddy fields is partially neutralized by the pampas grass canopy from pH 4.8 to 6.0, while the oak canopy increases the pH only to 5.4. As the canopy drip penetrates the soil, the soil solution tends to acquire a lower value (pH 4.6) in the upper 50 cm of the soil under oak than in the pampas grass (pH 5.5). Al concentrations, in the horizon under the oaks are significantly higher (38 μmol l^{-1}) than in the pampas

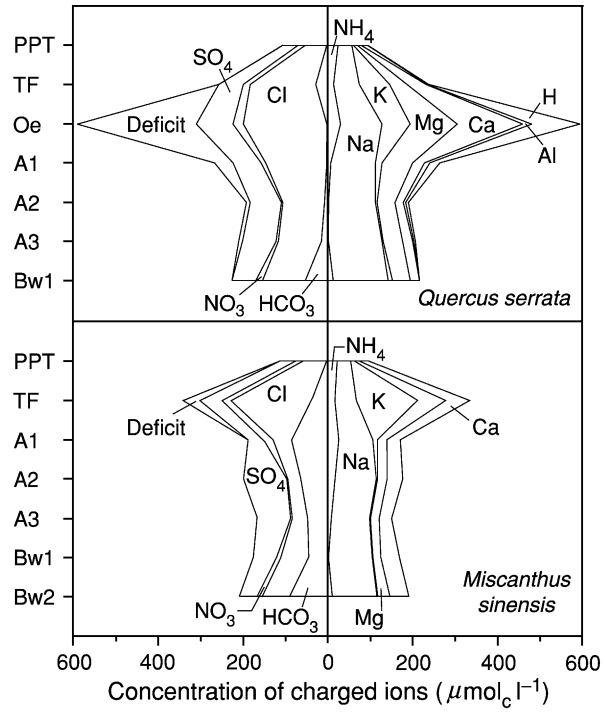


Figure 8 Charge balance for the soil solution of the Andisols developed under *Quercus serrata* and *Miscanthus sinensis*, Japan. The graph shows the concentrations of the charged species for precipitation (PPT), canopy throughfall (TF), and the soil horizon solutions. The anion deficit was assumed to be due to organic anions. (Reproduced with permission from Dahlgren RA, Ugolini FC, Shoji S, Ito T, and Slatten RS (1991) Soil-forming processes in Alic Melanudands under Japanese pampas grass and oak. (1991) *Soil Science Society of America Journal* 53: 1055.)

grass ($4\,\mu mol\,l^{-1}$). However, in both cases there has been no significant movement of Al into the B horizon. Iron concentrations are less than $1\,\mu mol\,l^{-1}$ in all horizons, except the Oe under the oaks, where a mean value of $0.4\,\mu mol\,l^{-1}$ has been recorded. Silicon concentrations under the oaks are high ($231\,\mu mol\,l^{-1}$) in the Oe due to biocycling. In the pampas grass plot, the highest levels are approximately $122\,\mu mol\,l^{-1}$ in the A horizon. Also, here the large amount of Si is due to the biocycling of silicon by the pampas grass. DOM concentrations reach a maximum of $3000\,\mu mol\,l^{-1}$ in the Oa of the oak site, but they decrease to less than $500\,\mu mol\,l^{-1}$ in the Al horizon. In the pampas grass site, the values are lower, approximately $1000\,\mu mol\,l^{-1}$, than under the oak. Under both species the mobility of DOM is limited to the surficial horizons. Uptake processes, in addition to the task of neutralizing the acid precipitation, mostly affect base cations (Na, K, Ca, and Mg). Calcium shows a higher concentration in the pampas grass, while Mg and Na are more abundant in the oaks. Anions show both differences and similarities between the two species. Chloride and sulfate show the highest concentrations. Sulfate reaches a minimum value in the B horizons, where it is retained by amorphous material.

Bicarbonate levels are higher under the pampas grass than under the oaks. In conclusion, the invasion of Japanese oak trees on a soil previously occupied by pampas grass has resulted in acidification of the soil solution in the upper 50 cm of the profile. This acidification has resulted in an increase in Al concentration and in a decrease in bicarbonate. Silicon tends to accumulate in the B horizon of the pampas grass, but with a net loss under the oaks. Also the oaks induce a greater cycling of Ca, Mg, K, and Cl. In spite of these differences, the process of andosolization, as characterized by *in situ* weathering in the B horizon and accumulation of organic carbon in the A, is still active under both species. The proton donors in the Oa of the oak site are the organic acids, while these acids have a lesser role in the pampas grass site, where carbonic acid prevails. Protons released by cation uptake dominate in the rooting zone of the oaks.

Impact of Burial by Volcanic Ash on Preeruption Soils

Regions in the proximity of active or dormant volcanoes are subjected to intermittent deposition of airfall tephra. These additions tend to rejuvenate the soil surface of the existing soils, bringing a high level of fertility and making these soils among the most productive in the world. The added volcanoclastic material varies in thickness, depending on many factors: volume of the ejecta, wind direction and velocity, and proximity to the source. Areas throughout the world that have received increments of airfall tephra display buried evidence associated with lithologic discontinuities. Although this phenomenon is widespread, there is a lack of information on the processes occurring when a new layer of fresh material is added to an existing soil. Dynamic pedology is the most suitable approach, because it provides an instantaneous record of the events and processes induced by the addition of tephra. Soils of the Pacific Northwest, USA, have been subjected, throughout the Holocene, to deposition of tephra produced by the volcanoes that dot the western part of the states of Washington, Oregon, and northern California. As recently as May 18, 1980, the eruption of Mount Saint Helens, western Washington, USA, provided an estimated $1.1\,km^3$ of airfall tephra that reached the soils along the fallout plume (Figure 9). In connection with this eruption, an experiment has been conducted by adding fresh tephra to soils that had received airfall tephra during the Holocene but had escaped the eruption of 18 May. Three metric tons of clean tephra have been

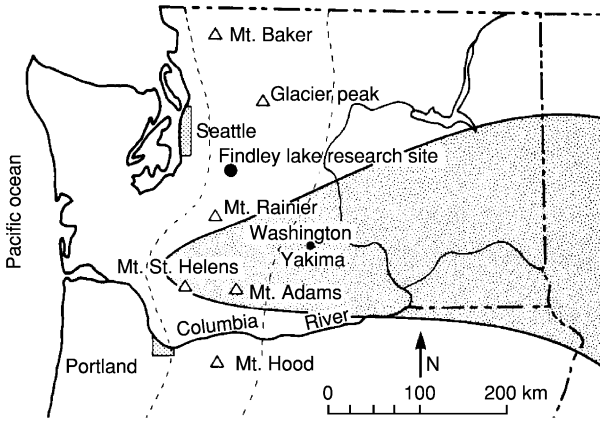


Figure 9 The shaded area indicates the approximate distribution of the 1980 Mt. St. Helens tephra. The location of the study, Findley Lake, Washington Cascades, is marked. (Adapted from Ugolini FC, Dahlgren RA, LaManna J, Nuhn W, and Zachara J (1991) Mineralogy and weathering processes in Recent and Holocene tephra deposits of the Pacific Northwest, USA. *Geoderma* 51: 279, with permission.)

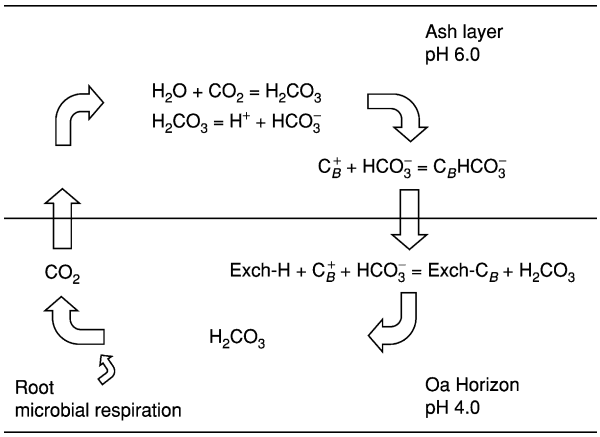


Figure 10 The CO_2–H_2CO_3–HCO_3 weathering and transport cycle between the added ash layer and the buried organic horizon (Oa). C_B, base cation. (Adapted from Dahlgren RA, Ugolini FC, and Casey WH (1999) Field weathering rates of Mt. St. Helens tephra. *Geochimica et Cosmochimica Acta* 63: 596, with permission.)

collected and applied in 5- and 15-cm increments to plots to be compared with those that had not received tephra. Hence a modern analog has been created of past burial episodes. The applied tephra contains soluble salts, mostly sulfates of Ca and Na formed while the ash was still airborne. These salts are readily leached once the tephra is exposed to the rainy and cool climate of the central Cascades, where the experiments have been carried out. The area receives 2300 mm of precipitation, two-thirds in the form of snow, and experiences a mean annual temperature of 5.5°C. The basic cations present in the leachates upon entering the buried forest floor displace H^+ and Al^{3+}. Consequently, solute entering the mineral soil (E horizon) have a low pH and a high concentration of soluble Al. Biological uptake, adsorption, and exchange reactions attenuate the pulse of basic cations. Calcium and K^+ are preferentially retained in the E horizon; Mg^{2+}, Na^+, and Cl^- are quickly removed from the solum. Sulfate is retained in the B horizon owing to ligand exchange reactions with imogolite/allophane material enhanced by the lowering of the pH. As the input from the fresh tephra of SO_4 decreases, most of the adsorbed SO_4 is desorbed from the B horizon. Al–bearing secondary minerals apparently control Al concentrations in the leachates collected below the B horizon. This *in vivo* experiment has revealed the mechanism of weathering in the newly added tephra layer. Carbonic acid is the major proton donor that liberates basic cations and Si and provides the conjugate base for their transport. Upon percolating through the acidic forest floor, the HCO_3^- gains a proton and is converted to H_2CO_3. At the low pH of the forest floor (approx. 4.0), volatilization of HCO_3 occurs, with the consequent diffusion of CO_2 upward into the tephra layer, where the pH (approx. 7.0) favors the formation of H_2CO_3, thus initiating a new weathering/transport cycle (Figure 10). This weathering scenario was observed during the first 4 months of monitoring the soil solution, but the process continued for 4 years. The information collected has allowed an estimation of field weathering rates, as well as of the weathering of the solid phase, after 10 years of exposure to the cool humid climate of the Cascades Mountains.

Testing New Theories of Soil Formation

From 1980 the classic theory of podzolization, also commonly known as the 'fulvate theory,' has been challenged by Farmer and his coworkers (Figure 11a). The fulvate theory holds that Al and Fe mobilization and transport occur through complexation with soluble organic acids. The new theory proposes that the migration of Al from the E horizon of Podzols occurs as positively charged, soluble hydroxyaluminosilicate sols called 'protoimogolite.' They are adsorbed/precipitated in the Bs and eventually transformed into imogolite (Figure 11b). Support for this process is the repeated presence of imogolite in the Bs horizons of Podzols/Spodosols derived from a variety of parent materials. However, in the absence of laboratory or field experiments that show the migration of protoimogolite from the E horizon, its presence in the Bs does not insure that it is derived from the E. It could have formed *in situ* in the Bs. Furthermore, previous experiments with aqueous solutions containing organic matter have shown the ability of

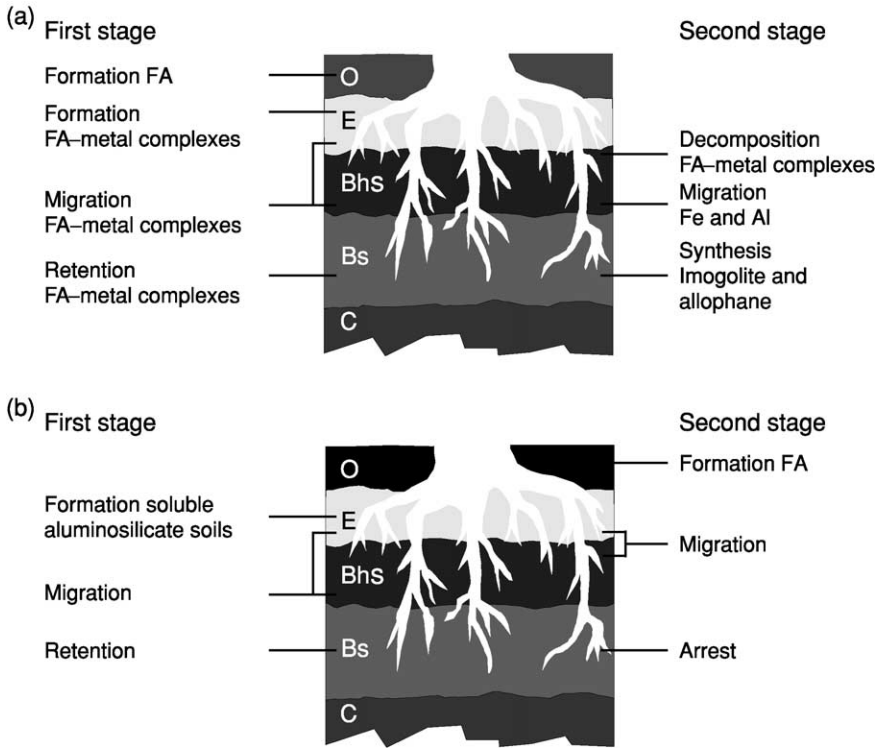


Figure 11 Schematic representation of the theories of podzolization: (a) fulvate theory; (b) protoimogolite theory. (Adapted from Ugolini FC and Dahlgren RA (1987) Mechanism of podzolization as revealed by soil solution studies. In: Righi D and Chauvel A (eds) *Podzols et Podzolization*. INRA, p. 196, with permission.)

these compounds to complex and mobilize Al and Fe. Also, soil solutions collected by lysimeters from a subalpine Podzol/Spodosol have evidenced the role of soluble organic matter in the process of podzolization. Nevertheless, to establish the validity of the proposed 'protoimogolite theory,' at least for the Pacific Northwest Podzols/Spodosols, it is crucial to answer a number of questions, such as: Does the Al migration occur in organic or inorganically complexed forms? Does protoimogolite migrate in the soil solution? Where is imogolite found in a subalpine Podzol/Spodosol? Is imogolite thermodynamically stable within the soil profile under the current weathering regimen?

To answer these questions it is necessary to collect soil solutions from the genetic horizons of a Podzol/Spodosol and to have available a method to fractionate the Al so as to establish the forms of this metal. Podzol/Spodosol profiles, located in the Cascades Mountains of Washington (USA), have previously been equipped with lysimeters and intensively studied. Soil solutions have been collected seasonally from this subalpine Podzol/Spodosol except in the summer because of low precipitation. Al fractionation is performed according to the technique proposed by Driscoll. This procedure partitions Al into three operationally defined fractions: strongly bound Al complexes, nonlabile monomeric Al, and labile monomeric Al. The labile monomeric Al includes aqueous Al as well as inorganic complexes. The Al^{3+} is estimated by speciating the labile monomeric fraction using GEOCHEM. The nonlabile monomeric fraction represents weakly held Al–organic complexes, while the strongly bound Al complexes appear to be a combination of colloidal Al or strong Al–organic complexes.

The first question is: In what form(s) does Al migrate from the E horizon? The answer is that between 73 and 100% of Al is organically complexed, mainly with the nonlabile monomeric form (Figure 12).

The second question asks whether protoimogolite migrates in the profile. The finding that Al in solution is organically complexed diminishes the role of inorganic Al forms and therefore of protoimogolite in the Al migration. However, an additional test has been carried out. Assuming, according to Farmer, that protoimogolite sols are present in the soil solution leaving the E horizon, being positively charged, they should have been retained on a cation exchange resin and hence decreased the Si concentrations of the solutions that had passed through the resin. However, statistical analysis of Si before and after passage through the cation exchange resin shows no change in Si concentration. This indicates the absence of

protoimogolite sols in the soil solution. Additionally, if the soil solution exiting the E horizon does contain protoimogolite, the value of the total Si, which includes also the Si in the aluminum silicate soils, should be higher than monomeric Si. However, the two values are the same; consequently, Si moves only in monomeric form and is not associated with Al. From these findings, apparently in these Podzols there is neither appreciable movement of inorganic Al nor migration of aluminum silicate soils.

To answer the question of whether imogolite is present in the Podzol under study, a search for this mineral has been made in all the inorganic horizons: E, Bhs, Bs, BC, and C. The methods used include selective dissolution, IR spectroscopy, scanning and transmission electron microscopy, and potentiometric titration. In the clay fraction, the minerals of short-range order represent less than 5% in the E horizon, while they range from 20 to 50% in the B and C horizons.

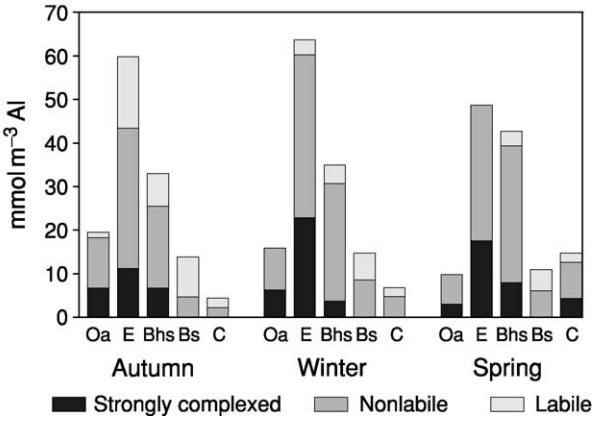


Figure 12 Aluminum fractionaton of soil solution from a Spodosol, Washington Cascades, into: strongly complexed, nonlabile monomeric and labile monomeric fractions. (Adapted from Dahlgren RA and Ugolini FC (1989) Aluminum fractionation of soil solutions from unperturbed and tephra-treated Spodosols, Cascade Range, Washington, USA. *Soil Science Society of America Journal* 53: 563, with permission.)

The final question is: Is imogolite thermodynamically stable in the soil studied? More specifically, is imogolite stable in the E horizon, where it should form according to the protoimogolite theory of podzolization? Chemical equilibria are used to determine the stability of imogolite and other minerals (Figure 13). The calculated saturation indices show that soil solution in the E horizon is undersaturated with respect to imogolite; consequently, the imogolite sols cannot be stable in this horizon. Soil solution from the Bs and C horizons appears to be in equilibrium with the hydroxyaluminum interlayer of 2:1 clay minerals and imogolite. The hydroxyaluminum interlayer seems to control Al activities, while imogolite controls H_4SiO_4 activities. From this study it appears that Farmer's proposed theory of protoimogolite to explain the process of podzolization is not applicable in Podzols/Spodosols developed in cool to cold forested environments, where organic acids tend to dominate the upper part of the profile where protoimogolite should form. Al is effectively complexed by the DOM and therefore not available for reacting with Si to form aluminum silicate sols.

Summary

Dynamic pedology, in the context of soil science, occupies a particular role. Based on collection, analysis, and interpretation of soil solution, dynamic pedology attempts to provide answers that traditional solid-phase investigations cannot give. It obtains an instantaneous picture of the ongoing processes and the impact of perturbations. However, a complete understanding of soil can be acquired by studying the liquid, the solid, and the gaseous phases. The adoption of dynamic pedology is restricted to those soils that experience percolative events, although collection of solution has been secured by irrigating the soil. Inherited limitations of dynamic pedology are obviously dependent on the collection of representative soil solution, which in turn is conditioned by the residence

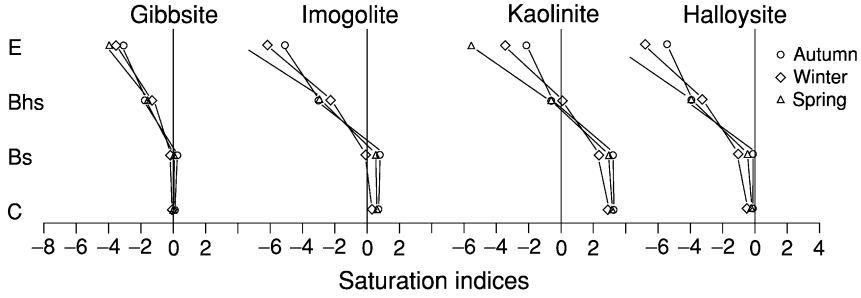


Figure 13 Saturation index of selected minerals calculated for the autumn, winter, and spring soil solutions collected from a Spodosol, Washington Cascades. (Adapted from Dahlgren RA and Ugolini FC (1989) Formation and stability of imogolite in a tephritic spodosol, Cascade Range, Washington, USA. *Geochimica et Cosmochimica Acta* 53: 1901.)

time of the water in the soil. Solution collected by gravitational means, zero-tension lysimeters, or tension lysimeters may differ. Seasonal variations in solution with tension lysimeters held at a constant of 10-kPa tension are few compared with solutions obtained with low (0–30 kPA)- and high-speed centrifuge (30–3000 kPa). Also the seasonal changes in solution composition determine the thermodynamic stability of minerals. Limitations in the use of dynamic pedology are therefore related to the sampling methodology and the time of sampling of the soil solution.

List of Technical Nomenclature

Concentrations:

μmol	Micromoles
$\mu mol_c l^{-1}$	Micromoles of charge per liter
mg l^{-1}	Milligram per liter
mmol m^{-3}	Millimoles per cubic meter

Length:

μm	Micrometer
cm	Centimeter
km	Kilometer
m	Meter
mm	Millimeter

Pressure:

kPa	Kilopascal

Temperature:

°C	Degrees Celsius

Volume:

l	Liter

Further Reading

Dahlgren RA and Ugolini FC (1991) Distribution and characterization of short-range-order minerals in Spodosols from the Washington Cascades. *Geoderma* 48: 391–413.

Dawson HJ, Ugolini FC, Hrutfiord BF, and Zachara J (1978) Role of soluble organics in soil processes of a Podzol, central Cascades, Washington. Soil Science 126: 290–296.

Driscoll CT (1984) A procedure for the fractionation of aqueous aluminum in dilute acid waters. *International Journal of Environmental Analytical Chemistry* 16: 267–284.

Duchaufour P (1982) *Pedology.* (Translated by Paton TR.) London, UK: Allen and Unwin.

Johnson-Maynard JL, McDanial PA, Ferguson DE, and Falen AL (1998) Changes in soil solution chemistry of Andisols following invasion of bracken fern. *Soil Science* 163: 814–821.

Likens GE, Bormann FH, Pierce RS, Eaton JS, and Johnson NM (1977) *Biogeochemistry of a Forest Ecosystem*. New York: Springer-Verlag.

Malcom RL (1990) Variations between humic substances isolated from soils, streams waters, and groundwaters as revealed by ^{13}C-NMR spectroscopy. In: MacCarthy P *et al.* (eds) *Humic Substances in Soil and Crop Sciences: Selected Readings*, pp. 13–35. Madison, WI: American Society of Agronomy Soil Science Society of America.

Mattigod SV and Esposito G (1979) Chemical modeling of trace metal equilibria in contaminated soil solutions using a computer program GEOCHEM. In: Jenne EA (ed.) *Chemical modeling of aqueous systems*, pp. 837–856. Washington, DC: American Chemical Society Publications.

Nuhn WW (1987) *Soil Genesis on the 1980 Pyroclastic Flow of Mt. St. Helens*. MS thesis, College of Forest Resources. Seattle, WA: University of Washington.

Pearson RW (1971) Introduction to symposium – the soil solution. *Soil Science Society of America Proceedings* 35: 417–420.

Schnitzer M and Kodama H (1977) Reactions of minerals with soil humic substances. In: Dixon JB and Weed SB (eds) *Minerals in the Soil Environment*, pp. 741–770. Madison, WI: Soil Society of America.

Schwab AP (2000) The soil solution. In: Sumner ME (ed.) *Handbook of Soil Science, sect B, Soil Chemistry,* vol. 3, pp. 85–122. New York: CRC Press.

Singer M, Ugolini FC, and Zachara J (1978) In situ study of podzolization on tephra and bedrock. *Soil Science Society of America Journal* 42: 105–111.

Stevenson FJ (1994) *Humus Chemistry, Genesis, Composition, Reactions*, 2nd edn. New York: Wiley.

Ugolini FC and Sletten RS (1991) The role of proton donors in pedogenesis as revealed by the soil solution studies. *Soil Science* 151: 59–75.

Ugolini FC, Zachara JM, and Reanier RE (1982) Dynamics of soil-forming processes in the Arctic. *Permafrost*. pp. 103–115. The 4th International Conference Proceedings of National Research Council of Canada, Ottawa. Ottawa, ON: NRC Research Press.

Ugolini FC, Dahlgren RA, Shoji S, and Ito T (1988) An example of andosolization and podzolization as revealed by soil solution studies, S. Hakkoda, NE Japan. *Soil Science* 145: 111–125.

Zabowski D and Ugolini FC (1992) Seasonality in the mineral stability of a subalpine Spodosol. *Soil Science* 154: 497–507.

PEDOMETRICS

I O A Odeh and A B McBratney, University of Sydney, Sydney, NSW, Australia

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

The term 'pedometrics,' first coined by McBratney in the late 1980s, 'is a neologism derived from the Greek roots pedos or $\pi\varepsilon\delta o\sigma$ (soil) and metron or $\mu\varepsilon\tau\rho o\nu$ (measurement).' It is used in a similar fashion to other words such as biometrics, psychometrics, econometrics, chemometrics, and, the oldest of all, geometry. Furthermore the term covers two main ideas – the 'soil' or 'pedo' part, which corresponds to that branch of soil science we call pedology, and the 'metrics' part, which is restricted to mathematical and statistical methods. Simply defined, pedometrics is the use of quantitative methods for the study of soil distribution and genesis, and as a sustainable resource. Another problem-oriented definition is 'soil science under uncertainty.' In this sense pedometrics deals with uncertainty in soil models that describe deterministic or stochastic variation, vagueness, and lack of knowledge of soil properties and processes (Table 1). Thus mathematical, statistical and numerical methods could be applied to resolve the uncertainty and complexity inherent in a soil system model, including numerical approaches to classification, which deals with supposedly deterministic variation.

Pedometrics is not new, as mathematical and statistical methods have been applied to soil studies since the 1960s and 1970s. However, it is now emerging as a technical branch of soil science complementing traditional pedology. Over time the use of computers has increased in both fields, and the difference between the two has decreased and, in some cases, overlapped. Due to new demands for quantitative soil information required for global-scale models, regional environmental planning, and field-scale agricultural land management, traditional pedology has become more quantitative through the increased use of computerized soil information systems. Concurrently, pedometrics has emerged as a collection of quantitative tools, which are increasingly being used to account for conceptual pedological models of soil variation. As of 2003, there is a strong and growing overlap and synthesis between 'traditional' pedology and quantitative pedology or pedometrics (as shown in Figure 1).

Methods and Problems

Pedometrics: Principles and Hypotheses

As the definition implies, pedometrics encompasses the whole quantitative approach to the study and description of soil, especially soil in the field. It also involves the use of mechanistic, stochastic, deterministic, and empirical models (Table 1). A more restricted form of pedometrics is one that involves applications of statistical and probabilistic modeling to soil variation. Pedometrics principles are therefore embodied in (geo-)statistical modeling of pedological processes and the associated phenomena as they influence soil variation, and the associated model uncertainty.

Pedometrics is therefore aimed at resolving a number of questions or hypotheses:

1. What is the pattern of soil distribution in multivariate space, i.e., soil taxonomy?
2. How does the soil vary spatially and temporally?
3. What are the utility and quality of soil?
4. What are the causal factors of soil, i.e., soil genesis?

Table 1 Types of quantitative models used for soil studies with examples and sources of uncertainty

Causality/ uncertainty	*Model*		
	Deterministic	*Stochastic*	*Nonstatistical*
Empirical	Generalized linear models, numerical taxonomy, Jenny functional relations and canonical ordination	Time series, spatial processes, temporal and spatial variation of soil properties	Fuzzy systems
Mechanistic	Flow and diffusion of soil plasmic materials, profile and landscape development	Indeterminate	Indeterminate
Uncertainty	Imprecise measurements, model uncertainty	Random process, probability	Vagueness, ambiguity, and fuzzy geostatistics

Source: McBratney AB and Moran CJ (1986) *Pedometric Research. CSIRO Australia Division of Soils Research Report.* Canberra, ACT: CSIRO Division of Soils.

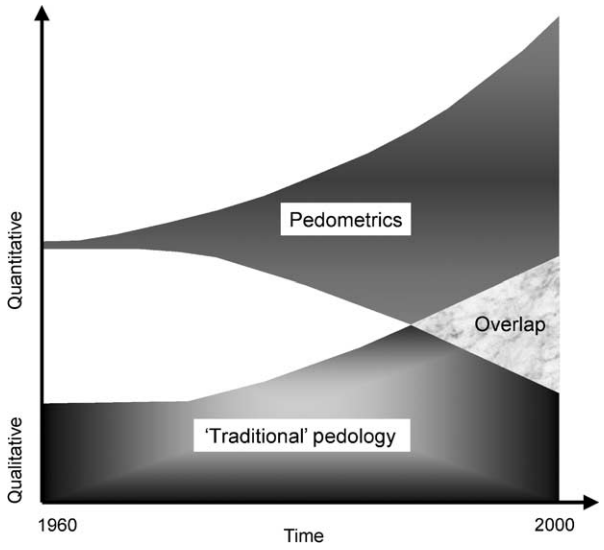


Figure 1 A time line of the growth of pedology and pedometrics. (Source: McBratney AB, Odeh IOA, Bishop TF, Dunbar MS, and Shatar TM (2000) An overview of pedometric techniques for use in soil survey. *Geoderma* 97: 293–327.)

5. What is the uncertainty of modeling the spatial pattern of soil phenomena, i.e., is a model a true representation of reality?

With the exception of (5), these hypotheses dovetail well with the requirements of conventional pedology. Although in the past developments in pedometrics have been restricted to statistics and probability, increasingly mechanistic models are being applied to model soil processes.

Soil Measurements and Properties

In order to model soil quantitatively, we need to perceive it, i.e., sense it or measure some quality or quantity of it. In other words, how do we numerically represent soil properties in order mathematically and/or statistically to model their variation? Ever since the advent of modern environmental science, technological advances have improved immensely the way we characterize or measure the quality and/or quantity attributes of soil. Soil measurement techniques include direct measurements, visual observations, and remote sensing (including proximal, airborne, and space-based modes of sensing).

Soil color and spectroradiometry Measurement of soil color has gained tremendously from advances in technology, particularly in spectroradiometry. The human eye is capable of seeing colors made up of combinations of reflected wavelengths throughout the visible portion of the electromagnetic spectrum.

The Munsell System of Color Notation, a color order system invented by Albert H. Munsell in 1905, is the most-used soil color system prior to the development of a quantitative means of measuring soil color. It specifies a limited number of 'standard' colors, within an asymmetrical, 'cylindrical' color space. While the hue (H) describes the similarity to a dominant color (red, yellow, and blue) or a combination of any two, the value (V) is the color intensity (lightness) and chroma (C) is the relative purity of the dominant wavelength (saturation). The Munsell color notations are semiquantitative, requiring further transformation for quantitative data analysis.

In 1931 the CIE, or Commission Internationale de l'Eclairage, standardized color order systems by specifying the light source, the observer, and the methodology used to derive the values for describing color. The CIE system defines how to map a spectral power distribution to a set of tristimulus coordinates which describe the color space: a luminance component Y and two other 'imaginary' components of the primary spectra, X and Z (Figure 2a). These in turn could be mapped to xy coordinates of chromaticity (Figure 2b). More recently, in 1978 the CIE has provided additional transformations (CIELUV and CIELAB color spaces) to make color representation on a continuous numeric color space more amenable for pedometric analysis.

Optical properties of the soil are primarily a function of its mineral composition, since the soil is formed from the transformation of rocks. A soil reflectance spectrum is therefore superimposition of spectra of the soil mineral components. Thus a soil type has its unique spectral reflectance curve. *In situ* or reference data are often collected at the time of image acquisition. One form of reference data is the ground-based measurement of the reflectance of surface features to determine their spectral response patterns. This might be done in the laboratory or in the field using a spectroradiometer. This device measures, as a function of wavelength, the energy coming from an object within its view. It could be used to prepare spectral reflectance curves of many soil types for classification or comparison purposes (e.g., Figure 3). Pedometricians are currently investigating the predictive power of such spectra. Such analysis is well known in chemometrics.

Field properties: hard and fuzzy descriptors? Other than color, many soil morphological properties, required in routine soil surveys, are described in somewhat vague terms. For example, soil structure grade is described as 'structureless' or 'weak' or 'moderate' or 'strong' in many national soil survey handbooks. To make the grades more amenable to quantitative

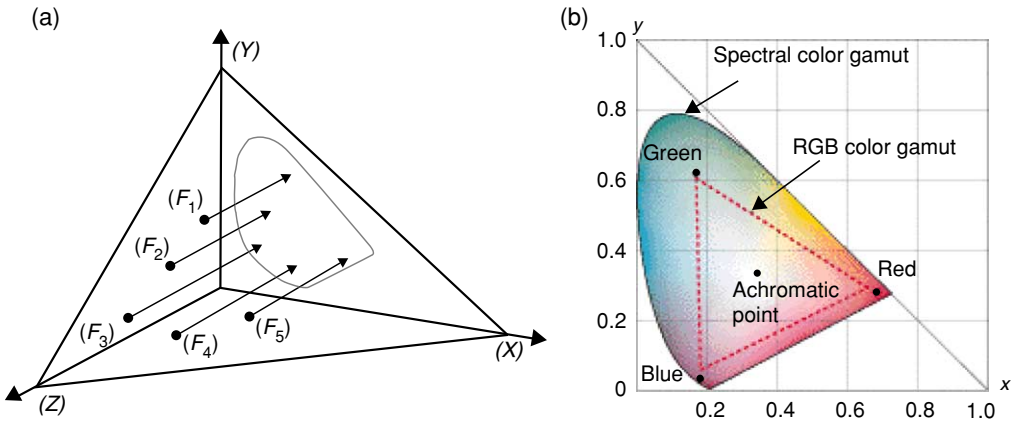


Figure 2 (a) Mapping of *XYZ* tristimulus values to (b) an *xy* chromaticity.

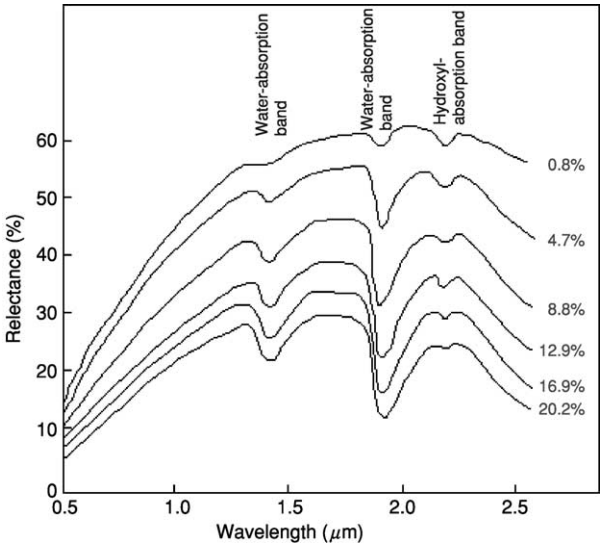


Figure 3 Spectral reflectance curves for Newtonia silt loam at various moisture contents. (Adapted from Bowers SA and Hanks RJ Reflection of radiant energy from soil. *Soil Science* (1965) 100: 130–138.)

Table 2 Fuzzy coding of structural type

Type	*Horizontality*	*Verticality*	*Flatness*	*Accommodation*
Platy	1.0	0.1	1.0	1.0
Lenticular	1.0	0.3	0.3	1.0
Prismatic	0.2	1.0	1.0	1.0
Columnar	0.2	1.0	0.9	0.9
Angular blocky	1.0	1.0	1.0	1.0
Subangular blocky	0.7	0.7	0.5	0.5
Granular	0.2	0.2	0.0	0.1
Massive	0.0	0.0	0.0	1.0
Single grain	0.0	0.0	0.0	0.0

Reproduced with permission from Odeh IOA, McBratney AB, and Chittleborough DJ (1991) Elucidation of soil–landform interrelationships by canonical ordination analysis. *Geoderma* 49: 1–32.

analysis (e.g., multivariate statistical analysis) these terms are usually coded as 0, 1, 2, and 3, respectively. But because of the vagueness in the linguistic characterization, fuzzy coding, which accounts for uncertainty or lack of clear boundary between the grades, has been used. Other morphological descriptors such as soil aggregate (structure) size classes, e.g., 'fine,' 'medium,' and 'coarse,' may be given linguistic variables which could be fuzzified to determine the degree of truth that a given soil layer is characterized by. Simple examples of fuzzy coding are presented in Table 2.

Pedotransfer functions The term 'pedotransfer function' (PTF) is generally defined as translating the raw soil data into more useful information. It can also be defined as predictive functions of certain soil properties from other easily, routinely, or cheaply measured properties. The most readily available data come from soil survey, such as field morphology, texture, structure, and pH. PTFs add value to the basic soil data by translating them into predicates of other more laborious and expensively determined soil properties. These functions fill the gap between the available soil data and the properties, which are more useful or required for a particular model or quality assessment. A simple illustration of how to apply PTFs to practical problems is shown in Figure 4.

Quantitative Soil Geography – Classification Models

Soil Classification and Soil Map Units

The pattern of soil in the landscape is as a result of effects of soil-forming factors, namely: climate, organisms, parent material, topography, and time

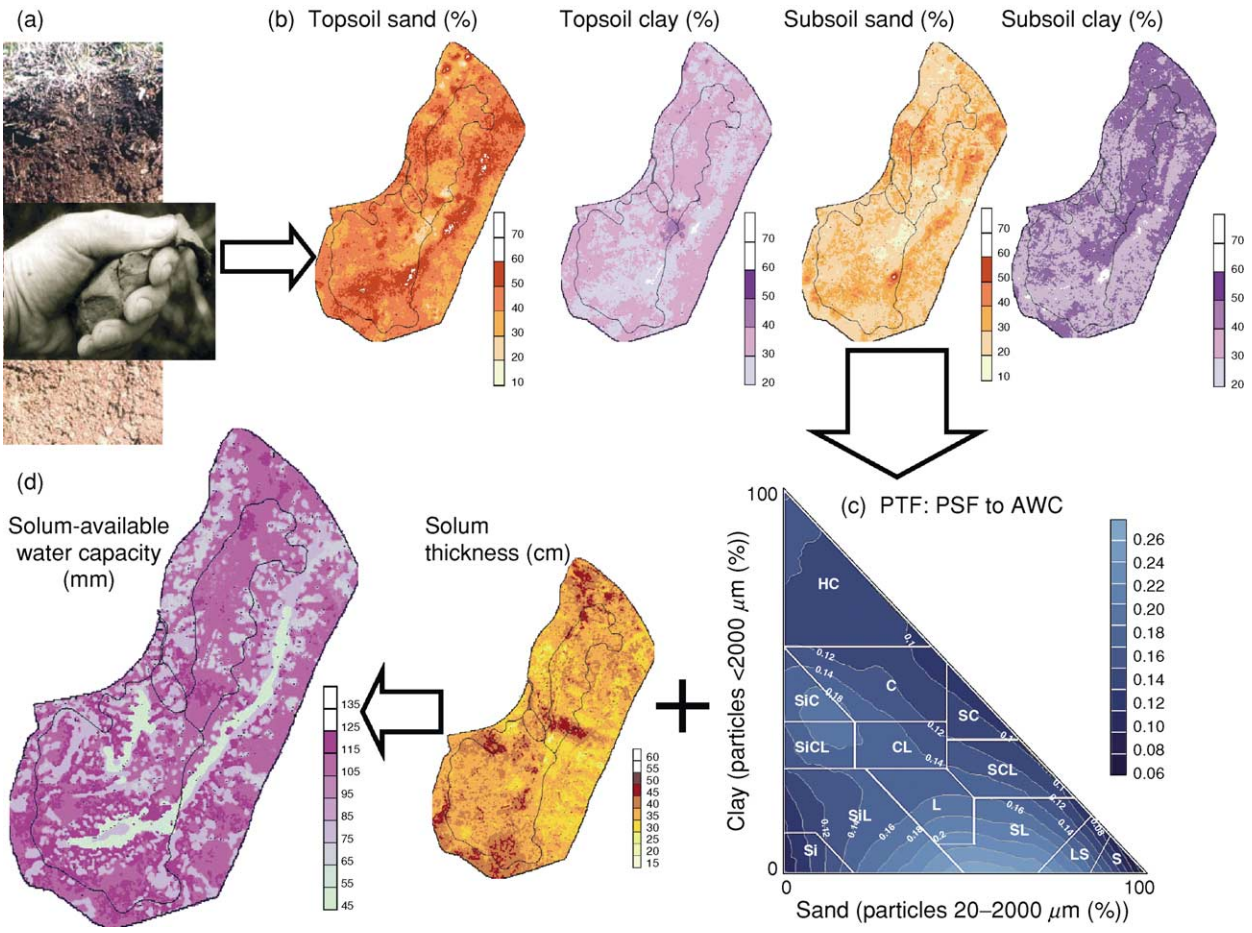


Figure 4 A simple example of pedotransfer functions (PTFs): (a) field texture was first determined by 'feel' method for topsoil and subsoil at each sample location; (b) particle-size fractions were then estimated from the textural classes and interpolated on to a fine grid; (c) horizon-available water capacity was in turn estimated through a PTF expressed as a function of particles sizes; (d) a combination of both horizons produces the solum-available water capacity. AWC, available water capacity.

(usually referred to as 'corpt,' but recently extended to 'scorpan' – see Eqn [2], below). These factors, particularly climate and parent material, produce broad patterns of soil in geographic space. The complex combination of these factors is what causes repetitive patterns of soil in the landscape, which form the basis for soil classification, identification, and mapping. In traditional soil classification, both the conceptual and real classes are qualitative and have no clear-cut boundary. Pedologists have, in the past, attempted to avoid these by applying numerical classification to complex soil data. Soil classification, in both geographical and taxonomic space, is a simple representation of complex, and sometimes repetitive patterns of soil in the landscape. Class identification, on the other hand, is aimed at matching classes that are conceptually described (usually in a classification system) with reality – represented by data obtained by soil morphological description and measurements.

A soil map unit is a collection of areas that are relatively homogeneous in soil constitution or miscellaneous areas that constitute different soil components or both. In a given survey region, a map unit differs in some respect from all others and can be uniquely identified on a soil map. A soil map delineation, on the other hand, is an individual area on the map, bounded by other delineations or the map boundary. Usually delineation contains the dominant components in the map unit name, but it may not always contain a representative of each kind of inclusion. Soil boundaries can seldom be shown with complete accuracy on soil maps; hence parts and pieces of adjacent polypedons are inadvertently included or excluded from delineations. This problem has been the stimulus for paradigm shifts in soil classification in the latter part of the 1990s; hence the applications of numerical hard and fuzzy classification systems in soil science.

Fuzzy classes and the breakdown of the classification paradigm Traditionally most classification systems are composed of mutually exclusive classes in order to conform to the discontinuous soil variation embedded in the traditional concept of soil map units. But soil variation is more continuous than discrete. The pioneer work in pedometrics, involving computer-based numerical classification, was designed to address this limitation among others. While the applications of numerical soil classification to soil studies are, to some extent, based on continuous representation of soil in space, their results are still interpreted in terms of 'hard' classes. There is also lack of any spatial coherence for the classes to be mapped. Recent advances are based on fuzzy sets for optimized prediction quality of the resulting classification, and which take cognizance of the continuous nature of soil variation and nonlinearity in the interattribute relationships.

The first application of fuzzy set theory in soil survey was principally for classification. Two different but complementary approaches to grouping individuals into fuzzy classes in soil science are: (1) fuzzy *c*-means (FCM, also known as 'fuzzy *k*-means'), and semantic import model (SI).

The FCM algorithm for improved predictive classification by providing for membership to an extragrade class is based on the objective function, defining the within-class sum-of-square errors, J_E, expressed as:

$$J_E(M,c) = \alpha \sum_{i=1}^{n} \sum_{j=1}^{c} m_{ij}^{\phi} d_{ij}^{2} + (1-\alpha) \sum_{j=1}^{c} m_{i*} \sum_{j=1}^{c} d_{ij}^{-2} \quad [1]$$

where c is the number of classes, n is the number of individuals or pedons; m_{ij} is the membership of an individual i in class j; φ is the fuzziness exponent $(1 < \varphi < \infty)$; d_{ij} is the character space between the feature value of an individual, i, and the feature centroidal value for class j; α is the parameter that determines the mean value of m_{i*}, which is the membership value of an individual, i, in the extragrade class.

The SI model was developed primarily for land evaluation. The need for using fuzzy set theory in land evaluation, such as defined in the Food and Agriculture Organization of the UN (FAO) framework, arises because basic soil information used for land evaluation is mostly described by seemingly vague terms such as 'poorly drained,' 'slightly susceptible to soil erosion,' and 'moderate nutrient availability.' Not even when these terms are defined precisely is the qualitative ambiguity removed. Usually, the land evaluator's aim is to produce a set of clearly defined classes of land qualities based on specified land use requirements. These subsequently provide the means of transferring information about the soil and its use. As land qualities are complex attributes that are derived from land characteristics such as topography, soil, water, or biological and human activity, subsequent Boolean logical operations in the process of land evaluation tend to throw away much useful information.

Spatial Prediction of Soil Classes

The corpt approach As previously stated, the 'corpt' methods are based on the empirical-deterministic models that originated from Han Jenny's *Factors of Soil Formation*. Jenny's state-factor equation, in its extended form, is expressed as:

$$S = f(s, c, o, r, p, a, n) \quad [2]$$

where S is some soil properties or soil type as a function of state factors: s as some other soil property at a point, c as climate, o as organisms, r as relief, p as parent material, a as time or age, and n as space or spatial position. Soil spatial variability is therefore considered as being causative realizations of the complex combinations of soil-forming processes as influenced by the soil-forming factors. The scorpan function (Eqn [2]) – in its original form, the corpt – earlier in the nineteenth century stimulated numerous studies, mainly quantitative prediction of soil attributes, which we shall treat in a later section. Until recently, because data on soil classes were nonquantitative, the scorpan methods were only restricted to predicting soil attributes measured on a continuous scale.

Models such as expressed in Eqn [2] can be derived for predicting soil classes or soil types using two different techniques: logistic regression and classification trees. Logistic regression is designed specifically for situations in which we have a dichotomous (nominal or ordinal) dependent variable (e.g., soil classes), in comparison to classical linear regression typically used for continuous dependent variables (e.g., soil pH value). Several problems arise when a dependent variable is binary: the error terms are nonnormal and their variance is nonconstant. Although the errors are not normal, this method still provides unbiased regression estimators that are approximately normal if the sample size is large. Classification tree algorithms, on the other hand, search for combinations of values of independent variables (e.g., Jenny's state factors or attributes derived from them) that best predict the value of the dependent variable (e.g., unordered factors such as soil classes). Prediction quality is based on criteria such as the within-group noises, variance, or statistical significance of the conditional frequency distribution of values for the dependent variable, conditioned on the answers

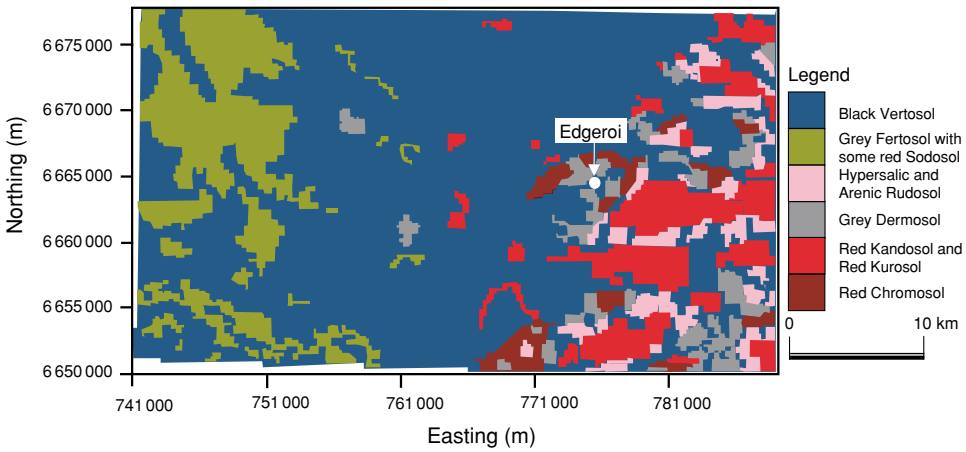


Figure 5 Predicted soil classes for the Edgeroi area in accordance with Australian Soil Classification. (Reproduced with permission from McBratney AB, Odeh IOA, Bishop TF, Dunbar MS, and Shatar TM (2000) An overview of pedometric techniques for use in soil survey. *Geoderma* 97: 293–327.)

to questions asked. A digital map of soil suborder classes, as produced by classification tree model, is shown in Figure 5.

Spatial Prediction of Soil Attributes

The scorpan approach The 'scorpan' approach (Eqn [2]) earlier in the nineteenth century stimulated numerous studies in quantitatively predicting soil attributes measured on a continuous scale. Much of the earlier studies, and indeed some recent examples, was based on general and bivariate-simple linear regression, although multiple polynomial regression models were occasionally applied. But due to nonlinearity in the correlations among many soil variables, and indeed of many of the soil variables with ancillary variables, robust methods such as generalized linear models (GLMs), generalized additive models (GAMs), and regression trees (RTs) have been developed and applied. Another development is the artificial neural network models, which are nonparametric modeling techniques that mimic the neural networks of the brain. The networks are composed of processing units, or neurons, which are organized into layers, i.e., input, hidden, and output layers.

But the problem is, while the classic models or the more robust methods may take care of the deterministic relations, they do not account for spatial autocorrelations of the soil properties, especially at the local level. To solve this problem, the pioneer pedometricians initiated the application of geostatistics (which was primarily developed for the mining industry).

Geostatistics Matheronian geostatistics is based on the theory of regionalized variables, which allows us to consider spatial variability of a soil property or even soil types, if quantified, as a realization of a random function represented by a stochastic model. The generic geostatistical method of spatial interpolation is termed 'kriging' in its various forms: simple, ordinary, lognormal, and disjunctive kriging.

In geostatistics, the semivariogram is a primary requirement for spatial prediction or kriging of a target geographical feature. The semivariogram can be obtained by processing a data set to produce empirical semivariances through the computation of the spatial correlation or covariance between pairs of samples at certain distances apart. The plot of the semivariance and the corresponding distance or lag at which the pairs are separated produce the semivariogram. The latter describes the magnitude, spatial scale, and the general form of variation of a given variable. An example of semivariogram is shown in Figure 6, with several spherical models based on different fitting methods.

The first major applications of kriging, in its ordinary and univariate form, for soil studies were in the early 1980s. Since then ordinary kriging has been widely used in various subfields of soil science, including soil reclamation, in soil classification, and soil pollution. Major limitations of the univariate geostatistics technique of kriging are due to the assumptions of stationarity, which are not often met by the field-sampled data sets and, of course, the often-cited requirement of large amounts of data to define the spatial autocorrelation. However, with increasing availability of ancillary information, the lack of adequate samples has been partially solved. The univariate usage of kriging is also limiting in situations of complex terrain where the soil-forming processes are themselves complex. In such situations there is the need to model both

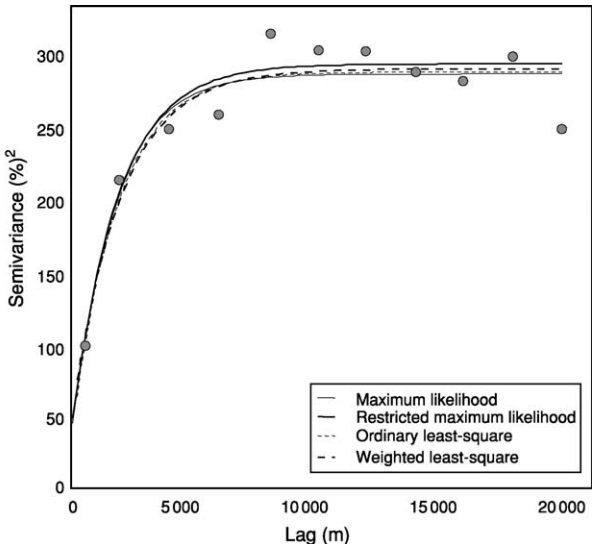


Figure 6 An experimental semivariogram fitted with spherical models based on different fitting techniques.

the structured and the spatially dependent components of the soil variable. Also there are economic and logistic reasons for including the ancillary variables influencing soil variability, especially if the latter are more readily and cheaply available. As both the soil and the exogenous factors are multivariate, the most obvious choices are appropriate combinations of multivariate or univariate analysis using the scorpan factors and the geostatistical methods. These combinations constitute the hybrid techniques.

The hybrid of scorpan and geostatistics The hybrid techniques for soil survey and mapping are based on various combinations of the geostatistical and multivariate or univariate scorpan methods. Let us suppose that a data vector describing a soil property is a random variable Z, determined at locations in a region, $X = x_1 \ldots, x_N$, and consisting of three components as:

$$Z = m + Z_1 + \epsilon \qquad [3]$$

where m is the local mean for the region, Z_1 is the spatially dependent component, and ϵ the residual error term, spatially independent. Now there may be situations where m is varying and dependent on some exogenous factors such as the scorpan factors. In other words it is deterministically related to some causative factors (in geostatistics parlance, the variable is said to exhibit a trend). Wherever trend exists, the ordinary univariate kriging is inappropriate. Several methods have been designed to accommodate the trend.

Universal kriging has been the commonly used method to accommodate the trend or 'changing drift,' as it is sometimes known, in a soil variable. Universal kriging is a combination of the standard model of multiple-linear regression and the geostatistical method of ordinary kriging, which is also analogous to combining scorpan methods with the univariate kriging, but only using the geographic coordinates as determining the drift (the n factor of the scorpan model). More recently, a more advanced approach, the intrinsic random function of order k (IRF-k), has been used to accommodate the varying nature of the trend in a regionalized soil variable. The term 'k' represents the order of polynomial trends: $k = 0$ means constant drift, and the IRF-k is equivalent to ordinary kriging system of equations. If $k = 1$, we have linear drift; $k = 2$ yields quadratic drift; but where there is no trend but deterministic relationships are with some known or readily available and inexpensive covariates (scorpan factors) or other easy-to-measure soil variables, co-kriging has played a major role in efficiently predicting the target soil variable. Universal co-kriging is also possible when considering the trend and covariation with one or more secondary variable.

Co-kriging is the multivariate extension of kriging that allows the inclusion of more readily available and inexpensive attributes in the prediction process. There are many instances in soil survey where the scorpan factors such as topography, time, and variable parent material, are easily discernible or are either readily available and/or are cheap to obtain. The most efficient way to predict the expensive-to-measure target soil variable, the variation of which is affected by the scorpan factors, is to use the factors in co-kriging the target soil variable, sampled at fewer locations, into dense grid nodes.

Regression-kriging (RK) is another hybrid method that combines either a simple or multiple-linear regression model (or a variant of GLM, GAM, and RTs) with ordinary, or simple, kriging of the regression residuals. The assumption here is that the deterministic component (m in Eqn [3]) of the target (soil) variable is accounted for by the regression model, while the model residuals represent the spatially varying but dependent component Z_1 in Eqn [3]). If the exogenous variables used in the regression equation are available at more dense locations than the target variable, the equation can then be used to predict the m on to those locations. The Z_1 can also be predicted to the same locations by simple kriging system of equations, and then added to the m to obtain Z^*. A variant of RK is kriging with uncertainty, which introduces regression residuals (as representing model uncertainty) into the kriging system used to predict the target soil variable. This reduces the extrema of the target soil variable and therefore produces a

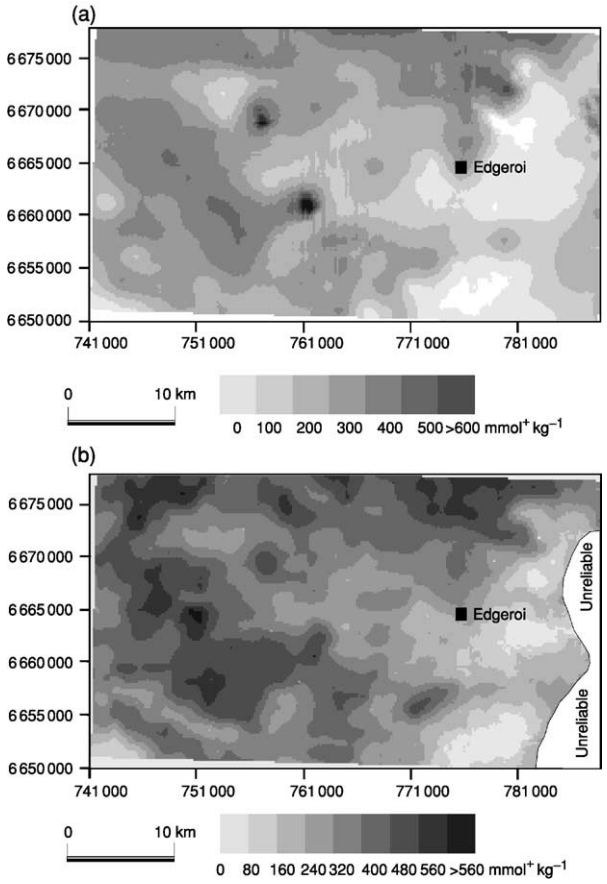


Figure 7 Predicted cation exchange capacity (CEC; millimoles per kilogram) for the Edgeroi area using (a) kriging with elevation as external drift and (b) regression-kriging. (Reproduced with permission from McBratney AB, Odeh IOA, Bishop TF, Dunbar MS, and Shatar TM (2000) An overview of pedometric techniques for use in soil survey. *Geoderma* 97: 293–327.)

smoother function of the predicted values. Another variant of RK, kriging with external drift (KED), is a hybrid technique that integrates the universality conditions into the kriging system using one or more of the ancillary drift variables. It is similar to universal kriging, but using an ancillary variable to represent the trend. As shown in Figure 7a, a regional digital map of cation exchange capacity (CEC) has been produced using KED with elevation as the external drift. Comparing this map with one produced using RK (Figure 7b) indicates a slightly more smoothed map produced by KED than the RK one.

Soil Utility and Quality

Land Evaluation, Soil Quality, and Soil Utility

Pedometrics has played a major role in advancing the role of land evaluation and, more recently, the quantification of soil quality for land management and sustainable use of land resources. Although 'soil quality assessment' is often used as a misnomer for 'land evaluation,' both could be regarded as the interpretative phase of soil survey. While land evaluation is concerned with the assessment of land performance when used for specified purposes, soil quality is defined as 'the capacity of a specific kind of soil to function, within natural or managed ecosystem boundaries, to sustain plant and animal productivity, maintain or enhance water and air quality, and support human health and habitation.' In considering productivity, environmental quality, and health as major functions of soil, this definition requires that values be placed on specific soil functions as they relate to the overall sustainability of alternate land-use decisions. Pedometric techniques (based on 'hard' and/or fuzzy techniques) are being used to quantify soil quality and associated uncertainty. For example soil quality can be assessed in terms of requirement for different specific uses such as shown in Figure 8. Many of the soil quality indicators are required at various scales and spatial extents in which pedometrics is playing a major role in spatially modeling them for incorporation into regional, catchment, and field-scale modeling. Another approach to soil quality analysis may be based on the concept of pedodiversity.

Recent Advances: Quantitative Pedogenesis

As stated above, soil variability is a function of factors of soil formation. Since the publication of Jenny's state factors, pedologists have attempted to develop pedogenetic (mathematical) functions that could explain soil variability and, in many instances, predict or even simulate the soil. A few examples are given here.

Process Models

The soil system is dynamic and regenerative under the influence of soil-forming processes. Since the soil is important for the survival of human civilization, it is imperative that researchers' interest in soil processes is ever-increasing. This has resulted in models of pedogenesis based on chemical and physical processes. The formulation of the process models and their applications are dependent on the scales in both the temporal and spatial dimensions. This modeling space–time continuum is determined by the complexity of the modeling process, which can be characterized in several ways. The first is based on computational complexity of the model, which ranges

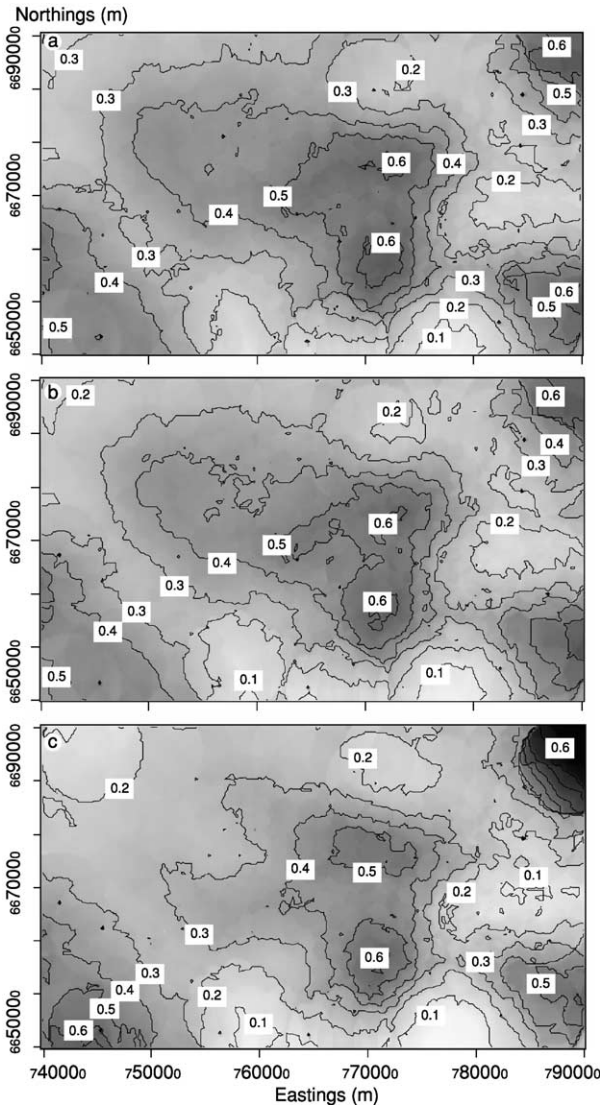


Figure 8 Interpolated suitability of (a) wheat, (b) sorghum and oats, and (c) dryland cotton. (Reproduced from Triantafilis J, Ward WT, and McBratney AB (2001) Land suitability assessment in the lower Namoi Valley of Australia using a continuous model. *Australian Journal of Soil Research* 39: 273–290.)

from purely qualitative (mental models) to highly quantitative, with the latter involving complex computer coding; the second is based on the complexity of the model structure, which distinguishes between mechanistic (highly complex) and empirical (simplified functional) models; and the last, but not least, is based on organizational hierarchy, which determines at which level a model is used to simulate the soil system. The hierarchical levels range from $i-4$ (molecular-level processes) to $i+6$ (global level). Each level is a subsystem of the level above it, therefore allowing room for upscaling.

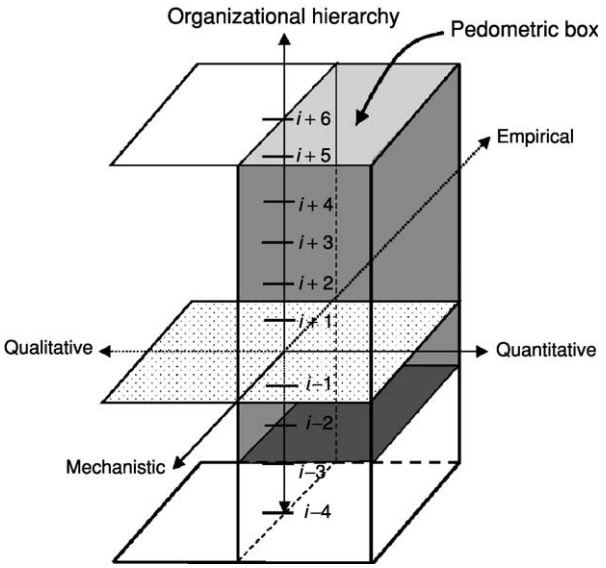


Figure 9 Pedometric box in the organizational hierarchy of pedogenetic modeling approach. (Adapted from Hoosbeek MR and Bryant RB (1992) Towards quantitative modeling of pedogenesis – a review. *Geoderma* 55: 183–210, with permission.)

The three characterizations above are well illustrated in Figure 9. It should be noted that pedometrics has played a dominant role in the quantitative half of the organizational hierarchy, particularly from and above the $i-3$ level, which is governed by material fluxes in the pores between the primary particles.

Recent developments in process modeling focus on mechanistic stochastic simulations, particularly at levels $i-3$ to i (Figure 9) and on how the latter can be upscaled to any of the higher hierarchies. For example: a mechanistic pedogenetic model, based on a digital elevation model (DEM), was used to simulate pedogenesis by a combination of several submodels: (1) physical weathering rate as exponential decline of soil thickness; (2) chemical weathering rate represented as a negative exponential function of both soil thickness and time; and (3) the movement of material as characterized by diffusion transport model. The upscaling result of such an analysis is illustrated in Figure 10, whereby, after 10 000 years, soil accumulation is predominant in the gullies compared with the ridges, where soil erodes.

Other emerging methods are geostatistical simulation models, e.g., indicator simulation, Gaussian simulation, etc. However, with increasing environmental concern, the development and applications of landscape dynamic models under different land-use scenarios may well dominate pedometric research in the first few decades of the twenty-first century.

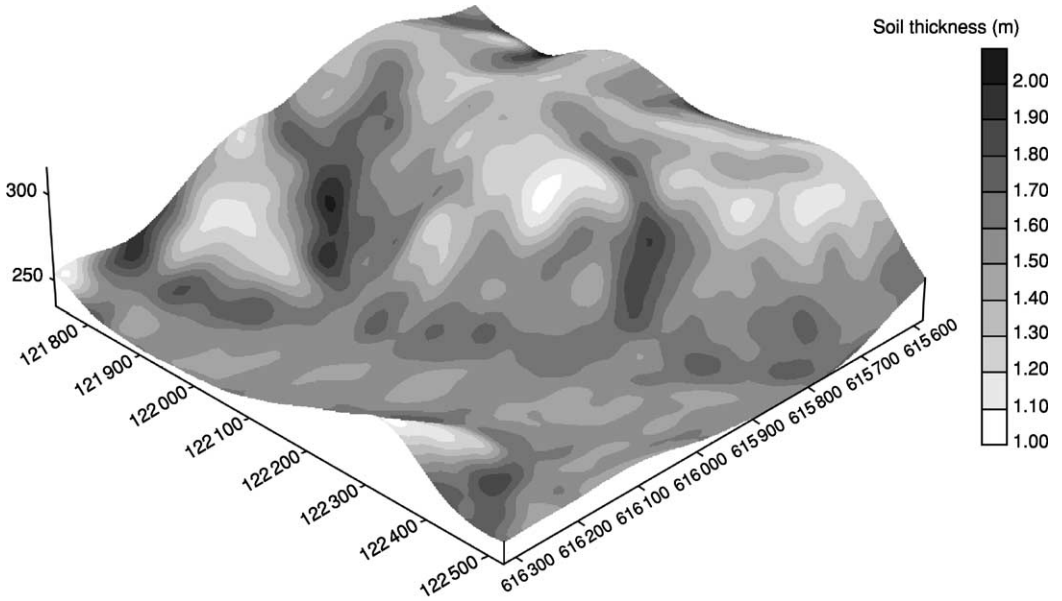


Figure 10 Soil formation in a landscape after 10 000 years. (Adapted from Minasny B and McBratney AB (2001) A rudimentary mechanistic model for soil formation and landscape development: II. A two-dimensional model incorporating chemical weathering. *Geoderma* 103: 161–179, with permission.)

See also: **Classification of Soils**; **Pedology:** Basic Principles

Further Reading

Addiscott TM (1998) Modeling concepts and their relation to the scale of the problem. In: Finke AF, Bouma J, and Hoosbeek RM (eds) *Soil and Water Quality at Different Scales*, pp. 239–246. Proceedings of the workshop *Soil and Water Quality at Different Scales*, 7–9 August 1996, Wageningen, The Netherlands. Dordrecht, The Netherlands: Kluwer Academic Publishers.

Arnold RW and Wilding LP (1991) The need to quantify spatial variability. In: Mausbach MJ and Wilding LP (eds) *Spatial Variability of Soils and Landform*, pp. 1–8. SSSA Special Publication No. 28, Madison, WI: Soil Science Society of America.

Burrough AB and McDonnell RA (2000) *Principles of Geographical Information Systems*. New York: Oxford University Press.

Collin M, McBratney A, Voltz M, and Walter C (eds) (2000) Developments in quantitative soil resource assessment (Pedometrics 1998). *Geoderma* 97(3–4): 147–424.

de Gruijter JJ (ed.) (1999) Special issue. (Pedometrics 1997). *Geoderma* 89(1–2): 1–175.

de Gruijter JJ, McBratney AB, and McSweeney K (eds) (1997) Fuzzy sets in soil science (Pedometrics 1994). *Geoderma* 77(2–4): 1–82.

de Gruijter JJ, Webster R, and Myers DE (eds) (1994) Development in spatial statistics for soil science (Pedometrics 1992). *Geoderma* 62(1–3): 1–326.

McBratney AB, Odeh IOA, Bishop TFA, Dunbar MS, and Shatar TM (2000) An overview of pedometric techniques for use in soil survey. *Geoderma* 97: 293–327.

McBratney AB, Mendonça Santos ML, and Minasny B (2003) On digital soil mapping. *Geoderma* 117: 3–52.

Odeh IOA and McBratney AB (eds) (2001) Estimating uncertainty in soil models (Pedometrics 1999). *Geoderma* 103(1–2): 1–229.

Wagenet RJ, Hutson JL, and Bouma J (1994) Modeling water and chemical fluxes as driving forces of pedogenesis. In: Bryant RB and Arnold RW (eds) *Quantitative Modeling of Soil Forming Processes*, pp. 17–35. SSSA Special Publication No. 39. Madison, WI: Soil Science Society of America.

Webster R (1994) The developments of pedometrics. *Geoderma* 62: 1–15.

Webster R and Oliver AO (2001) *Geostatistics for Environmental Scientists*. Chichester, UK: John Wiley.

PENMAN, HOWARD LATIMER

J L Monteith, Edinburgh, UK

© 2005, Elsevier Ltd. All Rights Reserved.

The amount of water needed to ensure that a crop will provide the maximum yield possible in a given environment depends in a complex way on three main factors: prevailing weather, soil type, and the nature of physiological responses to the environment in general. Throughout the twentieth century, steady progress was made in measuring, quantifying, and interpreting these responses in terms of mechanisms. Towards the end of the century, developments in computing encouraged the building of 'models' that were intended to predict how the growth and yield of crops would respond to environmental factors in general and to weather in particular. Unfortunately, many of these models grew so rapidly that they shared the fate of the dinosaurs! More compact models have lasted longer because they follow classical lines, employing standard procedures based on minimal assumptions. A notable example in agricultural science is the equation for estimating the water use of arable crops, published by H.L. Penman of Rothamsted Experimental Station in 1948.

Howard Latimer Penman was born in 1909 in Dunston-on-Tyne, a small town in the north-east corner of England where his father worked as a master carpenter in a shipyard. At Blaydon County Secondary School, Howard passed the Oxford Higher School Certificate in 1927 and was accepted by the University of Durham for a Joint Honours course in mathematics and physics. Years later, a fellow-student wrote: "He was very much a northerner – a Tynesider – and on a suitable occasion could amuse by lapsing into the vernacular or singing a Tyneside song. His general approach to physics was stimulating; he accepted little on authority and thought things out for himself."

In 1930, Penman was awarded a First Class Honours degree in physics and for a short time taught at a boys' school until he decided to make research his career. At this stage, he was fortunate to encounter E.G. Richardson who offered him a studentship to work on the absorption of sound in porous materials. The work was successful and original enough to be published in 1933 with Richardson as a co-author; it was the first of Penman's 104 scientific papers written over the following 43 years. A Senior Exhibition offered by the County Education Committee then enabled him to embark on a research project managed by Richardson to investigate the dependence on temperature of the dispersion of sound in a gas. When completed, this study gained him a PhD from Durham University. Outside the laboratory, Penman was an active member of the Student Representative Council, played soccer and cricket, and sang in two choirs, an activity that he enjoyed for most of his life.

In 1934, Penman successfully applied for a post in the photochemistry section of the Rayon Department at the Shirley Institute in Manchester, administered by the British Cotton Research Association. The reports he presumably wrote on this work did not provide material for general publication. In 1935, however, a short paper on 'The effect of temperature on supersonic dispersion in gases,' based on his PhD, was published in the *Proceedings of the Physical Society.*

In 1937, Penman took a crucial step in his career when he successfully applied for a post in the Soil Physics Department at Rothamsted Experimental Station in Harpenden, a small town about 30 miles (50 km) north of London. Established by Sir John Lawes in 1843, Rothamsted is one of the world's oldest and most distinguished centers of research on agricultural crops.

Three years later, Penman published two papers on 'Gas and vapour movement in the soil' in the *Journal of Agricultural Science*, followed in the same year and journal by an analysis of 'Meteorological and soil factors affecting evaporation from fallow soil.' The latter paper was Penman's first excursion into what he often referred to as 'muddy-boot physics.' In 1941, he collaborated with his colleague R.K. Schofield to produce a major analysis of many measurements of evaporation from fallow soil that predecessors had installed in tanks surrounding Rothamsted's 'Bear Pit.' This was a deep, circular brick-lined enclosure built many years previously by Rothamsted's founders, Sir John Bennet Lawes and Henry Gilbert. (It survived well into the twentieth century but was eventually filled in to satisfy safety regulations imposed by vandals ignorant of the historical importance of the site!) A formidable archive of rainfall and drainage records existed but no previous analysis had been successful because the physics of natural evaporation was not properly understood. To get a better feel for the scale and interaction of the processes involved, Penman conducted a set of laboratory experiments that he described in the *Journal of Agricultural Science*, also in 1941.

In 1942, Penman produced a paper that marked an important extension of his research horizon from soil to the complex ways in which crops respond to climate as well as to soil conditions. However, it was not until 1948 that he published, in the *Proceedings of the*

Royal Society, the classic paper that brought him world renown – 'Natural evaporation from open water, bare soil, and grass.' This was based on a highly original combination of two fundamental equations, one describing the heat balance of a crop stand in terms of micrometeorology and the other describing its water balance in terms of hydrology. The variables involved (as daily averages) were:

1. the net amount of heat available from radiation, both solar and terrestrial
2. air temperature at screen height
3. vapor pressure at screen height
4. windspeed.

One variable was common to both equations – the effective surface temperature of foliage. The core of Penman's contribution to agricultural science was his realization that this unmeasured and, at the time, virtually unmeasurable quantity, could be eliminated by solving the two equations simultaneously to derive a new third equation that achieved world renown as the Penman formula.

To subject his equation to extensive tests in the field, Penman set up parallel irrigation experiments at Rothamsted, which has a clay soil, and at its Woburn outstation, where the soil is sandy. Each spring, he mustered a working party from his staff to install irrigation pipes – hard manual labor but a welcome escape from the laboratory provided the weather was fine. In due course, two major series of papers describing the analysis of Woburn measurements were published in the *Journal of Agricultural Science*. When a working group of the Agricultural Research Council described such a study as 'speculative,' Penman responded in characteristically trenchant language: "The term is unfortunate in its association with gambles and guesses .... Why the Research Council for the country's chief industry should seem to need special pleading for this kind of work is puzzling."

Following successful local tests of his equation, Penman applied it first to the whole of the UK and eventually, when checks against hydrological measurements appeared successful, to the mainland of Europe. His work on crop evaporation is an outstanding example of how a relatively simple but rigorous general model can be applied to many types of mixed vegetation growing in a wide range of climates.

(When Penman's seminal paper was published, he had worked at Rothamsted for 11 years during which he published 14 papers – a fairly leisurely rate by the norms of today. By 1974, however, when he retired, he had published over 100 papers and articles in a variety of journals ranging from the *Proceedings of the Royal Society of London* to the *Transactions of the Hertfordshire Natural History Society Field Club*.)

In due course, the Penman equation was adopted by several government agencies and used both on relatively small scales to estimate the irrigation need of agricultural and horticultural crops and on much larger scales to assess the hydrological efficiency of catchments. As the scientific importance and practical value of Penman's work became recognized internationally, he was increasingly invited to attend conferences abroad and to serve on committees such as the Coordinating Council for the International Hydrological Decade.

In 1954, Penman succeeded R.K. Schofield as Head of the Physics Department at Rothamsted and continued the tradition of allowing his staff the freedom to identify and explore lines of work that challenged them. This treatment followed his own experience when, during his first interview with his head of department he was told to "find something interesting to do and get on with it!" Most, but not all of his staff, made good use of this invitation to tackle challenging problems of their own choice. The benefits were personal while the work was in hand and universal once it was published.

Penman did not live to know that, in 1990, an extended form of his equation was adopted by the Food and Agriculture Organization of the United Nations (FAO) as the international standard for estimating the water need of crops. This would certainly have given him great satisfaction, but he was a modest man and might well have concealed this!

Although the Bear Pit site was not ideal, Penman was confident that records, carefully preserved for many years, could be used to test formulae relating to evaporation to weather.

There are at least two ways in which such testing can be done: by empiricism based on the correlation of estimates with rates of evaporation measured at one or more sites over long periods; and by deriving appropriate formulae to describe the mechanisms involved. The first route was taken by an American, Warren Thornthwaite, who explored the correlation of evaporation rates with temperature as reported from standard weather stations. This statistical approach to the problem did not appeal to Penman, nor did Thornthwaite's suggestion that the total amount of water lost by evaporation from plants and directly from the soil on which they grew, should be referred to as 'evapotranspiration' – a clumsy hybrid banned from Rothamsted publications! Penman coined the term 'potential transpiration' to describe the rate at which water was lost from a stand of vegetation with an abundant water supply and he provided an unambiguous definition: "The loss of water from short, green vegetation, completely covering the ground and never short of water."

Penman's mechanistic formula for evaporation was founded on two major physical principles: the conservation of mass and the conservation of energy. Conservation of mass implied that the rate of evaporation from a uniform area of vegetation should be proportional to the change of vapor pressure with height and to a function of windspeed. Conservation of energy implied that the net input of heat from radiation must be balanced by losses of energy in the form of convection and evaporation.

The two equations describing these processes contain one common term: the surface temperature of the source for evaporation, but this could not be accurately measured in Penman's time, let alone recorded over long periods. Penman overcame this obstacle in an ingenious but essentially simple way: he solved the mass and energy balance equations simultaneously by eliminating surface temperature, the term common to both.

The success of this exercise led to the publication in the *Proceedings of the Royal Society of London* of 'Natural evaporation from open water, bare soil, and grass.' This paper rapidly achieved recognition as a major step advance with great potential benefits for both agronomists and for hydrologists. By a curious coincidence, the American geographer C.W. Thornthwaite, also in 1948, published, in the *Geographical Review*, a paper titled 'A rational approach to the classification of climate.' This included the proposal that rates of evaporation from both vegetation and water could be assessed from empirical correlations with air temperature. Once, at a meeting of meteorologists in the USA, Thornthwaite was asked to say what he thought of the Penman formula. He thought for a moment before replying, with the faint vestige of a smile: "What I really dislike about Penman's formula is that it's so empirical!" The non-American members of the audience were convulsed in laughter, leaving their American friends somewhat baffled!

The soil at Rothamsted is classified as 'clay with flints' in marked contrast to the light sandy soil at an outstation on sandy soil at Woburn. Penman was anxious to establish the extent to which crop growth and yield could be increased by irrigation but recognized that soil type would determine optimal schedules. Every spring, the younger members of his department were deployed to install irrigation spray-lines at both sites and every autumn they were housed for the winter. Practical conclusions from this work were brought together in a UK Ministry of Agriculture, Fisheries and Food Technical Bulletin (No. 4) which had several external co-authors, including L.P. Smith from the Meteorological Office, with whom Penman enjoyed a very fruitful working relationship. Details of measurements and their scientific implications were discussed in a major series of papers, published in the *Journal of Agricultural Science* in 1962.

Building on the success of local trials, Penman began to extend his expertise to the hydrology of catchments, particularly those that were forested, both in the UK and elsewhere in Europe. At the outset, he was convinced that because evaporation from vegetation was mainly limited by the supply of radiant energy, there should be little difference between annual totals for farmland and for forests. However, some of the early field work by staff at the Institute of Hydrology at Wallingford, Berkshire (whose Director, James McCulloch, had worked as a postgraduate with Penman) demonstrated unequivocally that forests lose significantly more water than crops exposed to the same climate, partly because they intercept more water and partly because they are aerodynamically much rougher. It was therefore inappropriate to use the same wind function for both types of vegetation, as Penman initially assumed.

Penman was very fortunate to have Ian Long among his staff, an experimental officer with unique skills in designing, building, and maintaining equipment. One of his most impressive achievements was the construction of a set of neutron probes for measuring the water content of soil as a function of depth; shortly after, this device became available commercially. In 1973, a report on extensive tests of this device occupied about 80 pages of the Rothamsted Annual Report for 1972. Under Penman's direction, Long acquired thermistors soon after they came on the market and used them for many years to provide long-term records of dry- and wet-bulb temperature.

Penman's success in estimating evaporation on a field scale encouraged him to move up to the catchment scale and eventually to the national scale by integrating catchment estimates. In 1950, a major paper on 'Evaporation over the British Isles' was published in the *Quarterly Journal of the Royal Meteorological Society*. This was followed in 1954 by 'Evaporation over parts of Europe,' presented to the General Assembly of the International Association for Scientific Hydrology.

Also in 1954, Penman collaborated with staff at the Ministry of Agriculture, Fisheries and Food and at the Meteorological Office to produce a Technical Bulletin on 'The calculation of irrigation need,' applicable to both agriculture and horticulture. Further afield, he contributed to a government publication on evaporation from Lake Volta.

In 1963, Penman initiated irrigation trials on the Rothamsted farm after getting permission for the drilling of a borehole. Using a neutron probe, changes of water content were monitored at regular intervals as

a function of depth and records for 13 seasons were eventually analysed. One major conclusion was that the soil water deficit at which growth became inhibited by drought was about 2.5 times larger than for the same crops grown on sandy soil at the Woburn farm.

The field study at Woburn, referred to above, ran for many years and major conclusions from this work were eventually published in a series of papers. The first three were published in the *Journal of Agricultural Science* in 1962, followed by a further group in 1970. Major sets of measurements made with the neutron probe were presented in Annual Reports of Rothamsted Experimental Station between 1969 and 1972 and a set of three major papers describing this work was published in the 1973 Report. Unfortunately, this valuable material did not find its way into the open literature.

Also in the 1960s, Penman began to use his formula to explore the hydrology of forested catchments in the UK and attended an International Symposium on Forest Hydrology at Pennsylvania State University to read a paper on the subject. This was followed by a review of the literature on evaporation from forests comparing field measurements with predictions from theory. On a much smaller scale, he became interested in the role that the stomata play in controlling the loss of water from leaves and collaborated with F.L. Milthorpe at the University of Nottingham to publish a paper on the diffusive conductivity of stomata on wheat leaves.

By this stage in his career Penman was much in demand as a speaker, partly because his work was of widespread interest to agronomists, ecologists, and meteorologists, but also because he lectured with admirable clarity and with occasional flashes of north-country humour. He also became involved in the management of scientific societies as a Councillor and Editorial Board member of the British Society of Soil Science, as Secretary of the London Branch of the Institute of Physics, as Secretary and eventually as President of the Royal Meteorological Society which elected him to Honorary Fellowship.

His help was sought by many national bodies: the Ministry of Agriculture for whom he helped advise the group that prepared a major report on 'The calculation of irrigation need'; the Institute of Water Engineers, the Meteorological Office, the Institute of Hydrology as it grew from a small unit to a major international centre. Further afield, he led the British delegation to meetings of the International Hydrological Decade and, in 1972, was host at the World Water Balance Symposium in Reading.

Away from his office, Penman's main interest was in music. During most of his time at Rothamsted, he was a member of the BBC Choral Society; but because this involved a fair amount of travelling and occasional nights away from home, he eventually joined the Bach Choir in nearby St Albans. However, he finally withdrew from this too because, as he explained "I almost sang a wrong note one day!" His repertoire included popular music, particularly folksongs in the strong Geordie accent that he enjoyed reverting to long after leaving Newcastle. Over the years, he accumulated an impressive collection of classical records, played on an early hi-fi system. This was a rare extravagance, for the habit of thrift acquired early in life remained with him even after he became a relatively prosperous Head of Department. He never owned a car, for example, and, twice a day, walked back and forth across Harpenden Common on the route between his very modest home in Rowan Way to his office at Rothamsted, which looked eastward over Harpenden Common. The fine view was sometimes partly obscured by smoke from Penman's pipe that became particularly dense when a new manuscript was on the stocks!

Penman's international reputation ensured that he had many opportunities to travel abroad, to lecture, and to discuss with colleagues a wide range of hydrological problems. These visits included Australia, Canada, Kenya, Lebanon, the Netherlands, the USA, and West Africa. These and many other countries have benefited immeasurably from his work. Throughout the world, water is an essential commodity but there are many regions where it is scarce so that prudent management and use are vital. Penman's formula made it possible to explore and provide answers to a wide range of major environmental problems such as the optimal application of water to crops in rain-scarce areas; the rate of loss of surface water from reservoirs; and, most recently, the hydrological implications of global warming. The economic value of the Penman formula for evaporation is therefore immeasurable.

See also: **Evaporation of Water from Bare Soil**; **Evapotranspiration**; **Penman–Monteith Equation**

Further Reading

Howard Latimer Penman (1909–1984) *Biographical Memoirs of Fellows of the Royal Society* 32: 379–404.

Penman HL (1940) Gas and vapour movement in the soil. *Journal of Agricultural Science*.

Penman HL (1940) Meteorological and soil factors affecting evaporation from fallow soil. *Journal of Agricultural Science*.

Penman HL (1948) Natural evaporation from open water, base soil and grass. *Proceedings of the Royal Society of London* 143: 120–145.

Penman HL (1950) Evaporation over the British Isles. *Quarterly Journal of the Royal Meteorological Society.*

PENMAN–MONTEITH EQUATION

R Allen, University of Idaho, Kimberly, ID, USA

© 2005, Elsevier Ltd. All Rights Reserved.

The Penman–Monteith (PM) evapotranspiration (ET) equation predicts the rate of total evaporation and transpiration from the earth's surface using commonly measured weather data (solar radiation, air temperature, vapor content, and wind speed). The PM equation follows a single-layer or 'big leaf' approach, where single-surface resistance and single aerodynamic resistance terms represent the transport properties of the cropped surface. The PM procedure provides scientists, engineers, and managers with the ability to evaluate effects of modifications to cultural practices and crop variety improvements on ET such as increased plant height, increased or reduced leaf size or density, reduced row spacing, and reduced stomatal conductance etc., by varying these parameters in the calculation algorithms. The PM method is useful in providing direction in determining changes in crop coefficients under new cropping conditions. Determination of crop coefficients is an expensive and time-consuming process, as it requires additional lysimeter or micrometeorological studies.

The PM equation has the form:

$$E \text{ or } \mathrm{ET} = \frac{\Delta(R_\mathrm{n} - G) + \rho_\mathrm{a} c_\mathrm{p}(e_\mathrm{s} - e_\mathrm{a})/r_\mathrm{a}}{\left(\Delta + \gamma\left(1 + \frac{r_\mathrm{s}}{r_\mathrm{a}}\right)\right)\lambda\rho_\mathrm{w}} \qquad [1]$$

where: Δ is the slope of the saturation vapor pressure vs. temperature curve; R_n is the net radiation flux at the surface; G is the sensible heat exchange from the surface to the soil (positive if the soil is warming); ρ_a is air density; c_p is specific heat of dry air; e_s is the saturation vapor pressure of the air at some height above the surface; e_a is the actual vapor pressure of the air; r_a is aerodynamic resistance to turbulent heat and/or vapor transfer from the surface to some height z above the surface; γ is the psychrometric constant (defined later); r_s is a bulk surface resistance that describes the resistance to flow of water vapor from inside the leaf, vegetation canopy, or soil to outside the surface; λ is the latent heat of vaporization, defined as the energy required to convert a mass of liquid water into vapor (having typical units of joules per kilogram); ρ_w is density of liquid water. All parameter units in eqn [1] must cancel so that the remaining units for evaporation (E) or ET (depending on application to evaporation from soil or ET from soil and vegetation) are presented as L T^{-1}, for example, in millimeters per hour or millimeters per day.

Surface resistance r_s may represent the resistance to vapor flow through plant leaf stomates and/or resistance to vapor flow within soil to the soil surface. For a wet (saturated) surface, r_s is by definition zero. When r_s is zero, eqn [1] reverts to the original Penman equation.

Derivation of the Penman–Monteith Equation

It is useful to review the derivation of the PM equation in order to understand how and why it functions. There are two fundamental approaches to predict E (or ET). These are the surface energy balance equation and the aerodynamic equation. Each equation uses a different approach and each has difficulties for solution when using typical weather data comprised of solar radiation, air temperature, vapor pressure, and wind speed. Penman combined these two approaches into the so-called combination or Penman equation that could be solved using standard weather data only. Monteith later improved on the Penman equation by including a surface resistance term and a more rigorous term for aerodynamic transfer. It is useful to review the combination of the two approaches.

The surface energy balance equation relates to the various ways in which net radiation energy from the sun and atmosphere is balanced at the surface by inputs or outputs of energy from nonradiative parameters. The vertical energy balance at the surface is the sum of sensible heat flux to or from the air and to or from the soil, along with latent heat flux, net radiation, and other miscellaneous fluxes. The major components of the vertical energy balance are expressed as:

$$R_\mathrm{n} = G + \lambda E + H \qquad [2]$$

where λE is the latent heat flux (positive during evaporation) and H is the sensible heat exchange from the surface to air (positive if the air is warming). In eqn [2], E (or ET) is expressed in terms of the energy required to convert the amount of liquid evaporated into vapor by multipling by λ. The miscellaneous terms involved in the surface energy balance, such as heat storage within foliage, photosynthesis, and respiration, are generally insignificant relative to magnitudes of R_n, λE, H, and G, and are commonly

neglected. Under some conditions, change in heat storage in a canopy may need attention, especially in forest canopies. If R_n and sensible heat flux densities H and G can be measured or estimated reliably, then the latent heat flux density, λE, can be computed from eqn [2].

$$\lambda E = R_n - G - H \quad [3]$$

The challenge to the solution of eqn [3] using standard weather data is in solving for H. The traditional aerodynamic expression H is:

$$H = \frac{\rho c_p (T_o - T_a)}{r_{ah}} \quad [4]$$

where T_o is surface temperature, T_a is air temperature, and r_{ah} is the aerodynamic resistance to heat transfer between some mean height within the surface to some height a few meters above the surface. Generally, r_{ah} is assumed equal to r_a of eqn [1].

Equation [4] is analogous to Ohm's law for electrical current flow, where T_o and T_a are equivalent to voltages or potentials, r_{ah} is equivalent to an electrical resistor, and H is analogous to current. Thus, eqn [4] behaves as a linear electrical circuit, so that if the difference between T_o and T_a is doubled, H will double. If resistance r_{ah} is halved (through a doubling of wind speed), then H will double.

The challenge of eqn [4] and thus eqn [3] is in the parameter T_o. T_o can be measured using infrared thermometers, if surface emissivity is known, or it can be measured using a fine thermocouple or other temperature sensor placed just beneath the surface skin. However, these measurements are not common due to the high operational and maintenance requirements of the instruments and difficulty in transfer of measured values to other locations and surface conditions, even those nearby.

The traditional equation for r_{ah} (and r_a) is:

$$r_{ah} = \frac{\left(\ln\left(\frac{z_u - d}{z_{om}}\right) - \psi_m\right)\left(\ln\left(\frac{z_T - d}{z_{oh}}\right) - \psi_h\right)}{k^2 u_z} \quad [5]$$

where: z_u is the height above the ground surface for the wind speed measurement, d is the zero plane displacement of the logarithmic wind profile, z_{om} is a roughness length governing the transfer of momentum from the surface, ψ_m is a correction factor for momentum transfer to account for buoyant instability or stability of the boundary layer, z_T is the height of the air temperature measurement above the ground surface, z_{oh} is an assumed roughness length governing the transfer of sensible heat from the surface, ψ_h is a correction factor for sensible heat transfer to account for buoyant instability or stability of the boundary layer, k is the von-Karman constant (0.41), and u_z is the wind speed measured at the z_u height. Equation [5] is sometimes criticized for its use of z_{oh}, commonly expressed as some constant fraction of z_{om}, whereas in reality the ratio of $z_{oh}:z_{om}$ can vary substantially, especially for dry, sparse vegetation. Other forms of eqn [5] can be written that rely on differences in T_a and wind speed at two different heights above the surface. These forms eliminate the need for z_{oh}; however, measurement of T_a and wind speed at two heights is not common and it is challenging to produce an unbiased pair of measurements. A variety of studies have investigated the effects of foliage density and geometry on apparent bulk roughness of vegetation.

The aerodynamic-based expression for λE is:

$$\lambda E = \frac{\rho c_p (e_o - e_a)}{\gamma (r_{av} + r_s)} \quad [6]$$

where e_o is vapor pressure 'inside' the surface, for example inside a leaf stomatal cavity or beneath the soil skin, e_a is vapor pressure of the air at some height z above the ground surface, γ is the pyschrometric constant, r_{av} is aerodynamic resistance to turbulent vapor transfer from the surface to some height z above the surface, and r_s represents any surface resistance to flow of vapor (defined after eqn [1]).

Aerodynamic resistance r_{av} is computed exactly as for r_{ah}, i.e., using eqn [5], except that z_T becomes z_v (height of vapor pressure measurement above the ground), z_{oh} becomes z_{ov}, which is an assumed roughness length governing the transfer of vapor from the surface, and ψ_h becomes ψ_v, a correction factor to account for the effect of instability or stability of the boundary layer on vapor flow. Generally, $z_T = z_v$ (air temperature and humidity sensors are usually collocated at the same height), and z_{ov} is assumed equal to z_{oh}, and ψ_v is assumed equal to ψ_h. Brutsaert and Allen have shown that the stability parameters ψ_v, ψ_h, and ψ_m can generally be set to zero when predicting ET from well-watered (moist) surfaces because of nearly neutral stability (buoyancy) for those conditions.

Again, the aerodynamic vapor transfer equation is analogous to Ohm's law, where e_o and e_a are equivalent to voltages or potentials, r_{av} and r_s are equivalent to electrical resistors in series, and λE (vapor flux) is analogous to electrical current.

Vapor pressure of air can be determined from paired measurements of relative humidity and air temperature, which are commonly available, or from measurement of dew-point temperature, also commonly available. Vapor pressure inside the soil

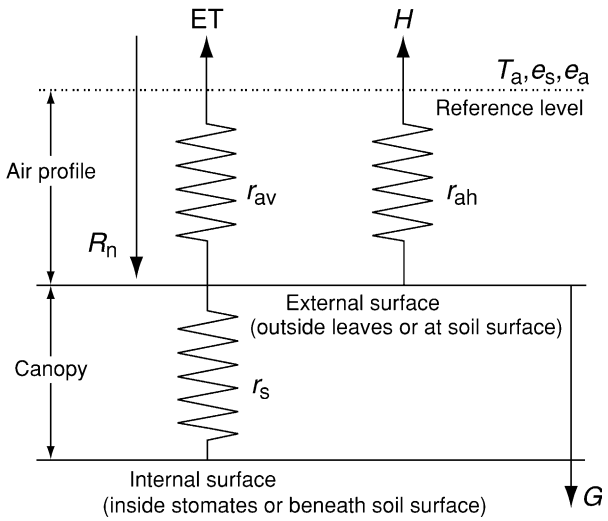


Figure 1 Schematic showing linkage between resistance terms in the Penman–Monteith equation relative to the surface and elevation of temperature and humidity measurements. ET, evapotranspiration; e_s, saturation vapor pressure at mean air temperature; e_a, actual vapor pressure of the air; G, heat exchange from surface to soil; H, heat exchange from surface to air; r_a, aerodynamic resistance; r_s, bulk surface resistance; R_n, net radiation flux at the surface; T_a, air temperature; r_{av}, aerodynamic resistance to vapor transfer; r_{ah}, aerodynamic resistance to heat transfer.

or plant surface is rarely measured, however, so that, as with eqn [3], eqn [6] is challenging to apply in routine practice. Hence the need for the Penman or PM combination equation.

Figure 1 is a schematic showing the relative locations of resistances r_a (or r_{ah} and r_{av}) and r_s, where r_s represents the internal resistance of the soil (if no vegetation) or of the soil, leaves, and canopy in parallel.

To create the PM equation, one needs to define what is known as the Bowen ratio, which is simply the ratio between H and λE. Expanded into the equations for H and λE and written for transfer from inside the surface to the air above:

$$\beta = \frac{H}{\lambda E} = \frac{\rho c_p (T_o - T_a)\gamma (r_{av} + r_s)}{\rho c_p (e_o - e_a)(r_{ah})} \quad [7]$$

Cancelling similar terms:

$$\beta = \frac{H}{\lambda E} = \gamma \frac{(T_o - T_a)(r_{av} + r_s)}{(e_o - e_a)(r_{ah})} \quad [8]$$

The energy balance equation can be rearranged and written as:

$$\lambda E = \frac{R_n - G}{1 + \beta} = \frac{R_n - G}{1 + \gamma \frac{(T_o - T_a)(r_{av} + r_s)}{(e_o - e_a)(r_{ah})}} \quad [9]$$

The next step in the derivation of the PM equation is to define the term Δ, which is defined as the slope of the saturation vapor pressure vs. temperature relation. Saturation vapor pressure $e^o(T)$ is a singular function of T and describes the ability of air to hold more vapor as temperature increases. The slope Δ of the $e^o(T)$ vs. T curve can be expressed as a linear approximation for a segment of the saturation curve between two temperatures T_o and T_a:

$$\Delta = \frac{e_o - e_s}{T_o - T_a} \quad [10]$$

where e_o, e_s, T_o and T_a were defined previously. Vapor pressure inside the surface, e_o, is assumed to be saturated and is therefore equivalent to e^o (T_o). e_s is the saturation vapor pressure associated with T_a, so that $e_s = e^o$ (T_a). In practice, Δ is often calculated as the slope of e^o (T) vs. T at T_a only, since T_o is usually not known.

Using the relationship from eqn [10] to define $(e_o - e_s)/\Delta = T_o - T_a$ and parsing $(e_o - e_s)$ into an equivalent expression of $(e_o - e_s) = (e_o - e_a) - (e_s - e_a)$ so that $T_o - T_a = [(e_o - e_a) - (e_s - e_a)]/\Delta$, this expression is substituted into eqn [9] so that:

$$\begin{aligned}\lambda E &= \frac{R_n - G}{1 + \beta} \\ &= \frac{R_n - G}{1 + \frac{\gamma}{\Delta}\frac{[(e_o - e_a) - (e_s - e_a)](r_{av} + r_s)}{(e_o - e_a)(r_{ah})}} \\ &= \frac{R_n - G}{1 + \frac{\gamma}{\Delta}\left[1 - \frac{(e_s - e_a)}{(e_o - e_a)}\right]\frac{(r_{av} + r_s)}{r_{ah}}}\end{aligned} \quad [11]$$

We define one more temporary term, λE_a, after Penman:

$$\lambda E_a = \frac{\rho c_p (e_s - e_a)}{\gamma r_{ah}} \quad [12]$$

where λE_a is similar to eqn [6], the aerodynamic expression for λE, except that r_s is omitted and $(e_o - e_a)$ is replaced with $(e_s - e_a)$. E_a is a fictitious term that allows the elimination of surface parameters T_o and e_s in the final PM equation. Term $(e_s - e_a)$ is known as the vapor pressure deficit (VPD) of the air, because it represents the difference between the saturation vapor pressure at air temperature $(e_s = e^o$ $(T_a))$ and actual vapor pressure of the air (e_a). Both e_s and e_a can be determined from commonly available humidity and air temperature data measured at some height above the surface, generally 1.5–2 m. Because:

$$\frac{e_s - e_a}{e_o - e_a} = \frac{\lambda E_a}{\lambda E}\frac{r_{ah}}{r_{av} + r_s} \quad [13]$$

Then eqn [11] becomes:

$$\lambda E = \frac{R_n - G}{1 + \frac{\gamma}{\Delta}\left[1 - \frac{\lambda E_a}{\lambda E}\frac{r_{ah}}{r_{av} + r_s}\right]\frac{r_{av} + r_s}{r_{ah}}} \quad [14]$$

Rearranging eqn [14] and setting $r_{av} = r_{ah}$ results in the PM equation:

$$\lambda E = \frac{\Delta(R_n - G) + \gamma\lambda E_a}{\Delta + \gamma\left(1 + \frac{r_s}{r_{ah}}\right)} = \frac{\Delta(R_n - G) + \rho c_p(e_s - e_a)/r_{ah}}{\Delta + \gamma\left(1 + \frac{r_s}{r_{ah}}\right)} \quad [15]$$

where, in eqn [15], E is expressed in terms of energy, rather than in terms of depth of liquid water as in eqn [1]. The PM equation has the extremely valuable advantage of requiring only commonly available weather data, namely solar radiation (used to calculate R_n), air temperature (used to calculate ρ, e_s and Δ), air humidity (used to calculate e_a and assist with R_n), and wind speed (for r_{ah}). Parameter G is generally estimated as a function of R_n or by direct measurement, and r_s is generally estimated as a function of amount of vegetation or amount of surface wetness if soil. The PM (and Penman equation where r_s is assumed zero and $1/r_{ah}$ is replaced by a more simple empirical 'wind function') are widely used around the world as a standard means to predict E and ET.

Other Parameters in the Penman–Monteith Equation

The psychrometric constant γ is computed as:

$$\gamma = \frac{c_p P}{\lambda\epsilon} \quad [16]$$

where $c_p = 1.013\,\mathrm{kJ\,kg^{-1}\,K^{-1}}$ for moist air, P is atmospheric air pressure, λ is the latent heat of vaporization in kilojoules per kilogram, and $\epsilon = 0.622$ is the ratio of molecular weights of water vapor to air. For P in kilopascals, γ will have units of kilopascals per degrees celsius. Mean atmospheric air pressure, P, can be computed routinely as a function of surface elevation using the ideal gas law as:

$$P = P_{os}\left(\frac{T_{os} - \alpha(\mathrm{elev} - \mathrm{elev}_{os})}{T_{os}}\right)^{g/(\alpha_a R)} \quad [17]$$

where: P_{os} and T_{os} are standard atmospheric pressure in kilopascals and absolute temperature in degrees Kelvin at elevation elev_{os}, and elev is elevation of the ground surface above mean sea level; α_a is the assumed constant adiabatic lapse rate, normally taken as $0.0065\,\mathrm{K\,m^{-1}}$ for saturated air or sometimes as $0.01\,\mathrm{K\,m^{-1}}$ for nonsaturated air; g is gravitational acceleration, $9.807\,\mathrm{m\,s^{-2}}$, and R is the specific gas constant for dry air, $287.0\,\mathrm{J\,kg^{-1}\,K^{-1}}$. Values for P_{os}, T_{os}, and elev_{os} are commonly taken for the standard atmosphere at sea level, which are 101.3 kPa, 288 K, and 0 m, respectively. Equation [17] is relatively insensitive to the value of α_a for elevations up to 3000 m.

Air density, ρ, can be computed as:

$$\rho = \frac{1000\,P}{T_v\,R} \quad [18]$$

where P is in kilopascals, R is $287.0\,\mathrm{J\,kg^{-1}\,K^{-1}}$, and T_v is virtual temperature, in degrees Kelvin. T_v can be computed as:

$$T_v = \frac{T}{1 - 0.378\frac{e_a}{P}} \quad [19]$$

where T is air temperature in Kelvin and e_a is mean vapor pressure of the air (kilopascals). Latent heat of vaporization, λ, in megajoules per kilogram can be computed as:

$$\lambda = 2.501 - 2.361 \times 10^{-3}\,T \quad [20]$$

where T is mean air temperature in degrees Celsius. If available, mean surface temperature or wet bulb temperature can be used to compute the value of λ, which better represents conditions at the evaporating surface.

The PM equation can be applied for short periods of 1 h or less, and is intended to represent nearly instantaneous fluxes of ET. However, the equation has traditionally been applied for daily and longer (up to 1 month) calculation time-steps owing to paucity of weather data. For daily and longer time-steps, e_s in the PM equation should be computed as:

$$e_s = \frac{e^o(T_{max}) + e^o(T_{min})}{2} \quad [21]$$

where T_{max} and T_{min} are maximum and minimum daily air temperature for the period and e^o is the saturation vapor pressure function. Actual vapor pressure of the air, e_a, can be computed from relative humidity data, from wet-bulb/dry-bulb pyschrometer data, or from dew-point temperature measurements:

$$e_a = e^o(T_d) \quad [22]$$

where T_d is mean daily or early morning dew-point temperature.

Aerodynamic and Surface Parameters

When used to predict ET from dense vegetation, aerodynamic resistance in the PM equation can

be calculated using parameters for d, z_{om}, and z_{oh} determined as a function of vegetation height:

$$d = 0.67H \quad [23]$$

$$z_{om} = 0.12H \quad [24]$$

$$z_{oh} \approx 0.1 z_{om} \quad [25]$$

where z_{om}, z_{oh}, d, and mean plant height, H, are in meters. For bare soils, common values for z_{om} range from 0.00 001 m for mud flats to 0.005 m for smooth soil to 0.01 m or more for very roughly ploughed soil.

Parameterization of Surface Resistance

The bulk surface resistance r_s for a crop is an integration of effects of transpiration from various locations within the vegetation and contributions from soil evaporation, along with transfers of radiation and sensible heat within the canopy. Simple approaches determine soil evaporation indirectly by estimating the distribution of net radiation below and between a plant canopy having partial ground cover. For crops that completely shade the ground, total surface resistance can be computed, where surface resistances from soil and vegetation are added in parallel according to leaf area and radiation extinction coefficients. For an immature row crop having exposed bare soil between plant rows and incomplete wetting of the soil surface during irrigation, r_s can be expressed through a combination of various surface resistances acting in parallel. The combination is made according to the fractions of surface C represented by each parallel resistance component (Figure 2):

$$\frac{1}{r_s} = \frac{C(1 - K_r^{LAI/C})}{r_{sc}} + K_r^{LAI/C}\left(\frac{C_{sW}}{r_{sssW}} + \frac{C_{sD}}{r_{sssD}}\right) + \frac{C_{eW}}{r_{sseW}} + \frac{C_{eD}}{r_{sseD}} \quad [26]$$

where C is the effective fraction of ground shaded by vegetation and is computed as a function of sun angle, row height and width, leaf density, and row orientation; and C_{sW}, C_{sD}, C_{eW}, and C_{eD} are the fractions of the soil surface that are shaded and wet, shaded and dry, sunlit and wet, and sunlit and dry ($C_{sW} + C_{sD} + C_{eW} + C_{eD} = 1$ and $C_{sW} + C_{sD} = C$). Resistance r_{sc} is the surface resistance of the canopy, r_{sssW} is the surface resistance of wet, shaded soil; r_{sssD} is the surface resistance of dry, shaded soil; r_{sseW} is the surface resistance of wet, sunlit soil; and r_{sseD} is the surface resistance of dry, sunlit soil. Parameter K_r is a radiation extinction coefficient for the vegetation canopy, having a value of approximately 0.7; and LAI is leaf area index.

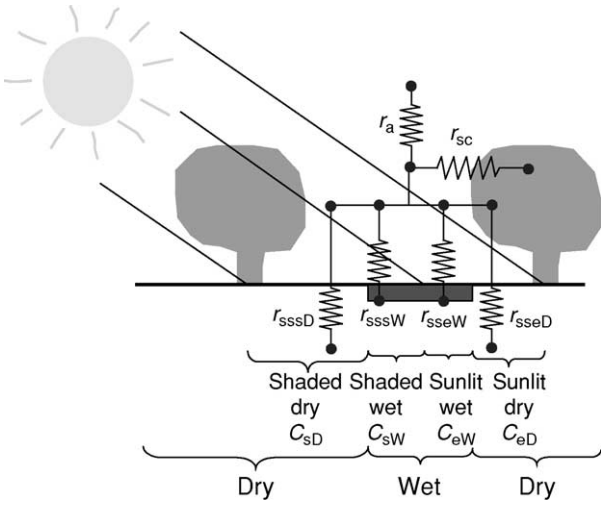


Figure 2 Schematic showing breakdown and definition of surface resistances for soil when vegetation is at partial cover. r_a, aerodynamic resistance; r_{sc}, surface resistance for the vegetation cover; r_{sssW}, surface resistance of wet, shaded soil; r_{sssD}, surface resistance of dry, shaded soil; r_{sseW}, surface resistance of wet, sunlit soil; r_{sseD}, surface resistance of dry, sunlit soil; C_{sD}, fraction of the soil surface that is shaded and dry; C_{sW}, fraction of the soil surface that is shaded and wet; C_{eW}, fraction of the soil surface that is sunlit and wet; C_{eD}, fraction of the soil surface that is sunlit and dry.

C has also been computed as LAI/4. Separate PM equations have been applied for evaporation and transpiration and then recoupled, similar to the procedure for multilayer ET models. For sun angles other than directly overhead, C can be estimated for a row crop as:

$$C = \frac{W + H\,|\sin(\eta + \beta)|\tan\theta_z}{S} \quad [27]$$

where H is mean plant height, θ_z is the solar zenith angle (zero = nadir (overhead)), η is the sun azimuth angle ($\eta = 0$ when sun is due south, η is negative when sun is east of south, and η is positive when sun is west of south), and β is the row orientation relative to north ($\beta = 0$ for north–south orientation, $\beta = 90°$ for east–west row orientation), W is row width, and S is row spacing. Equation [27] presumes a nearly rectangular row shape of height H and width W and has an upper limit of $C = 1$. In some situations, W may not be measured or available, but an estimate of LAI may be available. In this case, C may be approximated during incomplete row closure as:

$$C = \frac{\text{LAI}}{\text{LAI}_C} + \frac{H}{S}\,|\sin(\eta + \beta)|\tan\theta_z \quad [28]$$

where LAI_C is the LAI at complete row closure and shading of the ground. The value for C and, subsequently, for r_s changes with time of day and interacts with R_n, G, wind speed, and temperature. For 24-h computation time-steps, an average daily value for C is determined for computing soil moisture balances for shaded and unshaded soil. An average daily value for C for east–west rows can be determined by weighting eqn [26] for sunlight intensity and integrating between the limits of sunrise and sunset. For north–south row orientation, all soil is sunlit and shaded at various times of the day.

Surface resistance, r_{sc}, for densely growing vegetation is often computed in the 'single-layer' PM as a function of leaf area by assuming all leaves function as resistors in parallel:

$$r_{sc} = \frac{r_l}{LAI_{eff}} \qquad [29]$$

where r_l is the bulk stomatal (or surface) resistance of the vegetation per unit LAI (seconds per meter) and LAI_{eff} is the effective LAI contributing to ET. Parameter r_l is the inverse of g_l, the stomatal conductance per unit leaf area. The value of r_l has been taken as $100\,s\,m^{-1}$ for many well-watered agricultural crops when calculations are made on a 24-h basis. When ET calculations are made on an hourly or shorter basis, r_l for agricultural crops has a value of approximately $70–80\,s\,m^{-1}$, assuming that stability corrections are made or that buoyancy conditions are nearly neutral. The value for r_l increases with environmental stresses, including soil moisture deficit.

Surface resistance, r_{sc}, plays an important role in controlling the flux of transpiration from plants into the atmosphere and must be accurately predicted for routine application of the PM and multilayer ET models. r_{sc} represents the ET-averaged daily resistance for 24-h calculation time steps or the mean canopy resistance for hourly time steps for a fully developed canopy. Algorithms are used to modify r_{sc} as leaf area (LAI) and environmental parameters change. General, simple environmental stomatal regulation algorithms have been proposed, where bulk surface resistance is expressed in terms of multiplicative products of linear functions of solar radiation, air temperature, vapor pressure deficit, and soil water deficit.

Soil resistance is generally assumed to increase with drying of soil. Exponential functions describe the process relatively well. A moisture-based soil resistance function describes experimental data from a Portneuf silt loam soil near Kimberly, Idaho:

$$r_{ss} = 2000e^{-5x_s} \qquad [30]$$

where x_s is the remaining proportion of water in the evaporating layer of the soil surface and e is the natural number. x_s is computed as $x_s = W_{ei}/W_{emax}$, where W_{ei} is the total evaporable moisture near the soil surface on day i and W_{emax} represents the upper limit of evaporable water. A daily or more frequent water balance of the C_{sW}, C_{sD}, C_{eW}, and C_{eD} fractions of the soil surface layer need to be conducted to predict x_s, and thus r_{ss}, for each surface fraction. The evaporation from each fraction is predicted in some proportion to ratio of the inverse of r_{ss} for that fraction to the total bulk r_s for the soil–canopy complex. The soil surface layer is typically presumed to be 0.1–0.15 m thick. Equation [30] predicts values for r_{ss} ranging from approximately $50\,s\,m^{-1}$ directly after wetting to approximately $2000\,s\,m^{-1}$ for a relatively dry surface soil layer.

Multilayer ET Models

The PM equation is basically a single-layer model where the resistances for vegetation and soil are assumed to reside in parallel, i.e., there are no intermediate aerodynamic or convective resistances separating them or within them. More complicated, multilayer ET models have been employed that contain sophisticated parameter algorithms, intermediate resistances, and relatively complex calculation mechanisms than the PM. Multilayer models decompose the computation of r_s, r_{sc}, and r_a for discrete layers within the vegetation canopy. These models are generally capable of making more accurate predictions of ET than the PM, especially for short time-steps (hourly). Because of their increased complexity, multilayer models can provide more capability and flexibility in predicting ET under variable cropping and weather conditions without the need for calibration, depending on the physical basis of internal mechanisms. On the other hand, these models sometimes require substantial calibration of empirical coefficients internal to the model.

Net Radiation

Net radiation, R_n, can be measured using a net radiometer that measures downward and upward fluxes of shortwave and longwave radiation. The surface beneath the net radiometer must be very similar in wetness and vegetation type and amount to that being predicted in the PM equation. In many situations, R_n data are not available for a specific surface or condition, and therefore must be estimated from solar radiation data, air temperature, and possibly vapor pressure. R_n for row crops is often calculated by separately considering the reflectances (albedo) for

soil and vegetation. When calculated, the general formula for R_n is:

$$R_n = (1 - \alpha)R_s + R_{nl} \quad [31]$$

where α is the albedo of the surface, representing the crop-soil mixture, R_s is incoming shortwave radiation (measured by a pyranometer), and R_{nl} is net (incoming − outgoing) longwave radiation. R_{nl} is commonly predicted from air temperature and humidity data from a local weather station.

A general formula for α of a crop-soil mixture is:

$$\alpha = C\alpha_c + (1 - C)\alpha_s \quad [32]$$

where α_c is plant albedo, α_s is soil albedo, and C is the relative fraction of vegetation cover (defined previously) that considers effects of sun angle. Plant albedo can be computed by choosing typical values at the key growth stages. For example, values of α_c have been found to be 0.25, 0.23, 0.23, and 0.30 for beans at Kimberly, Idaho, at emergence, effective cover, ripening, and harvest. Values are 0.25, 0.15, 0.15, and 0.35 for wheat for the same stages. For overcast conditions the value for α_c increases, often to approximately 0.35.

Soil albedo in eqn [32] can be computed as a composite of wet soil and dry soil albedo:

$$\alpha_s = \frac{\alpha_{sM} C_{eW} + \alpha_{sD} C_{eD}}{C_{eW} + C_{eD}} \quad [33]$$

where α_{sM} is albedo of moist soil and α_{sD} is albedo of dry soil. The albedo of a moist, drying soil can be assumed to vary linearly with water content between the value for a fully wet soil and a fully dry soil over a given drying period:

$$\alpha_{sM} = \alpha_{sW} + (\alpha_{sD} - \alpha_{sW})(1 - x_{skin}) \quad [34]$$

where α_{sW} is albedo of wet soil and x_{skin} is the fraction of remaining water in the soil skin (top approx. 1 cm). Example values for α_{sW} and α_{sD} for the Portneuf silt loam near Kimberly are $\alpha_{sW} = 0.10$–0.20 and $\alpha_{sD} = 0.35$.

Soil Heat Flux

The magnitude of soil heat storage or release can be significant over a few hours, but is usually small from day to day because heat stored early in the day as soil warms is lost late in the day and at night as the soil cools. For application of PM to short periods, e.g., using hourly data, measurements or estimates for G are required.

One method commonly used in research is the use of heat flux plates installed at shallow depths below the soil surface (e.g., at 0.01 m). Measurements by soil heat flux plates, G_{zs} at some depth z_s are adjusted to represent G at the soil surface by incorporating the change in heat storage between the soil surface and heat flux plate:

$$G = G_{z_s} + C_s \int_0^{z_s} \frac{\partial T_s}{\partial t} dz \quad [35]$$

where $\partial T_s/\partial t$ is the change in temperature with time within the z_s thick layer above the heat flux plate. Temperature is usually measured with thermocouples or buriable thermistors.

Soil heat flux density for dense, short vegetation such as for grass, wheat, or alfalfa can be approximated during daylight periods as:

$$G = 0.4e^{-0.5\,LAI} R_n \quad [36]$$

where e is the natural number and G has the same units as R_n. Equation [36] predicts $G = 0.1\,R_n$ for an LAI = 2.8, which is typical for clipped grass. A low fraction of G/R_n (0.036) has been used during daylight for a deciduous forest. The American Society of Civil Engineers (ASCE) recommends $G/R_n = 0.04$ for full-cover alfalfa during daytime. For LAI = 0 (i.e., bare soil), eqn [36] predicts $G/R_n = 0.4$.

For nighttime under grass forage, the recommendation is $G = 0.5\,R_n$ for grass and 0.2 for full-cover alfalfa. In these reported ratios, the effects of seasonal variations are not addressed. The seasonal lag of soil temperature behind solar radiation can be expected to produce higher ratios.

Application of the Penman–Monteith Equation

Besides direct application of the PM equation to a complex surface of soil and vegetation or to soil only, the equation has been widely used to represent ET from a reference vegetation crop for use in irrigation management. The reference crop ET (ET_o) represents ET from either a cool-season clipped grass or a tall, full-cover alfalfa crop, depending on the parameters specified for aerodynamic and surface resistances.

When the PM equation is used as an ET reference, actual ET for a soil or soil–vegetation surface is computed by multiplying ET_o by a crop coefficient, K_c, where K_c represents the ratio of actual ET to ET_o. The value for K_c of soil is varied according to the moisture level of the soil. Figure 3 shows a comparison of hourly ET predicted by the ASCE version of the PM equation and ET measured by a precision-weighing lysimeter (essentially a huge flowerpot suspended

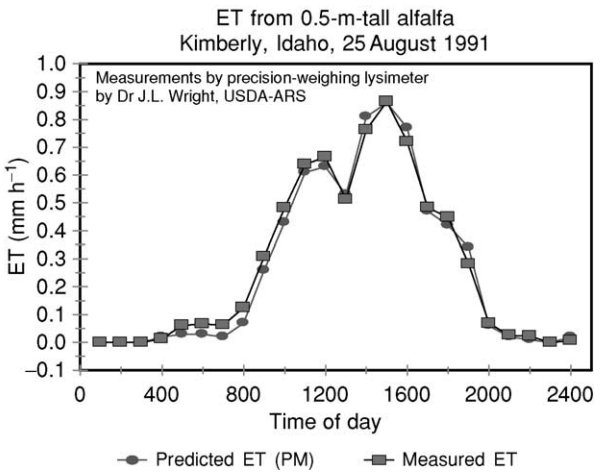


Figure 3 Comparison between evapotranspiration (ET) during hourly periods predicted by the Penman–Monteith (PM) equation (using parameters from the American Society of Civil Engineers) and ET measured by a precision-weighing lysimeter.

below ground via mechanical balances). The PM estimates, based on measured solar radiation, air temperature, humidity, and wind speed, agree very well with field measurements. Results are not always this good, however, and depend on the representativeness of values for roughness and for surface resistance.

See also: **Energy Balance**; **Evapotranspiration**; **Penman, Howard Latimer**

Further Reading

Allen RG, Pereira LS, Raes D, and Smith M (1998) *Crop Evapotranspiration: Guidelines for Computing Crop Water Requirements*. Irrigation and Drainage Paper 56. Rome, Italy: United Nations FAO.

ASCE (2002) *The ASCE Standardized Reference Evapotranspiration Equation*. Report of the Task Committee on Standardization of Reference Evapotranspiration, Environmental and Water Resources Institute of the American Society of Civil Engineers. Reston, VT: ASCE Publications.

Brisson N (1996) On the importance of soil evaporation and rainfall interception in irrigation scheduling. In: *Evapotranspiration and Irrigation Scheduling*, pp. 281–288. Proceedings of the International Conference of ASAE, San Antonio, TX.

Brutsaert W (1982) *Evaporation into the Atmosphere*. Dordrecht, the Netherlands: D. Reidel.

Choudhury BJ (1989) Estimating evaporation and carbon assimilation using infrared temperature data: vistas in modeling. In: Ghassem A (ed.) *Theory and Applications of Optical Remote Sensing*, pp. 628–690. New York: John Wiley.

Choudhury BJ, Idso SB, and Reginato RJ (1987) Analysis of an empirical model for soil heat flux under a growing wheat crop for estimating evaporation by an infrared temperature-based energy balance equation. *Agricultural and Forest Meteorology* 39: 283–297.

Clothier BE, Clawson KL, Pinter PJ, Moran MS, Reginato RJ, and Jackson RD (1986) Estimates of soil heat flux from net radiation during the growth of alfalfa. *Agricultural and Forest Meteorology* 37: 319–329.

Grant DR (1975) Comparison of evaporation from barley with Penman estimates. *Agricultural Meteorology* 15: 49–60.

Jensen ME, Burman RD, and Allen RG (1990) *Evapotranspiration and Irrigation Water Requirements*. ASCE Manuals and Reports on Engineering Practice No. 70. Reston, VA: ASCE Publications.

Lafleur PM and Rouse WR (1990) Application of an energy combination model for evaporation from sparse canopies. *Agricultural and Forest Meteorology* 49: 135–153.

Leuning R, Condon AG, Dunin FX, Zegelin S, and Denmead OT (1994) Rainfall interception and evaporation from soil below a wheat canopy. *Agricultural and Forest Meteorology* 67(3–4): 221–238.

Lhomme JP (1988) A generalized combination equation derived from a multi-layer micrometeorological model. *Boundary-Layer Meteorology* 45: 103–115.

Liu S, Riekerk H, and Gholz HL (1996) ETM evapotranspiration model: an introduction. In: *Evapotranspiration and Irrigation Scheduling*, pp. 329–334. Proceedings of the International Conference of ASAE, San Antonio, TX.

Monteith JL (1965) Evaporation and the environment. In: *The State and Movement of Water in Living Organisms*, pp. 205–234. XIXth Symposium of the Society for Experimental Biology, Swansea. Cambridge, UK: Cambridge University Press.

Penman HL (1948) Natural evaporation from open water, bare soil and grass. *Proceedings of the Royal Society of London*. Series A: Mathematical and Physical Sciences 193: 120–146.

Perrier A (1982) Land surface processes: vegetation. In: Eagleson P (ed.) *Land Surface Processes in Atmospheric General Circulation Models*, pp. 395–448. Cambridge, UK: Cambridge University Press.

Price DT and Black TA (1989) Estimation of forest transpiration and CO_2 uptake using the Penman–Monteith equation and a physiological photosynthesis model. In: Black TA, Spittlehouse DA, Novak MD, and Price DT (eds) *Estimation of Areal Evapotranspiration*, pp. 213–227. IAHS Publication No. 177.

Qualls R and Brutsaert W (1996) Effect of vegetation density on the parameterization of scalar roughness to estimate spatially distributed sensible heat fluxes. *Water Resources Research* 32(3): 645–652.

Seginer I (1973) Aerodynamic roughness of vegetated surfaces. *Boundary Layer Meteorology* 5: 383–393.

Shuttleworth WJ and Wallace JS (1985) Evaporation from sparse crops – an energy combination theory.

Quarterly Journal of the Royal Meteorological Society 111: 839–853.

Stewart JB (1989) On the use of the Penman–Monteith equation for determining areal evapotranspiration. In: Black TA, Spittlehouse DL, Novak MD, and Price DT (eds) *Estimation of Areal Evapotranspiration*, pp. 3–12. IAHS Publication No. 177.

Szeicz G and Long IF (1969) Surface resistance of crop canopies. *Water Resources Research* 5: 622–633.

Wallace JS, Roberts JM, and Sivakumar MVK (1990) The estimation of transpiration from sparse dryland millet using stomatal conductance and vegetation area indices. *Agricultural and Forest Meteorology* 51: 35–49.

Percolation *See* **Hydrodynamics in Soils**

Permafrost *See* **Polar Soils**

Permeability *See* **Hydrodynamics in Soils**

Persistent Organic Pollutants (POPs) *See* **Pollutants:** Persistent Organic (POPs)

PESTICIDES

R H Bromilow, Rothamsted Research, Harpenden, UK

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Pesticides are used worldwide and increase crop yield on average by 30% as well as improving crop quality. Modern organic (i.e., carbon-containing) pesticides have been used widely since the late 1940s, and so there is over 50 years of experience of their behavior in the environment and their impact upon it. The spectrum of compounds applied over this time has changed substantially, with, for example, the older organochlorine insecticides being phased out and with completely new types of chemical structures being regularly introduced. Though increasingly strict registration criteria, coupled with market forces, are leading to the discontinuation of the use of many older compounds, it is likely that several hundred active ingredients will remain available. In the European Union, for example, perhaps 300 of approximately 850 formerly approved compounds will still be marketed after the current review process.

The soil is the ultimate sink for much of the pesticide that is applied to crops or indeed to uncropped land such as railway tracks. Compounds such as nematicides (for control of plant-parasitic nematode worms) are usually incorporated into the topsoil. Many herbicides used in arable crops are applied either to bare soil prior to emergence of the sown crop or to the seed bed immediately postgermination, and so most of the compounds directly reach the soil surface. In contrast, insecticides and fungicides are

usually applied later in the season when the crop may have closed its canopy, and so most compound will be intercepted by the crop foliage; even so, some spray will drift down, and wash-off by rain may subsequently move more compound to the soil. The rates at which pesticides are applied vary widely: a typical arable field in the UK in 2003 received approximately $4\,kg\,ha^{-1}$ of pesticide active ingredients in total per annum, though an intensively sprayed crop such as potatoes averages $15\,kg\,ha^{-1}$.

To understand the behavior of so many diverse compounds, it is necessary to develop an overarching framework of understanding based on the physicochemical properties and degradative behavior of each compound. The behavior of new pesticides can then be assessed by comparison or benchmarking with current compounds of similar properties. Important aspects of pesticide behavior in soil to be considered are persistence, sorption, movement with its potential for contamination of water, uptake by plants, and possible impacts on soil biota and microbial processes.

Physicochemical Properties of Pesticides

The distribution of pesticides amongst the different environmental compartments, including the soil phases and the organisms living both in and on soil, is largely determined by the physicochemical properties of the compound. For example, vapor pressure determines the propensity of a pesticide to be lost by volatilization from a soil surface. If the soil is wet, then the Henry Law constant (the distribution coefficient between water and air) gives a better predictor:

$$H' = \frac{T}{273} \cdot \frac{\mathrm{VP}\,10^{-5}}{22.4\mathrm{WS}} \qquad [1]$$

where H' is the dimensionless Henry Law constant, VP is the vapor pressure (in pascals), and WS is the water solubility (in moles per liter) measured at temperature T (Kelvin). Organic compounds with high H' will have high concentrations in the vapor phase compared with the soil water. As diffusion in air is approximately 10^4 times faster than in water, then this region defines the boundary between movement predominantly via the air and via the water phase in soil (Figure 1).

Water solubility is an easily understood and oft-quoted property. However, lipophilicity, essentially a measure of the partitioning between water and lipid-like organic phases, is a better predictor of many processes, such as accumulation by organisms. To have comparable measurements, lipophilicity is

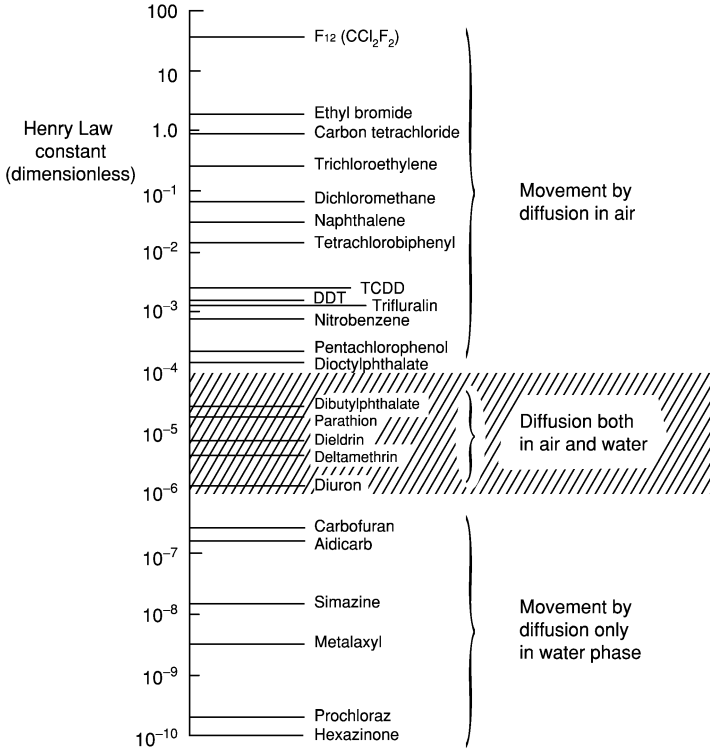


Figure 1 Pathways of movement of pesticides and environmental contaminants by Henry Law constants. Reproduced with permission from Trapp S and Mc Farlane JC (eds) (1995) *Plant Contamination, Modeling and Simulation of Organic Chemical Processes*. Boca Raton, FL: Lewis Publishers. Copyright Lewis Publishers, an imprint of CRC Press, Boca Raton, FL.

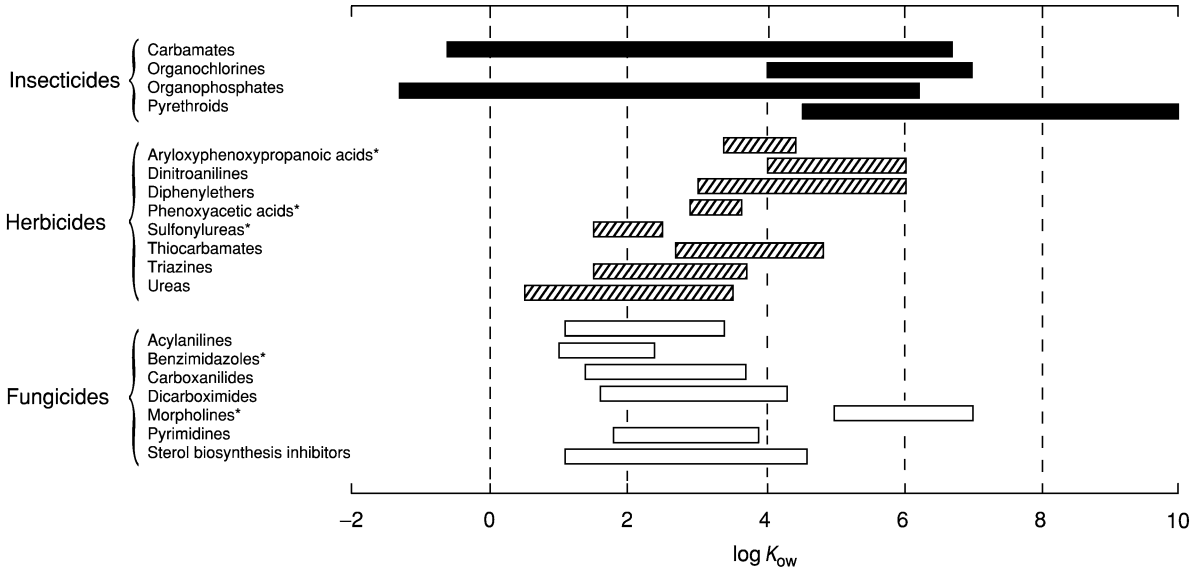


Figure 2 Ranges of 1-octanol/water partition coefficients (as log K_{ow}) for commonly occurring compounds in various classes of pesticides. Asterisks indicate ionizable compounds, whose log K_{ow} values are plotted for the undissociated molecule. Adapted from Trapp S and Mc Farlane JC (eds) (1995) *Plant Contamination, Modeling and Simulation of Organic Chemical Processes*. Boca Raton, FL: Lewis Publishers with permission. Copyright Lewis Publishers, an imprint of CRC Press, Boca Raton, FL.

normally assessed as the 1-octanol/water partition coefficient, K_{ow} (usually expressed as $\log_{10} K_{ow}$). Some pesticide classes span quite a narrow range of lipophilicity, whereas others span up to 7 orders of magnitude of K_{ow} (Figure 2). For compounds that are acids or bases, lipophilicity will depend on the balance of nonionized and ionized forms present; this is determined by the pH and the compound pK_a (the pK_a is the pH at which a particular acid or base function is 50% ionized). Ionized compounds are much less lipophilic (i.e., much more polar) than their undissociated forms.

Rates of degradation influence the depth of penetration into soil and degree of accumulation by biota; these rates generally need to be measured, for only approximate and often uncertain estimates can be made knowing the chemical structure.

Persistence in Soil

Loss Processes and Factors of Influence

Degradation is the main loss process for most pesticides in soil. Initial breakdown may be by biotic or abiotic processes, the latter including photolysis as well as chemical reactions such as hydrolysis and oxidation. However, the further extensive breakdown of complex fragments such as phenyl rings will be an essentially biotic process, with many species of microfungi, bacteria, and actinomycetes able to participate in the various steps. Thus a pesticide reaching soil will be subject to breakdown, leading to a sequence of metabolites or degradation fragments whose ultimate fate is mostly to be oxidized to CO_2 or incorporated into the polymeric matrix of the soil organic matter. This organic matter is itself constantly undergoing slow degradation, balanced by input of fresh organic matter such as crop residues such that an equilibrium level is reached over a number of years.

In soils, the rate of pesticide breakdown under constant conditions of temperature and moisture usually approximates to first-order kinetics, i.e., the rate of decomposition is proportional to the amount remaining:

$$\frac{dc}{dt} = -kc \qquad [2]$$

where c is pesticide concentration, t is the time, and k the rate constant. This is akin to radioactive decay, and by analogy a half-life can be proposed. The half-life ($t_{1/2}$) is estimated by plotting the logarithm of the amount of compound remaining against time, which should approximate to a straight line (Figure 3a). Integrating eqn [1], the rate constant k (units are per time) can then be calculated from:

$$k = \frac{\ln 2}{t_{1/2}} \qquad [3]$$

In the field, however, soil conditions of temperature and moisture can vary considerably over the lifetime of a pesticide, and this influences loss rates. Within

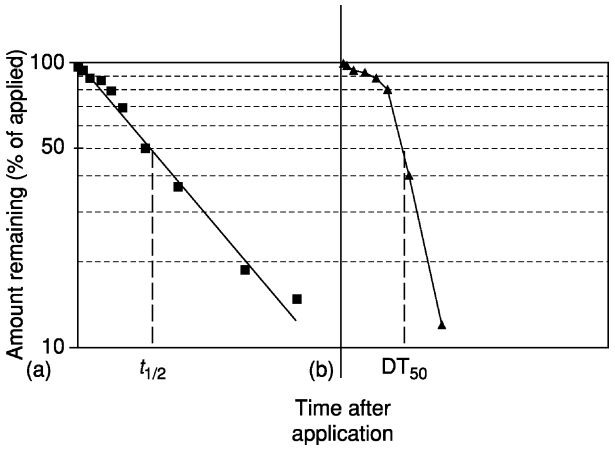


Figure 3 Kinetics of pesticide degradation in soil under constant conditions: (a) first-order kinetics showing the estimation of half-life; (b) microbial adaptation, showing a lag phase followed by rapid breakdown.

the normal range of ambient temperatures, a 10°C rise in temperature increases breakdown rates by two- to threefold both for abiotic and biotic degradation. In frozen soils, breakdown is usually very slow or negligible.

Most of the breakdown of pesticides by microorganisms is incidental to the general breakdown and turnover process of organic matter in soil, and so does not change the microbiota. The quantity of organic material present in the pesticide is normally a minute fraction of that being decomposed naturally in soil. However, occasionally certain pesticides or their metabolites, despite their low concentrations in soil, can be used as a food source by one or more species of microorganisms, and this is known as 'microbial adaptation.' The kinetics of breakdown then follow an initial lag phase, during which the population of the adapted organism increases until it reaches such levels that the pesticide is very rapidly degraded (**Figure 3b**). A second application would be immediately degraded. It should be noted that certain pesticides (e.g., preemergence herbicides and soil insecticides) need to persist in soil for a reasonable length of time if they are to be effective (e.g., several weeks), and such microbial adaptation with its concomitant rapid breakdown has thus limited the efficacy of some compounds, e.g., thiocarbamate herbicides and the carbamate insecticide carbofuran used to treat against corn rootworm (*Diabrotica* spp.).

Soil-moisture content influences both the availability of a pesticide for breakdown and the populations and activities of the microorganisms. It is thought that only compounds in the soil water are freely available to microorganisms for breakdown, and that sorbed compounds only become available as desorption into the soil water occurs to reestablish the sorption equilibrium after such breakdown. Thus, for example, the strongly sorbed herbicide paraquat is very persistent in soil, with a half-life of several years, but in culture medium without soil it can be rapidly degraded by microorganisms. As soils dry, the proportion of compound in the soil water will be decreased and also the activity of microorganisms will be increasingly limited, both of these processes causing a strong slowing of breakdown in drier soils.

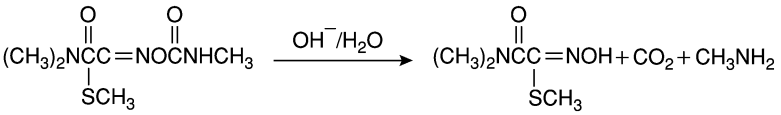
It should be noted that these two factors of temperature and moisture can often serve to cancel each other out in the field, because a sunny spell of weather causes both drying and warming of the surface layer of the soil, which often contains most of the pesticide. Thus breakdown in the field may approximate to first-order kinetics despite the changing environmental conditions caused by weather. However, losses of pesticide in the field can be biphasic, with a more rapid loss in the first few days of application than subsequently; this initial loss is usually attributed to surface processes such as volatilization or photolysis, and the first major rainfall event will move the compound sufficiently into the soil to block photolysis and reduce volatilization. Given the differing processes and variable conditions in the field, it is usual to describe pesticide persistence here by the DT_{50} or DT_{90} values, the disappearance time for 50% or 90%, respectively, of the parent compound; these are measured values with no implications as to the processes or kinetics involved.

Other factors can also influence breakdown rates of pesticides in soil. Increasing organic matter content encourages microbial activity but reduces pesticide availability in the soil water due to stronger sorption. Again these two factors are working in opposite directions, though, in poor sandy soils with little organic matter, degradation can be much slower than in more fertile soils.

Soil pH can also influence breakdown. For example, carbamoyloxime nematicides such as aldicarb and oxamyl are sensitive to hydrolysis at alkaline pH, and so these compounds are not very persistent in soils of pH >7.5 (**Figure 4a**). Conversely, the first generation of sulfonylurea herbicides such as chlorsulfuron (**Figure 4b**) are subject to acid hydrolysis and so do not persist in soils of low pH.

Microbial activity also usually diminishes with soil depth, resulting in breakdown in subsoil usually being slower than in topsoil. Nevertheless, microorganisms can survive to tens of meters depth, in numbers sufficient to cause slow degradation of pesticides in aquifers. Where aquifers are deep and the recharge takes decades, such degradation may remove pesticide traces before the water is abstracted. However,

(a) Oxamyl



(b) Chlorsulfuron

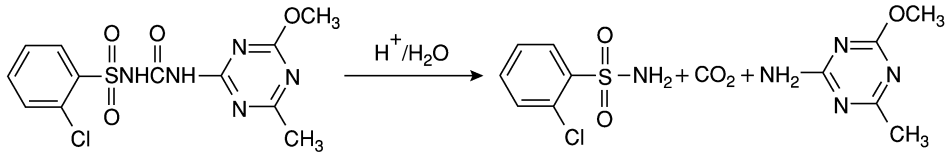


Figure 4 Hydrolysis of pesticides in soil: (a) alkali-catalyzed hydrolysis of oxamyl; (b) acid-catalyzed hydrolysis of chlorsulfuron.

shallow aquifers are inevitably more susceptible to contamination, and much monitoring is done to ensure that widespread contamination of groundwater does not occur.

Another factor in the breakdown process is whether soils are aerobic or anaerobic (i.e., lacking in oxygen). Anaerobic conditions are brought about by prolonged flooding, in which breakdown of soil organic matter uses up the oxygen. An example occurs in the cultivation of rice in paddy fields. This leads to strong reducing conditions, which can cause new degradation pathways to become predominant. Examples are the reduction of aromatic nitro compounds to anilines and of ketones to secondary alcohols. A further aspect is that ferric ions are reduced to ferrous ions under such conditions, and the latter can catalyze the breakdown of some pesticides such as the carbamoyloxime nematicides.

Methods for Studying Pesticide Breakdown

The breakdown pathways of pesticides are identified by incubating radiolabeled pesticide in soil under controlled laboratory conditions. The most usual radiolabel is ^{14}C, the pesticide being synthesized from an appropriate ^{14}C-labeled precursor such that the ^{14}C is in the portion of the molecule likely to be the most recalcitrant to degradation. Such core labeling ensures that most of the metabolites will contain the ^{14}C atom and so can be traced and isolated by thin-layer chromatography or high-pressure liquid chromatography. They can then be identified by techniques such as nuclear magnetic resonance spectroscopy and mass spectrometry. Usually such incubations are done in contained systems such that $^{14}CO_2$ or other volatile radiolabeled compounds produced can be trapped and estimated by liquid scintillation counting. After exhaustive extraction of the soil, typically with a succession of organic solvents increasing in polarity, ^{14}C incorporated in the soil organic matter can be estimated by combustion of the extracted soil with trapping of the evolved $^{14}CO_2$. Adding up the ^{14}C found in these various fractions at each time gives the mass balance which, if the experiments have been done satisfactorily, should be close to the dose applied to the soil at the beginning of the incubation.

Subsequent field studies are usually done with nonradiolabeled compound, with measurement of the parent compound and any major metabolites of interest by high-pressure liquid or gas-liquid chromatography. These techniques need to be very sensitive as some modern pesticides are applied at rates of only $10\,g\,ha^{-1}$ or even less.

Sorption of Pesticides by Soil

The sorption of pesticides to soil is a very important process in determining pesticide behavior in the environment. For example, strong sorption limits movement by leaching, and so such compounds would be unlikely to contaminate groundwater. It is not usually possible to distinguish between adsorption (a surface phenomenon) and absorption (whereby the pesticide penetrates the sorbing matrix such as soil organic matter), and hence use of the term 'sorption,' which covers both processes.

Sorption is typically measured by equilibrating sieved soil with an aqueous solution of the pesticide over a period of 4–24 h. The sorption coefficient ($K_d\,ml\,g^{-1}$) is then calculated from:

$$K_d = \frac{\text{concentration in soil } (g\,g^{-1})}{\text{concentration in water } (g\,ml^{-1})} \qquad [4]$$

Sorption of nonionized and acidic pesticides is largely by partitioning into the soil organic matter. This

process is viewed as essentially a reversible equilibrium, though detailed measurements usually reveal that desorption is somewhat slower than sorption. To allow for comparisons in sorption behavior over soils ranging in organic matter content, K_d can be converted into K_{om} (sorption coefficient on to soil organic matter; $ml\,g^{-1}$) or K_{oc} (sorption coefficient on to soil organic carbon; $ml\,g^{-1}$) by:

$$K_d = \frac{\%\ \text{organic matter}}{100} K_{om} = \frac{\%\ \text{organic carbon}}{100} K_{oc} \quad [5]$$

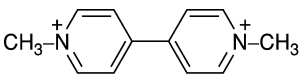
Such partitioning is largely controlled by the lipophilicity of a pesticide, and various relationships have been proposed to predict sorption from lipophilicity as assessed by the 1-octanol/water partition coefficient (K_{ow}) discussed above. Briggs proposed:

$$\log K_{om} = 0.52 \log K_{ow} + 0.62 \quad [6]$$

Though simplistic, this approach works reasonably well for most pesticides. However, certain compounds have functional groups that confer additional sorption by specific interactions. For example, compounds carrying a positive charge can be sorbed by displacing metal ions in the soil matrix, a process known as cation exchange. A particularly striking example occurs with the bipyridinium herbicides such as paraquat, which is a planar molecular carrying two positive charges (Figure 5a). This is strongly sorbed by both organic matter and clay particles; in soils containing the expanding-lattice clay montmorillonite, paraquat can be intercalated between the layers comprising the crystal lattice and is then very tightly bound. A further mechanism of binding is chelation to metal ions such as iron and aluminum, which is observed with oxalic acid, a simple dicarboxylic acid that is dissociated to the oxalate salt at soil pH (Figure 5b). Another example is the herbicide glyphosate, which carries both negative and positive charges (Figure 5c). Though very polar and water-soluble, it is sorbed quite strongly by soil, owing to chelation by the charged groups.

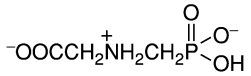
Movement of Pesticides in Soil

In soil, a pesticide is distributed between the soil solids, soil water, and air phases. Most pesticides, with the notable exception of fumigants, are not very volatile, and so movement of most compounds will be via the soil water rather than via the air phase. Movement can be vertically down the soil profile or laterally. Vertical movement, usually referred to as leaching, can lead to contamination of groundwater if the pesticide is weakly sorbed and persistent. Such leaching can be through the bulk of the soil if the soil is not structured as in many sandy soils; in strongly structured clay soils, movement of pesticides in soil water can occur through cracks or pores such as worm or root channels (often termed 'macropore flow' or 'bypass flow'). As such structured soils are often drained, macropore flow can lead to rapid movement of small amounts of pesticide to drainage, which then contaminates surface waters such as streams and rivers. Lateral movement to surface waters can also occur by surface runoff in regions subject to intense rainfall events. Several general classifications of pesticide mobility in soil based on sorption K_{oc} values have been proposed (and which can be related to K_{om}, K_d, and K_{ow}); one such example is given in Table 1. This is primarily of use for non-ionized pesticides, because sorption of ionizable pesticides may depend on soil pH, and shows that increasing lipophilicity decreases mobility.



(a) Paraquat

(b) Oxalate



(c) Glyphosate

Figure 5 Charged compounds undergoing strong sorption by soil: (a) paraquat; (b) oxalate anion; and (c) glyphosate.

Table 1 Classification of pesticide mobility in soil

Mobility class	*Log K_{ow}*	K_{oc} *($ml\,g^{-1}$)*	K_{om} *($ml\,g^{-1}$)*	K_d *($ml\,g^{-1}$)*[a]
Nonmobile	>5.3	>4000	>2320	>46
Slightly mobile	3.5–5.3	500–4000	290–2320	5.8–46
Moderately mobile	2.0–3.5	75–500	43–290	0.86–5.8
Mobile	0.6–2.0	15–75	9–43	0.18–0.86
Very mobile	<0.6	<15	<9	<0.18

[a] K_d values given for a soil with 2% organic matter. (Source: Hollis JM (1991) Mapping the vulnerability of aquifers and surface waters to pesticide contamination at the national/regional scale. In: Walker A (ed.) *Pesticides in Soils and Water*. BCPC monograph 47: 165–174. Silsoe, UK: National Soil Resources Institute, Cranfield University.)

Leaching of pesticides is often studied using lysimeter systems. These are undisturbed blocks of soil, typically 1 m deep and 0.5–1.0 m^2 in area, contained within a casing usually made of stainless steel or fiberglass. Such lysimeters can be taken from different soil types and then brought back to a central location for study. Radiolabeled pesticide is applied, and leachate from the lysimeters is monitored over a 2-year period. Any detected radioactivity is subjected to chromatography to identify it as parent compound or metabolite. In assessing the risk of pesticide leaching to groundwater, in the European Union, for example, the mean concentration of pesticide leached from such lysimeters would not be allowed to exceed 0.1 $\mu g\,l^{-1}$, the drinking-water limit for every pesticide. This is a strict criterion that precludes the registration of many uses of pesticides. At the end of the experiment, soil cores can be taken to measure remaining amounts of pesticide and metabolites, as well as to assess the amount of pesticide incorporated into the soil organic matter.

The behavior of pesticides in soil can be simulated by modeling techniques, and a number of such computer programs are now available. These vary in their complexity, and it can be difficult to parameterize the more detailed models. At a minimum, simulation of pesticide behavior in soil requires as input a sorption coefficient, degradation constant, climate data (temperature, rainfall), and soil properties to allow description of water flow through the soil. Such models can simulate well the behavior of the bulk of pesticide in soil, but it has proved difficult to obtain reliable estimates of variable processes such as losses to drainage via macropore flow.

Uptake of Pesticides from Soil by Plants

Plant roots and stem bases are in contact with the air and water in soil, and so can acquire pesticide present in these phases. Furthermore, crop plants typically transpire water amounting to several times their own weight each day, and so there is substantial mass flow to the roots of soil water and with it a potential burden of pesticide. In practice, the proportion of soil-applied pesticide found in plants is usually small, due in part to the pesticide being degraded and sorbed in soil and also often to it being in the uppermost layers of the soil above the main rooting zone. Nevertheless, pesticide uptake by plants can be important, examples being the carryover of herbicides from one crop to the next so causing phytotoxicity, and secondly the occurrence of pesticide residue in the edible portion of crops at harvest.

Only a few compounds move predominantly via soil air rather than via the water phase, these more volatile compounds being characterized as having the (dimensionless) Henry Law constant greater than 10^{-4}. This discussion will thus focus on uptake into plants via the soil water, the predominant pathway for most pesticides. Uptake can be divided into two processes: (1) uptake into roots, and (2) transport to shoots via the transpiration stream flowing up the xylem vessels. Entry into roots for nonionized pesticides is primarily an equilibration from the soil water into the aqueous phase of the root cells, with partitioning on to the lipid-like constituents of roots. This latter process is determined by pesticide lipophilicity and is the main uptake process for compounds of log $K_{ow} > 2$. Compounds of $\log K_{ow} > 4.5$ can be accumulated at concentrations greater than 100-fold (on a wet weight basis) compared with those of a bathing solution akin to soil water. As soils increase in organic matter content, stronger sorption reduces pesticide concentrations in the soil water and so uptake by plants is less. Translocation of such nonionized chemicals to shoots is, however, in no way related to uptake by roots. Such transport is a passive process, limited by the rate at which compounds can cross the root endodermis, which is a cylindrical sheath of cells surrounding the vascular tissues. Translocation efficiency is related to pesticide lipophilicity, being optimal at $\log K_{ow} \sim 1.8$, at which compounds move with approximately 80% of the efficiency of water, with more polar and more lipophilic compounds becoming increasingly less well transported.

Uptake behavior of ionized compounds is more complex. The most important group is composed of the weak acid herbicides; for these, uptake by roots and transport to shoots are proportional, increasing strongly in more acidic soils because the undissociated forms are much better taken up than their respective anions.

Effects on Soil Processes

The possible impact of pesticides on soil processes mediated by microorganisms, together with that on larger biota such as earthworms, has to be extensively studied before a pesticide is registered. Compounds that are applied to sterilize soil, known as 'soil sterilants' or 'fumigants,' are generally applied at a high rate and are volatile compounds such as methyl bromide (whose use will be phased out) or methyl isothiocyanate that are able to move in the soil air spaces throughout the top soil. These, as their name implies, kill most organisms in the treated zone, though recolonization can occur subsequently from the deeper soil layers that would be less affected.

Other pesticides, however, seem to have relatively little effect on microbially mediated soil

processes such as nitrification or respiration. Although microfungi are an important constituent of the soil microorganisms, fungicides at normal rates seem to have little effect on soil processes, even when thoroughly mixed into soil; and, in the field, only a shallow layer of topsoil would normally contain most of the fungicide reaching the soil surface following foliar sprays. Larger organisms such as earthworms and ground-dwelling beetles can be affected by some insecticides and nematicides, but such effects at the population level are usually short-lived.

Long-term trials over 20 years or more with repeated applications of single or combinations of pesticides applied at normal agronomic rates have found no deleterious effects on soil fertility, as assessed by crop yield, or on soil processes such as respiration and levels of microbial biomass carbon. These findings reinforce the more general observation that problems of soil fertility in worldwide agriculture have not occurred despite several decades of use of pesticides. This shows that the soil is quite robust, no doubt because its biota have been decomposing large amounts of complex organic compounds from plants for millions of years. These residues, and those from organic pesticides, are so degraded by the time that they are incorporated into soil organic matter that any original biological activity has been lost. There is thus no reason to believe that use of pesticides is incompatible with sustainable agricultural systems.

List of Technical Nomenclature

c	Concentration
DT_{50}, DT_{90}	Disappearance time (50%, 90%)
H'	Henry Law constant (dimensionless)
k	First-order rate constant (per time)
K_d	Sorption coefficient to soil (milliliters per gram)
K_{oc}	Sorption coefficient to soil organic carbon (milliliters per gram)
K_{om}	Sorption coefficient to soil organic matter (milliliters per gram)
K_{ow}	1-Octanol/water partition coefficient (dimensionless)
T	Temperature (Kelvin)
t	Time
$t_{1/2}$	Half-life
VP	Vapor pressure (pascal)
WS	Water solubility (moles per liter)

See also: **Pollutants:** Biodegradation; Effects on Microorganisms; Persistent Organic (POPs)

Further Reading

Briggs GG (1981) Theoretical and experimental relationships between soil adsorption, octanol–water partition coefficients, water solubilities, bioconcentration factors, and the parachor. *Journal of Agricultural and Food Chemistry* 29: 1050–1059.

Bromilow RH, Evans AA, Nicholls PH, Todd AD, and Briggs GG (1996) The effect on soil fertility of repeated applications of pesticides over 20 years. *Pesticide Science* 48: 63–72.

Führ F and Hance RJ (eds) (1992) *Lysimeter Studies of the Fate of Pesticides in the Soil*. Monograph No. 53. Farnham, Surrey: British Crop Protection Council.

Goring CAI and Hamaker JW (eds) (1972) *Organic Chemicals in the Soil Environment,* vols 1 and 2. New York: Marcel Dekker.

Gustafson DJ (1989) Groundwater ubiquity score. *Environmental Toxicology and Chemistry* 8: 339–357.

Hartley GS and Graham-Bryce IJ (1980) *Physical Principles of Pesticide Behaviour,* vols 1 and 2. London, UK: Academic Press.

Howard PH (ed.) (1991) *Handbook of Environmental Fate and Exposure Data for Organic Chemicals,* vols 1–3. Boca Raton, FL: Lewis Publishers.

Kearney PC and Roberts T (eds) (1998) *Pesticide Remediation in Soil and Water*. Colchester, UK: John Wiley.

Plimmer J (ed.) (2002) *Encyclopedia of Agrochemicals*. New York: John Wiley.

Tomlin CDS (2000) *The Pesticide Manual*, 12th edn. Farnham, Surrey: British Crop Protection Council.

Trapp S and McFarlane JC (eds) (1995) *Plant Contamination: Modeling and Simulation of Organic Chemical Processes*. Boca Raton, FL: Lewis Publishers.

Petroleum *See* **Hydrocarbons**

pH

N S Bolan and K Kandaswamy, Massey University, Palmerston North, New Zealand

© 2005, Elsevier Ltd. All Rights Reserved.

No other single chemical soil characteristic is more important in determining the chemical environment of higher plants and soil microbes than the pH. There are few reactions involving any component of the soil or of its biological inhabitants that are not sensitive to soil pH. This sensitivity must be recognised in any soil-management system (Brady NC and Weil RR (1999) *Nature and Properties of Soils*. Englewood Cliffs, NJ: Prentice-Hall).

Introduction

The acidity–alkalinity status of a soil is referred to as 'soil reaction,' which is often quantified by measuring soil pH. pH is considered to be the 'master variable' of a soil, which determines many of the physical (e.g., flocculation or deflocculation), chemical (e.g., adsorption or desorption), and biological (e.g., mineralization or immobilization) properties of soils. In this chapter, the fundamentals of soil pH will be discussed in relation to its measurement and practical implications.

Theoretical Basis of pH

Pure water dissociates into the hydronium ion $(H_3O)^+$, which is associated with acidity, and hydroxyl $(OH)^-$, which is connected with alkalinity:

$$H_2O + H_2O \rightarrow (H_3O)^+ + (OH)^- \qquad [1]$$

In aqueous media, the free acid cation is mostly present as the hydrated proton $(H_3O)^+$ but for convenience it is often referred to as a hydrogen ion or proton (H^+), and thus the dissociation of water is often represented simply as:

$$H_2O \rightarrow H^+ + OH^- \qquad [2]$$

This is a reversible reaction which obeys the law of mass action:

$$[H^+] \times [OH^-]/[H_2O] = K \qquad [3]$$

where K is a constant. Because the degree of dissociation of pure water is so small, the concentration of undissociated H_2O molecules always remains constant, so that the product $[H^+] \times [OH^-]$ gives a constant value (K_w), known as the dissociation constant or ionization constant of water. The K_w value for pure water at 21°C is 1×10^{-14}, with equal concentration of H^+ and OH^- ions $(1 \times 10^{-7}$ gEq):

$$K_w = [H^+][OH^-] = 1.0 \times 10^{-14} \qquad [4]$$

If a chemical substance is added to water, which either donates or consumes H^+ or OH^- ions, then the concentration balance between H^+ and OH^- ions is disturbed. The numerical product of the concentrations must, however, remain constant, so that if $[H^+]$ increases, $[OH^-]$ must decrease accordingly, and vice versa.

To simplify the arithmetical representation of eqn [4], Sörensen suggested that the equilibrium state could be defined simply by reference to the $[H^+]$ and, to avoid the use of cumbersome decimal fractions, he proposed the use of a term designated 'pH' (French *pouvoir hydrogène* or *puissance d'hydrogene*, or 'hydrogen power') which is defined as the $\log_{10}$ of the reciprocal of the $[H^+]$ measured in gEq:

$$pH = \log_{10}\frac{1}{H^+} \text{ or } -\log_{10}(H^+) \qquad [5]$$

Expressing eqn [4] as a negative logarithm yields:

$$pH + pOH = 14 \qquad [6]$$

where pH and pOH are the negative logarithms of the H^+ and OH^- ion activities, respectively.

Thus, the pH of an aqueous solution is defined as the negative logarithm of the H^+ ion activity in the solution, which is the same as the $[H^+]$ when the salt concentration in the solution is very low. This concept can be applied to a system in which the individual molecules and ions are uniformly dispersed throughout the solution within the limits set by variations due to molecular and ionic thermal movements.

The pH scale ranges from 0 to 14, with pH 7 as the neutral point. At pH 7, $[H^+]$ equals the $[OH^-]$. From pH 7 to 0, the solution is increasingly more acidic; from pH 7 to 14, the solution is increasingly more alkaline (basic; **Figure 1**). The $[H^+]$, which is the ion measured when determining pH, has a 10-fold change between each whole pH number. Thus, a solution with pH 5 will have 10 times more $[H^+]$ than a solution with pH 6.

Measurement of pH

pH is commonly measured either electrometrically using potentiometers or calorimetrically using pH-sensitive dyes.

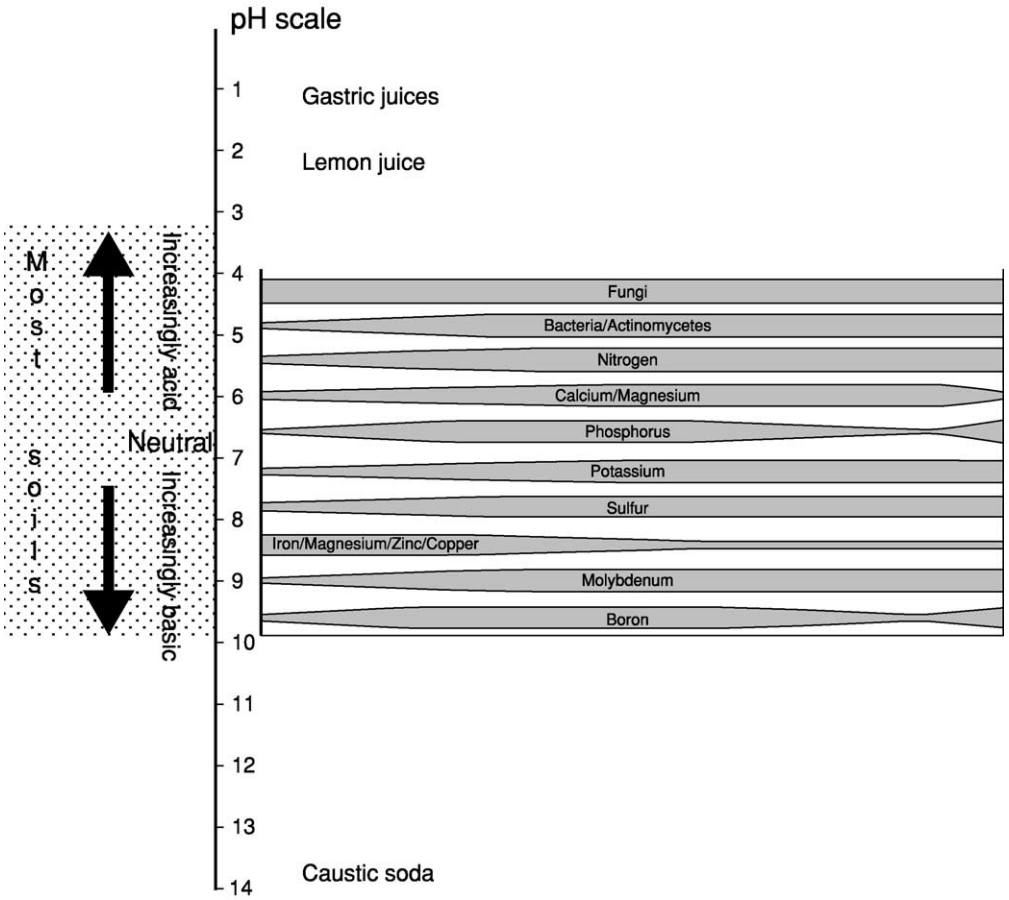


Figure 1 pH scale and the effect of pH on nutrient availability and microbial activity.

Potentiometric Method

This method involves the measurement of potential difference between an electrode whose potential is determined by the $[H^+]$ of the solution and a reference electrode whose potential is independent of $[H^+]$. The normal hydrogen electrode is the glass electrode, and the reference electrode is a calomel electrode in which calomel ($HgCl_2$) is bathed in a solution of potassium chloride (Figure 2).

The glass electrode consists of a thin glass bulb containing dilute HCl, into which is inserted an Ag–AgCl wire, serving as the electrode with a fixed voltage. The HCl solution is separated from the test solution by a membrane of special glass, usually a lithium silicate of particular composition. Differences in H^+ activity across this membrane cause a difference in electrical potential, which can be measured by a potentiometer.

When the glass bulb is immersed in a solution, a potential difference develops between the solution in the bulb and the solution outside the bulb. The electrical potentials developed by this electrode are the membrane potential, plus the potential of the Ag–AgCl–HCl reaction inside the electrode:

$$AgCl + e^- \leftrightarrow Ag + Cl^- \qquad [7]$$

This reaction is reversible and the Ag–AgCl potential ($E^0 = 0.222$ V at 25°C) is likely to remain constant. This potential is cancelled out when the electrode is standardized against a standard pH buffer solution.

The glass electrode develops a second potential at the membrane separating the standard HCl and the test solutions. The minute current flow required by the pH meter develops a double layer in the solutions, causes ion exchange at the inner and outer surfaces, and causes diffusion of ions across the glass membrane. The potentials across the glass membrane can be closely calibrated to the approximate value of the H^+ activity.

A typical reference electrode consists for a wire dipped into the liquid mercury, which makes electrical contact with the pH meter. Current flows from the electrode to the solution phase through the reversible reaction ($E = 0.268$ V at 25°C):

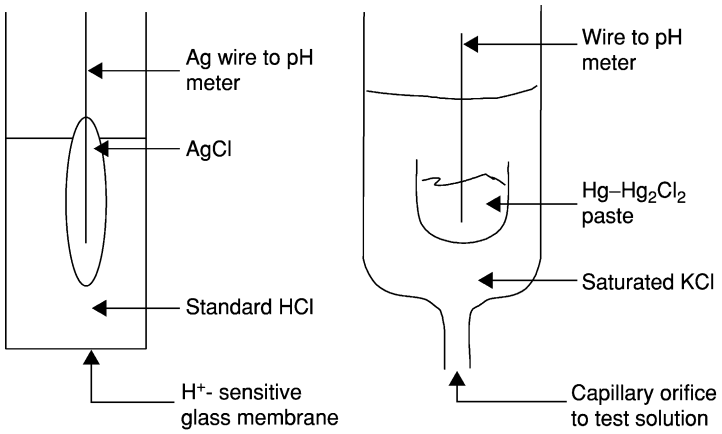


Figure 2 Hydrogen (glass) and reference (calomel) electrodes. Reproduced from Tan KH (1998) *Principles of Soil Chemistry*. New York: Marcel Dekker.

$$Hg_2Cl_2 + 2e^- \leftrightarrow 2Hg + 2Cl^- \quad [8]$$

The Cl^- activity is fixed by the KCl concentration (usually saturated KCl). As a result this potential also cancels out when the pH electrode system is standardized in a standard pH buffer. KCl diffusion through the orifice makes the electrical contact between the reference cell and the test solution. This KCl connection forms a 'salt bridge' between the test solution and the reference electrode. The reference electrode is often built in the same body as the glass electrode to form a combination electrode.

When measuring pH, the potential (E) in the glass electrode is formulated by the Nernst equation:

$$E = (RT/nF)\log(K/H^+) \quad [9]$$

where R is the gas constant, T the absolute temperature, n the valence, F the Faraday constant, K a constant, and H^+ the activity of H^+ ions. E is called the 'half-cell potential' and cannot be measured alone. If the glass electrode is placed against a reference calomel electrode, the potential difference between the two ($E - E_{ref}$) is measurable.

Before any pH measurement, the two electrodes have to be laced first in a solution of known pH. This is called 'standardizing' the electrodes and the pH meter. The overall potential of the total cell E_0 equals $E - E_{ref}$:

$$E_0 = [(RT/F)\log(K/H^+_{std})] - E_{ref} \quad [10]$$

If the two electrodes are now placed in the solution with the unknown $[H^+]$, the potential E_c is:

$$E_c = [(RT/F)\log(K/H^+_{test})] - E_{ref} \quad [11]$$

Subtracting eqn [10] from eqn [11] gives:

$$E_c - E_0 = (RT/F)[\log H^+_{test} - \log H^+_{std}] \quad [12]$$

$$E_c - E_0 = (RT/F)\log(H^+_{test}/H^+_{std}) \quad [13]$$

or

$$E_c - E_0 = (RT/F)(pH_{std}/pH_{test}) \quad [14]$$

where E is the measured potential and is converted to pH units by the scale of the pH meter. E_{ref} includes all other potentials, which are nullified by standardizing the systems with a standard pH buffer solution.

Since $RT/F = 0.0591$ at 25°C:

$$pH = (E_c - E_0)/0.0591 \quad [15]$$

This means that for a change of 1 pH unit, the potential will change by 0.0591 V.

The glass electrode has a uniform response to a wide range of H^+ activities and is less sensitive to interfering ions. The extent of interference is denoted by the selectivity ratio, the concentration ratio of the test ion to interfering ion at which the interfering ion exerts a significant potential at the membrane. The selectivity ratio of the pH glass membrane for H^+ over Li^+, the most serious interfering ion, is approximately 10^9. That is, Li^+ would cause a significant pH error in a pH 9 solution containing 1 mol l^{-1} Li^+ or in a pH 10 solution containing 0.1 mol l^{-1} Li^+. The selectivity ratio for Na^+, the next most serious interference, is approximately 10^{13}.

Determination of pH Using pH-Sensitive Dyes

Rapid measurement of pH can also be achieved colorimetrically with pH indicator dyes. Colorimetric methods are based on the change in color that takes place upon dissociation of a weak acid, or weak base organic dye; thus for a weak acid dye:

$$HI = H^+ + I^- \quad [16]$$

where raising the pH will cause a change in color as the undissociated dye (HI) changes to the dissociated form (I^-). Typically a mixture of dyes is used to yield color changes that take place over a wide range of pH. However, the color can only be measured accurately in a solution from which the solid particles are removed, and the color also depends on the concentration of various salts present in the solution. A dye solution is mixed with the solid to make a slurry, the dye is decanted and the color is visually compared with a color chart. The pH measured in this manner is not precise and can vary from the results obtained using the glass electrode.

Measurement of Soil pH

Because 'pH' is a term that is only defined for solution, in a strict sense, it cannot be applied to a solid-phase material such as soil. However, the chemical properties of the solid-phase components in soil define the pH of the soil solution (i.e., the solution in the soil pores). Further the pH of a soil dispersed in water is not a simple concept like that of a solution. When soil particles, which carry ions attached to them, are dispersed in solution, the ions are not uniformly distributed throughout the solution. The concept of the pH of a soil (or soil suspension) can therefore only be discussed in relation to the properties of the ionic atmosphere around the soil particles (i.e., soil–solution interface).

When a negatively charged soil particle is dispersed in water, the negative charge is neutralized by cations, some of which sit firmly on the clay surface, forming the Stern layer, and some dissociate into the dispersion medium to form the Gouy diffuse double layer. The thickness of this double layer depends in part on the ions dissolved in the dispersion medium, and the thickness increases with a decrease in the concentration of the ions, a decrease in the valency of the cations, and an increase in the size of the cation. The $[H^+]$ in the solution surrounding the soil particles is less than that close to the soil particles, due to the $[H^+]$ gradient in the double layer. As the double layer is made more compact by adding an electrolyte to the soil-water system, the $[H^+]$ gradient across the double layer is reduced, and the pH in the solution becomes almost equal to that close to the surface of the soil particle. Thus the pH of the bulk solution is greater than the pH just outside the Stern layer unless the salt concentration in the solution is high.

The pH of most soils ranges from 3 to 8.5. Soils do not normally have a pH below 3–4 unless they contain free acids (e.g., acid sulfate soils and acid mine drainage). Similarly soils normally only have a pH >8.45 if they contain enough exchangeable sodium (i.e., alkali soils). Further, in calcareous and alkali soils, the effective pH around plant roots can be determined by the concentration of carbon dioxide (CO_2) produced during respiration by roots and microorganisms.

Strictly speaking the pH of a soil should refer to the $[H^+]$ in the soil solution. This is very difficult to measure, and hence soil pH is often referred to measurements made in the soil suspension. Three common, standard methods for determining pH in soils involve the suspension in distilled water, $0.01\,mol\,l^{-1}$ $CaCl_2$, or $1\,mol\,l^{-1}$ KCl solutions at a soil-to-solution (weight-to-volume) ratio of 1:1 (in USA) or 1:2.5 (in New Zealand and Australia).

The pH of a solution in equilibrium with a soil varies with the composition and concentration of the salts in the solution. This is because cations in solution displace H^+ and Al^{3+} ions from soil surfaces. The measure of soil pH in water extract is expected to give a higher value than the pH of the soil solution at natural soil water contents. The excess quantity of water added to prepare the suspension results in a dilution of the salts in the soil solution. Since soils always contain some soluble salts, such as nitrates and chlorides, the apparent pH of the soil will depend on the amount of water added to the soil to make the suspension. Further, the amount of soluble salts in a soil varies continuously throughout the season, depending on factors such as the amount of drainage water percolating through the soil.

Soil pH values measured in $0.01\,mol\,l^{-1}$ $CaCl_2$ or $1\,mol\,l^{-1}$ KCl salt solutions are generally less than the values obtained using water, because the cations in these salts tend to displace H^+ and Al^{3+} from soil materials to soil solution. These salt solutions are sufficiently concentrated to overwhelm variations in salt concentration in soil solution, thereby minimizing the effect of salt concentration on pH in most soils. These solutions also eliminate the junction potential effect. Calcium is used in the determination of pH because it is the predominant soil solution cation in soils of temperate regions. The $0.01\,mol\,l^{-1}$ concentration, however, is somewhat higher than that found in most soil solutions, and thus the pH values obtained by this method are usually lower than the values of soil solutions.

For some subsurface horizons of highly weathered soils of the tropics, pH in salt solution can actually be greater than pH in water, especially in subsoils that are low in permanent-charge silicate clays and low in organic matter. Ion exchange in these soils is dominated by oxides and hydroxides of Fe and Al, and at acid pH values they are net positively charged, creating an effective sink for anion adsorption. Thus, on salt addition (i.e., $CaCl_2$ or KCl) more OH^- is

displaced by Cl^- than H^+ by cations, resulting in an increase in pH.

Factors Affecting Soil pH

The pH of a soil is affected by the concentration of CO_2 in the soil air, salt concentration (salt effect), and the presence of colloidal particles (suspension effect).

Carbon Dioxide

The higher the CO_2 concentration in the soil solution, the lower the pH, and the pH of a neutral or calcareous soil is very sensitive to small changes in CO_2 concentration. In calcareous soils, its magnitude can be calculated from the solubility of calcium carbonate and CO_2 in water, because in such soils the pH is basically determined by the system, calcium carbonate–carbon dioxide–water, which is given approximately by:

$$2\text{pH} = K + \text{pCa} + \text{pCO}_2 \qquad [18]$$

where pCO_2 is the negative logarithm of the partial pressure of CO_2 in equilibrium with the solution, pCa the negative logarithm of the activity of the calcium ions, and K is a constant whose value lies between 10 and 10.5, depending on the value used for the solubility constant of calcium carbonate. The decrease in pH is therefore approximately proportional to the logarithm of the partial pressure of CO_2.

Thus, the pH of a well-aerated calcareous soil containing exchangeable sodium or magnesium ions cannot exceed approximately 8.5 when in equilibrium with the atmosphere; but if the CO_2 concentration rises to 0.1% (i.e., a low value for most soils) the pH drops to 8, and if it rises to 1% (a common value for pasture soils) the pH will be less than 7.5. Many plants suffer from phosphate and minor element deficiencies if the pH rises much above 8, and hence they can only thrive on calcareous soils if they can maintain an adequate concentration of CO_2 around their roots. This is more easily achieved in soils of low porosity, such as clays, than in soils of high porosity, such as well-drained sands, and in organic matter-rich pasture soils than in organic matter-depleted arable soils.

Salt Effect

As discussed above, the pH of a soil depends on the salt concentration in the soil solution. This dependence of measured soil pH on salt concentration can be reduced by making all the measurements in a salt solution that is sufficiently strong to swamp the effects of the changes that occur naturally. But adding a salt solution to the soil will cause cation exchange to take place and, in particular, exchange of H^+ ions from clay surface with the cation of the added salt, which will give a pH that is lower than the soil pH under normal conditions. The added salt solution should therefore cause as little exchange as possible, and the use of $0.01\,\text{mol}\,l^{-1}$ $CaCl_2$ or $1\,\text{mol}\,l^{-1}$ KCl solution has been proposed for most arable soils on the grounds that they approximate to the salt concentration in the soil solution and do not give appreciable junction potential with the calomel electrode. The pH of a soil measured in these solutions is more constant and is much less dependent on the soil-to-solution ratio than the pH measured in a water suspension, and it is also presumably closer to the pH of the solution around the plant roots.

Suspension Effect

The difference in pH values between the suspension and its equilibrium solution is referred to as the 'suspension effect,' and this effect is identical to the Donnan e.m.f.

The electrode portion of a pH measurement circuit is presented in Figure 3, where each vertical bar represents a phase boundary and the double bar represents the liquid–liquid junction. Thus, very appreciable junction potentials can be created between a soil suspension and the calomel electrode so that the measured potential difference between the two electrodes cannot be interpreted solely in terms of the $[H^+]$ outside the glass electrode.

A junction or diffusion potential always develops when two dissimilar substances or solutions of different composition come in contact. It is inevitable that a diffusion potential is created at the liquid junction between a salt bridge and a solution of different composition. An example in this case is the saturated KCl of the reference electrode diffusing into the test solution. At liquid junctions, differences in the mobilities of the cation and anion create a diffusion potential which is also influenced by any differences in the ion activities that exist across the junction. The potential is minimal when the rates of K^+ and Cl^- diffusion into the test solution are equal. It is therefore appropriate that an electrolyte in which the mobilities of the cation and anion are almost identical is used in the salt bridge. Indeed, KCl is used as the reference-cell electrolyte because of the similarity of K^+ and Cl^- diffusion rates in water. Because the current at the liquid junction is mainly carried by K^+ and Cl^- ions, which have similar diffusion coefficients, the liquid

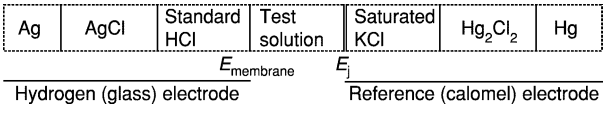


Figure 3 The electrode portion of a pH measurement circuit.

junction potential should be small. As long as the ionic strengths of the standard and unknown solutions are small compared with the concentrated KCl salt bridge solution, any liquid junction potential should, as well as being small, tend to be rather constant in magnitude and hence can usually be ignored.

Because factors involved in the liquid junction potential during pH calibration and subsequent pH measurement are likely to be different, the liquid junction potential contribution to the total e.m.f. of a cell will also be different. In a colloidal suspension, however, the colloid may cause K^+ and Cl^- to diffuse at different rates. Because of attraction and repulsion by the charged colloid, one ion moves ahead of the other. When a salt bridge makes contact through a suspension instead of clear solution, the situation is complicated by a reduced mobility of counter ions which arises from retarding effects created by the oppositely charged colloidal particles. This means that liquid junction potentials (E_j) are likely to be much greater in suspensions than in supernatant liquids. Accurate pH measurements require a negligible E_j, because such potentials are virtually unpredictable. The potentiometer measures all the potentials of the circuit and cannot distinguish between the H^+ ion potential at the glass membrane and the spurious E_j. The value of E_j at the interface between saturated KCl and a colloidal suspension of very low salt concentration can be as high as 240 mV, equivalent to more than 4 pH units (1 pH unit = 59 mV). However, values of E_j greater than 30 mV are uncommon in soils, because salt concentrations in even highly leached soils are high enough to suppress E_j. Therefore measuring pH in salt solutions of 0.01 mol l^{-1} or more virtually eliminates E_j.

Suspension effects can be reduced by introducing the liquid junction of the reference electrode into clear supernatant liquid and not into the suspension. The rates of K^+ and Cl^- diffusion are then unaffected by soil colloid particles. The pH glass electrode, on the other hand, can be placed either in the supernatant solution or in the colloidal suspension. The H^+ activity is the same in both phases, and the glass electrode is unaffected by the presence of the colloid. However, it has been suggested that when measuring soil pH the glass electrode should be in the suspension on the basis that the suspension is better buffered than the supernatant liquid.

Importance of Soil pH

The soil pH is considered as the 'master variable' which controls many other soil properties (Figure 1). pH greatly affects the solubility of minerals in soils. Most minerals are more soluble in acid soils than in neutral or slightly basic solutions. A low pH may indicate the deficiency of basic cations such as Ca^{2+} and Mg^{2+} in soils. Under acid soil conditions, the concentration of most heavy metals, such as lead, cadmium, and mercury, reach phytotoxic levels. Strongly acidic soils usually have high and toxic concentrations of soluble aluminum and manganese. Azaleas, tea, rhododendrons, cranberries, pineapple, blueberries, and several conifer timber species tolerate acidity and grow well in acid soils. In contrast, alfalfa, beans, barley, and sugarbeets grow well only in slightly acidic to moderately basic soils because of a high calcium demand or inability to tolerate soluble aluminum.

Soil pH can also influence plant growth through its effect of beneficial microorganisms. For example, most nitrogen-fixing legume bacteria are not very active in strongly acidic soils. Bacteria that decompose soil organic matter and thus release nitrogen and other nutrients for plant use are also hindered by strong acidity. Similarly earthworms are not active in acid soils, whereas fungi usually tolerate acidity better than do other microbes.

Soil basicity, although more difficult to manage than soil acidity, may also be undesirable for plants. Alkaline soils may reach pH values of more than 10 with increasing levels of exchangeable sodium. Plants on soils with pH >9 usually have reduced growth or even die. Some plants (halophytes) are tolerant of high salt concentration or pH. Soluble $Al(OH)_4^-$ may exist at these high pH values and is also toxic to plants. The major effect of basic pH is to reduce the solubility of all micronutrients (except chlorine, boron, and molybdenum), especially those of iron, zinc, copper, and manganese. Also, phosphate is often not readily available to some plants because of its precipitation in the soil solution by calcium or precipitation on solid calcium carbonate. Iron deficiency, associated with soils high in carbonates, has long been known, and is referred to as 'lime-induced iron chlorosis.'

Summary

Soil pH is one of the most important chemical properties which control many other soil physical, chemical, and biological properties. The pH of a soil depends on the CO_2 concentration in the soil air and the salt concentration in the soil solution, and these are constantly changing.

Further Reading

Arnold PW (1978) Surface–electrolyte interactions. In: Greenland DJ and Hayes MHB (eds) *The Chemistry of Soil Constituents*, pp. 335–404. New York: John Wiley.

Bloom PR (2000) Soil pH and pH buffering. In: Sumner ME (ed.) *Handbook of Soil Science*, pp. B333–352. New York: CRC Press.
Bohn H, McNeal BL, and Connor GA (1979) *Soil Chemistry*. New York: John Wiley.
Brady NC and Weil RR (1999) *Nature and Properties of Soils*. Englewood Cliffs, NJ: Prentice-Hall.
Hausenbuiller RL (1978) *Soil Science: Principles and Practices*, 2nd edn. Dubuque, Iowa: WC Brown.
Miller RW, Donahue RL, and Miller JU (1990) *Soils: An Introduction to Soils and Plant Growth*, 6th edn. New York: Prentice-Hall.
Russel EW (1973) *Soil Conditions and Plant Growth*, 10th edn. London, UK: Longman.
Sorenson SPL (1909) Enzyme studies. II. The measurement and importance of the hydrogen ion concentration in enzyme reactions. *Comptes Rendu du Travail Laboratoire de Carlsberg* 8: 1.
Sparks DL (1995) *Environmental Soil Chemistry*. New York: Academic Press.
Tan KH (1998) *Principles of Soil Chemistry*. New York: Marcel Dekker.
Wolt JD (1994) *Soil Solution Chemistry, Application to Environmental Science and Agriculture*. New York: John Wiley.

PHOSPHORUS IN SOILS

Contents
Overview
Biological Interactions

Overview

J T Sims, University of Delaware, Newark, DE, USA
P A Vadas, USDA Agricultural Research Service, University Park, PA, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Phosphorus (P) is essential for growth and development of living organisms. In agricultural systems, P inputs to soils are often required to replenish P removed by harvested crops or lost by soil erosion or runoff. Because soils bind P strongly, historically P inputs have been made more to maintain adequate or high levels of soil fertility and less from a concern that excessive P inputs promote P loss in runoff and erosion. However, all soils have a finite capacity to bind P. When this capacity is diminished through repeated P applications in excess of plant needs, soil P mobility and loss to the environment increase. Even when soil P fertility is relatively low, P loss by excessive erosion can be an environmental concern. The concern for P loss is its delivery to surface waters, where P inputs can increase the growth of algae and aquatic plants to undesirable levels. The result is often eutrophication, whose negative ecological and economic effects include anaerobic conditions during decomposition of algal biomass and subsequent deleterious effects on aquatic organisms, increased cost and complexity of water treatment for drinking quality, and impaired recreational water use. Thus, the challenge for soil P management is to maintain adequate soil P fertility for economical crop production and minimize P loss to the environment.

The Soil Phosphorus Cycle

Major Forms of Phosphorus in Soils

Total soil P ranges from 0.01 to 0.30% and occurs in three forms: P dissolved in the soil solution (solution P), organic P, and inorganic P (Figure 1). Solution P exists largely as orthophosphate ions (HPO_4^{-2}, $H_2PO_4^{-1}$), and can vary from 0.01 to 3.0 mg l^{-1}. Some organic P compounds are also present in solution. Organic P comprises between 30 and 65% of total soil P, depending on soil type, and includes phosphate esters, nucleic acids, phospholipids, and other esters. Inorganic P occurs as primary minerals (those derived directly from weathered parent material) and secondary minerals (those formed by precipitation of P with Al, Ca, and Fe), P adsorbed on to the surface of clay minerals, Fe, and Al oxyhydroxides, or Ca carbonates, and P physically occluded within secondary minerals.

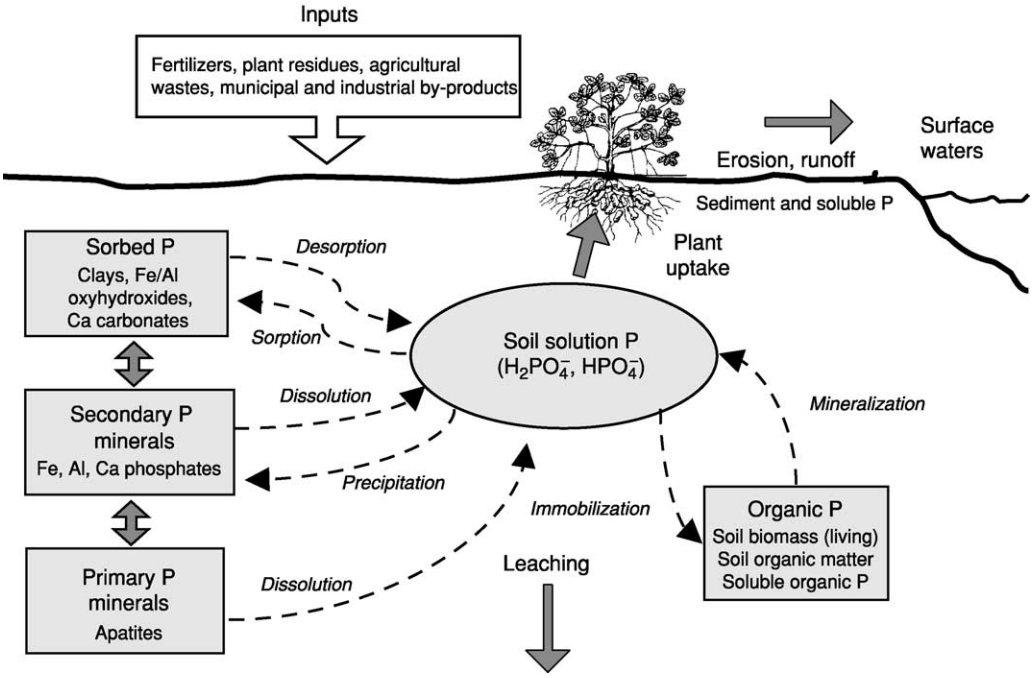


Figure 1 The soil P cycle. Reproduced with permission from Pierzynski GM (2000) Soils and Environmental Quality. Lewis Publications. CRC imprint.

Overview of Soil Phosphorus Cycling and Transformations

Solution P concentrations are largely controlled by mineralization and immobilization of organic P and adsorption and/or desorption and precipitation and/or dissolution of inorganic P. Mineralization is the release of organic P to solution, and immobilization is the conversion of solution P to organic P. Soil microbes require P for activity and growth. If organic materials that they oxidize for energy contain less P than required, microbes immobilize P from solution. If organic materials contain more than enough P, microbes mineralize P to solution. Factors that influence P mineralization and immobilization include the quantity of organic P, the C-to-P (C:P) and N-to-P (N:P) ratios of organic materials, temperature, moisture, aeration, pH, cultivation intensity, and P fertilization.

Adsorption is the chemical bonding of P to mineral surfaces, and desorption is release of P from these surfaces. Positively charged surfaces of clays and Fe, Al, and Ca minerals attract orthophosphate ions (e.g., $H_2PO_4^-$, HPO_4^{-2}) from solution, which adsorb to the surface by exchanging with OH^- and OH_2 groups. When P is bonded through one metal–O–P bond, it is considered labile and can be readily desorbed from the surface to solution. With two metal–O–P bonds, a stable six-member ring forms, desorption is more difficult, and P is considered nonlabile.

Precipitation and dissolution reactions are controlled by the solubility of secondary P minerals, the most common of which are Al–P, Fe–P, and Ca–P. A mineral's solubility represents the concentrations of its constituent ions that can be maintained in solution at equilibrium. When the concentration of solution P is greater than that supported by a mineral's chemical solubility, solution P will precipitate with metal cations (e.g., Al, Ca, Fe) and form P minerals. When the solution P concentration is less than a mineral's chemical solubility, P minerals dissolve into solution. Soil physical and chemical properties influencing P adsorption and/or desorption and precipitation and/or dissolution reactions include the nature and amount of soil minerals, soil pH, cation and anion effects, extent of soil P saturation, organic matter, reaction time and temperature, soil oxidation–reduction status, and P fertilization practices.

Phosphorus Additions to Soils

Because plants absorb P from the soil solution, solution P must be maintained at adequate concentrations, commonly reported as $0.2\,mg\,P\,l^{-1}$ but which may be as low as $0.03\,mg\,P\,l^{-1}$. Soils with relatively high concentrations of labile P can provide all P required by plants. Conversely, soils with little labile P require P additions to achieve optimum crop yields. When fertilizer P (inorganic or organic) is added to soil, solution P increases due to dissolution of inorganic P or mineralization of organic P. This increase disturbs the equilibrium between the solid and solution phases, and P moves from solution to

the solid phase, mostly through precipitation and adsorption. These initial reactions are rapid and leave the P in labile forms that are easily converted back to solution through dissolution and desorption. Over time, some P in the solid phase stabilizes and becomes nonlabile, so that movement back to solution is more difficult or even impossible. As plants absorb P from the soil solution or soluble P is lost in runoff, labile P dissolves or desorbs from the solid phase to solution.

If P additions consistently exceed plant uptake over time, concentrations of soil P increase. Eventually, labile P concentrations may be great enough to meet plant needs without P fertilization. The greater the concentration of labile P, the longer the soil will be able to maintain plant growth. When such soils are left unfertilized, P fertility levels will progressively decrease, through crop production or P loss, to the point when they no longer adequately supply plant P needs. New additions of P are then necessary.

Phosphorus Losses from Soils

Phosphorus transfer from soils to aquatic environments is a major nonpoint-source pollution concern. The major pathway for P loss is surface runoff and soil erosion, although P can also be transported by subsurface water flow to nearby streams, especially if artificial drainage is present (e.g., tiles, drainage ditches). As rainfall or irrigation water interacts with the surface soil, P desorbs or dissolves from soil particles, crop residues, fertilizers, or manure. Dissolved P can leave a field in surface runoff or be resorbed by soil or runoff sediment and then move as particulate P. The eroded sediment load in surface runoff determines particulate P loss. In areas where soils are vulnerable to wind erosion, soil P can also be transported with fine, wind-blown sediments and deposited in nearby waters. Years of extensive research show that P loss is minimal when management practices prevent surface accumulation of P and minimize surface runoff and erosion.

Because soil P is relatively immobile, most fertilizer P remains near the soil surface. However, leaching of P from the surface soil into the subsoil may occur in sandy soils and organic soils, and may be greater in soils where organic fertilizer is added than where inorganic fertilizer is added (Figure 2). Leaching usually occurs in soils where P additions have increased soil P to excessive concentrations. In such soils, P is more mobile because sorption sites have become increasingly saturated and solution P concentrations have increased. Leaching of P can be exacerbated by preferential water flow where soil water containing relatively high concentrations of P from surface layers flows rapidly through soils in cracks, fissures, biopores, or finger flow. Phosphorus that has leached from surface soils may be retained in subsoils, gradually leach further into groundwaters, or move by subsurface flow to nearby streams.

Soil Phosphorus Management

Soil Testing for Phosphorus

Quantifying the ability of a soil to provide plant-available P is usually the first step in P fertility management. Beginning in the 1940s, agronomic soil tests using chemical extractants were developed to estimate rapidly and economically the ability of a soil to supply P to plants over a growing season. Use of the four most common tests, Bray and Kurtz P-1, Mehlich 1, Mehlich 3, and Olsen P, is a well-established practice in the USA. The Bray and Kurtz P-1 test is an acid extractant used in the midwestern and north central USA for acid and neutral soils. The Mehlich 1 and Mehlich 3 tests are acid extractants, with Mehlich 1 used in southeastern and mid-Atlantic states for acid soils and Mehlich 3 used throughout the USA and Canada for both acid and basic soils. The Olsen P test is a basic extractant used in the north central and western USA for calcareous soils. Some other soil P tests used in the USA and Europe include ammonium bicarbonate-DTPA, Morgan and modified Morgan, ammonium lactate–acetic acid, the P_i soil test, and water and dilute salt tests. Soil testing is a reliable method to determine soil P fertility levels and fertilizer P requirements, and can identify soils where P has accumulated to environmentally unacceptable levels.

Characteristics of Phosphorus Fertilizer Sources

An effective P fertilizer must contain an adequate amount of available P so P will be released quickly and sufficiently enough to replenish soil P reserves (Table 1). Mined rock phosphate minerals, or apatites, are the only important raw material for inorganic P fertilizers. Although they contain 12–18% total P, very little of the P is water-soluble, which limits their fertilizer value in certain settings, such as highly acid soils. More common P fertilizers are produced when rock phosphate is acid-treated or heat-treated to break apatite bonds and increase the water-soluble P content. Phosphorus can also be applied to soils in organic fertilizers, including animal wastes, municipal wastes, composts, plant residues, and occasionally industrial by-products (Table 1). Organic fertilizers generally contain relatively low concentrations of P, ranging from 0.1 to 3.0%.

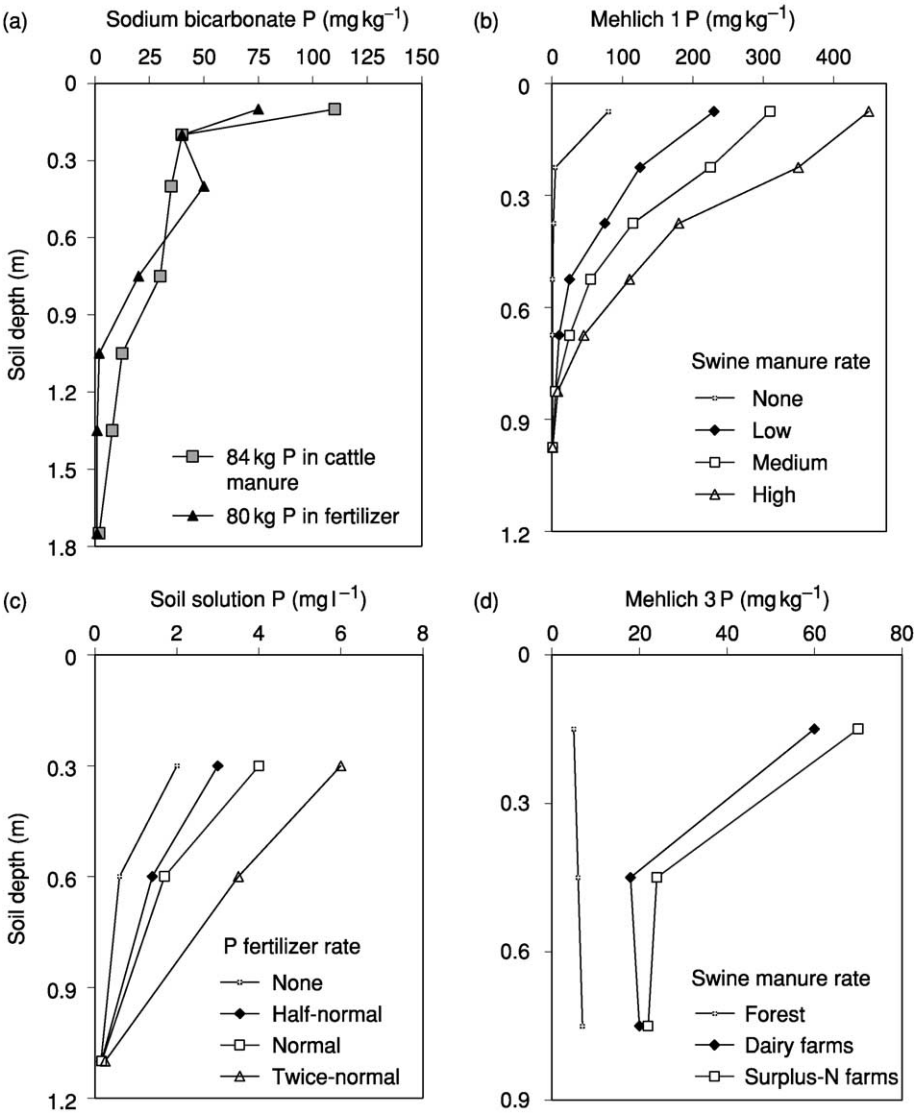


Figure 2 Examples of P-leaching for (a) soils where organic fertilizers and inorganic fertilizers are applied, (b) sandy soils, (c) organic soils, and (d) soils that have excessive concentrations of P. (Source: Eghball B, Binford GD, and Baltensperger DD (1996) Phosphorus movement and adsorption in a soil receiving long-term manure and fertilizer application. *Journal of Environmental Quality* 25: 1339–1343; Hortenstine CC and Forbes RB (1972) Concentrations of nitrogen, phosphorus, potassium, and total soluble salts in soil solution samples from fertilized and unfertilized Histosols. *Journal of Environmental Quality* 1: 446–449; King LD, Burns JC, and Westerman PW (1990) Long-term swine lagoon effluent applications on ''coastal'' Bermuda grass. II. Effect on nutrient accumulation in soil. *Journal of Environmental Quality* 19: 756–760; Simard RR, Cluis D, Gangbazo G, and Beauchemin S (1995) Phosphorus status of forest and agricultural soils from a watershed of high animal density. *Journal of Environmental Quality* 24: 1010–1017.)

Nutrient Management Practices for Phosphorus

The fundamental goal of agricultural P management is to identify the most profitable and environmentally sound approach to provide crops with P, to prevent P accumulation in soil, and minimize P loss from soil. Effective P management begins with a nutrient budget that calculates P inputs to and P outputs from a farm. When P inputs exceed outputs over time, the usual result is an accumulation of soil P and an increased risk of P loss to water. If P outputs exceed inputs, soil P fertility declines. A positive imbalance (P surplus) is economically undesirable and represents a waste of a limited natural resource in phosphate ores. A P deficit indicates that soil productivity is being slowly degraded and crop production will probably decline. Therefore, balancing farm-scale P inputs and outputs is critical to long-term P fertility management.

On farms where commercial fertilizers are the only P inputs, soil tests and realistic yield goals will assess soil P fertility, crop P needs, and P fertilizer requirements. Analyzing harvested crops will quantify P leaving the farm. In this situation, a P surplus

indicates unnecessary fertilization and a P deficit suggests farmers face economic or logistical limitations to maintaining soil P fertility. In animal-based operations where P inputs also include animal feed, soil testing and yield goals will still identify soil P fertility, crop P needs, and P fertilizer requirements; and a manure analysis will identify available P that can be used as fertilizer. Manure application rates should be based on crop nutrient requirements and manure nutrient concentrations. Crop and animal product nutrient analyses or estimations will quantify farm P outputs. Because animal-based agriculture can produce large quantities of manure and often relies exclusively on land application as the end use for manure, it is especially important that P inputs do not exceed P outputs to prevent the buildup of soil P to concentrations of environmental concern (Table 2). If P inputs significantly exceed P outputs, then management options that reduce P inputs, use on-farm P resources without promoting P losses, and increase P outputs can be investigated. Examples are reducing unnecessary P additions to animal feeds, increasing availability of P in feed with dietary amendments such as phytase, or exporting manures for alternative uses such as compost, fertilizer products, or bioenergy production.

Crop response to P additions in either fertilizers or manures is influenced by the rate, timing, and method of P application. At low rates of P fertilization, the water solubility of fertilizer P should be high to get the full benefit of the limited amount applied. In soils that sorb P strongly and quickly, higher rates of P fertilizer may be necessary to achieve the solution P concentration necessary for crop growth. In such soils, it may be more efficient to have smaller, more frequent P additions closer to the times of high crop P demand. If P fertilizers are broadcast and incorporated, much of the added P interacts with soil particles capable of P sorption. If soils are low in P or have high P sorption capacity, high rates of fertilizer P must be broadcast to increase P availability to desired values. For row crops, placing fertilizer P in concentrated bands immediately adjacent to the crop reduces the contact between soil and fertilizer P, allowing for use of lower fertilizer rates and enhanced P uptake by crops.

Loss of P from soils is also influenced by the rate, timing, and method of fertilizer or manure application. The greatest P loss occurs from fields where high rates of P are applied, where P is broadcast and

Table 1 Major inorganic and organic sources of phosphorus for crop production

Material	*P_2O_5 (%)*	*Total P available (%)*
Inorganic fertilizers		
Rock phosphate	25–40	14–65
Single superphosphate	16–22	97–100
Wet-process phosphoric acid	48–53	100
Triple superphosphate	44–53	97–100
Ammonium phosphates		
Monoammonium phosphate	48–62	100
Diammonium phosphate	46–53	100
Ammonium polyphosphate	35–62	100
Urea–ammonium phosphate	16–42	100
Nitric phosphates	14–28	80–100
Ammoniated normal superphosphate	14–21	97–100
Ammoniated triple superphosphate	44–53	96–100
Potassium phosphates		
Monopotassium phosphate	51	100
Dipotassium phosphate	41	100
Potassium polyphosphate	51	100
Organic fertilizers		
Beef manure	2.1	–
Dairy manure	1.4	–
Poultry manure	4.1	–
Swine manure	3.5	–
Aerobically digested sludge	7.6	–
Anaerobically digested sludge	8.3	–
Composted sludge	3.0	–

Adapted from Pierzynski GM, Sims JT, and Vance GF (1994) *Soils and Environmental Quality*. Boca Raton, FL: Lewis Publishers; Tisdale SL, Nelson WL, Beaton JD, and Havlin JL (1993) *Soil Fertility and Fertilizers*. New York, NY: Macmillan.

Table 2 Phosphorus budget for a typical poultry farm in Delaware, USA, and a typical dairy farm in New York, USA

	P input				*P output*				
	Chicks (kg)	*Feed (kg)*	*Fertilizer (kg)*	*Total (kg ha^{-1})*	*Birds (kg)*	*Crops (kg)*	*Total (kg ha^{-1})*	*P balance (kg ha^{-1})*	*P input in surplus (%)*
Poultry farm	90	9300	0		2500	1300		74	58
Dairy farm				14.5			8.4	6.1	42

Source: Sims JT and Vadas PA (1997) *Nutrient Management Planning for Poultry-Grain Agriculture*. Fact Sheet ST-11. College of Agricultural Sciences and Cooperative Extension. Newark, DE: University of Delaware.
Source: Rotz CA, Sharpley AN, Satter LD, Gburek WJ, and Sanderson MA (2002) Production and feeding strategies for phosphorus management on dairy farms. *Journal of Dairy Science* 85: 3142–3153.

unincorporated, and when runoff events occur soon after P application. The least P will be lost from fields where little P is applied, where P is incorporated into soil, where good soil conservation practices reduce erosion, and when runoff events occur well after application. Because these optimum conditions reduce the number of manure application opportunities available throughout the year, manure should be stored in suitable facilities until it can be applied.

Soil and Water Conservation Practices for Phosphorus

An integral component of sound P management is practices that conserve soil and manage water, thus reducing the risk of P loss. To minimize P loss via surface runoff and erosion, it is necessary to reduce rainfall impact on the soil, reduce surface runoff volume and velocity, and increase soil resistance to erosion.

Plant Cover, Crop Rotation, and Strip Cropping

Rainfall breaks up soil aggregates, allowing soil to move with runoff, and compacts and puddles the soil surface, decreasing water infiltration and increasing surface runoff. Crops can intercept up to 45% of rainfall, dissipating raindrop energy and reducing runoff and erosion. After planting, rapid crop canopy helps minimize soil exposure. Adequate soil fertility, optimum seedbed preparation and planting conditions, pest control, and optimum plant spacing all help promote seed germination, plant growth, and canopy closure. Crop rotations where new crops quickly follow preceding harvests minimize the duration of soil exposure. Rotations also help break insect, disease, or weed cycles, which impede plant growth and reduce plant cover. Strip cropping divides fields into long, narrow segments that help control runoff and erosion while growing crops in rotation. More heavily vegetated strips slow runoff and catch soil eroded from more exposed strips. Strips planted on the contour of slopes (contour strip-cropping) are especially effective in controlling runoff and erosion (Figure 3).

Companion Crops, Cover Crops, Green Manures, and Crop Residues

With companion crops such as small-grain and forage crops, the small grain starts first and provides quick plant cover. Harvesting the grain leaves residue and the forage to protect the soil. Cover crops grown during cold or dry seasons can provide soil cover when soils would otherwise be left exposed. Cover crops should germinate easily, grow quickly, provide sufficient plant cover, and be hardy against weather and pests to maximize soil coverage. Green manures act as cover crops and, when plowed under, can add nutrients to soil to increase fertility for succeeding crop growth and build soil organic matter content, which promotes the soil structure, permeability, and aeration needed to maximize rain infiltration and minimize surface runoff and erosion. Crop residues also act as mulch to increase soil coverage. Although plowing residues under may decrease soil coverage, it can promote residue decomposition and erosion-control benefits such as those from green manures.

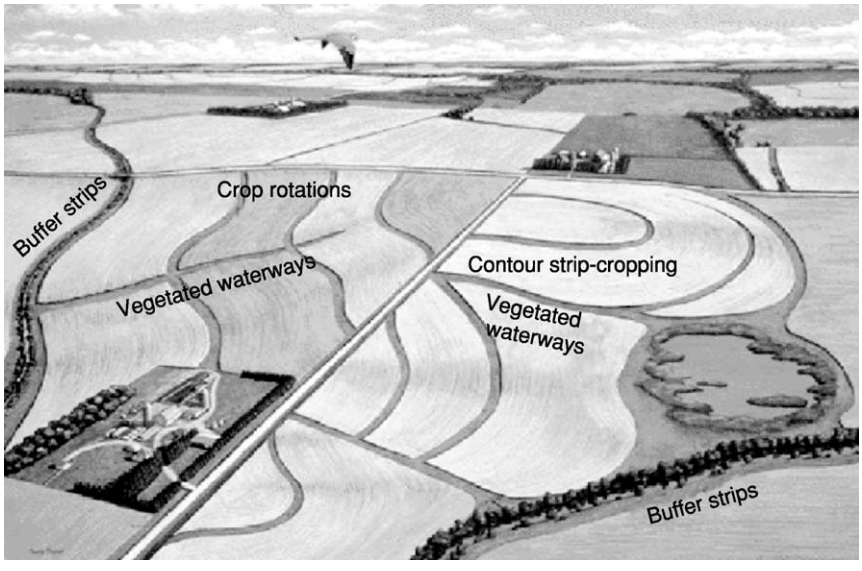


Figure 3 Soil and water conservation practices that can help reduce P export from agricultural fields. Courtesy of USDA Natural Resources Conservation Service.

Conservation Tillage, Contour Cultivation, and Deep Plowing

Conservation tillage, such as stubble mulch tillage, minimum tillage, reduced tillage, and no-tillage, leaves at least 30% of the soil surface covered by plant residue after planting to reduce erosion and surface runoff. Contour cultivation follows the contour lines of sloping fields. Contour ridges produced by tillage and planting form barriers that slow downhill water movement and decrease surface runoff and erosion. In applicable situations, deep plowing can break up pans to increase water infiltration and reduce runoff and erosion. It can also help dilute topsoil layers with high P concentrations into subsoils with less P.

Although reducing tillage will help decrease total P loss in surface runoff by reducing erosion, it may increase dissolved P loss. Because soils are not tilled, P from added fertilizers or manures accumulates at the surface, increasing the concentration of dissolved P in surface runoff. Also, in no-till situations where manures and fertilizers are left unincorporated, dissolved P loss in surface runoff soon after application can increase dramatically.

Terraces, Vegetated Waterways, and Buffer Strips

Terraces divide slopes so that surface runoff is intercepted and carried to a protective outlet. Terraces decrease erosion by shortening slope length, slowing runoff velocity, and trapping sediments. Surface runoff water concentrates in waterways that when bare or unstable are extremely erodible. Vegetated waterways are natural or constructed channels that, when properly established and maintained, transport surface runoff water at a nonerosive velocity from fields, prevent gully formation, and greatly decrease erosion. Vegetated buffer strips act primarily to control water pollution by preventing eroded soil from being carried into streams or ditches.

Innovations in Soil Phosphorus Management

The past two decades have seen an increased awareness of the importance of P loss from agricultural soils. Extensive research has been conducted to better understand sources and pathways of P loss, to identify areas or fields where risk of P loss is high, to develop management practices to minimize P loss, and determine the best means by which to implement such practices. A number of innovations in soil P management have evolved from this research.

Animal Feed Formulations

Inorganic P supplements are routinely added to animal feeds to ensure adequate dietary P. Because P is often added in excess of animal needs, significant amounts of P are excreted in manure and applied to soils, where they are a potential source of P loss. Adjusting diets to match P inputs with animal needs, or increasing the availability of native feed P, as with phytase additions to poultry feeds, can potentially reduce the amount of P that is excreted in manure and land applied. This would also help reduce P surpluses that exist in farm-scale nutrient balances (Table 2). However, research must determine whether dietary manipulations will indeed reduce P loss from soils where manure is applied. For example, adding phytase to poultry feeds to increase P availability to birds may increase the solubility of P in manure and thus increase dissolved P loss in surface runoff from fields where manure has been applied.

Soil and Manure Testing

Soils are important sources of P loss. As soil P concentrations increase due to long-term applications of fertilizer and manures, tests are needed to identify soils that pose a significant environmental risk. Good relationships exist between soil P concentrations as measured by traditional agronomic tests and dissolved P loss in surface runoff or subsurface flow. Research has also shown that some new soil tests, such as extractions with water, iron oxide-coated filter paper, and ion-exchange membranes, and a test for the degree of soil P saturation, can also estimate the risk of P loss from soil. Land-applied manures can too be significant sources of P loss. Simple manure tests, such as extractions with water, can estimate the potential for manure P loss. Although soil and manure tests can identify potential sources of P loss, they give no information about the transport of P. For example, manure applications to a given field increase soil P concentrations and thus the potential for P loss in runoff or erosion. However, if runoff from the field is rare, then transport of P from the field is unlikely. Tools such as the Phosphorus Index can help identify fields where the risk of P pollution from both source and transport factors is high.

Soil and Manure Amendments

The loss of P in runoff is related to the P concentration of both soils and manures. Amending soils, manures, or municipal biosolids with materials containing high amounts of Fe, Al, or Ca can transform the soil or manure P into less-available forms and thus reduce P loss in runoff. Such amendment materials include an Fe-rich by-product from paint production, Al-rich drinking water treatment residuals, aluminum sulfate (alum), and coal combustion by-products. While these amendments are effective in reducing P availability in

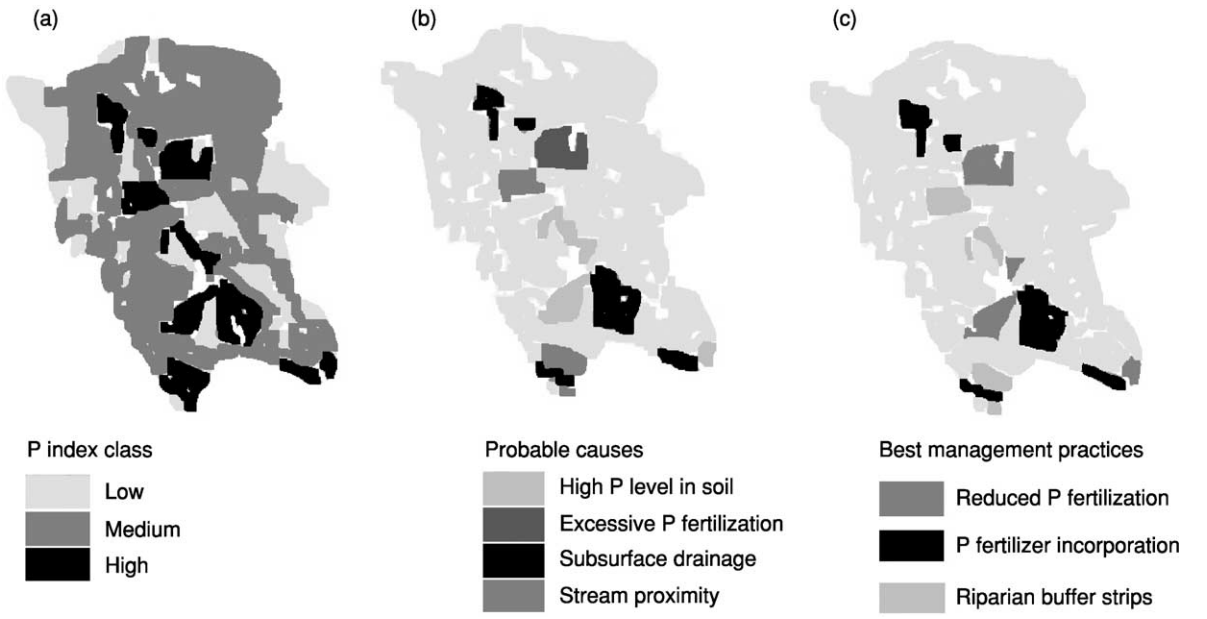


Figure 4 Examples of the P index for a watershed scale, showing (a) P-index ratings, (b) causes of P export, and (c) management practices that can help reduce the P export. (Source: Djodjic F, Montas H, Shirmohammadi A, Bergström L, and Ulén B (2002) A decision support system for phosphorus management at a watershed scale. *Journal of Environmental Quality* 31: 937–945.)

soils and manures, they may also contain high concentrations of salts or heavy metals that could have adverse agronomic or environmental effects.

Water-Quality Modeling and the Phosphorus Index

Water-quality simulation models such as EPIC, FHANTM, GLEAMS, AGNPS, SWAT, or ANSWERS are designed to predict P loss from agricultural areas for a variety of soil, hydrologic, or management conditions and are seen as relatively fast, cost-effective tools. Continued research with these models has led to their improvement to simulate P loss better. However, such models are often complex and require large amounts of input data that may be unknown or difficult to measure. They may also be limited in the number of soil, hydrologic, or management conditions they simulate. The Phosphorus Index (P Index) is a much simpler screening tool, developed by the Natural Resources Conservation Service (NRCS) and recently modified; it accounts for source and transport factors controlling P loss and ranks sites according to the risk of P movement (Figure 4). The P Index is designed to allow field personnel easily to identify agricultural areas or practices that are most vulnerable to P loss, thus allowing producers more flexibility in developing P-control strategies.

Further Reading

Abu-Zreig M, Rudra RP, Whitely HR, Lalonde MN, and Kaushik NK (2003) Phosphorus removal in vegetated filter strips. *Journal of Environmental Quality* 32: 613–619.

Akhtar MS, Richards BK, Medrano PA, deGroot M, and Steenhuis TS (2003) Dissolved phosphorus from undisturbed soil cores: related to adsorption strength, flow rate, or soil structure? *Soil Science Society of America Journal* 67: 458–470.

Andraski BJ, Mueller DH, and Daniel TC (1985) Phosphorus losses in runoff as affected by tillage. *Soil Science Society of America Journal* 49: 1523–1527.

Frossard E, Condron LM, Oberson A, Sinaj S, and Fardeau JC (2000) Processes governing phosphorus availability in temperate soils. *Journal of Environmental Quality* 29: 15–23.

Gburek WJ, Sharpley AN, Heathwaite AL, and Folmar GJ (2000) Phosphorus management at the watershed scale: a modification of the phosphorus index. *Journal of Environmental Quality* 29: 130–144.

Harrison AF (1987) *Soil Organic Phosphorus: A Review of World Literature*. Wallingford, UK: CAB International.

Heckrath G, Brookes PC, Poulton PR, and Goulding KWT (1995) Phosphorus leaching from soils containing different phosphorus concentrations in the Broadbalk experiment. *Journal of Environmental Quality* 24: 904–910.

Kleinman PJA, Sharpley AN, Wolf AM, Beegle DB, and Moore PA Jr (2002) Measuring water-extractable phosphorous in manure as an indicator of phosphorus in runoff. *Soil Science Society of America Journal* 66: 2009–2015.

Moore PA Jr, Daniel TC, and Edwards DR (2000) Reducing phosphorus runoff and inhibiting ammonia loss from poultry manure with aluminum sulfate. *Journal of Environmental Quality* 29: 37–49.

Pote DH, Daniel TC, Nichols DJ, Sharpley AN, Moore PA Jr, Miller DM, and Edwards DR (1999) Relationship between phosphorus levels in three ultisols and phosphorus concentrations in runoff. *Journal of Environmental Quality* 28: 170–175.

Sharpley AN, Chapra SC, Wedepohl R, Sims JT, Daniel TC, and Reddy KR (1994) Managing agricultural phosphorus for protection of surface waters: issues and options. *Journal of Environmental Quality* 24: 437–451.

Sharpley AN, Kleinman PJA, McDowell RW, Gitau M, and Bryant RB (2002) Modeling phosphorus transport in agricultural watersheds: processes and possibilities. *Journal of Soil and Water Conservation* 57: 425–439.

Sims JT, Edwards AC, Schoumans OF, and Simard RR (2000) Integrating soil phosphorus testing into environmentally based agricultural management practices. *Journal of Environmental Quality* 29: 60–71.

Vadas PA, Kleinman PJA, and Sharpley AN (2004) A simple method to predict dissolved phosphorus in runoff from surface applied manures. *Journal of Environmental Quality* (in press).

Biological Interactions

M D Mullen, University of Kentucky, Lexington, KY, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Phosphorus (P) is essential to life; it is a primary component of many important biomolecules such as adenosine triphosphate (ATP), deoxyribonucleic acid (DNA), phospholipids, the phosphopyridine nucleotides ($NADP^+$), and phytin (P storage molecule – inositol hexaphosphate). Plant requirements for P are, on average, between 0.2 and 0.4% of plant dry weight. While this is relatively modest relative to nitrogen (N) requirements, the challenge with P lies in maintaining enough soluble phosphate to prevent deficiencies of P in agricultural crops. Therefore the management of P in agriculture is of major interest.

Phosphorus enters the soil environment in a number of ways. It is present, in varying concentrations, as the primary mineral apatite from which the plant-available phosphate ion is weathered. Phosphorus is applied as inorganic fertilizers or organic materials such as animal manures and other biosolids, or is simply recycled from plant and animal materials in undisturbed soil ecosystems.

The P cycle in soils has both abiotic and biological components. The abiotic component is primarily concerned with the dissolution of P-containing minerals, the formation of phosphates of Al, Fe, Ca, and Mg, and the sorption of phosphates to clays and Al- and Fe-oxides in soil. The biological component is comprised primarily of mineralization of organic P sources and immobilization of soil solution orthophosphate during microbial metabolism. Soil organisms also produce a variety of acids and chelating agents which facilitate the solubilization of phosphate from the inorganic P in the soil. The organisms involved include bacteria and fungi living in the bulk soil and the rhizosphere, as well as mycorrhizal fungi which supply P to the plant. The biological transformations of P in the soil are very important in the availability of P to plants. However, mismanagement or overapplication of P-containing materials can result in excessive amounts of inorganic P, which can have negative environmental impacts. The introduction of excessive P into surface waters can result in eutrophication, or excessive algal productivity, which can ultimately reduce oxygen content in the water, degrading the quality of the habitat for aquatic organisms.

Forms of Soil Phosphorus

The original source of most soil P is from the primary mineral apatite, or rock phosphate (RP). Apatite is a highly insoluble Ca-phosphate compound with the general formula of $Ca_{10}(PO_4)_6X_2$, where X is F^- (fluoroapatite), OH^- (hydroxyapatite), or Cl^- (chloroapatite). In highly weathered, acidic soils, little or none of these apatite minerals remain in the soil. The P in these minerals is released, either chemically or by microorganisms and plant roots, then used by plants and microorganisms. Over time, this P is incorporated into organic matter or other relatively insoluble inorganic P minerals with Al and Fe in acidic soils, and with Ca in neutral to alkaline soils. In highly weathered soils, P is tied up by Al- and Fe-oxides. These compounds are only sparingly soluble (**Figure 1**). Generally, solubility of phosphorus from Al and Fe sources increases with pH levels in the range of about pH 6.5. As pH continues to increase, though, Ca-phosphate minerals are formed and these control soluble P concentrations. Therefore as pH increases above neutrality, soluble P concentrations decrease significantly. The low solubility of inorganic soil P makes the use of P fertilizer materials necessary in many high-yield agricultural systems. Additionally, the practice of liming acid soils to increase pH to about 6.5–7.0 has the practical effect of enhancing P availability for plant growth.

Total amounts of P in soils vary widely, depending on parent material and the extent of weathering. In sandy soils with low organic matter, there may be as little as 100 mg kg^{-1}, while soils high in primary

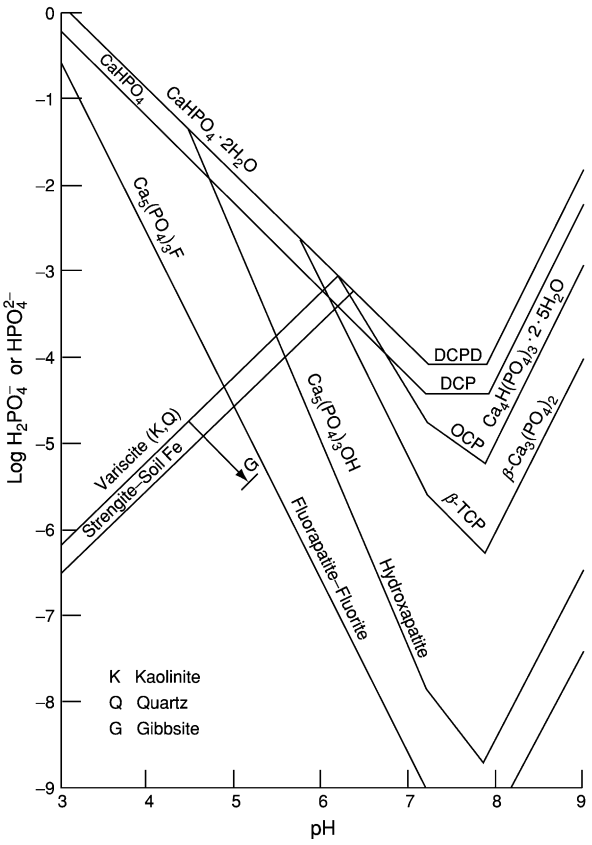


Figure 1 Solubility of orthophosphate from variscite ($AlPO_4 \cdot 2H_2O$), strengite ($FePO_4 \cdot 2H_2O$), and various Ca-phosphates as a function of solution pH. These equilibria assume a Ca^{2+} concentration of $10^{-2.5}$ mol l^{-1} and CO_2 at 0.03 kpa. DCPD, calcium phosphate dehydrate, brushite; DCP, calcium phosphate, monetite; OCP, octacalcium phospate; β-TCP, β-tricalcium phosphate. Reproduced with permission from Lindsay WL (1979) *Chemical Equilibria in Soils*. New York: John Wiley.

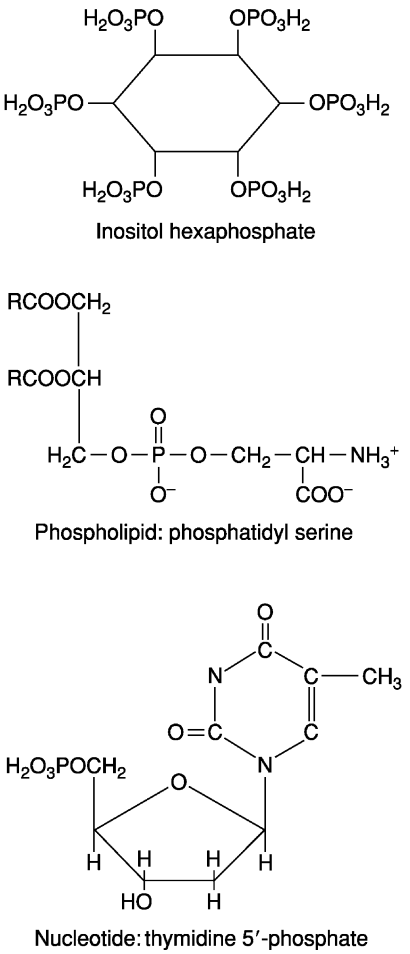


Figure 2 Representative forms of organic phosphorus inputs into the soil environment. Reproduced with permission from Mullen MD (1998) Transformations of other elements. In: Sylvia DM, Fuhrmann JJ, Hartel PG, and Zuberer DA (eds) *Principles and Applications of Soil Microbiology*, pp. 369–386. Upper Saddle River, NJ: Prentice-Hall.

P minerals (e.g., high-phosphate limestone parent materials), clay content, or organic matter may contain well over 2000 mg kg^{-1} soil. In most soils, inorganic P is the dominant form, although organic P increases with increasing soil organic matter.

The plant-available form of P in soils is orthophosphate, $H_2PO_4^-$ (dominant form at pH 2.2–7.2) or HPO_4^{2-} (dominant form at pH 7.2–12.4). Orthophosphate from soluble phosphate fertilizer materials or mineralized from organic sources rapidly forms precipitates as Al-, Fe-, or Ca-phosphates or is sorbed by Al- and Fe-oxides. Due to these reactions, soluble P is present at very low concentrations in soil solution, typically in the range of less than 0.1 to 1 mg l^{-1}. In nonfertilized soils the solution orthophosphate must be continuously replenished by chemical or biological dissolution of inorganic P compounds or by microbial mineralization of organic P.

Organic P can account for 3–90% of the total soil P, with 30–50% common in most soils. The P content of soil organic matter ranges from about 1 to over 3%. Total organic P in a soil is a function of organic matter content, so that as organic matter increases, so does organic P. Organic P also decreases with soil depth, as does organic C. The chemical structure of organic P in soil is not well known. The major types of organic P compounds which are added to soil by recycling plant, animal, and microbial remains in the soil environment are shown in Figure 2. Of the compounds shown, phytins, or inositol phosphates, are usually found in the greatest quantities, making up 10 to 50% of the total organic P content. The inositol phosphates are typically in a polymeric state, making these compounds relatively resistant to decomposition. The phospholipids and nucleic acids may account for up to 5% of the organic P fraction, while other identifiable forms of organic P are present in trace amounts only.

The microbial biomass represents an important part of the organic P pool in soil. This pool represents actively cycling P in the soil, and is part of the labile or readily available organic P. The microbial biomass is responsible for mineralization and immobilization reactions which convert organic P to inorganic P and inorganic P to organically bound P. The microbial biomass is also active in solubilization of relatively insoluble forms of Al-, Fe-, and Ca-phosphates. The biomass P fraction accounts for up to 5% of the total organic P pool in arable soils, while in undisturbed grassland and forest soils, the fraction may be as high as 20% of the total. Studies have shown that soil disturbance, e.g., tillage, reduces organic matter and total organic P content, and that the microbial biomass P fraction decreases more rapidly than the total organic P pool in general. Therefore, the labile organic P fraction is more susceptible to disturbance than the more stable, complex organic P forms. This indicates the importance of the microbial biomass to making P available for plant use in stable terrestrial ecosystems.

The Phosphorus Cycle

In some ways, the phosphorus cycle is not as complex as the nitrogen or sulfur cycles because P does not typically undergo oxidation–reduction reactions or exist in gaseous forms. However, as in the N and S cycles, the biological processes of mineralization and immobilization as well as plant uptake are present. Additionally, the microbial fraction is active in the solubilization of relatively insoluble inorganic P materials, making P available for plant or microbial uptake or other transformations.

Figure 3 shows the P cycle, illustrating the various compartments of P in the soil environment. The P cycle is divided into a geochemical subcycle and a biological subcycle, with the inorganic solution P pool serving as the central point in the overall cycle. This pool of inorganic solution P serves as the source of orthophosphate for plants and soil microorganisms. The remainder of this article will deal with the biological subcycle of P.

In the biological subcycle, soluble orthophosphate is either taken up by plants or immobilized into the microbial biomass. When organic sources of P, such as crop residues, manures, and municipal biosolids, are land-applied, three possible fates of the P can occur. The organic P, especially inositol phosphates, may be incorporated directly into stable humus, becoming relatively unavailable for plant and microbial use. If the residue has a relatively high P content, relative to C and N, a portion of the P will

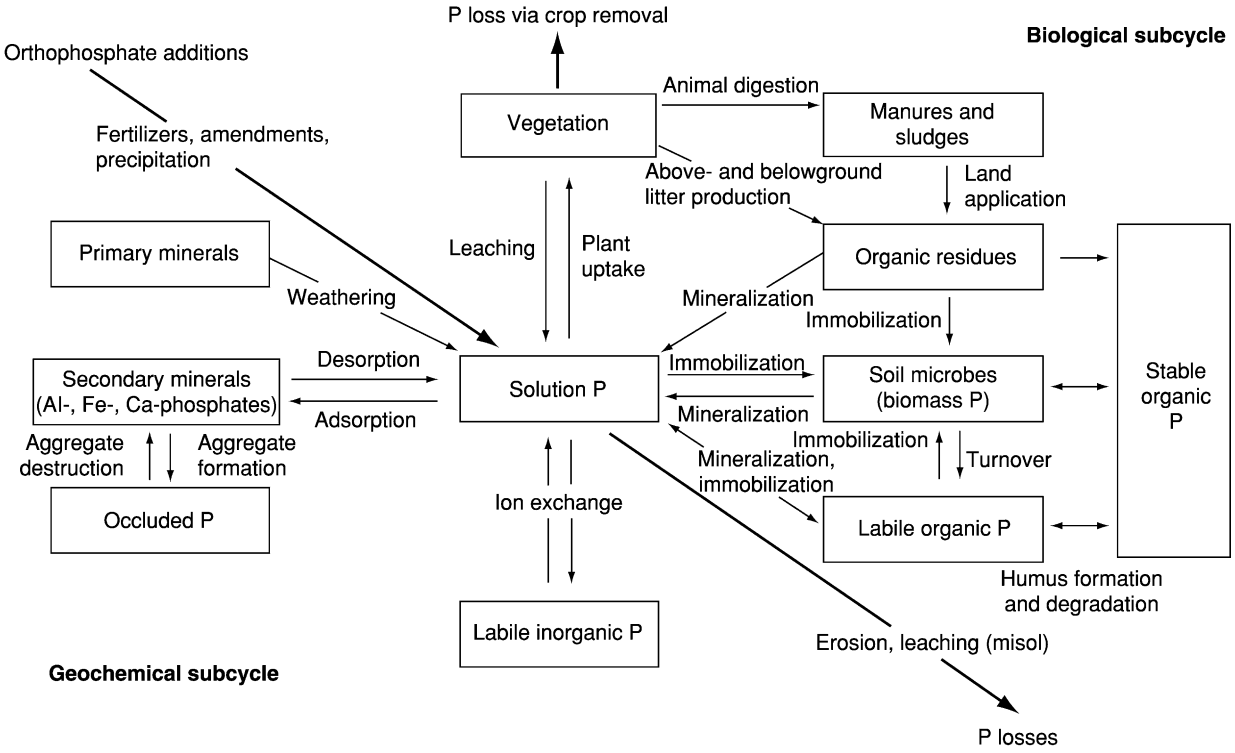


Figure 3 The phosphorus cycle showing inputs, losses, and transformations. This model illustrates the relationship of the chemical and biological P subcycles through the solution orthophosphate pool. Adapted with permission from Walbridge MR (1991) Phosphorus availability in acid organic soils of the lower North Carolina coastal plain. *Ecology* 72: 2083–2100.

be mineralized as orthophosphate during microbial decomposition of the residue. Finally, part of the P in the residue will be incorporated into the microbial biomass during decomposition. This immobilized P may become available as soluble P when the microbial biomass fraction dies back and is itself mineralized by other microorganisms.

While this model of the P cycle segregates biological activity from geochemical activity, it should be noted that the dissolution of apatite and other secondary P minerals can be enhanced by the activities of microorganisms and plants. Dissolution or solubilization of these minerals releases orthophosphate to solution P, where it can then be utilized by plants or microorganisms. Of course, a portion of this P will again precipitate as secondary P minerals, or be adsorbed by Al- and Fe-oxides and clay surfaces.

Mineralization and Immobilization of P

As briefly discussed above, key components of the biological cycling of P in soil are mineralization, the transformation of organic P into inorganic orthophosphate P, and immobilization, the transformation of inorganic P back into organic forms. These reactions are catalyzed by the active microbial biomass fraction in the soil, and are largely a consequence of heterotrophic microbial catabolic and anabolic processes. In order to acquire energy, electron-rich organic compounds are oxidized and degraded in order to produce energy. As these compounds are dismantled, inorganic nutrients are released. In order to build new cells, the biomass must acquire inorganic P and other nutrients. The extent to which P is either mineralized or immobilized during these reactions depends on the carbon-to-phosphorus (C:P) ratio of the compound being degraded. If the C:P ratio is 200:1 or less, then the residue is rich in P and the excess will be mineralized. Conversely, if the residue C:P is greater than about 300:1 there will not be enough P available in the residue to facilitate complete degradation. Therefore immobilization of orthophosphate from the soil by the microbial biomass will occur if decomposition of the residue and biomass growth is to continue. While immobilization decreases the amount of orthophosphate immediately available to plants, this is typically a very short-lived phenomenon and may indeed protect orthophosphate from being tied up in relatively insoluble inorganic compounds. After the added organic materials are mineralized, the microbial biomass will begin to die back to previous levels, resulting in the subsequent mineralization of the microbial P pool.

Mineralization is simply the enzymatic release of orthophosphate during the decomposition of organic materials that contain P. The phosphatase enzyme group is responsible for the hydrolysis of esters and anhydrides of orthophosphate-P. A number of different phosphatase enzymes are observed in the soil environment. The most commonly measured soil phosphatases are the phosphomonoesterases, which hydrolyze phosphomonoester bonds. Two phosphomonoesterases have been widely studied: acid phosphatase and alkaline phosphatase. Acid phosphatase is produced by microorganisms and plants and predominates in acid soils, while alkaline phosphatase is only produced by soil microorganisms and is dominant in alkaline soils (although it is present in acid soils as well). The phytase enzyme is also important. Phytins or inositol phosphates (Figure 1) can make up as much as 50% of the organic P in a soil. Therefore, hydrolysis of orthophosphate from this source can be important to plant nutrition. These extracellular enzymes are actively excreted into the soil by organisms, and also accumulate in soil, protected by sorption to organo-clay complexes in the soil. These enzymes are important not only for the degradation of freshly added organic residues, but also for the release of inorganic P from soil organic matter.

Microbial Solubilization of Inorganic P

Microorganisms are also important in the solubilization of P from mineral sources. Orthophosphate is typically present in soil solution at very low concentrations. This solution P is replenished either from mineralization of organic P sources, as discussed above, or from the dissolution of P from phosphate minerals. The primary P-containing mineral in soils is apatite or rock phosphate. In weathered, acidic soils, little apatite typically remains. In these soils, secondary inorganic P minerals in the form of aluminum and iron phosphates, such as variscite ($AlPO_4 \cdot 2H_2O$) and strengite ($FePO_4 \cdot 2H_2O$), dominate. In alkaline soils, the calcium phosphates dominate (e.g., $Ca_3(PO_4)_2$).

Most forms of inorganic P in soil have limited solubility (Figure 1). It is generally agreed that microorganisms can facilitate the enhanced dissolution of these compounds by at least two mechanisms: production of organic acids and excretion of protons during NH_4^+ assimilation. These mechanisms are performed by soil fungi and bacteria and appear to be potentially important, especially in soils receiving little or no phosphate fertilizer materials.

Phosphate solubilization has been observed for many different species of bacteria, fungi, and actinomycetes in laboratory studies. Bacterial species reported include those in the genera *Bacillus*, *Micrococcus*, and *Pseudomonas*. The most commonly observed P-solubilizing fungi are *Aspergillus* and *Penicillium* species, although a number of other

species have been reported to show P-solubilizing activity. In soil, P-solubilization activities are greatest in the rhizosphere. The plant root system provides ample quantities of readily utilized carbon which results in high microbial activity. The activity of P-solubilizing organisms in the rhizosphere can be potentially important in the P nutrition of plants. While not a direct topic of this chapter, it should be noted that mycorrhizal fungi are also active in solubilization and uptake of P which benefits the host plant. (For more detail) on mycorrhizal activities, see **Mycorrhizal Fungi**. In general, it appears that fungi may be more effective in soil than bacteria. Fungal hyphae can grow across distance in soil microenvironments, allowing for exploitation of P resources not available to relatively immobile bacteria.

Organic acids produced by bacteria and fungi have been shown to solubilize various forms of inorganic phosphates (Figure 4). Citrate, and other organic acids, can be quite effective in solubilizing only minimally soluble forms of P, such as $AlPO_4$. Organic acids are effective as P-solubilizing agents in liquid medium because they cause a reduction in pH in the medium, facilitating acid hydrolysis of the $AlPO_4$, and the released Al^{3+} is then complexed by the organic anion. This chelation or complexation of the released Al helps to minimize reprecipitation of Al-phosphates and the potential for Al-toxicity. While organic acids appear to be effective in solubilizing Al-phosphates, solubilization is typically more effective in the presence of microbial cells than that observed with organic acids alone. Also, some organisms solubilize Al-phosphates with no appreciable production of organic acids. Proton production during the assimilation of NH_4^+ or through respiratory activities appears to be a primary mechanism when organic acids are not produced.

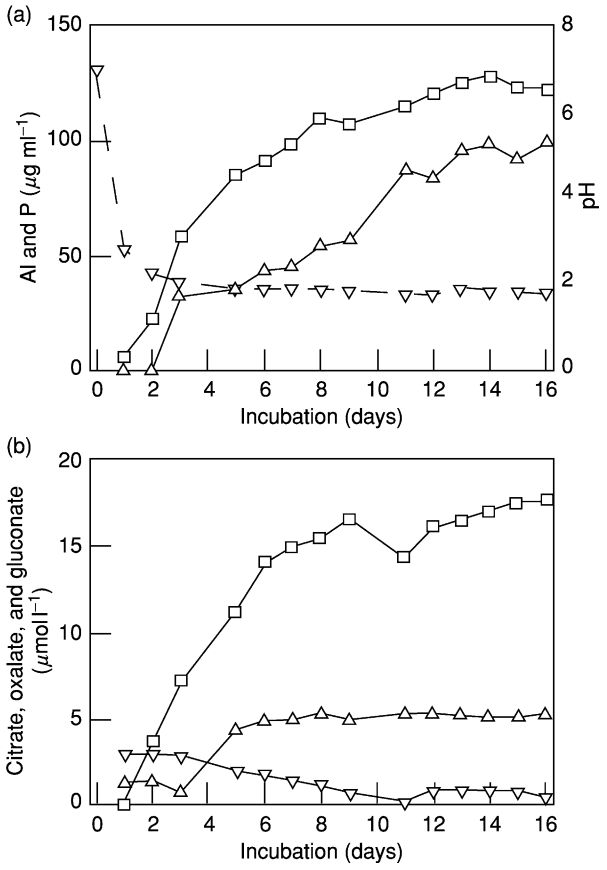


Figure 4 (a) Al (squares) and P (triangles) concentration and pH (inverted triangles) and (b) production of organic acids (squares, citrate; triangles, oxalate; and inverted triangles, gluconate) during the solubilization of $AlPO_4$ by the fungus *Aspergillus niger*. Reproduced with permission from Illmer P, Barbato A, and Schinner F (1995) Solubilization of hardly-soluble $AlPO_4$ with P-solubilizing microorganisms. *Soil Biology and Biochemistry* 27: 265–270.

Dissolution of Ca-phosphates, such as rock phosphate or $Ca_3(PO_4)_2$, also occurs by organic acid production and/or the production of protons during ammonium assimilation or respiration. Typically, a decrease in medium pH occurs during solubilization of P from these minerals. The pH decrease is most pronounced, and solubilization is usually highest, in the presence of ammonium N sources. While organic acids can be effective in solubilization of Ca-phosphates, they are not always produced by all organisms capable of the process. Again, production of protons by physiological processes is important in this regard.

P-Solubilizing Microorganisms and Plant Growth

Inoculation of crops with P-solubilizing microorganisms has met with mixed results. In the 1950s in the Soviet Union, researchers reported excellent results when *Bacillus megatherium* var. *phosphaticum* (also referred to as phosphobacterin) spores were applied to soils that were neutral to alkaline in pH and with high organic matter. The mode of action was thought to be the mineralization of organic P. Research in the USA was not able to verify the claims of significant crop yields and the practice was never adopted. However, other P-solubilizing microorganisms have been used with varying degrees of success in greenhouse and field situations.

Many greenhouse studies have shown positive effects of inoculation on yield and P uptake for several crops, including beans, peas, wheat, soybean, and sorghum. Numerous field studies have also been conducted. In some cases, there is little or no effect, while other research has shown positive effects of inoculation on yields and P uptake. Crops reported to respond

Table 1 Effect of the P-solubilizing fungus *Penicillium bilaji* inoculation on wheat dry-matter production and P uptake under field conditions with rock phosphate (RP) as a P source

Treatment	Dry matter		P uptake
	Grain ($g\,plot^{-1}$)	Straw ($g\,plot^{-1}$)	($mg\,plot^{-1}$)
P. bilaji + RP	207.7	402.7	834
P. bilaji	197.9	401.8	850
RP	186.8	388.3	771
Monoammonium phosphate	208.6	419.5	975
Control	180.6	378.9	742
Tukey's HSD (0.05)	23.6	42.3	61

HSD, honestly significant difference.
Adapted with permission from Kucey RMN (1988) Effect of *Penicillium bilaji* on the solubility and uptake of P and micronutrients from soil by wheat. *Canadian Journal of Soil Science* 68: 261–270.

positively to inoculation with P-solubilizing fungi and bacteria include wheat, canola, soybean, and others. An example of the impact of *Penicillium bilaji* on wheat growth and P uptake when used with rock phosphate is shown in Table 1. It should be noted that inoculation of soil or growth media with microorganisms can positively impact plant growth due to production of plant growth-promoting compounds or through suppression of deleterious microorganisms. However, any organism that positively impacts root growth and soil exploration will likely lead to better P availability for the plant.

Phosphorus and the Environment

Interest in the impact of P on the environment has greatly increased over the past decade. Overapplication of P-containing materials, whether inorganic fertilizers or organic sources such as manures and biosolids, can result in high levels of extractable soil inorganic phosphate. If these soils are close to surface waters, then this P can be exported to those waters either as soluble P in runoff water or as P adsorbed to eroding soil particles. In both cases, this can result in the elevation of the bioavailable P in these waters. This can potentially lead to eutrophication of the waters. Eutrophication occurs when excess nutrients, typically N or P, accumulate in the water, resulting in accelerated growth of algae. These algae can cloud the water, decreasing light penetration, and resulting in decreased growth of aquatic plants. Increased algal biomass production stimulates higher rates of degradation in the water column by bacteria. This accelerated decomposition then decreases available oxygen, often making the habitat unfit for many forms of aquatic life.

Many soils that have received regular applications of manure over time have become quite high in total and available P. If manure is applied to a crop to satisfy N requirements, too much P is typically applied, resulting in a P buildup. Although much of the P in these manures is added in the organic form, the result is typically a significant increase in the total amount of inorganic phosphate due to mineralization by soil microorganisms. Table 2 shows the impact of long-term manuring on total P, organic P, and inorganic P content of nine soils in Tennessee. Note that the inorganic P content is on average 2.4 times higher in the manured soils, and the proportion of inorganic P increased by 10.3%. Microbial mineralization of the manure P releases the phosphate and more of it tends to remain relatively labile. This may be due to increased soluble soil C which may then bind to P-sorption sites or with soil Al and Ca, reducing complexation. Microbial mineralization of these organic P sources over long periods of time can therefore result in potential offsite degradation of water quality.

Table 2 Inorganic (Pi) and organic (Po) P content in the 0–15-cm depth of nine Tennessee soils treated with animal manures. Soil textures range from sandy loams to silt loams. Four soils received swine manure, three received dairy manure, and two received poultry litter

	Total P ($mg\,kg^{-1}$)	Po ($mg\,kg^{-1}$)	Pi ($mg\,kg^{-1}$)	% Pi
Inorganic fertilizers	975	793	182	18.7%
Manured soils	1478	1049	429	29.0%

See also: **Eutrophication**; **Fertility**; **Fertilizers and Fertilization**; **Macronutrients**; **Mycorrhizal Fungi**; **Phosphorus in Soils:** Overview

Further Reading

Germida JJ and Siciliano SD (2000) Phosphorus, sulfur and metal transformation. In: Sumner ME (ed.) *Handbook of Soil Science*, pp. C95–C106. Boca Raton, FL: CRC Press.

Haygarth PM and Jarvis SC (1999) Transfer of phosphorus from agricultural soils. *Advances in Agronomy* 66: 195–249.

Hocking PJ (2001) Organic acids exuded from roots in phosphorus uptake and aluminum tolerance of plants in acid soils. *Advances in Agronomy* 74: 63–97.

Illmer P, Barbato A, and Schinner F (1995) Solubilization of hardly-soluble $AlPO_4$ with P-solubilizing microorganisms. *Soil Biology and Biochemistry* 27: 265–270.

Kucey RMN (1988) Effect of *Penicillium bilaji* on the solubility and uptake of P and micronutrients from soil by wheat. *Canadian Journal of Soil Science* 68: 261–270.

Kucey RMN, Janzen HH, and Leggett ME (1989) Microbially mediated increases in plant-available phosphorus. *Advances in Agronomy* 42: 199–228.

Kuo S (1996) Phosphorus. In: Sparks DL, Page AL, and Helmke PA *et al.* (eds) *Methods of Soil Analysis. Part 3. Chemical Methods*. SSSA Book Series no. 5, pp. 869–919. Madison, WI: Soil Science Society of America.

Lindsay WL (1979) *Chemical Equilibria in Soils*. New York: John Wiley.

Mullen MD (1998) Transformations of other elements. In: Sylvia DM, Fuhrmann JJ, Hartel PG, and Zuberer DA (eds) *Principles and Applications of Soil Microbiology*, pp. 369–386. Upper Saddle River, NJ: Prentice-Hall.

Oehl F, Oberson A, Probst M, Fliessbach A, Roth H-R, and Frossard E (2001) Kinetics of microbial phosphorus uptake in cultivated soils. *Biology and Fertility of Soils* 34: 31–41.

Pierzynski GM, Sims JT, and Vance GF (2000) *Soils and Environmental Quality*. Boca Raton, FL: CRC Press.

Stevenson FJ and Cole MA (1999) *Cycles of Soil: Carbon, Nitrogen, Phosphorus, Sulfur, Micronutrients*. New York: John Wiley.

Tabatabai MA (1994) Soil enzymes. In: *Methods of Soil Analysis. Part 2. Microbiological and Biochemical Properties*. SSSA Book Series no. 5, pp. 775–833. Madison, WI: Soil Science Society of America.

Walbridge MR (1991) Phosphorus availability in acid organic soils of the lower North Carolina coastal plain. *Ecology* 72: 2083–2100.

Whitelaw MA (2000) Growth promotion of plants inoculated with phosphate-solubilizing fungi. *Advances in Agronomy* 69: 99–151.

PHYTOTOXIC SUBSTANCES IN SOILS

M Qadir, International Center for Agricultural Research in the Dry Areas (ICARDA), Aleppo, Syria; and International Water Management Institute (IWMI), Colombo, Sri Lanka
S Schubert and D Steffens, Justus Liebig University, Giessen, Germany

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Plants are continuously exposed to their ambient soil environment during the whole span of their life-cycle. Sometimes the presence, application, and/or formation of certain substances in soils reach to the extent that they become phytotoxic – the property of a substance at a specified concentration that restricts or constrains the plant growth. With a wide range of characteristics and impacts, phytotoxic substances have long been accumulating in the terrestrial ecosystem, including soils. Once these substances enter the soil, they become part of a cycle that affects all forms of life. Soils act as a sink and a filter, protecting the groundwater from inputs of phytotoxic substances during initial stages of their addition. With continued uncontrolled applications, the concentrations of potentially harmful substances may increase to levels that become toxic to a wide range of plants, and in certain cases toxicity can occur in soil faunas and floras, higher animals, and humans.

The contribution of several different human activities to accumulation of phytotoxic substances has been several times greater than natural sources. Although the use of chemicals expanded in the nineteenth century, it was not until the middle of the twentieth century that the use of chemicals became widespread. In recent years, the quantities of many products in which phytotoxic substances are used have increased notably, enhancing the opportunity for contamination and phytotoxicity in soils. Although there are several forms of phytotoxic substances in soils, they can be grouped into five broad categories: (1) pesticides, (2) metals and metalloids, (3) soil acidity and aluminum and manganese toxicity, (4) soil flooding and manganese and iron toxicity, and (5) soluble salts and sodium toxicity.

Pesticides

Pesticides are substances or mixtures thereof intended to prevent, destroy, repell, or mitigate any pest, which may be an insect, rodent, nematode, fungus, weed, or other form of terrestrial or aquatic plant or animal

Table 1 Class, chemical group, and persistence time of some pesticides

Pesticide class	*Chemical group*	*Pesticide name*	*Persistence time*
Insecticide	Chlorinated hydrocarbons	DDT	3–15 years
Insecticide	Chlorinated hydrocarbons	Chlordane	2–5 years
Insecticide	Chlorinated hydrocarbons	Dieldrin	2–5 years
Herbicide	Triazines	Atrazine	1–2 years
Herbicide	Triazines	Simazine	1–2 years
Herbicide	Benzolics	Dicamba	2–12 months
Herbicide	Phenoxyalkyl acids	2,4-D	1–5 months
Insecticide	Organophosphates	Diazinon	1–12 weeks
Insecticide	Organophosphates	Malathion	1–12 weeks
Insecticide	Carbamates	Carbaryl	1–8 weeks

life that is injurious to growing plants, animal or human health, or the environment. The chemical revolution in agriculture was initiated when the insecticidal properties of dichlorodiphenyltrichloroethane (DDT) were discovered in 1939 along with the herbicidal effects of 2,4-dichlorophenoxyacetic acid (2,4-D) in 1941. Since their discovery, tens of thousands of such chemicals and multichemical formulations have been developed, tested, and put to use. In the past 60 years, synthetic organic pesticides have become a major element in modern agricultural production practices. The advent of pesticide use has coincided with the tremendous increases in agricultural productivity. Although the use of pesticides has provided many benefits to society by protecting crops, animals, and humans, it has also created several problems, including widespread accumulation of residues, with damage to plants, wildlife, fisheries, beneficial insects, and even humans.

The impact of pesticides on plants may be as dramatic and devastating as the impact on insects and animals. Exposure of a sensitive plant may produce remarkable growth distortions and yield loss. During the process of application, pesticides may move from the treatment area to deposit a residue affecting a nontarget species. In the case of some pesticides, the amount remaining after a prior treatment can be a limiting factor for the subsequent crop(s). Even though the quantity of pesticide would be small, it could be sufficient to limit the growth of a sensitive species.

Pesticides are commonly classified according to the target group used to control, deter, or destroy pest organisms: (1) insecticides, (2) fungicides, (3) herbicides, (4) rodenticides, and (5) nematocides. The first three are used in much larger quantities and therefore contaminate soils to a greater extent than rodenticides and nematocides. Whatever the mode of application of pesticides, a high proportion of such chemicals eventually moves into the soil. After finding their way into soils, these chemicals then meet with one or a combination of the following fates as they may: (1) vaporize into the atmosphere without a chemical change, (2) adsorb on soil colloidal complex that consists of humus and clay particles, (3) move with the percolating water to the deeper soil layers or to the groundwater, (4) undergo chemical reactions with other chemicals, (5) be broken down by soil microorganisms, and/or (6) be subjected to absorption by plants. The specific fate of different pesticides in soils is determined by their chemical structures and properties, various processes occurring in soils (retention, transformation, and transport processes and their interactions), climatic conditions, and different forms of anthropogenic activities dealing with the management of pesticides in soils. Some pesticides are not readily biodegradable and persist in soils for many years. Thus there is a large variation in the persistence time of different pesticides in soils (Table 1).

Metals and Metalloids

Metals are a group of chemical elements that exhibit all or most of the following physical qualities: (1) solid at ordinary temperatures, (2) opaque except in extremely thin films, (3) good electrical and thermal conductors, (4) lustrous when polished, and (5) crystalline when in the solid state. Metalloids such as arsenic (As) and selenium (Se) are elements with properties intermediate between those of metals and nonmetals. Some metals and metalloids such as zinc (Zn), copper (Cu), manganese (Mn), nickel (Ni), cobalt (Co), chromium (Cr), molybdenum (Mo), iron (Fe), and Se are essential for living organisms. However, these metals and those that are thought to be nonessential such as lead (Pb), mercury (Hg), cadmium (Cd), and As can pose risks when they build up in soils due to many forms of anthropogenic influences (Table 2). When their bioavailability is high, toxicity can result, with negative effects on plants.

Although metals and metalloids are usually present in soil parent materials, they have several different

Table 2 List of anthropogenic sources of some selected metals and metalloids

Chemical	*Symbol*	*Major uses and sources*[a]
Arsenic	As	Pesticides, animal feed additives, and manures from intensive animal production, desiccants, fossil fuel combustion, electronics, pigments and paints, medical uses
Cadmium	Cd	Electroplating, pigments for plastics and paints, batteries, sewage sludge, fossil fuel combustion, polymer stabilizers, waste disposal, impurities in fertilizers
Chromium	Cr	Chrome-plated metals, pigments, electronics, fossil fuel combustion, impurities in fertilizers, wood preservatives, stainless steel
Copper	Cu	Fly ash, wind-blown copper containing dust, impurities in fertilizers, warfare and military training, waste disposal, medical uses, electronics, sewage sludge, wood preservatives, pesticides
Lead	Pb	Fossil fuel combustion, batteries, pigments and paints, polymer stabilizers, printing, sports shooting, warfare, and military training, waste disposal, impurities in fertilizers, pesticides
Mercury	Hg	Pesticides, chlorine manufacture, medical uses, batteries, electronics
Nickel	Ni	Alloy manufacture, electroplating, sewage sludge, batteries, catalyst
Selenium	Se	Pigments and paints, printing, medical uses
Zinc	Zn	Metallurgical industries, electronics, pesticides, impurities in fertilizers, sewage sludge, pigments and paints, polymer stabilizers, printing, medical uses

[a]In addition to mining of ore bodies in the Earth's crust and smelting of scrap material, and weathering of soil minerals.

anthropogenic sources, including: (1) mining of ore bodies in the earth's crust and smelting of scrap material; (2) sewage sludge produced from the treatment of domestic and industrial waste waters; (3) combustion of fossil fuel consisting of coal and petroleum; (4) manufacture, use, and disposal of metal commodities in metallurgical industries; (5) manufacture, use, and disposal of semiconductors, cables, contacts, and other electrical components of electronic industries; (6) chemical and other industrial sources such as chlorine manufacture, batteries, pigments and paints, catalysts, polymer stabilizers, printing, medical uses, and additives in fuels and lubricants; (7) disposal of household, municipal, and industrial wastes; (8) flood or seepage waters; and (9) sports shooting, warfare, and military training. As an important anthropogenic source of metals and metalloids in soils, the disposal of domestic, municipal, and industrial wastes also contributes to significant quantities of certain substances that can produce harmful environmental effects. In some cases, such wastes may contain less decomposable organic substances. In addition to the nonagricultural sources, agricultural practices constitute important sources of metals which make a significant contribution to their total concentrations in soils in many parts of the world, especially in regions of intensive farming. The main sources are impurities in fertilizers, application of sewage sludge as a source of plant nutrients and organic matter, manures from intensive animal production, use of pesticides, irrigation with metal-rich waters, and corrosion of metal objects.

Atmosphere and water are important transport media for metals from different sources. Soils are often contaminated for up to hundreds of kilometers away from the site of emission. The metals are usually present in air as aerosol particles with a size range of 5–20 μm, but most are between 0.1 and 10 μm in diameter and have a mean residence time of 10–30 days. A high proportion of the metals in more-recent dust deposits are of anthropogenic origin. Water, on the other hand, is an integral component of the weathering processes relating to the release of metals from soil parent material. In addition, contamination of either surface water or groundwater impacts the quality of each other and subsequently the soils irrigated with such waters.

The total metal content of a soil is the result of inputs of metals from the abovementioned sources minus losses in metals removed in crop material, leaching, and volatilization. However, it is not the total amounts but the bioavailable concentrations that are crucial and depend on a number of soil and environmental factors, which also control the speciation of metals in soils. Thus the effect of any metal on plant growth depends not only on its chemical properties but also on its bioavailable concentration. The physiological age and species of the plant concerned are also important.

Soil Acidity and Aluminum and Manganese Toxicity

Soil acidity is usually expressed in terms of pH scale, which quantifies the activity of protons (H^+) present in a solution. Soil acidification can develop naturally when a soil profile is leached, but can be accelerated by certain farming practices and acid precipitation, which is commonly known as acid rain. Various amounts of sulfuric and nitric acids (H_2SO_4 and HNO_3) form in precipitation from certain sulfur (S) and nitrogen (N) gases produced by a number of

sources such as lightning, volcanic eruptions, forest fires, and the combustion of fossil fuels. Since N_2-fixing legumes release H^+ in the root zone, farming practices dealing with continuous cultivation of such legumes may increase the acidity of acid soils. This contribution is crucial when all or a major part of the harvested shoot is not added back to the soil as a measure to control soil acidity. On the other hand, the effects of acid rain are more pronounced on the acidity of water than on soil acidity. Soils generally are sufficiently buffered to accommodate acid rain with little increase in acidity on an annual basis. However, continued inputs of acid rain would have significant effects on the pH of acid soils, especially those that are weakly buffered. Acid rain is also a serious problem for soils that are already highly acidic, since a small increase in acidity could make them even less fertile. Other acidifying processes in soils include: (1) formation of carbonic acid (H_2CO_3) from carbon dioxide (CO_2) in the root zone; (2) dissociation of acids; (3) oxidation of N, S, and Fe compounds; (4) accumulation of organic matter; and (5) deprotonation of pH-dependent charges. The growth of sensitive plants may be adversely affected by high free-H^+ concentration in soils.

Soils become extremely acidic ($pH < 4$) when large amounts of reduced forms of S are oxidized to H_2SO_4 if the soils are exposed to oxygen when they are drained or excavated. With a $\geq$0.15-m-thick sulfuric horizon that is often present, such soils are commonly referred to as acid sulfate soils. These soils could be categorized into potential acid sulfate soils (H_2SO_4 may be produced), active acid sulfate soils (H_2SO_4 is being produced), or postactive acid sulfate soils (H_2SO_4 has been produced). The amounts of S present in these soils provide the basis for estimating the amount of liming material ($CaCO_3$) required to neutralize the H_2SO_4 that could be produced. In many cases, the volume of soil that needs to be treated and the amounts of $CaCO_3$ required are huge enough to make such neutralization an impractical solution to the problem.

Aluminum (Al) is the most abundant metal in the Earth's crust, comprising about 7% of its mass, and is found in soils primarily in the form of insoluble aluminum silicates and oxides. Although Al is a nonessential element, its toxicity occurs in acid soils, which is considered as the most common and severe problem associated with such soils. When soils become more acidic, phytotoxic species of Al such as Al^{3+}, $Al(OH)^{2+}$, and $Al(OH)_2^+$ are released into soil solution. Since most plant species are sensitive to micromolar concentrations of Al, toxicity of this metal may become a major limiting factor for plant productivity on acid soils. Aluminum toxicity has been recognized as a major constraint for the production of several crops in acid soils for the past 80 years. When plants grow and develop under acid soils in natural conditions, they are continuously exposed to Al and the damage is cumulative. As a general phenomenon, Al toxicity primarily restricts root growth, with an effect on plant metabolism through interference with the uptake, transport, and use of several nutrients by plants. However, it is not known whether these observed effects are the primary cause of Al toxicity or they are only indirect or secondary effects. In addition to plants, many bacteria such as those that are involved in certain transformations in N cycle are also adversely affected by high Al levels associated with low soil pH. In contrast to Al, Mn is an essential plant nutrient, which is readily transported from roots to shoots and, therefore, symptoms of Mn toxicity first appear on shoots. Excessive levels of Mn in soils also interfere with the uptake of other mineral nutrients such as calcium (Ca^{2+}) and magnesium (Mg^{2+}). Plant species and genotypes within a species vary greatly with regard to their susceptibility to Mn toxicity. Manganese toxicity is not as widespread as Al toxicity.

Soil Flooding and Manganese and Iron Toxicity

With the exception of rice (*Oryza sativa* L.) and Kallar grass (*Leptochloa fusca* (L.) Kunth.), periodic flooding during the growing season adversely affects growth and productivity of crops. Soil-flooded conditions are also expressed through terms such as 'soil saturation,' 'waterlogging,' 'anoxia,' and 'hypoxia.' Lack of oxygen has been found to be the major problem associated with soil flooding that impacts plant growth. A great deal of information during recent decades has accumulated from research on the molecular, biochemical, and physiological responses of a number of plant species to flooding of the growth medium. The use of molecular plant-breeding and genomic transformations to improve flooding tolerance in crops has been rated as a promising solution to this problem.

In addition to oxygen stress, reduced forms of certain elements can also cause phytotoxicity in flooded soils. For instance, the reduced forms of Mn and Fe are far more soluble than their oxidized forms. Therefore, the toxicity of Mn and Fe is greatly increased by low-oxygen conditions, particularly in acid paddy soils. Other growth-reducing and yield-loss factors in flooded soils could arise from root rot disease, N deficiency, nutrient imbalance, and/or accumulation of high levels of CO_2 in the root zone. Such levels of CO_2 in flooded soils could cause severe leaf chlorosis and reduction in plant biomass in flood-sensitive crops such as soybeans (*Glycine max* L.).

Soil Salinity and Sodium Toxicity

Accumulation of soluble salts can occur in soils to levels that can adversely affect crop growth and yield. Salt-affected soils mostly exist under arid and semiarid regions. Major cations in salt-affected soils are Na^+, Ca^{2+}, Mg^{2+}, and, to a lesser extent, potassium (K^+). The major anions are chloride (Cl^-), sulfate (SO_4^{2-}), bicarbonate (HCO_3^-), carbonate (CO_3^{2-}), and nitrate (NO_3^-). Salt-affected soils are generally classified into two main categories, saline and sodic. Soil salinity is generally characterized by measuring electrical conductivity of the saturated soil paste extract (EC_e) and expressed in units of deci-siemens per meter. In some parts of the world, saline soils are also known as white alkali soils. Soil sodicity is evaluated in terms of the relative amounts of Na^+ on the cation exchange sites (exchangeable sodium percentage, ESP) or in the soil solution (sodium adsorption ratio, SAR), given the accompanying levels of salinity. Other names for sodic soils include 'black alkali soils,' 'solonetz,' and 'slick-spots.'

The presence of voluminous data indicates that there are no single values of EC_e and ESP (or SAR) for saline and sodic soils, respectively. In general, an EC_e of $4\,dS\,m^{-1}$ is used as the boundary between saline and nonsaline soils; the soils of higher EC_e are called 'saline.' Actually, salt-sensitive plants are affected at half of this salinity ($2\,dS\,m^{-1}$) and highly tolerant plants at about twice of this salinity ($8\,dS\,m^{-1}$) or at an even greater value. Similarly, an ESP of 15 (an ESP of 6 in Australia) is considered as the critical limit for soil sodicity; the soils of higher ESP show dispersive behavior in most cases and suffer from serious physical problems when wetted. Different terminology for alkali soils is used in some parts of the world to express sodicity hazard. Alkali soils are characterized by high sodicity (ESP of more than 15) and pH (pH >8.3), and contain soluble CO_3^{2-} and HCO_3^- of Na^+, i.e., C_{Na}: $(C_{Cl} + C_{SO_4})$ ratio greater than 1. These soils contain Na^+ and $CO_3^{2-} + HCO_3^-$ as the dominant ions and tend to have low salinities and high pH values, which cause an increase in swelling and dispersion. On the other hand, the pH of sodic soils can be either more than or less than 7, and such soils can be either saline or nonsaline. Therefore, use of the term 'alkali' provides a practical distinction in terms of soil management.

Deterioration of salt-affected soils occurs through changes in the proportions of soil solution and exchangeable ions and soil reaction (pH) that lead to osmotic and specific ion effects together with imbalances in the availability of plant-available nutrients. Excess Na^+ in soils accelerates the development of structural problems created by certain physical processes (slaking, swelling, and dispersion of clay minerals) and specific conditions (surface crusting and hard-setting) that may affect water and air movement, plant-available water-holding capacity, root penetration, seedling emergence, runoff, erosion, as well as tillage and sowing. These chemical and physical changes have a bearing on the activity of plant roots and soil microbes, and ultimately on crop growth and yield.

Excess soluble salts in salt-affected soils originate either through the weathering of parent minerals (causing fossil or primary salinity and sodicity) or from man-made activities involving improper use and management of land and water resources (contributing to secondary salinity and sodicity). Secondary salinity and sodicity of arable land resources are as old as the history of human settlement and irrigation. A well-documented case occurred in Mesopotamia where, between 4000 and 2000 BC, secondary salinization affected the land in the valleys of the Euphrates and Tigris as a result of improper irrigation practices. History repeated itself in the twentieth century in many countries and regions. Although the vast river systems in arid and semiarid regions with extensive irrigated agriculture supply water for irrigation, they also contain appreciable quantities of salts. The inappropriate use of such irrigation systems accelerates the development of secondary salinity and sodicity, which plague these regions. Dryland (or nonirrigated) salinization is commonly due to rising water tables over essentially saline subsoils caused by clearance of trees or other deep-rooted crops, usually to make land available for rain-fed agriculture. The area under human-induced salinity and sodicity has increased substantially, particularly in the twentieth century.

Remediation Strategies for Phytotoxic Substances

There is a long history of accumulation and control of phytotoxic substances in soils. The concentrations of such substances in soils are increasing because the inputs are not balanced by the output in terms of removal from the plough layer. Although the accumulation of phytotoxic substances in soils provides a direct stimulus for cleanup, mostly public and political pressure to reduce these substances in soils only occurs when critical levels are reached. The past focus to remove phytotoxic substances from soils was on the use of traditional physical and chemical methods. Physical remediation is the process of correcting the problem by a number of physical means. For example, containment of pollutants could be done by covering the polluted site with a layer impermeable to rain

water. Clay or cement has been used in the past to provide an impermeable cap preventing water from reaching the pollutants and leaching them away. Chemical remediation involves the use of a variety of chemicals for different phytotoxic substances. For example, $CaCO_3$ is used to neutralize H_2SO_4 in acid sulfate soils, while gypsum ($CaSO_4 \cdot 2H_2O$) has been a commonly used amendment to ameliorate excess Na^+ in sodic soils. The enormous cost associated with the physical and chemical means is one of the major reasons for slow progress in the removal of phytotoxic substances from the soils. Therefore, efforts have been made to minimize the cost of amelioration of such soils. In recent decades, strategies assisted by certain microorganisms and higher plants have shown the potential to remediate the soils at relatively low costs. These strategies are gaining an increasing interest among policy makers and scientists. Biological remediation, commonly known as bioremediation, deals with cleaning soils with the assistance of certain microorganisms that have the capability of breaking down some specific phytotoxic substances. Persistent pesticides such as DDT and dioxin have been reported to succumb to the attack of a white rot fungi species, *Phanerochaete chrysosporium*. Phytoremediation is a plant-based approach, which involves the use of suitable crops, grasses, shrubs, and trees that have the potential to grow on contaminated soils with the ability to improve such soils in several different ways.

Phytoremediation of metals and metalloids can be done through phytoextraction, which involves the use of pollutant-scavenging plants to transport and concentrate metals and metalloids from the soil into the aboveground shoot. The efficiency of phytoextraction for a given metal ($Phyto_{Metal}$) is the product of the dry-matter yield of shoot (S_{DM}) and concentration of the metal in shoot (S_{CM}), as given in the following equation:

$$Phyto_{Metal} = S_{DM} \times S_{CM} \quad [1]$$

The plants able to accumulate high concentrations of metals are known as hyperaccumulators (Table 3). The concentrations of metals accumulated in such plants may be 100 times greater than those occurring in nonaccumulator plants growing on the same substrates. Following the harvest of the metal-enriched plants, the weight and volume can be reduced by ashing. Metal-enriched plants can be disposed of as hazardous material or, if economically feasible, used for metal recovery. Phytoremediation can be done through reduction in the bioavailability of metals in the soil (phytostabilization) and volatilization of pollutants such as Hg and Se from the foliage (phytovolatilization).

Table 3 Some important metal hyperaccumulators

Element	*Metal-accumulator plant species*
Cadmium	*Thlaspi caerulescens*
Cobalt	*Haumaniastrum robertii*
Copper	*Haumaniastrum katangense*
Lead	*Thlaspi rotundifolium* subsp.
Manganese	*Macadamia neurophylla*
Nickel	*Berkheya coddii*
Selenium	*Astragalus pattersoni*
Thallium	*Iberis intermedia*
Uranium	*Atriplex confertifolia*
Zinc	*Thlaspi calaminare*

In contrast to the phytoremediation of metal-contaminated soils, phytoremediation of salt-affected soils works through plant roots to enhance leaching of soluble salts and Na^+ from the root zone to the deeper soil layers. The enhanced partial pressure of carbon dioxide (P_{CO_2}) in the root zone during cropping is considered as a major factor contributing to the amelioration of calcareous sodic soils. Aqueous CO_2 helps in the dissolution of native soil calcite ($CaCO_3$) to provide Ca^{2+} for the Na^+–Ca^{2+} exchange at the cation exchange sites. In addition, H^+ released by the roots of certain phytoremediation crops such as N_2-fixing legumes may react with soil $CaCO_3$ to provide Ca^{2+}. Phytoremediation of sodic soils ($Phyto_{Sodic}$) is considered as a function of the following factors:

$$Phyto_{Sodic} = \sum R_{P_{CO_2}} + R_{H^+} + R_{Phy} \quad [2]$$

where $R_{P_{CO_2}}$ refers to P_{CO_2} within the root zone, R_{H^+} is H^+ release in the root zone of N_2-fixing crops, and R_{Phy} refers to physical effects of roots in improving soil aggregation and hydraulic properties of the root zone. The collective effects of these factors ultimately lead to soil improvement, provided drainage is adequate and leaching occurs. The salinity–sodicity combination in soil solution during phytoremediation remains suitable to maintain adequate soil structure and aggregate stability, which prevent reduction in soil permeability to air and water.

Phytoremediation of sodic soils has been shown to be advantageous in several economic, environmental, and agronomic aspects: (1) no financial outlay to purchase chemical amendments, (2) financial and other benefits from plant species grown during phytoremediation, (3) promotion of soil aggregate stability and creation of macropores that improve soil hydraulic properties, (4) better plant nutrient availability in soil during and after phytoremedition, and (5) more uniform and greater zone of amelioration in terms of soil depth. In general, phytoremediation

Table 4 Some plant species used for phytoremediation of salt-affected soils

Common name	*Botanical name*
Alfalfa	*Medicago sativa* L.
Barley	*Hordeum vulgare* L.
Bermuda grass	*Cynodon dactylon* (L.) Pers.
Kallar grass	*Leptochloa fusca* (L.) Kunth
Millet rice	*Echinochloa colona* (L.) Link
Para grass	*Urochloa mutica* (Forssk.) T. Q. Nguyen
Rhodes grass	*Chloris gayana* Kunth
Rice	*Oryza sativa* L.
Sesbania	*Sesbania bispinosa* (Jacq.) W. Wight
Sordan	*Sorghum* × *drummondii* (Steud.) Millsp. & Chase
Sorghum	*Sorghum bicolor* (L.) Moench

works well on coarse- to medium-textured and moderately salt-affected soils. On such soils, its performance has been found to be comparable with soil application of chemical amendments such as gypsum. Several plant species of agricultural significance have been successfully grown as a phytoremediation tool on a variety of salt-affected soils (Table 4). Phytoremediation may face certain limitations under conditions of very high levels of salinity and/or sodicity, where the growth of phytoremediation crops is likely to be variable and patchy. Under these conditions, the use of chemical amendments is probably necessary.

See also: **Acidity**; **Pesticides**; **Sodic Soils**

Further Reading

Alloway BJ (1995) *Heavy Metals in Soils*. London, UK: Chapman and Hall.

Cheng HH (1990) *Pesticides in the Soil Environment: Processes, Impacts, and Modeling*. Madison, WI: Soil Science Society of America.

De la Fuente-Martinez JM and Herrera-Estrella L (1999) Advances in the understanding of aluminum toxicity and the development of aluminum-tolerant transgenic plants. *Advances in Agronomy* 66: 103–120.

Ghassemi F, Jakeman AJ, and Nix HA (1995) *Salinisation of Land and Water Resources: Human Causes, Extent, Management and Case Studies*. Wallingford, UK: CABI Publishing.

Hillel D (2000) *Salinity Management for Sustainable Irrigation: Integrating Science, Environment, and Economics*. Washington, DC: The World Bank.

Marschner H (1995) *Mineral Nutrition of Higher Plants*. London, UK: Academic Press.

McGrath SP, Zhao FJ, and Lombi E (2002) Phytoremediation of metals, metalloids, and radionuclides. *Advances in Agronomy* 75: 1–56.

Mengel K and Kirkby EA (2001) *Principles of Plant Nutrition*. Dordrecht, the Netherlands: Kluwer Academic Publishers.

Ponnamperuma FN (1972) The chemistry of submerged soils. *Advances in Agronomy* 24: 29–96.

Qadir M and Oster JD (2002) Vegetative bioremediation of calcareous sodic soils: history, mechanisms, and evaluation. *Irrigation Science* 21: 91–101.

Qadir M and Schubert S (2002) Degradation processes and nutrient constraints in sodic soils. *Land Degradation and Development* 13: 275–294.

Raskin I and Ensley BD (1999) *Phytoremediation of Toxic Metals: Using Plants to Clean Up the Environment*. New York: John Wiley.

Selim HM and Sparks DL (2001) *Heavy Metal Release in Soils*. Boca Raton, FL: CRC Press.

Terry N, Zayed AM, de Souza MP, and Tarun AS (2000) Selenium in higher plants. *Annual Review of Plant Physiology and Plant Molecular Biology* 51: 401–432.

PLANT–SOIL–WATER RELATIONS

R A Feddes and J C van Dam, Wageningen University and Research Centre, Wageningen, The Netherlands

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Water for evapotranspiration from land surfaces is supplied mainly by the soil. The soil water reservoir balances the episodic excesses of water supply from rainfall against the more smoothly varying atmospheric demand for evapotranspiration. The role of soil moisture within the soil–plant–atmosphere system depends on the soil moisture reservoir size and the availability of water in that reservoir, which, in turn depends on the texture and structure of the soil and the characteristics of the root system.

One key function of plant roots is their ability to link the soil, where water and nutrients reside, to the organs and tissues of the plant, where these resources are used by the canopy. Hence roots serve to connect the soil environment to the atmosphere by providing a first link in the pathway for fluxes of water and solutes through the plant to the atmosphere. Fluxes along

the soil–plant–atmospheric continuum are regulated by aboveground plant properties such as the regulation of plant transpiration by the leaf stomata, which interact with the atmosphere, and plant root properties such as depth, distribution, and activity of roots, as well as soil properties such as water potential, water content, and hydraulic conductivity.

Plant root systems show a remarkable ability to adapt to soil depth and to changes in availability of water and nutrients and the chemical properties (e.g., salinity) in soils. Root response to soil properties in turn affects the uptake of soil water and nutrients and the storage of carbon belowground. Root distribution may change when ecosystems respond to greenhouse warming and carbon dioxide fertilization.

Liquid Water Flow Inside the Plant

Water is absorbed in the *liquid state* by the roots from the soil. The general flow pattern is shown in Figure 1: soil → root cells → endodermis → xylem → leaf cells → stomata → air (the latter as vapor flow). Most of the transport (±80%) through the root and leaf cells takes place along the microcapillaries situated in the cell wall. However, the roots contain endodermis cells in which the continuous system of microcapillaries in the cell walls is interrupted by the so-called Casparian strips. Consequently most of the water has to pass through the cells itself. This means that the permeability of the endodermis cell and the semipermeable membrane (plasmalemma) can be an important determinant in controlling the water flow. Hence a low soil temperature and/or poor soil aeration will induce a low metabolic activity of the endodermis cell, which may cause a high resistance of the semipermeable membrane against water uptake. The resistance of the xylem vessels is quite low compared with the large resistances encountered in transport from cell to cell.

Figure 2 shows a daily cycle of transpiration and water uptake which is regularly observed in summer. Before sunrise a plant is usually at full turgor, there is no transpiration, and stomata are fully closed. After sunrise, radiation increases and stomata open. During midday hours the transpiration demand increases, the water head in the leaves drops, because plant roots cannot take up water quickly enough, turgidity declines, and stomata will partially or completely close. This leads to a decrease in transpiration and an increase in leaf temperature. After closing the stomata, water head and turgidity will be restored, which may lead to opening of the stomata again in the afternoon. During nighttime there is complete recovery of water head and turgidity under nonevaporation conditions.

Accurate mathematical description of water uptake by crops with a nonuniform root system is complicated. Therefore it has been customary to describe water flow through the entire soil–plant–atmosphere system with an electrical analog. This model assumes that the water flux density v_z (cubic centimeters per square centimeters per second) through the rooted soil zone and the root–stem–leaf-stomata path is proportional to the total water potential/head difference Δh_{total} (centimeters) and inversely proportional to the total resistance R (seconds) of the system. Thus,

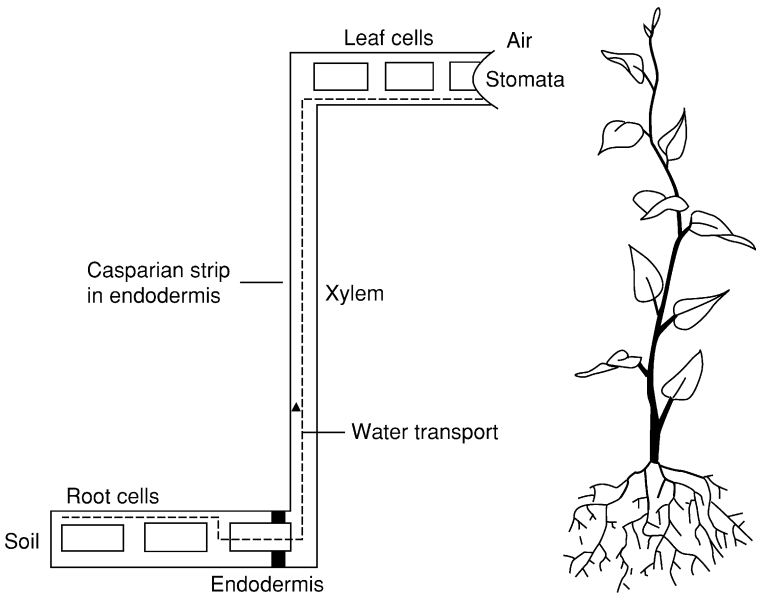


Figure 1 Liquid transport through the plant tissues.

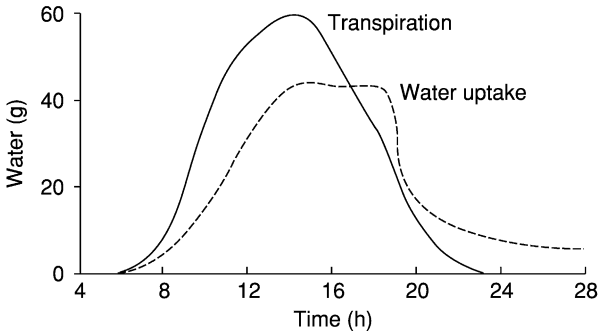


Figure 2 Daily cycle of plant transpiration and root water uptake.

considering a one-dimensional, steady-state flow in a series network, the liquid flow equation is:

$$T = -\frac{\Delta h_{\text{total}}}{R_{\text{total}}} = -\frac{(h_{\text{root}} - h_{\text{soil}})}{R_{\text{soil}}} = -\frac{(h_{\text{leaf}} - h_{\text{root}})}{R_{\text{plant}}} = \frac{h_{\text{leaf}} - h_{\text{soil}}}{R_{\text{soil}} + R_{\text{plant}}} \quad [1]$$

where T (centimeters per second) is the transpiration rate, h_{soil}, h_{root}, and h_{leaf} (centimeters) are pressure heads in the soil, at the root surface and in the leaves, respectively, and R_{soil} and R_{plant} (seconds) are liquid flow resistances of the soil and the plant, which are assumed to be constant; hence R_{plant} does not include stomatal resistance. When the transpiration demand of the atmosphere on the plant system is too high or when the soil is rather dry, R_{soil} and R_{plant} influence h_{leaf} in such a way that transpiration is reduced by closure of the stomata.

Equation [1] can be applied to the root system as a whole by measuring T, h_{soil}, and h_{leaf} during two periods, thus obtaining two equations with two unknowns, from which R_{soil} and R_{plant} can be computed. The relative importance of R_{soil} and R_{plant} is an important object of study. Under wet conditions R_{soil} is close to zero. Generally one can state, except for very dry soil, that $R_{\text{plant}} > R_{\text{soil}}$. Most of the plant resistance is concentrated in the roots, to a lesser extent in the leaves, and a minor part in the xylem vessels. Some researchers have concluded that R_{plant} increases with progressive drying of the soil and decreases when the transpiration rate is higher (**Figure 3**). It must be emphasized that this is a controversial issue, as other researchers think that R_{plant} depends neither on soil water content nor on transpiration rate.

The uptake of water and nutrients by plant roots is not always a passive process, it can be actively regulated by the plant. Water often discharges from the cut surface of active roots even if they are growing in unsaturated soil. If a manometer is attached to such a root, the pressure is usually high (0.1–0.2 MPa

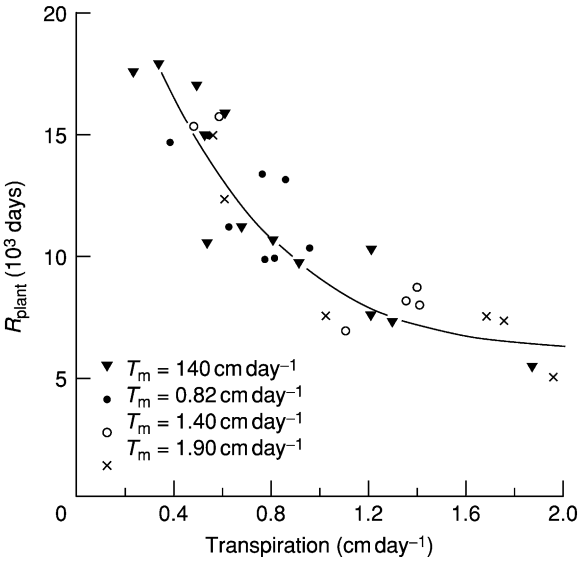


Figure 3 Plant resistance R_{plant} versus transpiration rate T as computed from an Italian ryegrass column experiment, showing the inverse proportionality of R_{plant} with T. (Reproduced from Hansen GK (1974) Resistance to water flow in soil and plants, plant water status, stomatal resistance and transpiration of Italian ryegrass as influenced by transpiration demand and soil water depletion. *Acta Agriculturae Scandinavica* 24: 83–92.)

(1–2 bar)), sometimes even more. This root pressure arises as a result of osmotic effects. The root surface can be considered as a semipermeable membrane: permeable for water in two directions while solutes are held back inside the root. Accumulation of solutes in the root xylem causes a very low osmotic potential. As a result water flows into the roots and a positive root pressure develops in the root xylem, resulting in the discharge from cut surfaces of active roots. Such a root pressure can be high enough to carry water up to a height of 10–20 m. Therefore it has been commonly thought that this active process is the only mechanism that supplies water to the tops of trees. However, if there is a demand for water by the atmosphere, the root pressure is soon unable to satisfy it. The driving force is then no longer the pumping action of the roots, but the sucking action of the leaves, after they have lost water to the atmosphere. This sucking action is more powerful than the force from a mechanical pump could ever produce, as it arises from the chemical potential of water in the presence of solutes. The positive pressure in the xylem vessels of the root then changes to a negative pressure, which may be lower than −1 MPa (−10 bar). The result is that water flows by a passive process from the soil to the xylem vessels in the vascular cylinder. In conclusion: root water uptake due to the osmotic effect plays no essential role in the water economy of plants.

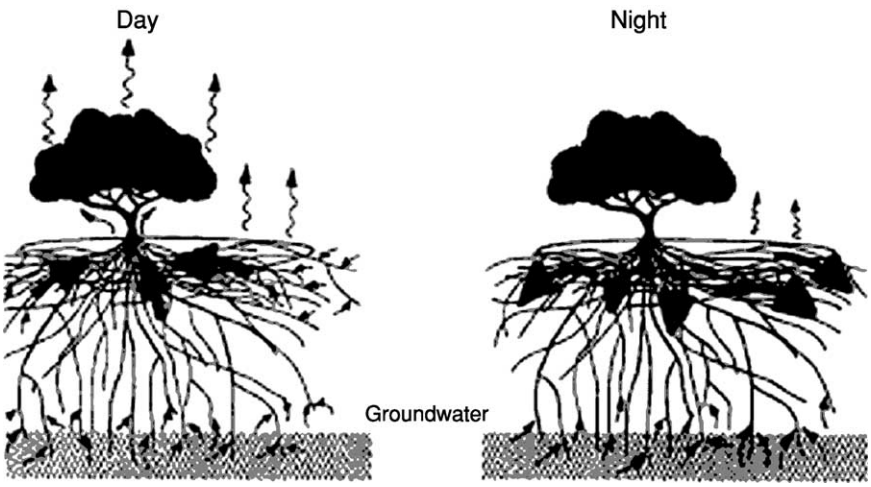


Figure 4 Mechanism of hydraulic lift by plants during night and subsequent uptake during daytime. (Reproduced from Dawson TE (1977) Determining water use by trees and forests from isotopic, energy balance and transpiration analysis: the roles of tree size and hydraulic lift. *Tree Physiology* 16: 263–272.)

In tall trees the xylem vessels are so strong that they do not collapse when the internal equivalent pressure (a potential) is appreciably lower than the external pressure: possibly more than 1.5 MPa (15 bar!) below the atmospheric pressure of 0.1 MPa (1 bar). Water has a high tensile strength, so when the chains of water molecules are drawn toward the leaves, they hold together against the frictional drag imposed by the walls of the conduit. As a result of this cohesion, cavitation does not occur, not even at very low equivalent pressure. In other words, the water is prevented from boiling. The plant exploits this cohesion of water, to transport water through the xylem vessels at very low equivalent pressures. It is this cohesion that enables trees to grow higher than the critical 'pump lift height' of 10 m.

A potentially important effect that happens in the real world but that models overlook is the marked influence plant roots can have on the distribution and redistribution of soil water via the processes of 'hydraulic lift' (Figure 4). During the day, water is absorbed at all depths and passes into the main transpiration stream. During the night, in the case of hydraulic lift, deep-rooted herbs, grasses, shrubs, and trees take in water from deeper, moist soil layers and exude that water during the night into the drier, upper soil layers. For example, sugar maple trees can lift hydraulically 100 l of water through their root systems into the upper soil layer during one night. This water is then absorbed the next day and transpired. But neighboring plants may also utilize this source of water. The result of hydraulic lift is usually a decline in groundwater table depth as well as stream discharge, compared with vegetation systems where hydraulic lift is absent.

Measurement of Root Water Uptake

Root water uptake can generally be derived by applying the water conservation equation to a given volume of soil. Commonly it is assumed that, in an unsaturated soil, water flows only in the vertical direction z. Consider a soil volume of unit cross-sectional area in the horizontal plane, bearing vegetation, with a lower boundary at $z = 0$ (for example, at the groundwater table or a level with constant pressure head) and an upper boundary at the soil surface at $z = z$ (positive upwards). This soil volume is partly explored by roots and partly without roots. The water conservation equation can then be written for every height z and for a short time interval $t_2 - t_1$ as:

$$\int_0^z \bar{S}_z \mathrm{d}z = \bar{v}_0 - \int_0^z \frac{\partial \theta}{\partial t} \mathrm{d}z - \bar{v}_z$$
$$\bar{v}_{\mathrm{root}} = \bar{v}_{\mathrm{total}} - \bar{v}_{\mathrm{soil}} \qquad [2]$$

where $\bar{S}_z$ (cubic centimeter per square centimeter per day) is the time-averaged volume of water taken up by the roots per unit bulk volume of the soil in unit time and considered positive from the soil into the roots, $\bar{v}_0$ (centimeters per day) is the time-averaged volumetric flux density through the lower boundary, θ (cubic centimeters per cubic centimeter) is the volumetric soil water content and $\bar{v}_z$ (centimeters per day) the time-averaged volumetric flux density through the upper boundary. The latter can be calculated from Darcy's law as:

$$\bar{v}_z = -k(h)\left(\frac{\partial h}{\partial z} + 1\right) \qquad [3]$$

where $k(h)$ is the hydraulic conductivity of the soil (centimeters per day) and h is the soil water pressure

head (centimeters). For the calculation of root water uptake patterns, the $k(h)$ curve and the profiles of θ and h have to be known. When the moisture-retention curve is used, it suffices to measure either θ or h and infer the unknown one via the retention curve. The flux at the bottom $\bar{v}_0$ should be derived either from measurements (e.g., using a lysimeter) or by means of eqn [3]. Application of eqn [2] gives the integrated root water uptake over a given depth interval in the root zone. Differentiation of eqn [2] gives the water uptake rate by the roots at depth z.

An example is shown in Figure 5, where a calculation of water extraction by roots at depth z is given for red cabbage on clay for a period of 7 days. The mean pressure head in this period is given in Figure 5a; the time-averaged cumulative withdrawal patterns of both crop and soil, $\bar{v}_{total}$, soil only, $\bar{v}_{soil}$, and crop only, $\bar{v}_{roots}$ are presented in Figure 5b, while the time-averaged extraction rate at depths z, $S(z)$, is given in Figure 5c. From Figure 5b and c it follows that the effective rooting depth can be found at the depth where the cumulative withdrawal function of total flow and soil flow intercept. From Figure 5c it can be seen that for the considered period the maximum extraction rate occurs at a depth of 0.30–0.35 m below surface.

The main conclusions of many reported root water uptake measurements on different vegetation can be elucidated with Figure 6, which shows at various times the root water uptake pattern of a red cabbage crop growing on sticky clay in the presence of a 90-cm-deep groundwater table. The magnitude of the root extraction rate is generally small at the top of the profile (unless the soil is wet just after rainfall or irrigation). Downward the extraction rate increases to a certain maximum and next decreases to zero at the root zone bottom. As the soil dries, the zone of maximum root

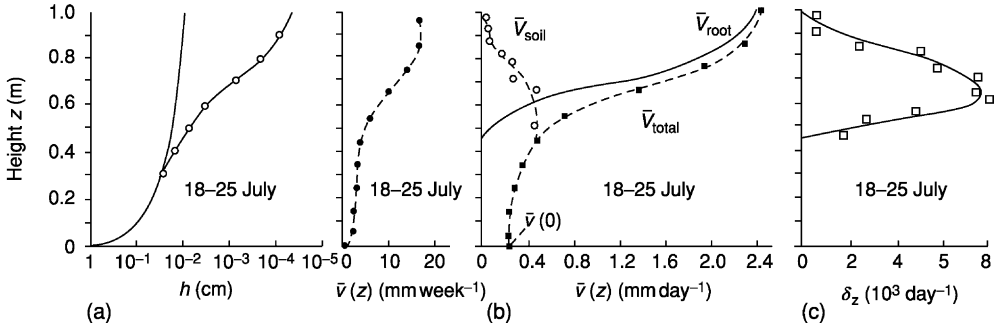


Figure 5 Profiles of: (a), pressure head h; (b), time-averaged cumulative water withdrawal of both a cabbage crop and clay soil, $\bar{v}_{total}$, of soil only, $\bar{v}_{soil}$, and of roots only, $\bar{v}_{root}$; and (c) time-averaged extraction rate S for red cabbage on clay for the period 18–25 July 1967. (Reproduced with permission from Feddes RA (1971) *Water, Heat and Crop Growth*. Dissertation. Wageningen, The Netherlands: Wageningen University Press.)

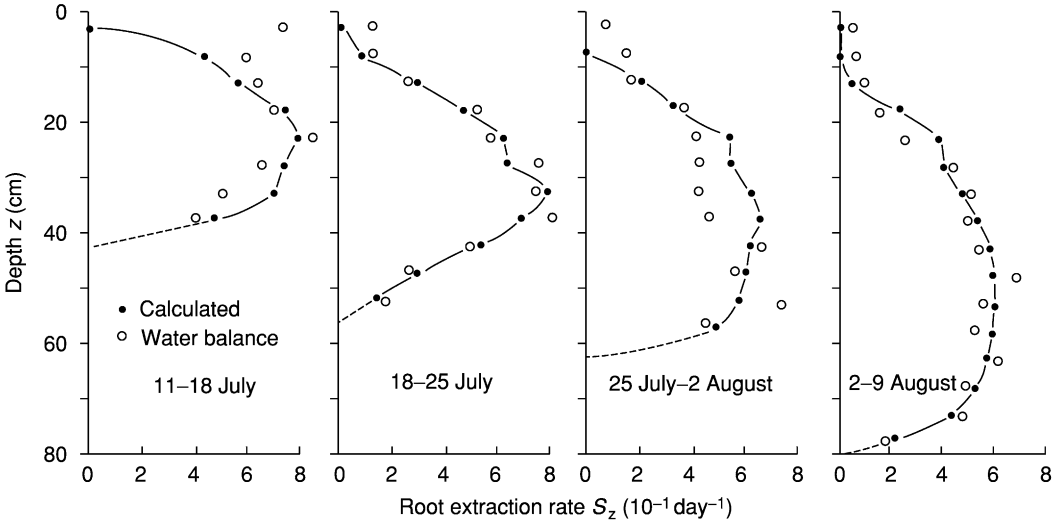


Figure 6 Example of measured variations of root water uptake with depth and time of red cabbage grown on a clay soil with groundwater table at 90–110 cm depth and obtained from water balance studies over 4 consecutive weeks. (Reproduced with permission from Feddes RA (1971) *Water, Heat and Crop Growth*. Dissertation. Wageningen, The Netherlands: Wageningen University Press.)

water uptake moves from shallow to larger depths in dynamic correspondence with the downward progression of roots into deeper, moist soil. The maximum extraction rate appears to depend on the demand the atmosphere exerts on the plant system, on the depths to which the roots penetrate, and on the soil moisture pressure-head distribution. Later in the season, water uptake from the upper layers becomes relatively less important. Most of the water is absorbed from the zone of low tension near the water table. Thus, a relatively small part of the root system can be responsible for most of the plant water uptake.

We may conclude that root water uptake depends on a number of factors such as soil hydraulic conductivity, rooting depth, root density distribution, soil moisture pressure head, demand set by the atmosphere ('potential' transpiration) on the plant system, and the presence of a water table. This list indicates that it is not simple to model water uptake by roots, nor to generalize on the effect of a single modification at the root zone.

Modeling Root Water Uptake in Ecologic, Hydrologic, and Atmospheric Communities

Ecologic Communities: Detailed Radial Root Model

A detailed, plant-based approach is to consider the convergent radial flow of soil water toward and into a representative individual root, taken to be a line or narrow-tube sink. Uniform thickness and absorptive properties along its length are assumed. The root system as a whole can then be described as a set of such individual roots, assumed to be regularly spaced in the soil at definable distances that may vary within the soil profile. The flow equation is cast in cylindrical coordinates and is solved for the distribution of soil-water pressure heads, water contents, and fluxes from the root outward.

In the single-root radial model, the root is viewed as a hollow cylinder of uniform radius and infinite length, having uniform water-absorbing properties. The governing Darcy flow equation can be written in radial coordinates as:

$$\frac{\partial\theta}{\partial t} = \frac{1}{r}\frac{\partial}{\partial r}\left[rk(h)\frac{\partial h}{\partial r}\right] \qquad [4]$$

where r is radial distance from the center of the root (centimeters). For steady-state conditions ($\partial\theta/\partial t = 0$) with water flowing from a distance, r_2, to a root with radius r_1, the solution under the assumption of constant k is:

$$q_{root} = \frac{2\pi k(h_{soil} - h_{root})}{\ln(r_2/r_1)} \qquad [5]$$

where q_{root} (cubic centimeters per centimeter per day) is the rate of water uptake per unit length of root, h_{soil} (centimeters) is the soil-water pressure head and h_{root} is the pressure head in the soil at the root surface (centimeters). Equation [5] is in hydrology known as the steady-state flux toward a well per unit length of well. For a discrete soil layer of thickness Δz and a rooting length per unit volume of soil of L_{root} (centimeters per cubic centimeter), the water uptake rate Δv_{root} (centimeters per day) can be written as

$$\Delta v_{root} = L_{root}\Delta z q_{root} \qquad [6]$$

Within this soil layer located at depth z, eqn [6] can be written as:

$$\Delta v_{root} = \frac{2\pi}{\ln(r_2/r_1)} L_{root} k(h_{soil} - h_{root} + z)\Delta z \qquad [7]$$

or similarly:

$$\Delta v_{root} = BL_{root} k(h_{soil} - h_{root} + z)\Delta z \qquad [8]$$

where $B = 2\pi/\ln(r_2/r_1)$ and represents a dimensionless geometry and root distribution factor.

Due to root shrinkage in a drying soil, the surface area of roots in contact with soil decreases and the root membrane resistance increases, and thus a kind of 'contact' resistance may develop. This resistance can be taken into account by assuming that root water uptake is proportional to the relative saturation, θ/θ_{sat}, of the soil. Hence:

$$\Delta v_{root} = BL_{root} k\frac{\theta}{\theta_{sat}}(h_{soil} - h_{root} + z)\Delta z \qquad [9]$$

Does the concept of contact resistance also hold for heavier-textured soils? **Figure 7** shows computed hydraulic conductivities of a soil root system according to eqn [9] and conductivities of the bulk soil. At low soil-moisture contents, both conductivities are comparable, but at higher moisture contents the hydraulic conductivity of the soil root system becomes not higher than 10^{-6} cm per day, far less than the hydraulic conductivity of the bulk soil. Hence, the use of soil hydraulic conductivity is not very realistic to describe root water uptake. It seems better to take an 'effective conductivity,' which represents the conductivity of the soil–root interface and the root tissue. Moreover, in modeling studies, h_{root} is often taken to be constant over the root zone, while water uptake is assumed to be potential until $h_{root} = -16\,000$ cm (wilting point).

In addition to an expression such as eqn [9], a second, similar type of equation can describe the flux over the root surface into the root using the difference between h_{root} and the pressure head inside the root,

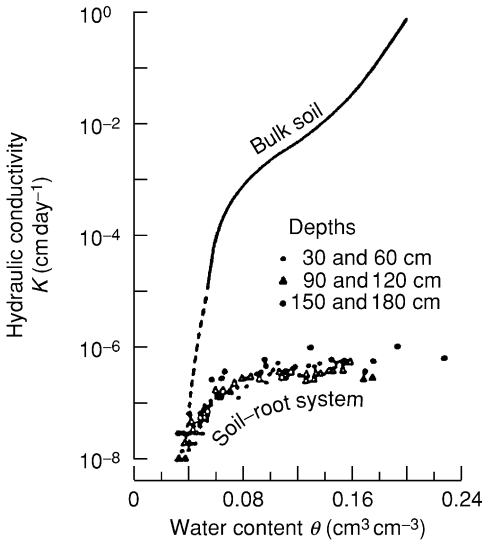


Figure 7 A comparison of the hydraulic conductivity of a soil–cotton root system obtained at various depths with the soil hydraulic conductivity (K) for various soil moisture contents. (Reproduced from Taylor HM and Klepper B (1975) Water uptake by cotton root systems: an examination of assumptions in the single root model. *Soil Science* 120: 57–67.)

h_{plant} (assumed to be uniform over the whole root system), and the hydraulic conductivity of the root. By equating the flux densities toward and into the root, h_{root} and h_{plant} can be derived by a number of iterations.

Hydrologic Communities: General Root System Model

The second, more hydrologically oriented approach is to regard the root system as a diffuse sink that penetrates each layer of soil uniformly, though not necessarily with a constant strength throughout the root zone. Root water uptake can then be represented as a sink term that is added to the vertical water flow equation through the soil. This macroscopic way of solving the root water uptake problem is to combine the continuity equation of water flow with a sink term representing water extraction by plant roots:

$$\frac{\partial\theta}{\partial t} = -\frac{\partial v}{\partial z} - S \qquad [10]$$

where θ (cubic centimeters per cubic centimeter) is the soil water content, t (days) is time, v_z (centimeters per day) is the Darcian soil water flux density (eqn [3]) taken positively upward, and S (cubic centimeters per cubic centimeter per day) is the root water uptake rate, which depends on depth, time, soil-water pressure head, and root density.

Combination of eqns [3] and [10] results in the physics-based soil water flow equation, known as the Richard equation:

$$\frac{\partial\theta}{\partial t} = C(h)\frac{\partial h}{\partial t} = \frac{\partial\left[k(h)\left(\frac{\partial h}{\partial z}+1\right)\right]}{\partial z} - S(z) \qquad [11]$$

where C is the differential water capacity ($d\theta/dh$) (per centimeter).

Under optimal moisture conditions, the maximum possible root water extraction rate $S_p(z)$, integrated over the rooting depth, is equal to the potential transpiration rate, T_p (centimeters per day), which is governed by atmospheric conditions. In many applications a homogeneous root distribution over the rooting depth is assumed, which gives:

$$S_p = \frac{T_p}{|z_{root}|} \qquad [12]$$

where T_p is the potential transpiration rate (centimeters per day) and $|z_{root}|$ is the root zone depth (centimeters). The potential root water extraction rate at a certain depth, $S_p(z)$ (per day), may for nonhomogeneous root distributions be determined by the root length density $L_{root}(z)$ (centimeters per cubic centimeter) as fraction of the total root length density over the rooting depth $|z_{root}|$ (centimeters), according to:

$$S_p(z) = \frac{L_{root}(z)}{\int_{-z_{root}}^{0} L_{root}(z)\partial z} T_p \qquad [13]$$

So far root water uptake under optimal soil-water conditions, S_p, has been considered. Under nonoptimal conditions, i.e., too dry, too wet, or too saline, a reduction is applied. A much-used reduction function for water stress is shown in Figure 8. Water uptake greater than h_1 (oxygen deficiency) and less than h_4 (wilting point) is set equal to zero; between h_2 and h_3 (reduction point), water uptake is maximal. Between h_1 and h_2, a linear variation is assumed, and between h_3 and h_4 either a linear (Figure 8) or hyperbolic variation is assumed. The value of h_3 depends on the demand of the atmosphere and thus varies with T_p.

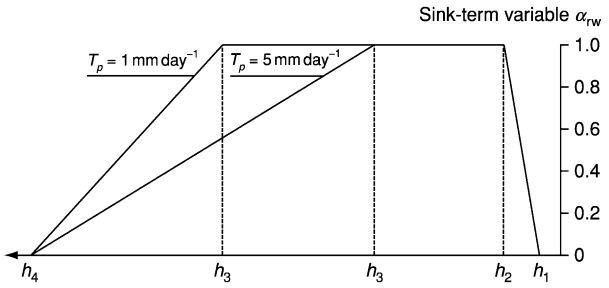


Figure 8 Dimensionless sink term variable α as a function of the soil-water pressure head h. (Source: Feddes RA, Kowalik PJ, and Zaradny H (1978) *Simulation of Field Water Use and Crop Yield*. Simulation Monographs. Wageningen, The Netherlands: Pudoc.)

For an indication of these h_1–h_4 values for five crops, see Table 1.

For many crops the reduction in crop yield due to salinity can be linearly related to the electrical conductivity of the soil solution. Crops tolerate an increase in the soil salinity up to a threshold value above which relative yields Y/Y_m and thus relative transpiration T/T_m show an approximately linear decrease with increasing salt concentration or electrical conductivity EC_e (Figure 9).

In order to simplify parameter calibration and take advantage of existing experimental data, we may assume the water and salinity stress to be multiplicative. This means that the actual root water flux density, $S(z)$ (per day), is calculated from:

$$S(z) = \alpha_{rw}(z)\alpha_{rs}(z)S_p(z) \qquad [14]$$

where α_{rw} (−) and α_{rs} (−) are the reduction factors due to water and salinity stresses respectively. Integration of $S_a(z)$ over the rooting depth then yields the actual transpiration rate T_a.

Numerical solution of eqn [11] requires knowledge of the soil hydraulic functions of which various databanks exist in literature for all kind of soil textures. Also data are needed for the initial moisture profile and during the simulation period of the top and bottom boundary conditions of the flow system. At the top the vegetation plays a dominant role in the partitioning of the various fluxes. Hence one needs in principle a coupling of the soil water balance model with a daily vegetation growth model. Only in this way can a proper prediction of vegetation development and growth in dependency of the actual prevailing soil water conditions be obtained, thus assuring proper feedback.

Table 1 Critical pressure head values of the sink-term function depicted in Figure 8 for some main crops

Crop	h_1	h_2	$h_{3,high}$	$h_{3,low}$	h_4
Potatoes	−10	−25	−320	−600	−16 000
Sugar beet	−10	−25	−320	−600	−16 000
Wheat	0	−1	−500	−900	−16 000
Pasture	−10	−25	−200	−800	−8000
Corn	−15	−30	−325	−600	−8000

Reproduced with permission from Wesseling JG (1991) Meerjarige simulaties van grondwateronttrekking voor verschillende bodemprofielen, grondwatertrappen en gewassen met het model SWATRE (in Dutch). Report 152. Wageningen, The Netherlands: Winand Staring Centre Press.

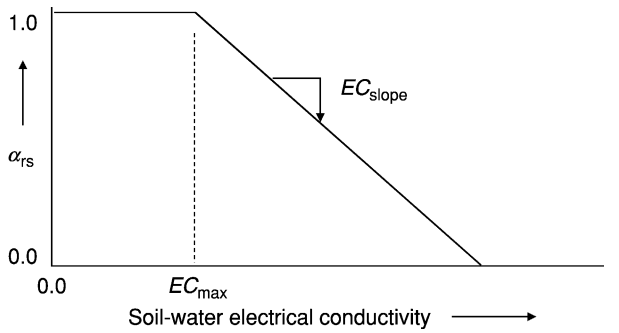


Figure 9 Reduction coefficient for root water uptake, α_{rs}, as a function of soil-water electrical conductivity, EC. (Source: Maas EV and Hoffman GJ (1977) Crop salt tolerance – current assessment. *Journal of the Irrigation and Drainage Division* 103: 115–134.

Atmospheric Communities: Large-Scale Root System Model

In global circulation models (GCM), land surface parameterizations are often based on the concept of one big leaf per grid element. However, at the resolvable scale of GCM models, land surfaces are very heterogeneous. On the other hand, various soil–vegetation–atmosphere-transfer (SVAT) schemes have been developed for use in GCM and numerical weather-prediction models. Their weakest component, however, remains their link with the lower boundary. SVAT models, which sometimes may be overparameterized, use a variety of different methods to represent the relationship between roots, soil moisture, and transpiration. Moreover SVAT parameters are generally highly variable in space and difficult to measure. Different SVAT models driven by the same meteorological forcing of air temperature, humidity, wind speed, incoming solar radiation, long-wave radiation, and rainfall can produce remarkably different surface energy and water balances. The question in this context is therefore: What is the role of roots?

With agrohydrologic models, the effect of root distribution on the course of actual transpiration in time can be evaluated. Take, for example, a grass vegetation (covering the soil completely), with a rooting depth of 80 cm growing on a loamy sand containing 10% clay (less than 2 μm). At the bottom, free drainage prevails; and initially throughout the profile the soil-water pressure head $h = -200$ cm, implying a rather wet soil. If at the soil surface a potential transpiration rate $T_p = 4$ mm day^{-1} is applied for two different relative root density distributions: root 1, where most of the roots are located in the topsoil, and root 2, with a uniform root distribution (Figure 10a), the results of this simulation are as shown in Figure 10b. The moisture content of the top soil layer decreases more rapidly than the moisture content of deeper soil layers. Hence root 1 produces an earlier onset of moisture stress than root 2, after 30 days showing an actual transpiration rate that is about half that of root 2. This clearly demonstrates that roots influence the behavior of a land surface model, the role of

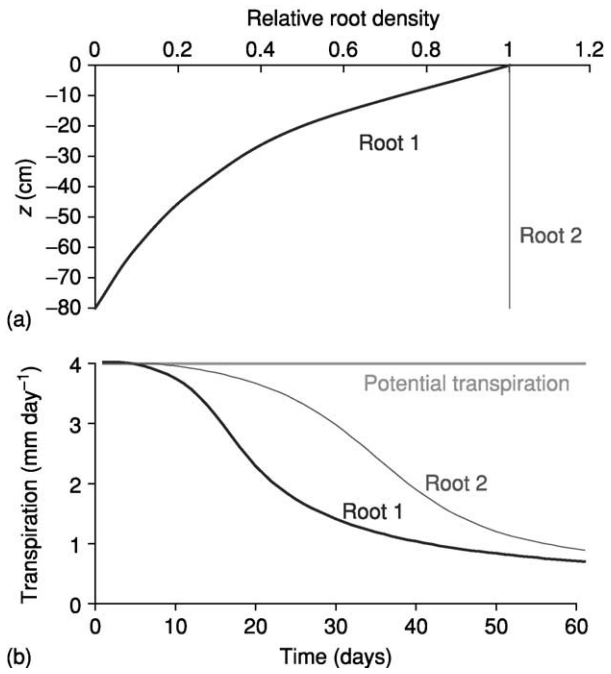


Figure 10 (a) Two different relative root density distribution functions adopted for grass, with a rooting depth of 80 cm, growing on loamy sand; (b) simulated actual transpiration rate in time for the two root-distribution functions, taking $T_p = 4$ mm per day as the upper boundary condition.

roots being particularly important when soil moisture limits evapotranspiration.

See also: **Crop Water Requirements**; **Desertification**; **Irrigation:** Environmental Effects; **Plant–Water Relations**; **Root Architecture and Growth**; **Soil–Plant–Atmosphere Continuum**

Further Reading

Dane JH and Topp GC (eds) (2002) *Methods of Soil Analysis, part 4, Physical Methods*. Soil Science Society of America Book Series, No. 5. Madison, WI: Soil Science Society of America.

Feddes RA (1981) Water use models for assessing root zone modification. In: Arkin GF and Taylor HM (eds) *Modifying the Root Environment to Reduce Crop Stress*, pp. 347–390. ASAE Monograph 4. St. Joseph, MI: ASAE.

Feddes RA, Kowalik PJ, and Zaradny H (1978) *Simulation of Field Water Use and Crop Yield. Simulation Monographs*. Wageningen, The Netherlands: Pudoc.

Gardner WR (1960) Dynamic aspects of water availability to plants. *Soil Science* 89: 63–73.

Homaee M (1999) *Root Water Uptake Under Non-uniform Transient Salinity and Water Stress*. PhD Thesis. Wageningen, The Netherlands: Wageningen University Press.

Molz FJ (1981) Models of water transport in the soil-plant system: a review. *Water Resources Research* 17: 1245–1260.

Raats PAC (1974) Steady flows of water and salt in uniform soil profiles with plant roots. *Soil Science Society of America Proceedings* 38: 717–722.

Scott HD (2000) *Soil Physics. Agricultural and Environmental Applications*. Ames, IA: Iowa State University Press.

Tanji KK (1990) *Agricultural Salinity Assessment and Management*. Manuals and Reports, No. 71. New York: American Society of Civil Engineers.

van Dam JC, Huygen J, Wesseling JG *et al.* (1997) *Theory of SWAP, Version 2.0. Simulation of Water Flow, Solute Transport and Plant Growth in the Soil–Water–Atmosphere–Plant Environment*. Report 71, Department of Water Resources. Wageningen, The Netherlands: Wageningen University Press.

Warrick AW (ed.) (2002) *Soil Physics Companion*. New York: CRC Press.

Willigen P de and van Noordwijk M (1995) A model for interactions between water and nutrient uptake. In: Kabat P, Marshall B, van den Broek BJ, Vos J, and van Keulen H (eds) *Modelling and Parameterization of the Soil–Plant–Atmosphere System: A Comparison of Potato Growth Models*, pp. 135–153. Wageningen, The Netherlands: Wageningen Press.

PLANT–WATER RELATIONS

C Gimenez, University of Cordoba, Cordoba, Spain
M Gallardo and R B Thompson, University of Almeria, Almeria, Spain

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Water constitutes the largest single chemical component of plants, yet the volume within the plant is very small in relation to the total volume transpired. Adequate plant water status requires that root uptake of soil water satisfies the atmospheric demand (i.e., transpiration requirements). When water uptake is insufficient, a plant water deficit can develop. Various parameters are used to measure plant water status, the most common being water potential. Numerous plant processes are affected by declining plant water status, the most sensitive being cell expansion, which affects new growth. Stomatal closure is another sensitive process. Plants' responses to declining soil water content are mediated by lowered leaf water potential and/or by chemical agents produced in roots. Some parameters of plant water status can be used for irrigation-scheduling purposes.

Water in Plants

In herbaceous plants, water normally constitutes more than 90% of fresh weight, although in rare cases it can be less than 70%. In woody plants, over 50% of fresh weight consists of water. Of the total water content of plants, 60–90% is located within cells, the rest (10–40% of total water) is mainly in cell walls. The water in cell walls forms a continuum with specialized transport cells throughout the plant.

The total volume of water within a plant is very small in relation to the total volume of water transpired during its lifetime. Even on a daily basis, the volume of water within plants is insufficient to buffer appreciably daily transpiration requirements on a warm, sunny day. The very large amounts of water transpired by plants, in relation to that retained within plant tissue, can be viewed as the 'cost' that plants incur as a consequence of stomatal opening to allow CO_2 absorption for photosynthesis.

A primary function of water contained in cells is the maintenance of cell and tissue turgor. Cell turgor is essential for cell enlargement and therefore for optimal plant growth. Other primary functions are the transport of solutes and participation in metabolic activities. Because of its high dielectric constant, water acts as a solvent for many mineral and organic solutes, enabling their transport within cells and throughout the plant. Also, water is directly involved in chemical reactions in cells such as CO_2 reduction in photosynthesis. Cooling is another primary function of water in plants. Because of the high energy requirement for water vaporization (10.5 kJ mol^{-1} at 25°C), water evaporating from leaf surfaces (during transpiration) cools the leaf, thereby avoiding excessive daytime heating from incoming solar radiation.

The water status of plants is a primary determinant of plant growth and development, and therefore of crop productivity in agricultural systems, and of plant survival in natural systems. Almost every plant physiological process is directly or indirectly affected by plant water content. For example, cell enlargement is dependent on the level of cell turgor, photosynthesis is directly inhibited by insufficient water, and stomatal control of transpiration and CO_2 absorption is dependent on the water status of stomatal guard cells.

The water status of plants is the sum of the interaction of various atmospheric, plant, and soil factors. The availability of soil water, the atmospheric demand (determined by radiation, humidity, temperature, wind), the capacities of the root system to absorb water and of the plant to transport absorbed water to transpiring leaves, and stomatal responses for regulating transpiration, can all appreciably influence plant water status.

Plant water status is commonly characterized by its water potential (Ψ). Water potential is a measure of the free energy status of water, which, because of its applicability to each component of the soil–plant–atmosphere system, enables water movement between these components to be considered. It theoretically represents the work involved in moving one mole of water from a selected point within the plant (or soil) to a reference point of pure water at the same temperature and at atmospheric pressure. Ψ varies from zero at the reference point to negative values within the plant and soil. It is normally measured in units of pressure, with megapascals (MPa) being most commonly used.

Leaves are the plant organs where most of the exchange of CO_2 and H_2O between the plant and the atmosphere occurs. The pathway for the inward diffusion of CO_2 is much the same as that for the outward diffusion of H_2O vapor. The outward diffusion of H_2O vapor from the saturated surfaces within the

plant to the drier atmosphere follows a gradient of the partial pressure of H_2O vapor. To maximize CO_2 fixation by photosynthesis, stomata must remain open for as long as possible during daylight periods. This also maximizes the period of water loss by transpiration.

The water status of leaves (considered as Ψ) is the balance between the water lost to the atmosphere by transpiration (T) and the water absorbed by the plant from soil, which is a function of soil water potential (ψ_{soil}) and the combined resistance to water movement within the roots and shoots (r). These relationships are described by the equation:

$$\Psi = \psi_{soil} - T\ r \qquad [1]$$

Even in saturated soils (where $\psi_{soil} = 0$), Ψ is negative when transpiration occurs. During daylight periods, absorption of water lags behind transpiration owing mainly to the high resistance to water flow from soil into root xylem tissue. As atmospheric evaporative demand increases during the morning, transpiration increases, which lowers the water potential of cells from which water is evaporating. Within the plant, water then moves from nonevaporating parenchyma cells of leaves, which have a higher Ψ, toward the evaporating cells, establishing a Ψ gradient. This gradient is transmitted throughout the plant–soil system, enabling continuous water movement. In the afternoon, transpiration decreases on account of reduced atmospheric evaporative demand. However, water uptake by roots continues until parenchyma cells fully rehydrate, and their Ψ equals soil Ψ, which usually occurs during the night. At this stage, plant and soil water are in equilibrium, and absorption by roots ceases. (*See* **Soil–Plant–Atmosphere Continuum**.) In some species under certain climatic conditions (high nighttime vapor pressure deficit, wind), transpiration can occur at night; when it does, it is generally relatively small compared to daytime transpiration. However, it can be sufficient to prevent nighttime equilibration of plant and soil water potentials.

Plant Water Status Indicators

Direct Indicators of Plant Water Status

Relative water content The water content of plant tissue is the fraction of the total fresh tissue weight that is water. It can be considered either on a fresh- or dry-weight basis. As there can be considerable variation between different plant tissues and organs in the maximum water content, the normalized parameter relative water content (RWC) is used. RWC is the water content of plant tissue or a plant organ relative to its maximum water content when it is fully hydrated (i.e., saturated). RWC provides a measure of the degree of hydration of tissue, which is considered to have an important controlling role on numerous plant functions.

Relative water content is expressed as:

$$\mathrm{RWC} = (\mathrm{FW} - \mathrm{DW})/(\mathrm{SFW} - \mathrm{DW}) \qquad [2]$$

where FW is the fresh weight (grams), DW the dry weight (grams), and SFW is the saturated fresh weight (grams).

Water potential and components The most commonly used parameter to characterize plant water status is water potential (Ψ). Total water potential (Ψ) has four components: the osmotic potential (ψ_s), pressure potential (ψ_p), matric potential (ψ_m), and gravitational potential (ψ_g):

$$\Psi = \psi_s + \psi_p + \psi_m + \psi_g \qquad [3]$$

Osmotic potential (ψ_s) results from dissolved solutes in cell sap and is proportional to solute concentration and inversely proportional to cell water volume. ψ_s in plants is always negative and decreases as solutes concentrate during plant dehydration. Pressure potential is a measure of tissue turgor produced by the diffusion of water into the protoplast of cells enclosed by largely inelastic cell walls. Matric potential (ψ_m) arises from the action, on water, of electrostatic forces of attraction associated with cell wall and colloidal surfaces, and of capillary forces associated with narrow transport vessels. In plants it is considered to be negligible. Gravitational potential (ψ_g) results from gravitational forces acting on the water within plants. On account of the gradient of ψ_g with height of 0.01 MPa m^{-1}, ψ_g is normally negligible compared with the other components of plant Ψ; it is only considered to be significant in very tall trees. In most situations, total plant water potential is considered to be the sum of the pressure potential (ψ_p) and osmotic potential (ψ_s).

As both pressure and osmotic potential are dependent on tissue water content, there are relationships between RWC and ψ_s and ψ_p, and consequently between RWC and Ψ. These relationships are schematically represented in Figure 1. In a fully hydrated plant, RWC equals 1, ψ_p is positive, ψ_s is negative, and Ψ (the sum of ψ_p and ψ_s) equals zero. As RWC progressively declines, both ψ_p and ψ_s decline, and consequently Ψ becomes more negative. When RWC has declined sufficiently for ψ_p to equal zero, the plant loses turgor and wilts.

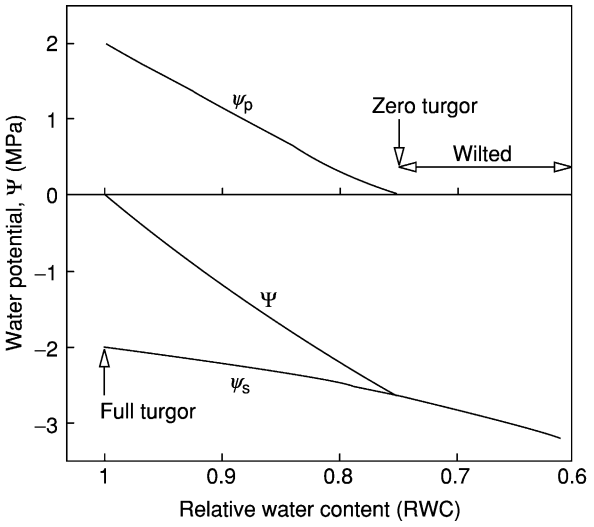


Figure 1 Höffler–Thoday diagram illustrating the relationships between total water potential (Ψ), pressure potential (ψ_p), osmotic potential (ψ_s), and relative water content (RWC) as a cell or tissue loses water from a fully turgid state. Full and zero turgor, and wilting are indicated by arrows. (Adapted from Jones HG (1992) *Plants and Microclimate*. Cambridge, UK: Cambridge University Press, with permission.)

Several methods are used to measure the total water potential of plant tissue. Thermocouple psychrometry and hygrometry are the most accurate methods, but both require very stable environmental conditions. Consequently, whilst they can be used under controlled conditions, they are very difficult to use in field studies. The pressure-chamber method is suitable for routine and rapid field measurement.

Pressure potential (ψ_p) can only be measured directly by the pressure microprobe, which is inserted into cell protoplasm. This technique is limited to relatively large cells and is generally restricted to laboratory measurement. In practice, the most commonly used procedure is to measure both Ψ and ψ_s, and then to calculate ψ_p as the difference between Ψ and ψ_s. Osmotic potential (ψ_s) is commonly measured by either thermocouple psychrometry or hygrometry after previously freezing and thawing the sample to break cell membranes and reduce ψ_p to zero; normally a correction is made for the dilution by apoplastic water. Another method for measuring ψ_s is by determining a moisture-release curve (also called pressure–volume curve) with a pressure chamber.

Indirect Indicators of Plant Water Status

Stomatal conductance Stomatal conductance (g_l) is a measure of the degree of stomatal opening and can be used as an indicator of plant water status. Stomatal conductance is related to leaf Ψ by feedback processes. Reductions in g_l prevent further decreases in Ψ by reducing transpiration; also, reductions in Ψ can induce stomatal closure, resulting in lowered g_l. Stomatal responses are discussed more fully in the section Stomatal Responses, below. Stomatal conductance can be measured with both dynamic and steady-state diffusion porometers.

Stem diameter variations Transpiring plants and trees undergo diurnal variation in stem or trunk diameter. During the morning, when plants are most vigorously transpiring, some water from stem tissue, mostly from phloem tissue, is incorporated into the transpiration stream, producing a measurable reduction in stem diameter. During the afternoon, as transpiration slows, this tissue begins to rehydrate, with a corresponding increase in stem diameter. Rehydration continues throughout the night as some water uptake continues from soil, when transpiration has usually completely ceased. Commonly, stem diameter has a maximum daily value just before dawn, and a minimum daily value in the early afternoon. The magnitude of the diurnal contraction is dependent on atmospheric evaporative demand and plant water status. Under given atmospheric conditions, the magnitude of daily stem contraction increases with increasing plant water deficit. Consequently, stem contraction can be used as an indicator of plant water status. In some fruit-tree and vegetable species, variations in stem diameter have been shown to be very sensitive measures of plant water status.

In young herbaceous plants and young trees, well-watered healthy plants show a trend of increasing daily maximum stem-diameter values, which is an indication of growth. In such plants, a reduction in the rate of growth is apparent in a change in the slope of daily maximum stem-diameter values, which may be due to a lack of water or also to inadequate crop management or changed environmental conditions. In more mature plants, daily maximum stem-diameter values tend to be more constant. Sensors incorporating sensitive pressure transducers, known as linear variable differential transducers (LVDT), connected to data loggers enable continuous monitoring of stem diameter.

Sap flow Sap-flow sensors measure transpiration flow as the ascent of sap within xylem tissue; measurements can be made in stems, trunks, branches, or tillers. Given that transpiration is sensitive to plant water status, with the effect being mediated by stomatal opening (see Stomatal Responses, below), sap flow can be used as an indicator of plant water status.

Sap flow rates can be out of phase with transpiration because of capacitance effects in stems or branches arising from the storage of water. Continuous data recording enables the time course of transpiration to be followed.

Two different techniques are used to measure sap flow; both use heat as a tracer. One is the stem (or trunk) sector heat-balance method, in which a section of the entire stem circumference is electrically heated, and the axial and radial heat-loss measured. The mass flow rate of sap is calculated as a function of the heat dissipated by the ascending sap. The other method is the heat-pulse method in which the heater and temperature sensor probes are placed inside the trunk in a radial direction. The sap velocity is calculated as a function of the time required by the flowing sap to transport heat to a particular location.

Leaf or canopy temperature When plant water status is adequate, canopy temperature (i.e., average temperature of total leaf surface) is appreciably less than the air temperature, through the cooling effect of evaporation. As stomata partially close in response to increasing plant water stress, the energy balance of the plant is altered; less heat is dissipated through the evaporation of water, and consequently canopy temperature increases. With increasing plant water stress, the difference between canopy and air temperature lessens and, under conditions of severe water stress, may become positive. Canopy temperature is most commonly measured with infrared thermometers.

Other Methods

As plant water status affects overall plant growth, measurable parameters such as fruit growth and expansive growth of leaves or stems have been used as indicators of plant water status. Visual characters such as leaf rolling, color change, and visible wilting have also been used. However, visual characteristics are normally an indication that the water stress is severe.

Water Deficits in Plants

Traditionally, 'plant water deficit' or 'plant water stress' has been defined as being when plant water status is reduced sufficiently to affect normal plant functioning (e.g., plant growth, stomatal conductance, rate of photosynthesis). It is not possible to define water deficit in terms of absolute values on account of the complex interaction of atmospheric, plant, and soil factors involved. Also, preconditioning effects from the prior history of the plant can modify plant response at a given water potential. Consequently, a commonly used approach is to express the measured values of the stressed plants in relation to those of well-irrigated plants which have experienced the same environmental conditions and previous general history.

Plant water deficits occur when any plant process is affected by:

1. Limited water absorption by roots because of dry, cool, or poorly aerated soil;
2. High evaporative demand, on account of low relative humidity, high air temperature, high wind speed, high radiation, or combinations of the four;
3. A combination of limited water absorption and high evaporative demand.

Generally, plant water deficits can be considered as being induced by either insufficient available soil water, or a high atmospheric evaporative demand. Plant water deficits induced by lack of soil water may continue for days, possibly weeks, until they are either alleviated by rain or irrigation or the plant dies. Those that are solely due to high atmospheric evaporative demand ('midday' water deficits) are much more transient, occurring for no more than several hours in the middle part of the day.

When not alleviated, water deficits induced by lack of soil water become progressively more intense. The time course of leaf and soil Ψ during a theoretical drying cycle of several days is shown in Figure 2. As soil continues to dry and soil Ψ to decrease, there is increasingly less available soil water for root uptake.

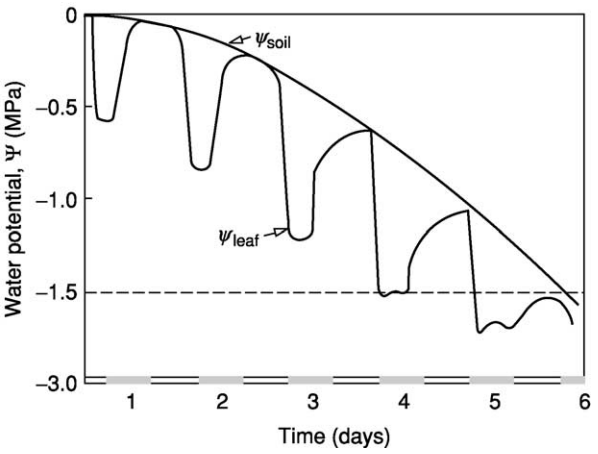
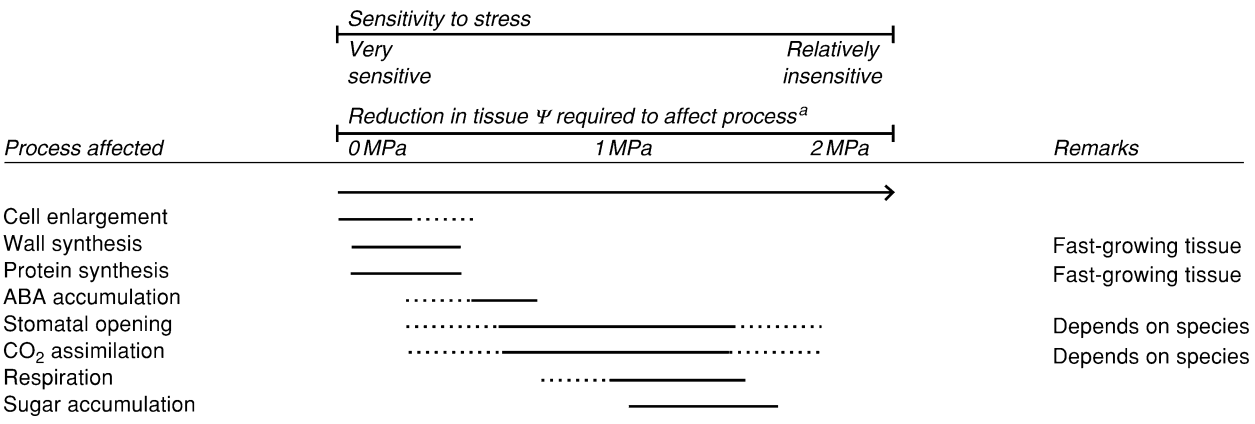


Figure 2 Time course of soil water potential (ψ_{soil}) and leaf water potential (ψ_{leaf}) during a drying cycle. Dashed line represents soil water potential at which permanent wilting of the plant occurs. (Adapted from Kramer PJ (1983) *Water Relations of Plants*. New York: Academic Press, with permission.)

Table 1 Plant process affected by water stress and their sensitivity. Length of the horizontal lines represent the range of stress levels within which a process becomes first affected. Dashed lines signify deductions based on more tenuous data



[a]With Ψ of well-watered plants under mild evaporative demand as the reference point.
Source: Hsaio TC (1973) Plant responses to water stress. *Annual Review of Plant Physiology* 24: 519–570.

This results in a general tendency for plant Ψ to decrease over time. Eventually, the plant is unable to absorb sufficient water for plant Ψ to equilibrate with soil Ψ at night. When plants are unable to recover cell and tissue turgor, which occurs at a threshold soil-water potential value (usually approximately −1.5 MPa), permanent wilting of the plant occurs.

Midday water deficits can even occur in plants growing in moist soil or nutrient solutions. They occur when high atmospheric evaporative demand causes high midday transpiration rates that exceed the rate of water uptake by roots. Where these situations are recurrent, daytime growth is inhibited by these water deficits, and most growth occurs at night, as long as nighttime temperatures are not limiting. Midday water deficits are not restricted to hot and dry environments; they have also been observed in the humid tropics and in greenhouses. Recurrent midday water deficits can reduce yield in cropping situations. In intensive greenhouse crop-production systems, fog systems are commonly used to prevent midday water deficits.

Plant processes affected by water deficits, and an indication of their sensitivity, are shown in Table 1. Processes such as cell enlargement, cell wall synthesis, and protein synthesis are very sensitive, being affected by relatively small reductions in Ψ. Other processes, such as stomatal opening and CO_2 assimilation, generally require larger reductions in Ψ. As plant water deficits intensify, an increasing number of plant processes are affected (Table 1), and there can be complex interactions between them. Two of these processes, expansive growth and stomatal responses, are discussed more fully below.

Expansive Growth

The plant process most sensitive to water deficit is cell enlargement, which affects expansive growth. The effects of plant water deficits on expansive growth are very important, because they result in reduced expansion of the plant assimilation surface (i.e., leaf area). The main factor controlling cell enlargement and therefore leaf expansion is turgor pressure (ψ_p). The relationship between ψ_p and cell expansion can be represented by the Lockhart equation:

$$dV/V\,dt = E_g(\psi_p - \psi_{p,th}) \qquad [4]$$

where the relative change in cell volume (V), with time, is dependent on ψ_p above a threshold value ($\psi_{p,th}$) below which no cell expansion occurs. E_g is the gross extensibility of the cell wall (i.e., its capacity to expand). Increases in cell wall extension during growth are irreversible. Values of $\psi_{p,th}$ are normally only slightly less than ψ_p values of nonstressed plants, so even small reductions in ψ_p can result in reduced growth. Neither E_g nor $\psi_{p,th}$ is constant: they can be influenced by previous exposure to water stress and by growing conditions. With the progressive imposition of water stress, plants can adapt by either reducing $\psi_{p,th}$, increasing E_g, or doing both in order to maintain growth at lower ψ_p.

Stomatal Responses

When leaf water status is adequate (i.e., well-watered plants) and atmospheric evaporative demand is low, stomatal opening is determined primarily by light conditions and the low CO_2 partial pressure of substomatal cavities. Under these conditions, the

stomata are fully open during daylight periods, maximizing assimilation of CO_2, thereby ensuring optimal rates of photosynthesis.

When plant water deficits develop and leaf Ψ decreases, partial stomatal closure occurs to reduce water loss from the plant, with a consequent reduction in photosynthesis. In many species, Ψ must decline to a threshold value before stomatal closure commences. Threshold Ψ values for stomatal closure vary with species, leaf age, previous exposure to radiation, the stress history of the plant, and environmental conditions. If the plant has previously been subjected to a progressive, mild water stress, it can be preconditioned so that the threshold Ψ value is lowered, thereby enabling assimilation to proceed at lower Ψ. Stomata can also close in response to low air humidity, even when leaf water status is adequate.

When recovering from water stress, stomata respond relatively slowly. Unlike Ψ, which rapidly recovers to nonstressed values once water stress is alleviated, stomatal conductance can take several hours, even days, to recover to nonstressed values.

Stomatal closure in response to plant water deficits is not solely controlled by turgor pressure of leaf cells. The process is complex, with the additional involvement of solute movement. When water deficits develop, stomatal guard cells simultaneously lose both water and cell solutes, particularly K^+ ions. It seems that water stress promotes the efflux of K^+ ions from guard cells, resulting in the loss of cell turgor, which induces stomatal closure. In the contrary situation, when water stress is relieved, synchronized changes in K^+ concentration and water content of guard cells have been detected prior to stomatal opening.

Root Signals

Traditionally it has been accepted that drying soil has a hydraulic effect on plant water relations and consequently on cell expansion and leaf gas exchange. The basic theory is that a reduction in soil water potential causes a reduction in plant pressure potential, inducing reduced cell expansion and stomatal closure. In recent decades, it has become clear that stomatal closure, and also leaf expansion, are controlled by a mechanism or mechanisms additional to changes in pressure potential of leaf cells. Reductions in stomatal conductance (g_l) of plants, in drying soil, have been related more strongly to changes in soil water status than to leaf water status. These observations indicate that plants can 'sense' that soil in the root zone is drying and can communicate this information to the leaves by a means other than reduced leaf water status. The evidence to date suggests that a chemical signal, abscisic acid (ABA), is produced in roots in drying soil and is transported to leaves in the xylem sap, where it induces stomatal closure before reductions in leaf water potential occur. Although ABA is produced in both roots and leaves, it is accepted that root tips are the major source of ABA produced in response to drying soil. The sensitivity of stomata to ABA is mediated by the nutritional and water status of the plant and can be genetically determined. The triggering of stomatal closure by chemical root signals has been observed in many agricultural species.

Much of the research work to date on chemical root signals has been conducted on the effects of ABA on stomatal responses. However, other plant growth processes sensitive to water stress, such as leaf expansion and leaf initiation, are also influenced by chemical root signals. For example, in some species, even when leaf water potential is artificially maintained, soil drying induces reductions in leaf elongation.

The relative degree of control of hydraulic and chemical root signals over stomatal conductance and leaf expansion, in field-grown plants, is the subject of considerable research and debate. In some studies, changes in ABA concentration in xylem sap do not correlate well with changes in stomatal conductance, suggesting that there may be chemical agents other than ABA that induce stomatal closure. As yet, no other such agents have been identified. Additionally, there is controversy associated with some data from studies conducted under controlled conditions, such as the induction of ABA production by mechanical impedance to root growth in potted plants, where the nature of the experimental conditions may contribute to ABA production.

It is now accepted that plant processes such as stomatal conductance and leaf expansion that are sensitive to water stress are affected by both chemical root signals and changes in leaf water status. Currently, research is being conducted to investigate possible interactions between hydraulic and chemical signals.

Applications of Plant–Water Relations to Irrigation Scheduling

The most common practical application of plant–water relations is to assist in irrigation management. Different plant water status indicators have been proposed for use in irrigation scheduling (IS). The

advantage of plant-based measurements over alternative IS methods based on soil-moisture monitoring or estimation of evapotranspiration (ET) requirements is that the actual unit of production (the plant) is being assessed, rather than using an environmental parameter (e.g., the soil) or a mathematical estimate (e.g., of ET).

Generally, indicators of plant water status are suitable for determining the timing of irrigation (i.e., when to irrigate), but not the amount of water to apply. An additional method, for example that of the Food and Agriculture Organization of the UN (FAO) for estimating crop evapotranspiration requirements (*See* **Crop Water Requirements**), can be used to estimate the amounts to apply. Interpretation of data of plant water status indicators requires comparison with previously established threshold values that define the need for irrigation (e.g., a maximum allowable daily stem-diameter contraction). Plant water status data are normalized to take into account variations in water status due to evaporative demand, i.e., to distinguish effects mostly due to insufficient soil water from those mostly due to atmospheric demand. Ideally, measurements of plant water status are considered in relation to those obtained from fully irrigated plants.

In some crops such as cotton, the timing of irrigation has been determined by predawn or midday measurements of leaf water potential. To use leaf Ψ for this purpose, the relationship between leaf Ψ and growth is established previously, for the individual species. In some fruit-tree species, stem Ψ has been used in preference to leaf Ψ, because variations in leaf Ψ, due to rapid changes in evaporative demand, are avoided. Stem Ψ is measured in leaves that have been previously covered to prevent transpiration.

Until recently, plant water status measurements were generally based on discrete manual measurements (e.g., leaf water potential, stomatal conductance), which limited the amount of data that could be obtained on account of the time and labor required for each measurement. Recent developments in sensor and data technology have enabled continuous monitoring of plant water status. Data from infrared thermometers, LVDT (stem-diameter) sensors and sap-flow gauges can be recorded every few minutes, providing detailed information on the dynamics of plant water status. Automatic programmers for irrigation management are being increasingly used; these systems can be integrated with continuously measured parameters of plant water status so that crops can be automatically irrigated when the selected indicator of plant water status reaches a defined limit. Discrete measurements of plant water status cannot be used in this way.

Different methods based on temperature difference between air and canopy have been developed to schedule irrigation. The most widely used is the crop water stress index, which relates canopy (T_c) and air (T_a) temperature to the vapor pressure deficit (VPD),

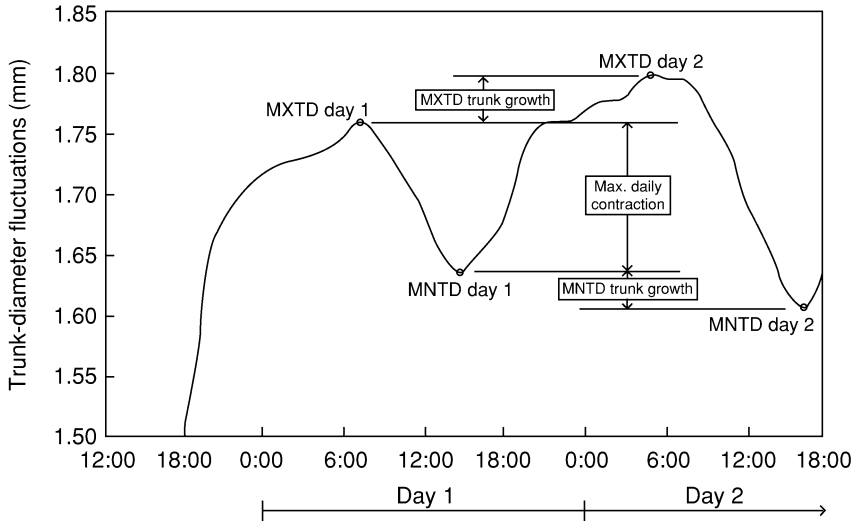


Figure 3 Parameters that can be derived from trunk-diameter measurements, including maximum daily trunk contraction, and trunk growth expressed as daily differences in maximum and minimum daily trunk diameters (MXTD and MNTD, respectively). (Adapted from Goldhamer DA and Fereres E (2001) Irrigation scheduling protocols using continuously recorded trunk diameter measurements. *Irrigation Science* 20: 115–125, with permission.)

and compares this with $T_c - T_a$ for a well-watered crop at the same VPD.

Trunk-diameter measurements with LVDT sensors have been proposed for IS in a number of fruit-tree species. Good correlations between variations in trunk diameter and changes in tree water status were observed in these species. Applying trunk- or stem-diameter measurements for IS requires firstly the definition of parameters derived from the data, and then the development of criteria for applying the derived parameters, as discussed previously. Maximum daily contraction, the difference between daily maximum values, and difference between daily minimum values (Figure 3) have been proposed as the most sensitive parameters to detect water stress in some fruit trees. As of 2004, irrigation protocols are being developed for fruit trees and vegetable species that involve: (1) selecting the derived parameter most suitable for an individual species and particular growth stage, and (2) relating the derived parameters to reference values of well-watered crops and normalizing them for VPD. A consideration when using this method is the number of replicates required to account for between-plant variability.

Sap-flow sensors have also been used for irrigation scheduling in fruit trees. A method for the automatic control of irrigation has been developed, in which sap-flow sensors determine the amount of water consumed by the plant, at short time intervals. An irrigation system can be programmed to replenish this water at specified times or time intervals. Variability between plants is an issue when scaling from single plants to crops.

See also: **Plant–Soil–Water Relations**; **Soil–Plant–Atmosphere Continuum**; **Water-Use Efficiency**

Further Reading

Davies WJ and Zhang J (1991) Root signals and the regulations of growth and development of plants in drying soil. *Annual Review of Plant Physiology* 42: 55–76.

Goldhamer DA, Fereres E, Mata M, Girona J, and Cohen M (1999) Sensitivity of continuous and discrete plant and soil water stress monitoring in peach trees subjected to deficit irrigation. *Journal of the American Society of Horticultural Science* 124: 437–444.

Hsiao TC (1973) Plant responses to water stress. *Annual Review of Plant Physiology* 24: 519–570.

Hsiao TC (1990) Measurements of plant water status. In: Stewart BA and Nielsen DR (eds) *Irrigation of Agricultural Crops*, pp. 243–279. Madison, WI: ASA–CSSA–SSSA.

Hsiao TC and Bradford KJ (1983) Physiological consequences of cellular water deficits. In: Taylor HM, Jordan WR, and Sinclair T (eds) *Limitations to Efficient Water Use in Crop Production*, pp. 227–265. Madison, WI: ASA–CSSA–SSSA.

Jones HG and Tardieu F (1998) Modelling water relations of horticultural crops. *Scientia Horticulturae* 74: 21–46.

Kramer PJ and Boyer JS (1995) *Water Relations of Plants*. San Diego, CA: Academic Press.

Smith DM and Allen SJ (1996) Measurement of sap flow in plant stems. *Journal of Experimental Botany* 47: 1833–1844.

Turner NC (1997) Further progress in crop water relations. *Advances in Agronomy* 58: 293–338.

Poiseuille's Law *See* **Hydrodynamics in Soils**

POLAR SOILS

J C F Tedrow, Formerly with Rutgers University, New Brunswick, NJ, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Soil science, traditionally allied with food and fiber production, has been viewed primarily through the eyes of the farmer. It is therefore understandable why soils of the polar regions, with negligible agricultural potential, were the last to receive serious scientific attention. Early investigators of polar soils showed all land in northern Siberia as being mantled with Tundra soil. The tundra is described as developed on a treeless plain of the north where conditions are swampy and the land is underlain by permanently frozen ground (permafrost). Investigators tend to use the 50°F July isotherm as the approximate southern limit of the arctic, a term used somewhat interchangeably with the word 'tundra.' Subsequent investigators of the arctic lands have convincingly shown that not only are there a variety of soils present locally, depending upon relief, substrate, and physiographic setting, but also there are major regional differences.

Climate

Climate of the northern polar regions is characterized by long, cold winters and short, cool summers. Mean monthly temperature for the warmest month in the vicinity of the tree line ranges from approximately 50 to 54°F but, in the northern extremities of ice-free land such as northern Greenland, Svalbard, and Franz Josef Land, the temperatures of warmest month decrease to 32–34°F. Precipitation values approximate 25 cm annually in the main polar belt but decrease somewhat northward.

Vital activity in polar soils is limited to 2–4 months per year, depending upon region, site, relief, vegetation, snow cover, and related factors. Annual snowfall in the polar regions generally is 30–60 cm, with much redistribution by wind action.

In the far north, the ice-free sectors have a desert-like aspect. Naturalists have long had a variety of names for these northern sectors, such as 'high polar,' 'cold steppe,' 'arctic wastes,' 'rock desert,' 'polar tundra,' 'barren-ground tundra,' 'fell field' (desert-like), 'desert glacial,' 'nival zone,' and 'dry periglacial zone.' The far north (polar desert) has a number of desert affinities such as saline soil conditions, and the presence of desert pavement and wind-abraded rocks.

Biotic Factors

One of the distinctive characteristics of the polar landscapes is the absence of trees. The northern edge of forested land marks one of the most striking vegetative changes anywhere on earth. There are, however, a number of problems associated with plotting a circumpolar tree line in detail, because there are interruptions by high mountains, glaciers, and open bodies of water. Further, over long periods the tree line may shift as evidenced by the presence of tree stumps well north of the present tree line. Also, there are differences between the physiognomic forest line and the tree species line.

The tundra zone is characterized mainly by wet cotton-grass meadows and wet sedge meadows, commonly referred to as 'typical tundra' (**Figure 1**). There are also extensive bogs present. The uplands and mountain sectors, however, have drier, upland and barren plant communities.

The polar desert zone supports a discontinuous, sparse vegetative cover (**Figure 2**). Here the vascular plants do not compete for "open space" as they do within the tundra zone. The plant cover consists mainly of lichens, mosses, liverworts, and a scattering of arctic-alpine herbaceous plants.

The subpolar desert zone, also known as midarctic or arctic tundra, is a blend of high arctic and tundra vegetative elements. **Figure 3** shows the location of the three soil zones of the arctic: tundra, polar desert, and subpolar desert. This is a generalized depiction and does not take into account factors such as littoral climates, high mountains, and special geologic conditions.

Permafrost

Permafrost, also known as perennially frozen ground, is generally considered to be a thickness of soil, surficial deposits, or bedrock at a variable depth below the surface of the earth in which the temperature of the matrix has existed below the freezing point of water for a period of at least several years. Permafrost therefore may consist of frozen soil, frozen bedrock, and clear ice. Permafrost underlies nearly all of the polar lands, including adjacent offshore sectors. The thickness of permafrost varies greatly, with a few locations in Siberia having a depth of more than 1200 m. In Siberia permafrost extends hundreds of miles south of the tree line (**Figure 4**) and still much further into high mountain areas of the world.

Figure 1 Wet meadow in the tundra zone, northern Alaska. © J Tedrow.

Figure 2 Polar desert on Cornwallis Island. Most of the landscape is barren except the dark depression, which is mantled with Tundra (Aquiturbel) soil. © J Tedrow.

Of special interest to soil scientists is the rate and depth of thaw in polar soils. Seasonal thaw usually begins in snow-free sectors in early June, but by late July or early August seasonal thaw has largely reached its maximum. Depth of seasonal thaw may be less than 30 cm in the far north, but within the northern fringes of the forested land it may extend to a depth of 150–250 cm or more.

Permafrost generally exists as massive ice or as frozen soil with the pore spaces in the soil filled with ice. There are, however, many locations in mountainous areas and elsewhere where ice is present only at the contact points of the soil particles, a condition referred to as dry permafrost.

Soil Development and Classification

Systems of Classification

Up until the mid-1900s, there was very little attention given to detailed classification of polar soils. Tundra soil was recognized, but descriptions were vague and incomplete. The most reliable information was from Russian sources, but even these

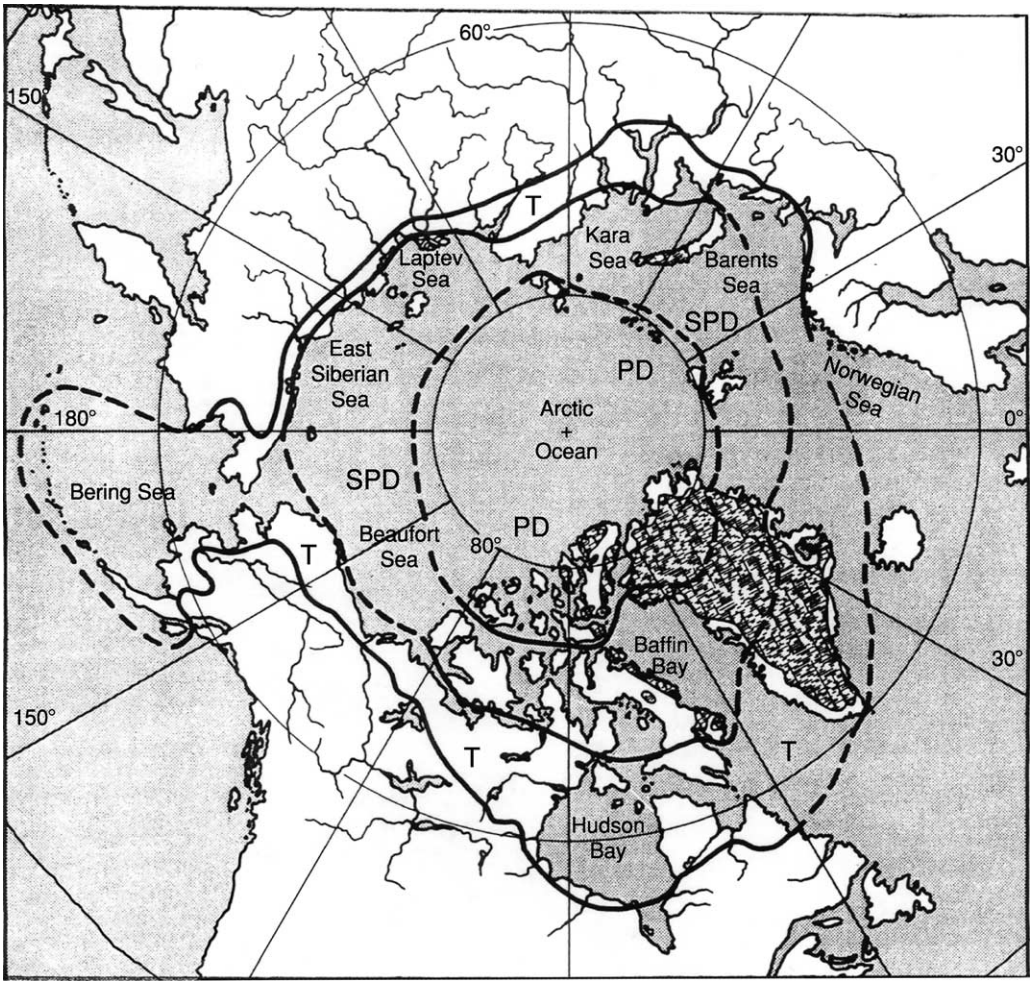


Figure 3 Delineation of soil zones of the northern polar lands. PD, Polar Desert zone; SPD, Subpolar Desert zone; T, Tundra zone. © J Tedrow.

were preliminary and generally carried out just north of the forested land. During more recent years there have been several proposals advanced for classifying polar soils. Table 1 shows four systems that have been proposed. Other schemes of classification have also been proposed by Russian, French, and New Zealand (for Antarctica) investigators. In *Soil Taxonomy* (US), soils of the polar regions are collectively referred to as Gelisols (soil order). Gelisols are divided into Histels, Turbels, and Orthels. The Canadian system is similar in principle to that of *Soil Taxonomy* but there are differences in nomenclature. Soils are first described under the Cryosol order followed by Organic Cryosols, Turbic Cryosols, and Static Cryosols. The Food and Agriculture Organization of the UN (FAO) system is similar to that of *Soil Taxonomy* and also of Canada. The Tedrow system uses a somewhat different approach in that the polar regions are first divided into soil zones, then genetic soil divisions. There have been a number of systems proposed by Russian investigators, most of which follow similar lines in that the polar regions are first divided into bioclimatic zones, then genetic soil characterization.

Tundra Zone

The tundra zone is characterized by a preponderance of hydric mineral soils. These were, for years, the benchmark soils of the northern polar regions. A variety of soils are now recognized within the tundra zone, including the mineral soils that have impeded drainage. The soils are generally silty, but there are wide ranges in texture, from sandy loams and sands to heavy bentonitic clays. Earlier writings have indicated that, with frequent freezing and thawing, individual mineral particles become reduced to silt size, but such a phenomenon has been discounted largely.

Figure 5 shows an idealized Upland Tundra (Molliturbel) soil. The buried organic-rich layer, as shown at the 24- to 26-inch depth, usually has an age of 8000–11 000 years BP, which puts the time of burial within the Hypsithermal Interval. Pollen from these buried layers is generally rich in Ericaceae, Cyperaceae, *Betula*, and *Salix*.

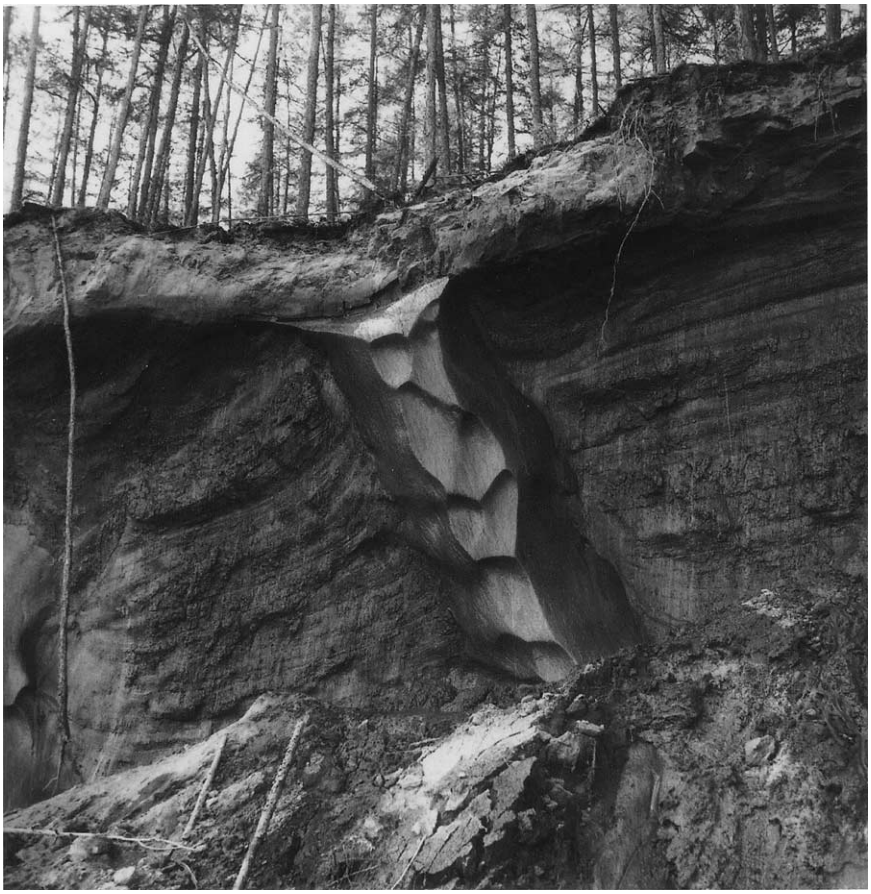


Figure 4 Ice wedge in permafrost along the Aldan River, Siberia. The top 150 cm active layer above the ice wedge is composed of well-drained silt and fine sand. © J Tedrow.

Clay minerals in Tundra (Molliturbel) soil generally consist of illite and kaolinite. The pH values of wet mineral soils in the tundra zone generally are within the 4.5–6.5 range, but, where there is a dusting of carbonate-bearing minerals from sources such as vegetation-free floodplains, the soils are alkaline.

Within the tundra zone there are extensive sectors of Bog soils (Histels). Organic deposits having a thickness of greater than 10 m have been recorded. The bogs are generally acidic.

There are many small, isolated areas of well-drained soils within the tundra zone. These soils, designated as Arctic Brown (Haplorthels), develop on sites where there is free drainage to a depth of at least 60 cm. The soils develop a brownish color in the solum, much as they do within the northern forested zone. These soils are usually acid, but on wind-swept spurs the soil may be alkaline.

Also, within the tundra zone there are a few Rendzina (Umbrothels) soils formed on limestone bedrock. On the other hand, where the soils from on acid bedrock such as a granite, a Podzol-like (Spodic Psammoturbel) soil may develop. In mountainous sectors of northern lands, the soils may be shallow and rocky, corresponding to a Lithosol (Orthels).

Polar Desert Zone

Polar Desert soils (Haploturbels) generally form on positive relief elements and on sandy/gravelly plains. Desert pavement is usually present where there is an abundance of gravel in the substrate. Maximum Polar Desert soil development is usually confined to open-textured material underlain by a dry permafrost. A browning effect may be present within a solum with up to 3% of easily reducible iron. Mud coatings and travertine formation may be evident in some locations.

The organic matter content of Polar Desert soil is generally less than 1%, but, where a vascular plant cover is present, the organic matter content increases accordingly. It is believed that most of the organic matter in Polar Desert soil is contributed by algae, including diatoms. The organic matter in Polar Desert soil is quite stable. In northern

Table 1 Approximate comparison of soil taxa between *Soil Taxonomy*, the World Reference Base for Soil Resources, the Canadian system, and Tedrow's system

Soil taxonomy[a]	*World reference base*[b]	*Canadian*[c]	*Zonal system*[d]
Gelisol (soil order)	**Cryosol** (soil unit)	**Cryosol** (order)	**Soils of the Cold Regions**
Histels (suborder)	Histic Cryosols (soil subunit)	Organic Cryosols (great group)	Organic (bog) soil (great soil group)
Folistels (great group)	(No equivalent)	(No equivalent)	
Glacistels	Glacic Histic Cryosols	Glacic Organic Cryosols	
Fibristels	Fibric Histic Cryosols	Fibric Organic Cryosols	
Hemistels	Mesic Histic Cryosols	Mesic Organic Cryosols	
Sapristels	Humic Histic Cryosols	Humic Organic Cryosols	
(Terric Fibristels) (subgroup)	(No equivalent)	Terric Fibric Organic Cryosols	
(Terric Hemistels)	(No equivalent)	Terric Mesic Organic Cryosols	
(Terric Sapristels)	(No equivalent)	Terric Humic Organic Cryosols	
Turbels	Turbic Cryosols	Turbic Cryosols	
Histoturbels	Stagnic & Gleyic Turbic Cryosols	Gleysolic Turbic Cryosols	Meadow Tundra soils
Aquiturbels	Stagnic & Gleyic Turbic Cryosols	Gleysolic Turbic Cryosols	Meadow Tundra soils
Anhyturbels	Salic, Gypsic & Calcic Turbic Cryosols	(No equivalent)	Ahumisols, Ahumic (subxerous) soils
Molliturbels	Mollic Turbic Cryosols	Brunisolic Turbic Cryosols	Upland Tundra soils, Soils of the Hummocky Ground, Grumusols
Umbriturbels	Umbric Turbic Cryosols	Brunisolic Turbic Cryosols	Upland Tundra soils, Soils of the Hummocky Ground, Grumusols
Psammoturbels	Regic Turbic Cryosols	Regosolic Turbic Cryosols	Regosols, some Polar Desert soils
Haploturbels	Regic Turbic Cryosols	Orthic Turbic Cryosols	Polar Desert soils, Soils of the Solifluction Slopes
(Glacic subgroups)	Glacic Turbic Cryosols	(No equivalent)	Meadow Tundra and Upland Tundra soils
(Sulfuric Aquaturbels)	Thionic Turbic Cryosols	(No equivalent)	Meadow Tundra and Upland Tundra soils
(Andi, Molli-, & Umbriturbels)	Andic Turbic Cryosols	(No equivalent)	Upland Tundra soils, Soils of the Hummocky Ground, Grumusols
(Vitrand-, Molli-, & Umbriturbels)	Tephric Turbic Cryosols	(No equivalent)	Upland Tundra soils, Soils of the Hummocky Ground, Grumusols
(Spodic Psammoturbels)	Spodic Turbic Cryosols	(No equivalent)	Podzol-like soils
Orthels	Haplic Cryosols	Static Cryosols	
Historthels	Stagnic & Gleyic Haplic Cryosols	Gleysolic Static Cryosols	Meadow Tundra soils
Aquorthels	Stagnic & Gleyic Haplic Cryosols	Gleysolic Static Cryosols	Meadow Tundra soils
Anhyorthels	Salic, Gypsic & Calcic Haplic Cryosols	(no equivalent)	Ahumic soils (xerous, ultraxerous)
Mollorthels	Mollic Haplic Cryosols	Brunisolic Static Cryosols	Arctic Brown soils, Subarctic Brown soils, Rendzina
Umbrorthels	Umbic Haplic Cryosols	Brunisolic Static Cryosols	Arctic Brown soils, Subarctic Brown soils, Rendzina
Argiorthels	(No equivalent)	(No equivalent)	Subarctic Grey Wooded soils
Psammorthels	Regic Haplic Cryosols	Regosolic Static Cryosols	Regosols, some Polar Desert soils
Haplorthels	Regic Haplic Cryosols	Regosolic Static Cryosols	Arctic Brown soils, Subarctic Brown soils
(Glacic subgroups)	Glacic Haplic Cryosols	(No equivalent)	Meadow Tundra and Upland Tundra soils
(Sulfuric Aquorthels)	Thionic Haplic Cryosols	(No equivalent)	Meadow Tundra and Upland Tundra soils
(Andi, Molli-, & Umbrorthels)	Andic Haplic Cryosols	(No equivalent)	Upland Tundra soils, Soils of the Hummocky Ground, Grumusols
(Vitrand-, Moll-, & Umbrorthels)	Tephric Haplic Cryosols	(No equivalent)	Upland Tundra soils, Soils of the Hummocky Ground, Grumusols
(Spodic Psammorthels)	Spodic Haplic Cryosols	(No equivalent)	Podzol-like soils

(*Continued*)

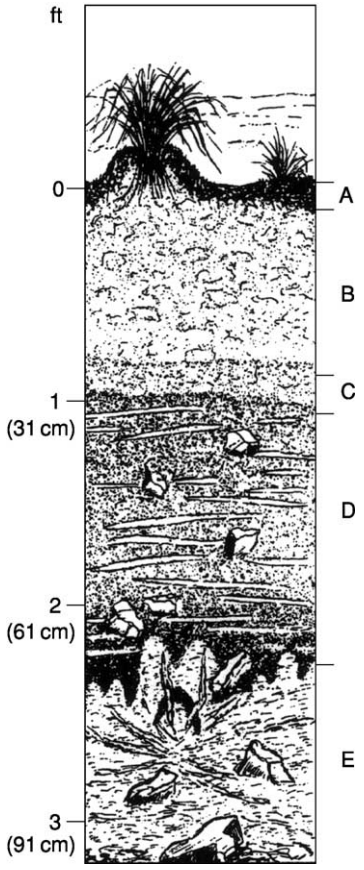


Figure 5 Idealized profile of Tundra (Molliturbel) soil. A, dark brown to black organic matter, consisting of partially decomposed sedges, sphagnum, and heaths. This horizon may vary from a thin, discontinuous one to an organic mat some 15 cm thick. Usually very loose, fibrous, wet, and strongly acid. B, light olive-brown (2.5Y 5/4) silt loam. Usually very wet, but loose and friable when dried. Nearly always mottled but in varying degrees. Strongly acid. C, dark gray (2.5Y 4/0) silt loam, mottled, very wet. The bottom of this layer approximates the permafrost table. D, very dark gray (2.5Y 3/0) silt loam, permanently frozen. Considerable organic staining in the ice; shreds of organic matter intermixed throughout the horizon. The bottom of the layer is very uneven and is much darker in color than the central portion of the horizon. The upper part of the horizon is also darker than the central portion. The horizon varies from frozen mineral soil to ground ice. Pieces of peaty material are commonly present. E, frozen gray mineral matter interspersed with ground ice. No evidence of organic staining. In many instances the material is nearly clear ice. Weakly acid to calcareous, depending upon parent material. © J Tedrow.

Greenland a sample of topsoil yielded an age of 3300 years BP.

The pH of Polar Desert soils varies widely from approximately 4.4 to as high as 7.9. Likewise electrical conductivity ranges from less than 10 to 66 $m\Omega\ cm^{-1}$. Salt crusts of thenardite (Na_2SO_4) and other minerals form on some of the well-drained to xeric sites. Figure 6 shows an idealized profile of Polar Desert soil. It is common to have only minor differentiation of the soil horizons other than that induced by the distribution of iron coatings on soil particles.

Within the polar desert zone, Tundra, Bog, Rendzina, and other soils are also present. These soils have similar characteristics to those found in the tundra zone (Table 1).

Subpolar Desert Zone

At one time many polar investigators included the subpolar desert zone as a part of the polar desert zone, but for specialist purposes it is appropriate to recognize a separate zone. The subpolar desert zone has been well recognized by naturalists, particularly plant geographers, who have used a number of climate and phytographic criteria to delineate the zone. Terms such as 'arctic tundra,' 'stony-sedge-moss-lichen tundra,' 'arctic desert,' and 'semipolar desert,' among others have been proposed. The subpolar desert zone is transitional and has few, if any, unique properties of its own. Conditions are somewhat analogous to those in parkland lying between the forest and prairies. There is no single criterion for delineating the subpolar desert zone, instead separation is based on a combination of factors, including soil, plants, and climate. The occasional Polar Desert soils within the subpolar desert zone usually contain little brunification and horizon differentiation, as is the case within the polar desert soil zone. In the subpolar soil zone, Tundra, Bog, Rendzina, and other soils are present (Table 1).

Patterned Ground

In the polar environment, there are at least two opposing forces operating. One involves pedogenic forces – those resulting in the development of genetic soil horizons. A second set of processes is frost-related and, in effect, consists of destructive forces. 'Patterned

Source: JG Bockheim from Bockheim JG and Tamocai C (2000) Gelisols. In: Sumner ME (ed.) *Handbook of Soil Science*, pp. 256–264. Boca Raton, FL: CRC Press.

[a]Soil Survey Staff (1999) *Soil Taxonomy*. USDA–NRCS, Agricultural Handbook 436. Washington, DC: US Government Printing Office.

[b]Dudal R (1974) Soil map of the world (1). UNESCO, FAO, Rome.

[c]Tarnocai C (1994) Genesis of permafrost-affected soil. Proc. Meeting on classification, correlation, and management of permafrost-affected soils. USDA. Lincoln, Nebraska, USA.

[d]Tedrow JCF (1977) *Soils of the Polar Landscapes*. New Brunswick, NJ: Rutgers University Press.

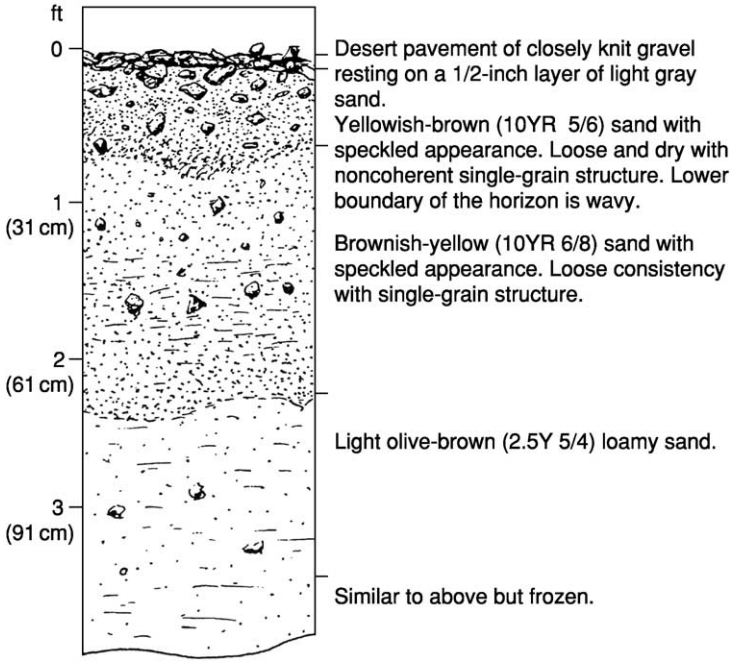


Figure 6 Idealized profile of Polar Desert (Pasammorthel) soil. © J Tedrow.

ground' is a group term for symmetrical forms such as circles, polygons, nets, steps, and stripes. Various terms have been used to describe the frost structures, such as 'rutmark,' 'råmark,' 'frost polygon,' 'strukturboden,' 'polygonboden,' 'ice-wedge polygon,' 'stone polygon,' 'soil circle,' 'mud circle,' 'stone ring,' 'stone stripe,' 'hummocky ground,' 'frost boil,' 'ice mound,' 'palsa,' 'pingo,' and many others.

One of the more common forms of patterned ground is the ice-wedge polygon. The ground develops a series of ice wedges that result from annual soil temperature amplitudes. Year after year the ice wedges tend to increase in size, exerting exceptionally strong lateral forces that result in a buckling of both the edges and centers of the polygons (Figure 7). Changes of elevation of the soil surfaces may result in the center of the polygon being either raised or depressed. Change in elevation induced by frost action will result in some parts of the polygon being better drained, but other parts will become more poorly drained. This complex phenomenon of changes in drainage induced by frost action has received very little attention from pedologists. In addition to hydric changes in the soil induced by polygon formation, another set of cryogenic processes ruptures and deforms the soil horizons.

Apart from polygonal ground formation, another set of processes involves frost stirring, resulting in the development of stone rings (sorted circles). Figure 8 shows a form of stone rings in which the pattern of deformation has resulted from an upwelling of the mineral material in the form of a core of raw, earthy material and an outward displacement of the coarser rock fragments. Figure 8 shows lichen crusts on the outer rock fragments, whereas the central part of the stone ring is free of lichens, indicating a less stable condition.

Nonsorted circles are common throughout the polar regions, especially where the soils are of medium to clayey texture. Figure 9 shows a slight dome-shaped pattern with small, secondary desiccation cracks (zellenboden).

On sloping land, there are many elongated forms of patterned ground that result from at least two processes: frost sorting and solifluction (the slow gravity driven downhill movement of saturated soil). Patterns consist of stone garlands, stone stripes, soil stripes, and others. Where solifluction processes are intense, the genetic soil horizons become highly deformed and, in some locations, completely obliterated.

Cold Desert Zone

The antarctic continent, also referred to as the cold desert zone, was the last sector of the globe to receive the attention of pedologists. Only since World War II has there been significant pedologic information forthcoming from Antarctica. Soil maps of the globe traditionally provide information only to about 60° S, which excludes all of Antarctica (Figure 10). One of the interesting aspects of cold desert zone pedology is that there is virtually no effective organic component in the soils except for a few special conditions such as

Figure 7 Ice-wedge (nonsorted) polygons in Prince Patrick Island. © J Tedrow.

Figure 8 Stone rings (sorted circles) in Prince Patrick Island. © J Tedrow.

occur at the penguin rookeries, where there are local accumulations of guano. Late in the nineteenth century, pedology was recognized as an independent scientific discipline, and it was agreed that organic matter was an essential component of the soil system. While virtually all pedologists who have worked in Antarctica speak of soils in the traditional sense, plenary bodies have largely ignored the question of antarctic soils. The position of plenary bodies was understandable when, in earlier times, very little information was available concerning Antarctica, but considerable pedologic information is now available.

Of the 2–3% of Antarctica that is free of ice and snow, only a small percentage has surficial deposits adequate for any semblance of soil formation. The remainder of the ice-free land consists of bedrock and boulder fields. Permafrost is believed to be present throughout the cold desert zone, but most

Figure 9 Nonsorted circles in northern Alaska. © J Tedrow.

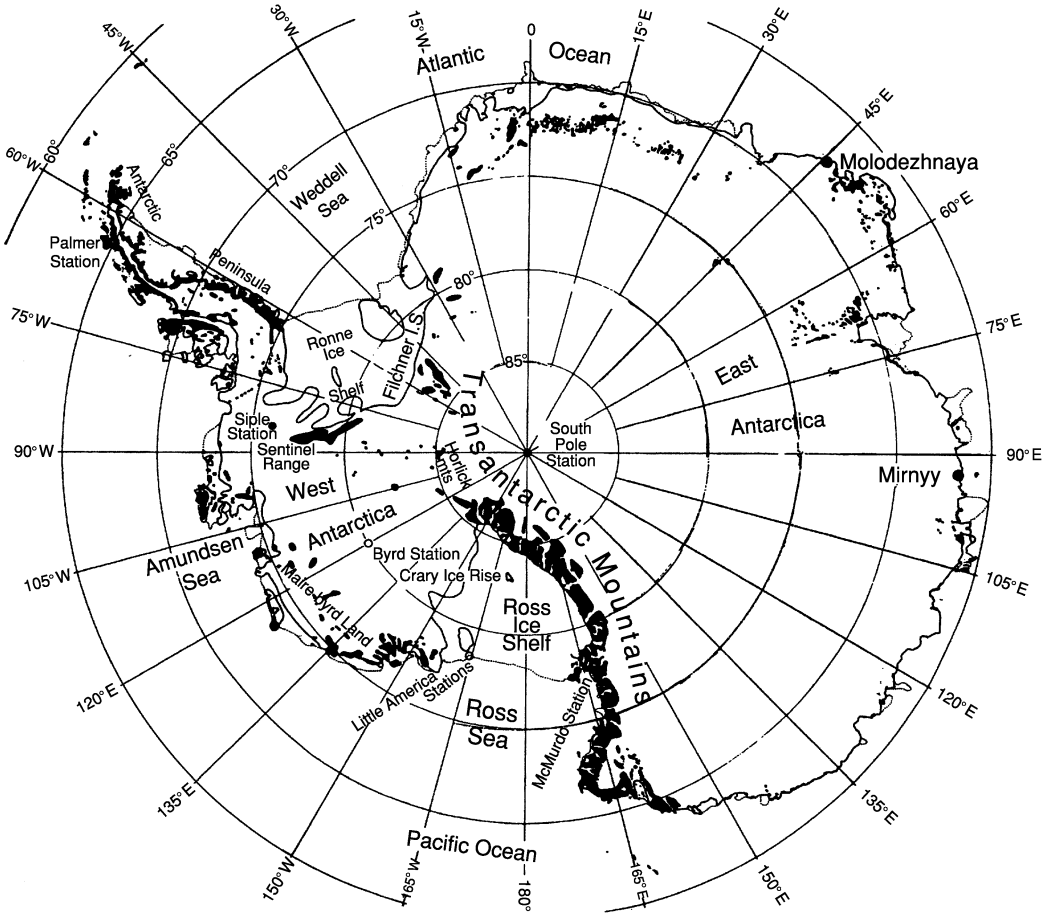


Figure 10 Map of Antarctica. The dark areas are largely without a snow-ice cover. © J Tedrow.

of the surface deposits are subject to dry frost conditions.

Physical and chemical weathering in the cold desert zone is intense. Boulders exhibit cavernous weathering forms in which shells of exfoliation can be pried off. Wind polishing of rocks is intense with the abrasion caused by the blowing of ice particles. Ice in the cold desert at very low temperatures may have a hardness like that of quartz. Desert varnish is also evident in many places, especially on older land surfaces.

Although air temperatures in the cold desert zone may be well below freezing, solar radiation causes the rock surfaces to attain elevated temperatures on wind-free, sunny days. The temperature gradient in rocks during sunny days results in both physical and chemical weathering. Garnets split, feldspars weather, and clay substances accumulate in the fractured areas. Accumulation of calcite, gypsum, magnesium sulfate, and other minerals is widespread. Also, highly soluble forms of calcium, magnesium, sodium, potassium, chloride, sulfate, and nitrate have been reported. There are a number of forms of patterned ground present in antarctic deserts, with the most common form being of a polygonal nature.

Table 2 shows two classification schemes proposed for the cold desert zone, both of which follow the same pathways in principle. Figure 11 shows an idealized soil of the cold desert (Subxerous-Xerous). This depiction is one of maximum soil development on till composed mostly of sandstone, conglomerate, and dolerite. Horizonation is somewhat blurred, with no structural aggregates being present. The 'active' layer is approximately 51 cm thick but, because of the ultra dry, loose condition, depth to the frost table can only be approximated. Cold desert soils, for all practical purposes, are without terrestrial life. Algae including lichens are occasionally present and, in rare

Table 2 Comparison of several proposed schemes for classifying cold desert soils of Antarctica

Campbell and Claridge[a]	*Tedrow*[b]
Frigic soils	*Soils of the cold desert zone*
Zonal soils	*Zonal soils*
Ultraxerous	Ultraxerous
Xerous	Xerous
Subxerous	Subxerous
—	Ahumisol
—	Protoranker
Intrazonal soils	*Intrazonal soils*
Evaporite soils	Evaporite soils
Algal peat soils	Algal peat soils
Avian soils	Ornithogenic (Avian) soils
Hydrothermal soils	Hydrothermal soils
Azonal soils	*Azonal soils*
Recent soils	Regosols
	Lithosols

[a]Campbell IB and Claridge GGC (1969) A classification of frigio soils – the zonal soils of the antarctic continent. *Soil Science* 107: 75–85.
[b]Tedrow JCF (1977) *Soils of the Polar Landscapes.* New Brunswick, NJ: Rutgers University Press.

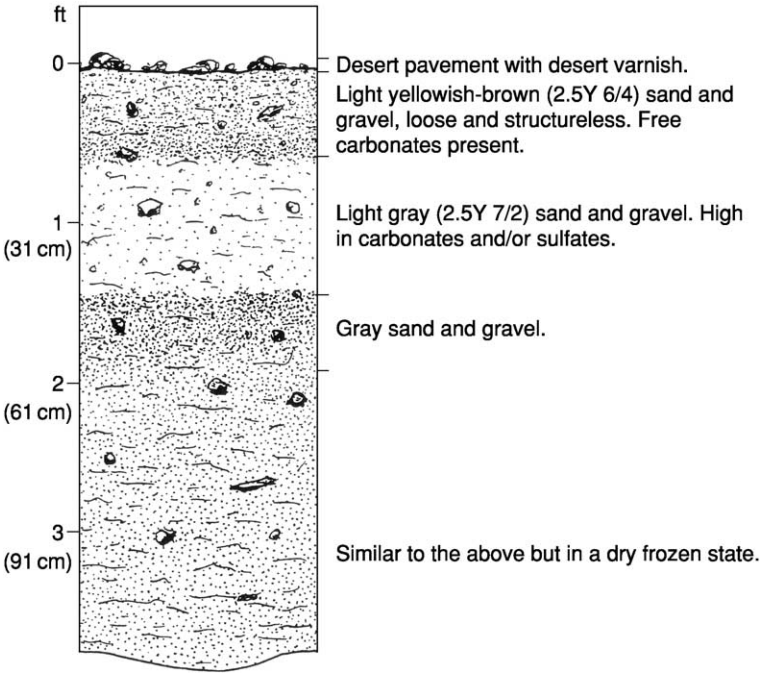


Figure 11 Profile of Cold Desert (Gelisol) soil. © J Tedrow.

situations, protozoa, rotifers, and nematodes can be found.

There is evidence that there is ionic migration in the soil, even when it is in a dry, frozen state. Buildup of salts is common in the soil as well as in bedrock and boulders. Nearly all soils of the cold desert are alkaline, with pH values being in the 7–10 range. In special situations, however, acid soils have been reported.

Closed basins, depressions, and other sites commonly have saline soils and accumulations of salt encrustations. Saline lakes have been reported in the dry valleys of Victoria Land and in the Mirnyy area. Salt content of water in a closed lake basin in Wright Valley is 34%. The salt consists mainly of calcium and chloride. Clay minerals in cold desert soils consist primarily of micas, kaolinite, vermiculite, montmorillonite, and chlorite. Some of the clay-size fraction also contains quartz, feldspar, and amorphous material.

The cold desert soils can be divided into Ultraxerous, Xerous, and Subxerous varieties (**Table 2**). In the Molodezhnaya sector of Antarctica, some of the soils show advanced weathering and have been designated as Ahumisol. There are also a very few, extremely small locations, where there are clumps of vascular plants, in which the soils have been designated as Protoranker. Along the margins of some saline lakes, there are shallow accumulations of algae. Also, in the penguin (and skua) nesting areas, the soils, rich in bird droppings, feathers, and bones, have been designated as Ornithogenic or Avian. In the area of volcanic vents such as at the top of Mt. Erebus, the soils have been designated as Hydrothermal.

See also: **Factors of Soil Formation:** Climate; **Freezing and Thawing:** Cycles; Processes; **Pedology:** Basic Principles

Further Reading

Aleksandrova VD (1980) *The Arctic and Antarctic: Their Division into Geobotanical Areas*. Cambridge, UK: Cambridge University Press.

Bockheim JG (2002) Gelisols: the permafrost-affected soils. In: *Encyclopedia of Soil Science*. New York: Marcel Dekker, Inc.

Brown J, Ferrians OJ, Heginbottom JA, and Melnikov ES (1997) *Circum-Arctic Map of Permafrost and Ground-Ice Conditions*. US Geological Survey. Map CP-45. Washington, DC: US Goverment Printing Office.

Campbell IB and Claridge GGC (1969) A classification of frigio soils – the zonal soils of the antarctic continent. *Soil Science* 107: 75–85.

Campbell IB and Claridge GGC (1987) *Antarctica: Soils, Weathering Processes and Environment*. New York: Elsevier.

Glazovskaia MA (1984) *Soils of the World*. Moscow, Russia: Moscow University Press.

Glinka KD (1927) *The Great Soil Groups of the World and Their Development*. Ann Arbor, MI: Edwards Bros.

Gorodkov BN (1939) Peculiarities of the Arctic topsoil. (in Russian). *Izv. Gosud. Geogr. Obsch*. 71: 1516–1532.

Goryachkin SV, Karavaeva NA, and Targulian VO (1998) Geography of Arctic soils: current problems. *Eurasian Soil Science* 31: 467–476.

Jahn A (1975) *Problems of the Periglacial Zone*. National Science Foundation (TT72-54001). Washington, DC: National Science Foundation.

Kimble M (ed.) (2004) *Cryosols–Permafrost-Affected Soils*. Berlin: Springer-Verlag.

King CAM (ed.) (1975) *Periglacial Processes*. New York: Halsted Press.

Linell KA and Tedrow JCF (1981) *Soil and Permafrost Surveys in the Arctic*. Oxford, UK: Oxford University Press.

Rieger S (1983) *The Genesis and Classification of Cold Soils*. New York: Academic Press.

Soil Survey Staff (1999) *Soil Taxonomy*, 2nd edn. US Department of Agriculture, Agricultural Handbook 436. Washington, DC: US Government Printing Office.

Tedrow JCF (1977) *Soils of the Polar Landscapes*. New Brunswick, NJ: Rutgers University Press.

Tedrow JCF (1991) *Pedogenic Linkage Between the Cold Deserts of Antarctica and the Polar Deserts of the High Arctic*. Antarctic Research Series, vol. 53, pp. 1–17. Washington, DC: American Geophysical Union.

Ugolini FC (1986) Pedologic zonation in the well-drained soils of Arctic regions. *Quaternary Research* 26: 100–120.

Ugolini FC (1986) Processes and rates of weathering in Cold and Polar desert environments. In: *Rates of Chemical Weathering of Rocks and Minerals*, pp. 193–235. New York: Academic Press.

Washburn AL (1980) *Geocryology: A Survey of Periglacial Processes and Environments*. New York: John Wiley & Sons, Inc.

Williams PJ and Smith MW (1989) *The Frozen Earth: Fundamentals of Geocryology*. Cambridge, UK: Cambridge University Press.

POLLUTANTS

Contents
Biodegradation
Effects on Microorganisms
Persistent Organic (POPs)

Biodegradation

P B Hatzinger, Shaw Environmental, Inc., Lawrenceville, NJ, USA
J W Kelsey, Muhlenberg College, Allentown, PA, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Industrialized societies annually release large quantities of synthetic chemicals into Earth's natural environments, both intentionally, in the case of pesticides, and inadvertently, through the production, transport, refining, and combustion of fossil fuels, waste streams from manufacturing, chemical spills, disposal of products by consumers, and discharges from a wide variety of other sources. Currently, there are more than 75 000 chemicals registered in the US Environmental Protection Agency (US EPA) inventory of chemical substances manufactured or processed in, or imported into, the USA. The number of potential chemical contaminants is substantially higher, as the EPA inventory does not include pesticides, radioactive materials, food additives, drugs, or cosmetics. Many of these compounds ultimately end up as contaminants in the atmosphere, water, sediment, or soil. To add to the complexity of the problem, the compounds of concern vary widely in their characteristics. A contaminant can belong to any of a number of disparate classes of compounds that differ in important properties such as charge, polarity, water solubility, pK_a, molecular weight, and aromaticity. These chemical characteristics subsequently determine a contaminant's fate in the environment and its hazard to humans and other organisms.

One of the most important processes determining the fate of a given chemical is its susceptibility to biological degradation. There have been more than five decades of research into chemical biodegradation, including microbial mechanisms and the chemical and environmental factors that govern its rate and extent in soils.

Biodegradation

Biodegradation is the partial or complete breakdown of a chemical pollutant resulting from the physiological reactions of microorganisms. In general, microorganisms carry out such processes to obtain necessary carbon, nutrients, and energy for their survival and growth. Through the millennia, bacteria have experienced extreme competition (primarily from each other) for limiting resources. As a result, they have evolved unrivaled metabolic capabilities, utilizing nearly every thermodynamically favorable reaction for energy and nearly every organic molecule for carbon. For example, whereas humans gain energy only through the oxidation of specific organic substrates using oxygen as an electron acceptor (i.e., through aerobic respiration), bacteria utilize virtually all natural and many man-made organics, numerous inorganic molecules (e.g., Fe(0); Fe(II), H_2S, S(0) H_2, NH_4), and light for energy. Moreover, besides oxygen, specific microorganisms can utilize a variety of chemicals, including an assortment of synthetic pollutants and pesticides, as terminal electron acceptors. Specific examples of microbial biodegradation processes under aerobic and anaerobic conditions are provided in subsequent sections.

The microbial degradation of pollutants can yield many different products. Generally, the biodegradation of a large organic molecule results in smaller, less complex organic molecules that may be more or less toxic than the parent compound. In some cases, an organic contaminant is completely converted to an inorganic product such as CO_2. A transformation of this type is referred to as 'mineralization' and is associated with the elimination of toxicity of the parent compound. Although plants, animals, and microorganisms can all metabolize some pollutants, the mineralization of organic contaminants is generally carried out by bacteria and fungi only. These microorganisms are responsible for nearly all of the relevant biotransformations of environmental pollutants.

Whereas biodegradation occurs naturally in soils and other environments without any intervention, bioremediation involves the application of natural processes to remove pollutants from contaminated sites. In short, bioremediation is the use of microorganisms to clean up unwanted compounds, although the specific technology employed will depend on the

nature of the site and the pollutant of concern. Contaminants may be treated *in situ* through the addition of nutrients to stimulate indigenous microorganisms (biostimulation) or by the inoculation of specific microorganisms capable of degrading the compound (bioaugmentation). Alternatively, contaminated soils can be excavated and treated *ex situ* using biological reactors, composting, or other approaches.

The biodegradation of soil pollutants is critical because of the impact the resultant transformations can have on human and environmental health. For example, in the absence of appreciable degradation, compounds released into soil may leach to groundwater. Contamination of groundwater poses a potentially serious threat to drinking-water supplies and public health. Moreover, compounds present in groundwater have a tendency to persist much longer than those in soils due to the limited supplies of oxygen and inorganic nutrients and low microbial numbers generally found in subsurface aquifers. Soil pollutants and pesticides can also pose a threat to human and nonhuman health if they are bioconcentrated by soil organisms. Biological uptake directly from soil can adversely affect both individual organisms and food webs. A contaminant may be sufficiently concentrated in soil to harm the organism that contacts it initially. However, certain compounds, even if initially present at low concentration, can be accumulated and magnified through herbivory and predation. A biomagnified compound has the potential to damage organisms at higher trophic levels, despite their very limited contact with the contaminated soil. Alternatively, some compounds are so potent that they are a threat at very low concentrations. Microbial transformations can lower the risk associated with a particular pollutant because they reduce or eliminate the toxicity of the compound.

Aerobic Transformation Processes

Many different synthetic pollutants and pesticides have been shown to serve as growth substrates for specific bacteria and fungi under aerobic conditions, providing both carbon and energy to these organisms. Often, these compounds are completely mineralized by one organism or by a series of different organisms to CO_2 and cell constituents. During this process, the novel pollutant is generally modified by specific enzymes to one or more metabolites that are common in cellular metabolism (e.g., pyruvic acid, acetyl-CoA). Those metabolites are then converted to energy and biomass through the common metabolic pathways. Aerobic mineralization is generally a growth-linked process whereby microbial numbers increase as the chemical concentration decreases. Growth-linked biodegradation is desirable in that it can bring about a rapid loss of a contaminant from soil. Many different organic contaminants are subject to mineralization in soils, including most components of crude oil and gasoline, and numerous pesticides.

A variety of organic chemicals are biotransformed through a nongrowth-linked process termed 'cometabolism.' Cometabolism is the fortuitous transformation of a molecule by an enzyme synthesized within a cell for another purpose. By this process, the cell modifies the chemical structure, but gains no energy, carbon, or inorganic nutrients from the reaction. Several groups of aerobic bacteria present in soils, including methanotrophic bacteria (which oxidize methane for energy), nitrifying bacteria (which oxidize ammonia for energy), and propane-, phenol-, and toluene-oxidizing bacteria, have been observed to transform a variety of different compounds through cometabolic reactions. Each of the above groups of bacteria produces nonselective monooxygenase or dioxygenase enzymes that are responsible for these reactions. For example, methanotrophic bacteria produce methane monooxygenase (MMO), an enzyme that normally functions to oxidize methane for energy (Figure 1). However, MMO has a broad substrate specificity and has been observed to catalyze the initial oxidation of numerous alkanes, alkenes, and aromatics as well as the degradation of an array of chlorinated aliphatics, such as *cis*- and *trans*-1,2-dichloroethylene (*cis*- and *trans*-1,2-DCE), trichloroethylene (TCE), and 1,2-dichloropropane. As previously noted, bacteria do not gain carbon or energy from the cometabolic oxidation of most compounds; thus, an additional cosubstrate is often required for bacterial growth (e.g., toluene for toluene

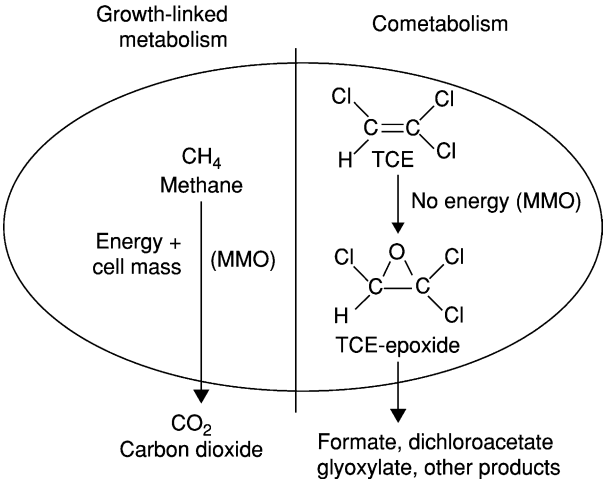


Figure 1 Representation of an individual methanotrophic bacterium degrading methane to carbon dioxide through growth-linked metabolism and degrading trichloroethylene (TCE) to TCE epoxide via cometabolism. MMO, methane monooxygenase.

oxidizers, methane for methanotrophs, etc.). Although cometabolic reactions sometimes result in the accumulation of persistent intermediates, in many instances, the initial oxidative reaction produces readily degradable compounds that are then mineralized. For example, as shown in Figure 1, the initial cometabolic attack of MMO on TCE creates an unstable TCE epoxide, which further degrades to a variety of growth substrates, such as glyoxylate, formate, and dichloroacetate.

Anaerobic Transformation Processes

A wide variety of pollutants and pesticides are subject to biodegradation under anaerobic conditions. As described for aerobic degradation, both mineralization and cometabolic reactions also occur in anaerobic environments. For some compounds, biodegradation will only occur under such conditions, since the organisms capable of carrying out the process are obligate anaerobes (i.e., they die in the presence of oxygen). Under anaerobic conditions, the degradation of specific pollutants and pesticides can be coupled to the reduction of terminal electron acceptors other than O_2, such as NO_3, Mn(IV), Fe(III), SO_4, or CO_2. This process is termed 'anaerobic respiration.' The gasoline constituent toluene, for example, has been reported to be biologically degraded with each of these different electron acceptors. For a given substrate, the energy yield during anaerobic respiration with the aforementioned compounds is less than for the same reactions using oxygen ($O_2 > NO_3 > Mn(IV) > Fe(III) > SO_4 > CO_2$) as an electron acceptor. Therefore, the anaerobic reactions tend to occur more slowly, and microbial growth yields are lower than for aerobic reactions.

If a large quantity of a readily biodegradable pollutant is added to a soil, it is not uncommon for the dominant, terminal electron-accepting reaction to progress from aerobic respiration (with O_2 as the primary electron acceptor) all the way to methanogenesis (with CO_2 as the primary acceptor), passing through each dominant electron acceptor in the order shown in Figure 2. Some microorganisms (e.g., facultative anaerobes) can utilize more than one electron acceptor, although rarely are these reactions performed simultaneously by a given organism. For example, many denitrifying bacteria grow aerobically in the presence of oxygen, then anaerobically on nitrate when oxygen levels are reduced. Other microorganisms, including many sulfate-reducing bacteria, are obligate anaerobes and are capable of using only one of the common electron acceptors (i.e., sulfate) for growth. As electron acceptors are depleted in sequence, the dominant metabolic reactions (and in many instances, the dominant microorganisms) accounting for the chemical losses shift. The reduction

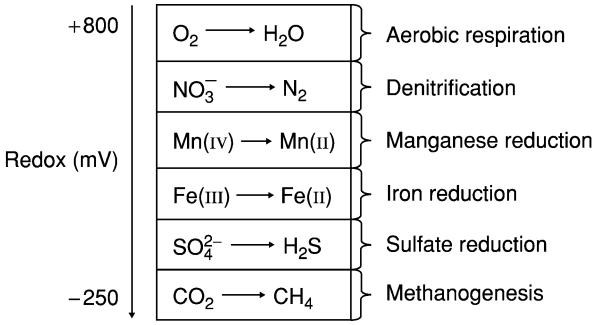


Figure 2 Sequence of common electron-accepting reactions. Reactions are given in general order of occurrence from top to bottom, with the reactions at the top taking place at the highest redox potential and providing the highest energy yield to microorganisms during degradation of a given organic compound.

potential declines accordingly as the sequence of terminal electron acceptors progresses from oxygen to carbon dioxide. In the absence of available external electron acceptors, some organic molecules are also subject to fermentation, whereby small amounts of energy are gained by an organism through internal oxidation–reduction reactions. However, when all common electron acceptors are consumed in a contaminated soil, the biodegradation reaction is likely to cease, even though biodegradable compounds are still present. It is for this reason that oxygen is frequently added via sparging or other means to enhance biodegradation in environments that have experienced heavy contamination with organic pollutants. It may also be necessary to add inorganic fertilizers in such instances, as the supply of macro- and micronutrients in the soil may be inadequate to sustain microbial degradative activity.

Some groups of chlorinated contaminants, including polychlorinated biphenyls (PCBs), chlorophenols, chlorinated solvents, and certain organochlorine pesticides can serve as alternate electron acceptors, by a process known as reductive dechlorination (also called halorespiration). This process, which occurs only in the absence of oxygen, is a critical step in the degradation of these persistent pollutants. During reductive dechlorination, bacteria transfer electrons from a growth substrate (primarily hydrogen) to the chlorinated compounds, causing the release of chloride ion and the production of various organic daughter products. For example, during reductive dechlorination of perchloroethylene (PCE), the daughter products are TCE, *cis*-1,2-DCE, vinyl chloride (VC), and ethylene, respectively (Figure 3). The occurrence of these degradation products (particularly *cis*-1,2-DCE and VC) suggests that reductive dechlorination of PCE or TCE is occurring in an environment. A similar reaction has been noted for PCBs and several organochlorine pesticides (e.g., 1,1,1-trichloro-2,2-bis

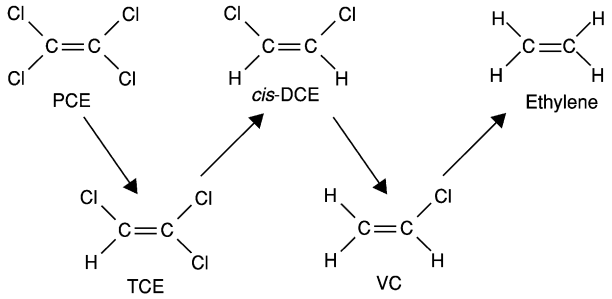


Figure 3 Biodegradation pathway of perchloroethylene (PCE) during reductive dechlorination. *cis*-DCE, *cis*-1,2-dichloroethylene; VC, vinyl chloride.

(4-chlorophenyl)ethane (DDT); pentachlorophenol (PCP); and 3,5,6-trichloro-4-amino-2-pyridinecarboxylic acid (picloram)) under strongly reducing conditions. In soils, reductive dechlorination of PCBs is observed as a shift in the representative PCB congeners found within an environment (there are 209 PCB congeners based on possible positions and numbers of chlorine atoms on the biphenyl molecule) from more chlorinated to less chlorinated congeners. Other chemical pollutants such as perchlorate (ClO_4; an oxidant used in solid rocket fuel) and hexahydro-1,3,5-trinitro-1,3,5-triazine (RDX; an explosive), have also been observed to serve as terminal electron acceptors in soils.

Chemical Properties Affecting Biodegradation

All natural organic compounds are biodegradable. The occurrence and rates of biodegradation of these compounds vary widely depending on environmental conditions and chemical complexity (e.g., lignin biodegradation is extremely slow, even under ideal conditions), but it is generally believed that no natural organic materials are completely resistant to microbial destruction. Given hundreds of millions of years of evolution and the constant competition among microorganisms for limiting nutrients and energy, it is not surprising that microbial enzymes have coevolved in nature in step with new organic compounds. In contrast to nature, however, rapid advances in organic chemistry have allowed mankind to develop and produce an overwhelming array of new organic compounds in a short time period. A majority of these thousands of new compounds have been introduced to the Earth's environment for the first time since the mid-1940s. Many of these chemicals, particularly those that are similar in structure to naturally occurring organics, have proven to be readily biodegradable. Other, more novel synthetics, however, are highly resistant to microbial attack.

Some of the key factors that influence the degradability of a molecule include the type, number, and location of molecular substituents, the type of bonds, the degree of branching, and the molecular weight of the compound (Figure 4). The addition of new organic or inorganic substituents to a common molecular structure is one means of creating novel synthetic compounds. Small differences in the type and location of these substituents can determine whether a molecule is readily biodegradable or highly persistent. Pollutants that are resistant to biodegradation owing to their chemical structure are termed 'recalcitrant' molecules. A number of synthetic organic compounds and pesticides have been developed by adding halogens (particularly chlorine, fluorine, and bromine) to organic compounds. Because the halogen–carbon bond is very stable and resistant to microbial attack, this type of substitution tends to make a molecule recalcitrant, with increasing halogenation leading to increasing persistence. For example, organochlorine pesticides, such as DDT, toxaphene, 2,4,5-trichlorophenoxyacetic acid (2,4,5-T), pentachlorophenol, chlordane, aldrin, hexachlor, and metoxychlor, were once widely used. Many of these pesticides have now been banned from use in the USA and most European countries due to their long-term persistence and negative impacts in various ecosystems. These molecules proved to be highly resistant to aerobic biodegradation and persistent or very slowly degraded under anaerobic conditions. Halogenated pollutants such as PCBs, polybrominated biphenyls (PBBs), freon propellants and refrigerants (e.g., CCl_3F, CCl_2F_2), and chlorinated solvents (e.g., PCE, carbon tetrachloride) have also proven to be persistent contaminants.

Other molecular traits that may decrease a chemical's biodegradability include high molecular weight, low water solubility, and extensive branching. As molecular weight increases to more than approximately 500–600, chemical biodegradability tends to decrease markedly. This reflects, in part, the inability of large molecules to enter the microbial cell. In some instances, extracellular enzymes produced by bacteria or fungi can initiate an attack on a large molecule, breaking it into smaller pieces that can then be metabolized further. For example, it is large size that prevents the biodegradation of many organic polymers that constitute plastics, such as polyethylene and polystyrene. If reduced to a molecular weight of less than 500 from their typical molecular weights of a few thousand to more than 100 000, many of these polymer structures are biodegradable. Molecular branching also tends to reduce chemical biodegradability. The molecular branches tend to sterically interfere with the common degradative enzymes used by microorganisms. For example, many alkanes

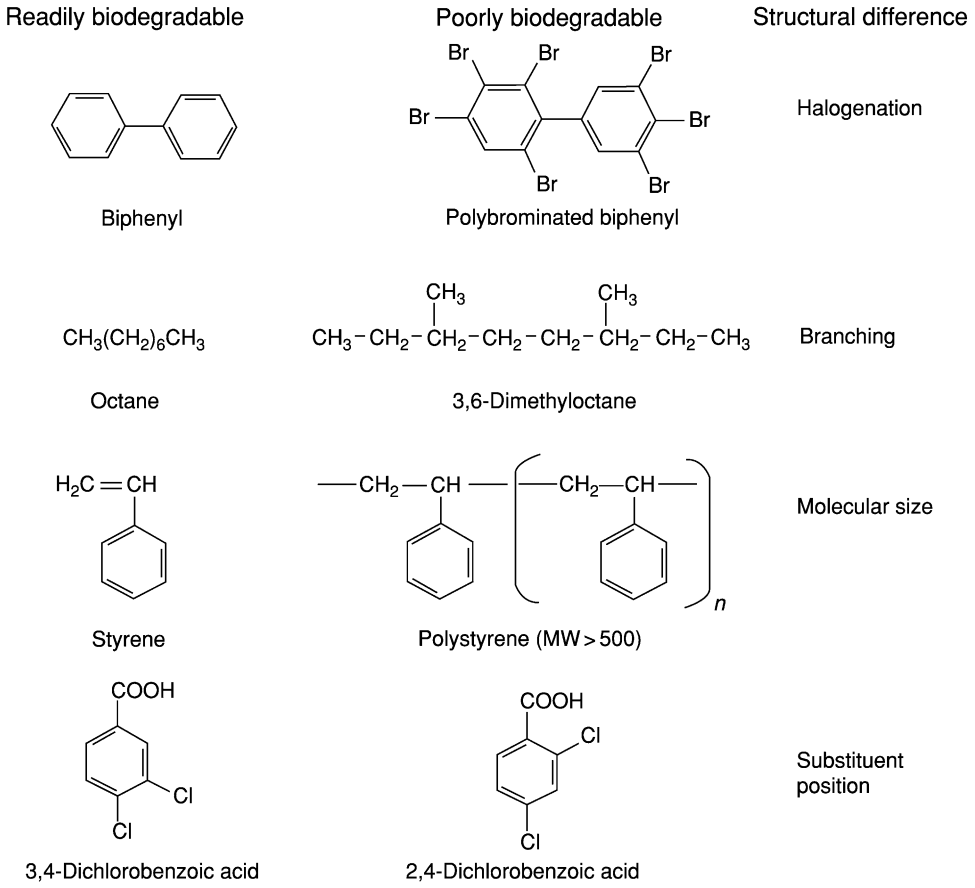


Figure 4 Influence of molecular structure on contaminant biodegradability.

are transformed by bacteria through the sequential oxidation and cleavage of two carbon molecules at a time, in a process known as beta-oxidation. During this process, the alkane is sequentially oxidized (usually at the terminal carbon) to an alcohol, an aldehyde, and then an aliphatic acid. The acid is then cleaved to form acetyl-CoA, which the cell can utilize for energy. This beta-oxidation process continues until the alkane is completely degraded to acetyl-CoA. Branching of the alkane, particularly branching with one or more methyl groups, can block the two-carbon cleavage of beta-oxidation, causing an otherwise biodegradable molecule to become recalcitrant. For this reason, branched alkanes, such as pristaine and phytane, tend to be among the slowest molecules to biodegrade during a crude oil spill. These molecules are frequently employed as internal standards to assess the biodegradation of other, more labile, components of crude oil after an environmental release.

Environmental Factors Affecting Biodegradation

A compound that is readily metabolized in a laboratory may escape breakdown and persist unexpectedly in the field, even if an organism capable of causing its degradation is present. Unless pollutants are accessible to microorganisms and environmental conditions are conducive to microbial growth, contaminants will not undergo biotransformations and may remain in soil for extended periods. Thus, the extent of chemical biodegradation will depend in large part upon the physical, chemical, and biological properties of a contaminated site. A number of important factors influencing the growth of microorganisms and therefore biodegradation are described below.

Bioavailability

Pollutants released into soil will not be biodegraded if they are biologically unavailable. Even the most labile compound may persist if it does not come into contact with microorganisms. Bioavailability of a compound can be reduced if the compound is sorbed to soil particles, entrapped within the soil matrix, or dissolved in a nonaqueous solvent (Figure 5).

Sorption Chemical compounds can bind to solid surfaces through the actions of a variety of attractive forces. In some instances, a compound will adhere to the outer surfaces of soil particles in a process known

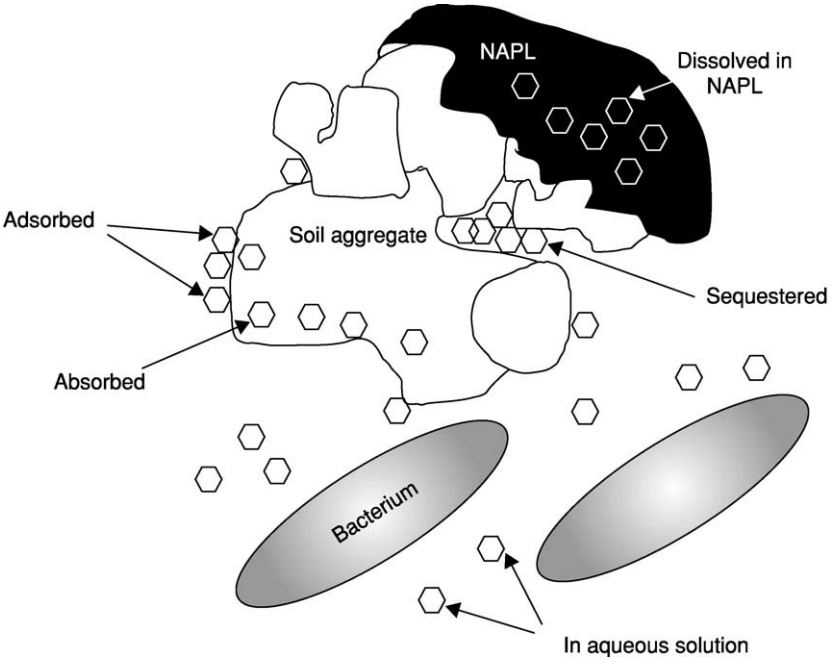


Figure 5 Representation of different potential states of an organic pollutant (cyclohexane) in a soil environment. (Not drawn to scale.) NAPL, nonaqueous-phase liquid.

as adsorption. Absorption refers to retention within the interior matrix of a solid particle. In either case, a compound partitions from a liquid phase (usually from water) onto or into a solid phase. Typically, the nonspecific term 'sorption' is used to describe this phenomenon. Since compounds must generally be in an aqueous phase and enter the microbial cell to be subject to biodegradation, sorption often increases a pollutant's resistance to biological attack. After a given period of time in a soil, a pollutant establishes equilibrium between the sorbed and the aqueous phase. The extent of sorption at equilibrium, as quantified by the pollutant's sorption coefficient (K_d), reflects the characteristics of both the pollutant and the soil matrix. When comparing pollutants of a single structural class (e.g., polycyclic aromatic hydrocarbons, PAHs), those molecules with higher K_d values, which tend to be primarily present in soils in the sorbed phase, generally biodegrade more slowly than those with lower K_d values. In order to enhance pollutant desorption, some bacteria and fungi have been observed to produce extracellular biosurfactants. The release of these surfactant molecules can increase the rate and extent of pollutant desorption and, subsequently, accelerate contaminant biodegradation.

Clay minerals and soil organic matter contain the principle sites to which pollutants can sorb. Charged compounds such as cationic pesticides (e.g., diquat) are often retained by clay minerals. Nonpolar organic molecules, on the other hand, tend to associate with soil organic matter through both adsorption and absorption. For example, PAHs such as phenanthrene, anthracene, and pyrene are subject to sorption and diminished availability as a result of their interactions with humic materials.

Entrapment within the soil matrix After their release into soil, some compounds can slowly move from positions within the soil that are accessible to organisms into positions that are inaccessible. These compounds can experience a time-dependent decline in bioavailability termed 'aging.' Nonpolar organic pollutants, for example, can diffuse deeply into hydrophobic regions of soil organic matter and, as a result, evade microbial cells that might otherwise degrade them. Other compounds may diffuse into soil nanopores, thus becoming physically protected from metabolism by microorganisms. Although the details of the mechanisms involved vary, a compound that has moved into remote sites is said to be sequestered in soil.

The degree to which sequestration occurs depends upon a number of variables, including the properties of the soil and the contaminant involved. Although much work needs to be done to isolate the soil components most responsible for the entrapment of pollutants, soil organic matter content appears to be positively correlated with the extent of chemical sequestration. Soil structure also probably plays a key role in the sequestration of some contaminants. Pollutants may be more protected from biodegradation

in soils possessing highly developed aggregates than in less structured soils. The chemistry of a contaminant is clearly critical also. Preliminary work has suggested an inverse relationship between a compound's water solubility and its potential for sequestration, but the relative importance of this and other properties remains largely unknown. Pollutants for which sequestration has been observed include the PAHs phenanthrene and anthracene, and various pesticides, such as 1,2-dibromoethane, DDT, and chlordane.

Nonaqueous-phase liquid A number of relevant environmental contaminants can be present in soils dissolved in non-aqueous solvents. For example, crude oil is a nonaqueous-phase liquid (NAPL) composed of a number of compounds, many of which have very low water solubility. Since microorganisms are typically active in the aqueous phase only, pollutants in NAPLs can escape biodegradation. The extent to which an individual compound dissolved in a NAPL will be degraded depends upon a number of factors, including the rate of partitioning between the nonaqueous and aqueous phases, the equilibrium concentration of the pollutants in the aqueous phase, the nature of the interface between the two phases, and the toxicity of the NAPL to the microorganisms responsible for degradation. Although individual compounds within a NAPL tend to be degraded at different rates, in general, pollutants in a nonaqueous phase can be extremely persistent.

Nutrient Availability

Heterotrophic microorganisms in soil require a number of macro- and micronutrients, an energy source, and an electron acceptor to grow. Often, a pollutant provides a carbon or energy source, but all other requirements must be fulfilled if biodegradation is to occur. A biodegradable compound released into a nutrient-poor soil (i.e., low available nitrogen or phosphorus) can persist for an extended period of time, because microorganisms that possess the capacity to metabolize the contaminant cannot appreciably grow. In some instances, as during the biodegradation of the explosives 2,4,6-trinitrotoluene (TNT) and RDX under nitrogen-limiting conditions, specific organisms can degrade an organic compound to obtain a scarce noncarbon nutrient (in this case, nitrogen). Some organic phosphorus-containing compounds, such as triethylphosphate (TEP), can also be utilized by microorganisms to obtain phosphorus under conditions in which the compound is unavailable for microbial growth. During bioremediation of soil contaminants, the addition of nitrogen, phosphorus, or other limiting nutrients is often conducted to stimulate the growth of indigenous or added microorganisms.

pH

Some bacteria and fungi can survive and grow under very acidic or alkaline conditions, but the optimum pH for most soil microorganisms is near neutrality. Unless organisms have evolved to live under extreme pH conditions, both high and low pH can lead to the denaturing of microbial enzymes and the hydrolysis of other crucial cell components. It is therefore not surprising that biodegradation tends to be fastest near pH 7 and slows considerably as the pH becomes appreciably more alkaline or acidic. In addition to its direct effect on cells, pH will also influence the chemistry of ionizable pollutants in soil. The sorption and subsequent bioavailability of cationic pesticides (e.g., paraquat, diquat), basic pesticides (e.g., atrazine), and acidic pesticides (e.g., 2,4-D, dicamba) will depend in part on the pH of the soil solution in which they are present. To enhance microbial growth and biodegradation, bioremediation protocols sometimes include chemical treatment of the contaminated soil (e.g., by lime addition) to adjust pH to the optimum for the biodegradative organisms.

Temperature

The growth of microorganisms is highly dependent on temperature, and the prevailing temperature in soil is an important factor controlling the rate and extent to which a pollutant is biodegraded. Most microorganisms in soil are mesophilic, with temperature optima ranging from approximately 25 to 35°C, and growth possible from 15 to 45°C. Therefore, optimal temperatures for biodegradation in soils tend to be within the best growth range for these organisms, with rates increasing as temperatures increase within this range. However, it should not be implied that biodegradation processes cease below 15°C or above 45°C. There are many bacterial species that are psychrophilic, with growth optima less than 15°C, and those that are thermophilic, with growth optima exceeding 40°C and sometimes with extremes as high as 65°C. These organisms can be responsible for contaminant biodegradation in extreme environments such as in polar climates (for psychrophiles) or in compost (for thermophiles). In fact, thermophiles have been applied in numerous *ex situ* bioremediation applications due to their ability to rapidly degrade some contaminants.

Water Content

The moisture content of soil influences the growth and metabolism of microorganisms. As with the other abiotic factors discussed above, tolerance to desiccation varies among organisms. Environments characterized by low water content tend to support only

limited microbial activity. The optimal activity of most microorganisms occurs at soil moistures ranging from 60 to 80% of the soil's water-holding capacity (WHC). Biodegradation can be very limited in dry soils, and pollutants may persist in arid or drought-affected regions longer than in wet areas. Conversely, excessive moisture can limit oxygen diffusion in soil and cause a shift from aerobic to anaerobic conditions. As previously noted, the absence of available oxygen can severely inhibit the disappearance of some pollutants and pesticides, although such conditions may be necessary to promote the degradation of other compounds (e.g., by reductive dechlorination).

Presence of Appropriate Microorganisms

Microorganisms capable of performing a given biodegradation reaction are frequently indigenous in soil. For example, bacteria that aerobically degrade the common gasoline components benzene, toluene, ethylbenzene, and xylenes (BTEX) are ubiquitous in most surface soils. If these contaminants persist for long times in a given soil, it is likely that environmental factors such as nutrient availability, oxygen levels, contaminant availability, or the presence of toxic cocontaminants are impeding the biodegradation process. For some contaminants, however, biodegradative populations are less universally distributed and/or are present only at very low densities. For example, the gasoline oxygenate methyl *tert*-butyl ether (MTBE) has been observed to biodegrade readily in some aerobic soils and aquifers but to be highly persistent in others. In environments where the contaminant persists, the absence of a biodegrading population has been cited as a key factor limiting its destruction. Bioaugmentation has been used at some sites to promote MTBE biodegradation. Similarly, the absence of a specific dehalorespiring microorganism (*Dehalococcoides ethenogenes*) has recently been cited as the key factor determining the extent of biodegradation of PCE and TCE in some anaerobic environments (see **Figure 3**). If this organism is absent, dechlorination of PCE (or TCE) appears to stall at *cis*-DCE or VC, rather than progressing all the way to ethylene. As with MTBE, bioaugmentation has been used to add this organism to specific sites in order to enhance contaminant biodegradation.

Predation

Bacteria are subject to predation by protozoa. Such grazing can influence the biodegradation of soil pollutants, with both inhibition and stimulation possible. Predation by soil protozoa can reduce the number of cells metabolizing a contaminant and cause a decline in the rate at which the compound disappears. Alternatively, due to the positive effect predation has on nutrient availability, grazing may stimulate the growth of microorganisms. Through their metabolism of microbial cells, protozoa release available nitrogen, phosphorus, and other nutrients back into soil and, consequently, stimulate the surviving microbial cells to grow faster. The dynamics of predator–prey relationships in soil are complicated and are not completely understood. Whether predation decreases or increases biodegradation rates depends upon a number of factors, including the starting density of the bacterial prey and predator organisms, the size of the cells, and the properties of the soil in which the organisms live.

Presence of Other Compounds

Microorganisms metabolizing a pollutant can be affected by the presence of other soil contaminants. In particular, it is not uncommon for a cocontaminant (such as a heavy metal or low molecular weight hydrocarbon) to be present in toxic concentrations in a heavily polluted soil, thereby inhibiting the biodegradation of a target compound by killing the biodegradative population. An organism may also preferentially use a compound other than the pollutant as a carbon or energy source and, as a result, biodegradation of the selected contaminant can be slowed. Alternatively, the presence of a second compound can stimulate microbial activity and accelerate biodegradation. For example, bioremediation of a compound that is cometabolized can be enhanced by the addition of a substrate on which the cometabolizing microorganism can grow. The addition of biphenyl to soils has been observed to enhance the rate of degradation of some of the less-chlorinated PCBs. In some instances, however, the growth substrate will actually inhibit degradation of the cometabolized compound owing to competitive enzyme inhibition. The presence of toluene above a given concentration, for instance, can reduce the rates of TCE cometabolism by specific toluene-oxidizing strains. For bioremediation applications, intermittent addition of the growth substrate may be necessary to promote degradation of a cometabolized substrate.

Summary

Biodegradation is an important process by which chemical pollutants may be removed from soil. Microorganisms have a remarkable ability to metabolize an extremely diverse range of synthetic compounds and can be invaluable in the remediation of contaminated environments. The rate and extent of biodegradation of a given contaminant in soil depends on a number of variables, including chemical structure, contaminant bioavailability, and an array

of environmental factors, such as nutrient availability, pH, temperature, soil moisture, predation, and the presence of cocontaminants.

See also: **Pesticides**; **Pollutants:** Effects on Microorganisms; Persistent Organic (POPs)

Further Reading

Alexander M (1999) *Biodegradation and Bioremediation*, 2nd edn. New York: Academic Press.

Alexander M (1981) Biodegradation of chemicals of environmental concern. *Science* 211: 132–138.

Atlas R and Bartha M (1998) *Microbial Ecology: Fundamentals and Applications*, 4th edn. Wokingham, UK: Benjamin/Cummings Science Publishing.

Bollag J-M and Liu S-Y (1990) Biological transformation processes of pesticides. In: Cheng HH (ed.) *Pesticides in the Soil Environment: Processes, Impact, and Modeling*, pp. 169–211. Madison, WI: Soil Science Society of America.

Chaudhry GR (ed.) (1994) *Biological Degradation and Bioremediation of Toxic Chemicals*. Portland, OR: Timber Press, Inc.

Dragun J (1998) Organic chemical biodegradation in soil. In: *The Soil Chemistry of Hazardous Material*, 2nd edn., pp. 664–742. Amherst, MA: Amherst Scientific Publishers.

Focht DD (1997) *Principles of Biodegradation*. Dordrecht, the Netherlands: Kluwer Academic Publishers.

Howard PH, Boethling RS, Jarais WF, Meylan WM, and Michalenko EM (1991) *Handbook of Environmental Degradation Rates*. Chelsea, MI: Lewis Publishers, Inc.

Maier RM (2000) Microorganisms and organic pollutants. In: Maier RM, Pepper IL, and Gerba CP (eds) *Environmental Microbiology*, pp. 363–400. New York: Academic Press.

Matsumura F (1982) *Biodegradation of Pesticides*. New York: Plenum Publishing Corp.

Wackett LP and Hershberger CD (2001) *Biocatalysis and Biodegradation: Microbial Transformation of Organic Compounds*. Washington, DC: ASM Press.

Effects on Microorganisms

M E Fuller, Shaw Environmental, Inc., Lawrenceville, NJ, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Many human activities have resulted in the intentional or unintentional release of pollutants into the soil environment. Some of these compounds are natural or derived from natural sources (i.e., petroleum hydrocarbons, metals), and would hence not be considered pollutants *per se* in their native environment. Their presence in soil is considered to be polluting because they are not native to soil or their quantities are much higher than normally occurs in soils. In contrast, many pollutant compounds are partially or totally man-made (i.e., pesticides, explosives), possessing no naturally occurring analogs.

While many of the pollutants that have been introduced into the soil environment have been shown to be biologically degraded by soil microorganisms, some pollutants have also proven to be toxic to various members of the soil microbial community. The adverse affects of toxic pollutants can be manifested as changes in the structure (composition) of the microbial community, alterations in the functions (activity) of the microbial community, or quite often both.

The effect of pollutants in terms of overall soil 'health' or ecosystem functioning has undergone considerable theoretical and empirical expansion since 1979. A conceptual framework for assessing pollutant effects is presented in **Figure 1**. This framework was developed based on side-effects testing of agricultural pesticides but is widely applicable to all classes of pollutants. The main aspects shown in **Figure 1** are that pollutant effects can range from negligible to critical (compared with a no- or preexposure control) depending on both the magnitude and the duration of the observed deficit in a given soil parameter. The soil parameter can be an aspect of either the microbial community structure (i.e., specific microbial populations) or function (i.e., nitrogen-cycling activities). Under this conceptual framework, large parameter deficits can be acceptable if they are short in duration, and small, long-term parameter deficits can also be acceptable.

Types of Effects

Microbial Community Composition

The introduction of a pollutant into the soil ecosystem can result in several types of general and specific changes in the microbial community composition. These effects are summarized in **Figure 2** and are a function of the concentration of the pollutant, the duration of the exposure, and the ability of the compound to be biodegraded.

General If the pollutant is very toxic or is present at very high concentrations, it can reduce the abundance of all microorganisms present in the soil due to cell lysis or death. In contrast, if the pollutant is introduced at a moderate concentration and can also serve as a carbon and/or energy source for soil microbes, an

overall increase in the abundance of microorganisms (bacteria, fungi, and protists) may be observed.

The general effects of pollutants on microbial community composition can be assessed using a variety of methods. Systematic laboratory experiments are designed to determine the abundance of total microorganisms before and after exposure to the test pollutant, which is usually introduced into the soil at several concentrations. Field samples collected from inside a contaminated area and from an adjacent uncontaminated area (matched as closely as possible in terms of soil type, vegetation, etc.) also allow the general effects of pollutants to be evaluated.

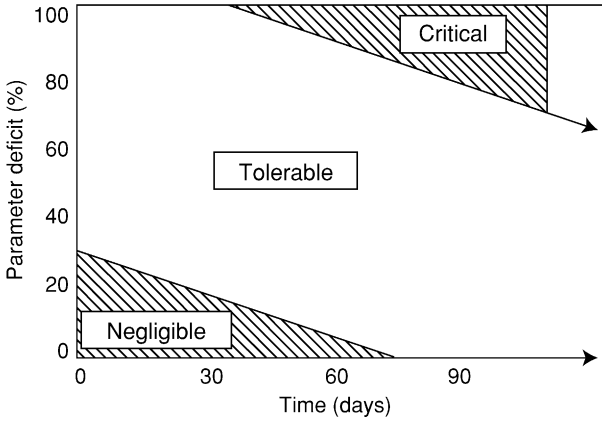


Figure 1 A conceptual framework for assessing the severity of pollutant effects based on the magnitude and duration of the observed effects. Adapted from Domsch K-H, Jagnow G, and Anderson T (1983) An ecological concept for the assessment of side effects of agrochemicals on soil microorganisms. *Residue Reviews* 86: 65–105, Figure 10, with permission. © 1983, Springer-Verlag, New York, Inc.

Microbial numbers or microbial biomass can be determined by a number of methods, including, but not limited to: spread or pour plate and/or most probable number (MPN) methods, using general microbiological media to enumerate total culturable heterotrophs; epifluorescent direct microscopic counts of total acridine orange- or 4′,6′-diamidino-2-phenylindole (DAPI)-stained cells; chloroform fumigation/extraction procedures to quantify microbial biomass carbon (converted to number of cells based on a factor); extraction and analysis of the total phospholipid fatty acid (PLFA) concentration (converted to number of cells based on a factor); and extraction of nucleic acids combined with quantitative polymerase chain reaction (PCR) techniques using Bacteria-, Fungi-, and Protista-specific primers.

Specific Some pollutants exert toxic effects on specific classes of soil microbes. The pollutants may kill or reduce the number of active (vegetative) cells, while spores and other resistant microbes may be relatively unaffected. In many cases, reductions in the abundance of a specific microbial population are due not to direct killing of specific cells, but to inhibition of specific cellular activities and functions to the extent that the cell cannot remain viable or reproduce. Similar to general effects, if the pollutant is present at moderate levels, the abundance of specific

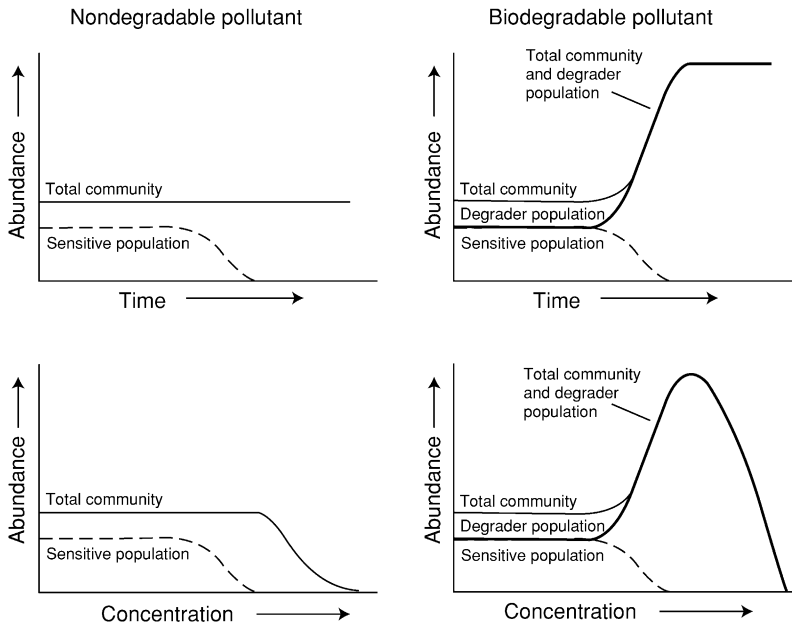


Figure 2 Expected changes in soil microbial community structure in response to biodegradable and nonbiodegradable pollutants with respect to duration and magnitude of exposure.

microorganisms able to use the pollutant as a carbon and/or energy source may increase substantially compared with their preexposure level.

Effects of pollutants on specific microbial populations are assessed in a manner similar to that described above. Most of the enumeration methods have variants to allow the concentration of specific types of microorganisms to be determined. Plate counts and MPN techniques can be performed using media tailored to certain types of organisms (i.e., purified agar incubated in the presence of hydrocarbon vapors to enumerate hydrocarbon degraders; anaerobic liquid media with nitrate or sulfate as the sole terminal electron acceptor to enumerate nitrate- or sulfate-reducing organisms, etc.). Epifluoresecence microscopy can be done using fluorescently labeled deoxyribonucleic acid (DNA) probes directly to count specific cells of a given microbial strain in contrast to total cells. The profile of different fatty acids that is generated using most PLFA methods can be interpreted to determine the proportion of different microbial populations (Gram-positive versus Gram-negative; sulfate reducers; methanogens; anaerobes, etc.) in a given sample. Lastly, the PCR methods can be adapted for use with genus- and even strain-specific primers/probes, allowing very small subsets of the overall microbial community to be enumerated.

Microbial Community Function

Pollutants have the ability to decrease (or increase) different activities or functions of the soil microbial community. Some of these changes are in direct response to the pollutant itself, while others may be the indirect result of changes brought about by the pollutant (i.e., changes in pH). In addition, the changes may be in general activities common to almost all microbial cells, or they may be activities that only a small subpopulation of the microbial community is capable of performing. Redundancy in the microbial community (i.e., more than one microbial strain able to perform the same activity) can be a major determining factor for whether the pollutant causes short- or long-term changes in the overall function of the soil microbial community. However, the heterogeneous nature of microbial distribution does limit the potential for redundancy to mitigate the adverse effects of pollutants at the microscale, and is most likely to be observed at the meso- and macro-scale (i.e., soil ecosystem *in toto*).

General Carbon mineralization (conversion of carbonaceous material to carbon dioxide) is the most general function of soil microbial communities that can be affected by exposure to pollutants. The majority of microorganisms in most soils are chemoheterotrophic, metabolizing available organic material as their carbon and energy source for growth and cell maintenance. Pollutants can interfere with carbon mineralization by direct interaction with specific degradative enzymes, but often act by disrupting or decoupling the cellular systems needed for energy generation and flow (i.e., production of adenosine triphosphate (ATP), regeneration of nicotinamide adenine dinucleotide (phosphate) (NAD(P)), and flavine adenine dinucleotide (FAD), etc.). Biodegradable pollutants will generally increase the net carbon mineralization (respiration) of a soil because of overall increases in microbial activity and microbial cell densities. However, nonbiodegradable pollutants can also cause increased respiration by triggering stress responses in the microbial community whereby more energy (and hence, carbon throughput) is needed to cope with the pollutant stressor.

Changes in carbon mineralization are usually determined by measuring the basal respiration of soil in the presence of the pollutant and comparing it with the same soil without the pollutant present. In many cases, substrate-induced respiration is also measured by adding a labile carbon source such as glucose and measuring the resulting respiratory production of carbon dioxide by soil in the presence and absence of the pollutant. Any pollutants (or pollutant concentrations) that cause significant decreases in carbon metabolism are also likely to affect adversely the input of fresh carbon (i.e., plant root exudates, plant biomass), reducing overall carbon cycling and raising concerns about long-term soil health. Finally, if the pollutant is biodegradable, its effects will probably decrease over time as its concentration is reduced.

Specific The cycling of nitrogen in the soil is by far the most studied microbial activity with respect to pollutant effects. The nitrogen cycle is shown in Figure 3. The consensus that has emerged from decades of research is that ammonification and denitrification activity is widely distributed throughout the soil microbial community, while ammonium oxidation, nitrate oxidation, and nitrogen fixation activity is concentrated in only a few microbial genera. This underlies the observation that nitrification, and especially ammonium oxidation, is very sensitive to many types of pollutants. Two bacterial genera, *Nitrosomonas* and *Nitrobacter*, are believed to be the predominant autotrophic nitrifying populations in soil, although some heterotrophic nitrifiers also exist. These two strains are highly sensitive not only to pollutants, but also to a range of environmental conditions, especially pH. Nitrification, therefore,

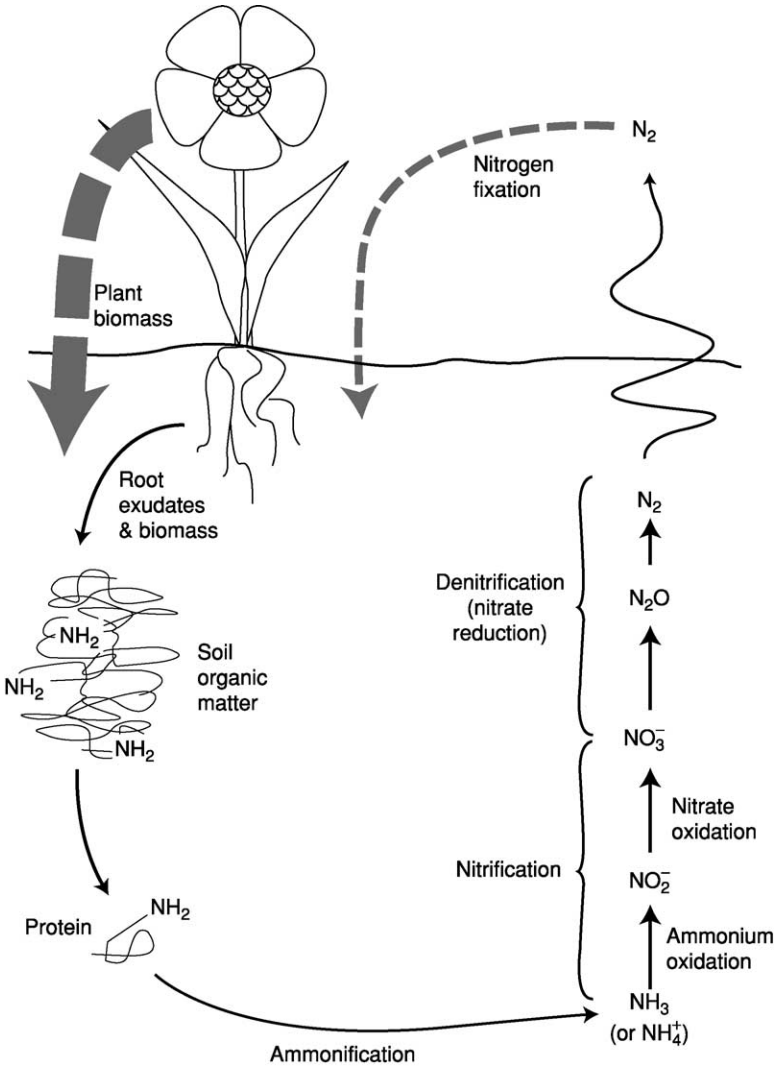


Figure 3 Soil nitrogen cycle.

may be thought of as a 'sentinel' soil activity, serving as the first indicator of soil ecosystem disturbance by pollutants. Although nitrogen-fixation ability is limited to a few symbiotic and free-living bacterial genera, nitrogen-fixing activity *per se* seems to be less sensitive to most pollutants than nitrification.

The effects of pollutants on the microbial activities associated with the cycling of other nutrients such as sulfur, phosphorus, etc., have not been extensively studied. This is probably due to the perceived lesser importance of these elements compared with carbon and nitrogen to the overall health of the soil microbial community. It has been assumed that these elements are present in excess what is needed by the soil biota, while carbon, and more so nitrogen, are the limiting nutrients. Because the activities involved in the cycling of phosphorus, sulfur, etc., appear to be widely distributed within the microbial community, only in extreme cases would dramatic pollutant effects be expected to be observed (i.e., massive decimation of total microbial biomass).

A final category of specific effects of pollutants relates to changes in the activities of soil enzymes. Some of the more generic enzymes include: dehydrogenase, acid and alkaline phosphatase, phosphodiesterase, arylsulfatase, amidases, urease, and protease. Many soil enzymes are released from microbial cells into the open soil environment, with the goal of liberating the soil nutrients associated with soil organic matter (SOM). Whereas disruption of carbon mineralization (respiration) is a general effect of pollutants on soil microbial community function, the effects of pollutants on soil enzymes disrupt other necessary aspects of carbon cycling in soils (i.e., conversion of large, extracellular organic molecules to smaller molecules available for cellular uptake and metabolism).

Several reports have focused on the effects of pollutants on soil enzymes, but the exact nature of the effects have been less-rigorously investigated (i.e., are the pollutants directly interacting with the enzymes, or with the organisms that are producing and releasing the enzymes?). Again, due to the relatively widespread and generic activities of these enzymes, pollutant effects seem to be relatively small except in cases of excessive and/or prolonged contamination. Generally, dehydrogenase and urease have been shown to be more inhibited by pollutants, while phosphatase has often been observed to be stimulated.

Effects of Different Classes of Pollutants

The effects that have been observed to be caused by different types of soil pollutants are presented in Table 1. This table, and the discussion in this section, reflects the general consensus of the current literature and hence should be considered as a guide regarding potential pollutant effects, in contrast to a comprehensive summary of all the available data. Indeed, in many cases the published literature presents contrasting results, pointing to the need for continued research with more standardized (and environmentally relevant) experimental procedures. Many of the effects are highly dependent on other physiochemical soil parameters (i.e., pH, redox, cation exchange capacity, organic matter, or clay content) which modulate the bioavailability of the pollutants. Information on the effects of a specific compound should be sought out in the peer-reviewed literature.

Petroleum Products

The effects of petroleum-based products, ranging from short, straight-chain hydrocarbons to highly condensed polynuclear aromatic hydrocarbons (PAH), have been studied largely in response to the reality of widespread pollution by these compounds due to accidental releases at gas stations and oil refineries. However, there are considerable 'natural' releases of these compounds into the environment from seeps (e.g., the La Brea tar pits in California, USA; undersea seeps in the Gulf of Mexico; etc.).

The most general effect of petroleum products is an increase in microbial biomass and changes in the community structure due to utilization of the compounds as carbon and energy sources (Table 1). Respiration is also usually increased, but effects on other activities are minimal.

Halogenated Compounds

The compounds in this class include compounds used in chlorinated degreasing and dry-cleaning such as perchloroethylene (PCE) and trichloroethylene (TCE), polychlorinated biphenyls (PCBs), and other halogenated industrial chemicals. Compounds containing chlorine and bromine have been more closely studied than fluorinated and/or iodinated chemicals.

Halogenated compounds enter the environment through accidental releases from industry, leaking waste drums, and landfill leachate. Very high levels of contamination usually coalesce into the groundwater below the soil as dense, nonaqueous-phase liquids (DNAPL), but vapors from these areas may still rise up through the subsurface and affect the soil community. In general, very few effects are observed from contamination with moderate levels of halogenated aliphatic compounds (Table 1). However, chlorinated aromatic compounds have been observed to affect some aspects of the microbial community structure (biomass and diversity) and function (especially nitrification).

Pesticides

Pesticides (including fungicides, nematocides, insecticides, and herbicides) are the only class of compounds that are intentionally introduced into the environment on a regular basis. The use (and misuse) of pesticides, the development of new pesticides, and the registration of new pesticides (particularly in Europe), has been one of the primary driving forces behind the study of the effects of pollutant on soil microorganisms. The major focus for pesticides has been on 'side-effects testing,' during which the effects of the compound(s) on soil organisms other than the target 'pest' organism are examined. Initially, broad measurements of general effects on soil functions that were deemed important indicators of overall soil health (especially from an agricultural perspective) were studied. More recent work, however, has employed molecular methods to examine effects down to the level of single microbial species and how pesticides may influence long-term soil health.

In a broad study of more than 50 pesticides, the recommended agricultural doses caused no long-term effects on the structure or function of the soil microbial community. Interestingly, when effects are observed, it is often attributed to the pesticide breakdown products rather than the parent compounds.

Explosives

These compounds have only recently been recognized as potentially significant pollutants. The major compounds in this class include: the nitroaromatics 2,4,6-trinitrotoluene (TNT), 1,3,5-trinitrobenzene (TNB), 2,4,6-trinitrophenol (picric acid), and *N*-methyl-*N*,2,4,6-tetranitroaniline (tetryl); the nitramines hexahydro-1,3,5-trinitro-1,3,5-triazine (RDX),

Table 1 Summary of effects of different classes of pollutants on soil microorganisms

	Parameter affected									
	Community structure		*Community function*							
Pollutant class	*Total biomass*	*Specific populations*	*Respiration*	*Nitrogen fixation*	*Ammonification*	*Nitrification*	*Denitrification*	*Soil enzymes*	*Sulfur cycling*	*Phosphorus cycling*
Hydrocarbon										
Aliphatic	+	+	+	None/−	?	−	?	?	?	?
Aromatic	+	+	+	None	None/−	−	None/−	None	?	?
PAH	None/+	+	+	None	?	−	?	?	?	?
Halogenated										
Aliphatic	None	None	None	?	None	−	None	?	?	?
Aromatic	−/None/+	+	−	?	?	?	−	?	?	?
Pesticides[a]	None	None/+/−	None/−	None/+/−	None/+	None/−	−	?	−/None	None
Explosives	None/−	+	None/−	−	?	+/−/None	?	?	?	?
Metals	−/None	+/−/None	+/−	−/+	?	−	−	−	−	−

[a]Applied at recommended rates for agricultural use.

+, increase in parameter; −, decrease in parameter; ?, no or not enough data available; PAH, polyaromatic hydrocarbon.

and octahydro-1,3,5,7-tetranitro-1,3,5,7-tetrazocine (HMX); and the nitroesters nitroglycerin (NG), nitrocellulose (NC), and pentaerythritol tetranitrate (PETN). Gross pollution with explosives is most often associated with manufacturing and training activities, and historical soil contamination for 50 years or longer has been reported at many military installations where these activities have occurred.

The toxic effects of these compounds in the environment have only recently been initiated, and reports of effects on soil (micro)organisms are sparse. Most of the explosives are sparingly soluble and, when mixed with the binders or waxes used in actual explosives formulations, their bioavailability is quite limited. The summary in Table 1 indicates that some effects have been observed, but more research is clearly needed.

Metals

Metals enter the soil environment primarily through atmospheric deposition downwind of smelting operations or via leaching from mine spoils. Some metals are essential for microbial cells, some are needed in trace quantities, and others are nonessential and/or toxic. The most toxic metals are the heavy metals cadmium, lead, and mercury, but high concentrations of copper, zinc, chromium, iron, manganese, nickel, cobalt, arsenic, and silver can also exert adverse affects on microorganisms.

The toxicity can be manifest by replacement of essential metal reaction centers of enzymes, reaction of metal ions with the sulfhydryl and phosphate groups of proteins, or displacement of key structural elements. Metal toxicity is highly dependent on the pH and redox state of the soil, as these strongly influence the oxidation state and solubility of metals and hence their bioavailability to soil microbes. Microorganisms themselves have also developed a broad range of resistance mechanisms to counteract the effects of metals. These include facilitated transport of the metal ions out of the cell, or complexation of the metal with reactive molecules (glutathione, etc.) that render it less bioavailable.

See also: **Bacteria:** Soil; **Biodiversity**; **Carbon Cycle in Soils:** Dynamics and Management; **Enzymes in Soils**; **Fungi**; **Heavy Metals**; **Hydrocarbons**; **Microbial Processes:** Community Analysis; **Nitrogen in Soils:** Cycle; **Organic Residues, Decomposition**; **Oxidation–Reduction of Contaminants**; **Pesticides**; **Phosphorus in Soils:** Biological Interactions; **Pollutants:** Biodegradation; Persistent Organic (POPs); **Pollution:** Industrial; **Quality of Soil**; **Remediation of Polluted Soils**; **Sorption:** Metals; Organic Chemicals; **Sulfur in Soils:** Biological Transformations; **Waste Disposal on Land:** Municipal

Further Reading

Domsch K-H, Jagnow G, and Anderson T-H (1983) An ecological concept for the assessment of side effects of agrochemicals on soil microorganisms. *Residue Reviews* 86: 65–105.

Hicks R, Stotzky G, and van Voris P (1989) Review and evaluation of the effects of xenobiotic chemicals on microorganisms in soil. *Advances in Applied Microbiology* 35: 195–253.

Lal R (1982) Accumulation, metabolism, and effects of organophosphorus insecticides on microorganisms. *Advances in Applied Microbiology* 28: 149–200.

Lal R and Saxena D (1982) Accumulation, metabolism, and effects of organochlorine insecticides on microorganisms. *Microbiological Reviews* 46: 95–127.

McGrath SP, Chaudri AM, and Giller KE (1995) Long-term effects of metals in sewage sludge on soils, microorganisms and plants. *Journal of Industrial Microbiology* 14: 94–104.

Nies DH (1999) Microbial heavy-metal resistance. *Applied Microbiology and Biotechnology* 51: 730–750.

US Environmental Protection Agency (1996) *Terrestrial (Soil-core) Microcosm Test*. Ecological effects test guidelines. OPPTS 850.2450. Washington, DC: Government Printing Office.

US Environmental Protection Agency (1996) *Soil Microbial Community Toxicity Test*. Ecological effects test guidelines. OPPTS 850.5100. Washington, DC: Government Printing Office.

van Straalen NM (2002) Assessment of soil contamination – a functional perspective. *Biodegradation* 13: 41–52.

Persistent Organic (POPs)

D Johnson, Rothamsted Research, Harpenden, UK

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Organic pollutants are ubiquitous environmental contaminants and, for many of the compounds, the environmental burden has risen, largely as a result of anthropogenic activity in the nineteenth and twentieth centuries. Soil can be a major environmental sink for many of the organic pollutants released into the atmosphere. Once in a soil, nonionic organic pollutants can often be highly persistent. The primary rate-limiting factors determining the long-term persistence of persistent organic pollutants (POPs) in soil are thought to be fundamental sorption/desorption mechanisms that control the distribution of chemical between the solid and aqueous phases of the soil. Soil POPs are potentially available to many organisms, including soil microflora, macrofauna

(earthworms are particularly important), and grazing animals. However, the availability of soil POPs tends to decline with time. The bioavailability of POPs to soil organisms is important in determining the potential hazard presented by a contaminated soil and may also influence the efficiency of bioremediation of contaminated soils.

Persistent Organic Pollutants

During the nineteenth and twentieth centuries humans have produced and released vast quantities of organic compounds into the biosphere. These are widely distributed throughout the environment and many are thought to be carcinogenic and mutagenic. This has led to the prioritization of organic pollutants by the United States Environmental Protection Agency (USEPA) and the European Community (EC). These lists include polycyclic aromatic hydrocarbons (PAHs), polychlorinated biphenyls (PCBs), polychlorinated dibenzo-*p*-dioxins and furans (PCDD/Fs), chlorobenzenes (CBs), phthalate acid esters, and other miscellaneous compounds. These compounds enter the environment through a wide variety of processes including inefficient combustion, dry and wet deposition from the atmosphere, widespread pesticide use, and wastewater drainage from industrial and domestic sources into sewers. Hence, the soil is a major sink for POPs. Once in a soil, high-molecular-weight non-ionic organic compounds may persist for long periods.

POPs are characterized by their relatively high octanol/water partition coefficient (log Kow) and low water solubility. Thus, once in a soil they tend be resistant to microbial degradation and prone to bioaccumulation. Atmospheric transport processes such as 'global distillation' result in an almost ubiquitous contamination of soils, often thousands of kilometers from major POP sources. The highest soil concentrations of POPs are usually found near industrial point sources, urban areas, and in organic-rich soils such as those that have received sludge amendment. The recent decline in the manufacturing industry in many western economies has left large areas of soil contaminated by POPs and requiring remediation before they can be redeveloped.

Legislation relating to organic compounds in soil is largely based on toxicological criteria with the ultimate aim of protecting wildlife and human health. Thus, it is important that the 'availability' of compounds to transfers from the soil to the organisms is accurately understood in a quantitative manner. Recent research on the fate and behavior of organic chemicals in soils suggests that the availability of an organic compound from the soil into the food chain may be highly dependent upon the form in which a compound is present in a soil. This is influenced by a number of factors, including its residence time within the soil, physicochemical properties of the compound and soil, and how it entered the soil. Desorption of sorbed compounds into the soil solution enhances biodegradation, leaching, volatilization, plant uptake and other chemical/photochemical transformations. Whilst these processes are generally enhanced by the chemical being present in the bulk soil solution, chemical degradation, photochemical transformations and biodegradation of soil-sorbed compounds are also possible. Furthermore, soil-sorbed compounds may also be lost from areas of soils by the erosive action of wind or running water, uptake by invertebrates or direct ingestion of soil by grazing livestock.

The persistence and relative significance of each loss pathway in a given soil depends on the physicochemical properties of the organic chemical and environmental conditions. However, regardless of the main loss pathway, the overall loss of organic chemicals from soils is often biphasic. This means that a short period of rapid initial loss is most often followed by a longer period of slow dissipation. Loss processes in soils exhibit similar patterns. The primary rate-limiting factors governing this biphasic behavior are thought to be fundamental sorption/desorption mechanisms that control the distribution of chemical between the solid and aqueous/gaseous phases of soil and hence, the supply of chemicals available to the various loss pathways. The availability of organic compounds in a soil will affect both the toxicological threat and the potential to remediate a contaminated site. Further details relating to specific technologies available for contaminated soils can be found in the article Remediation of polluted soils (*See* **Remediation of Polluted Soils**).

Physicochemical Properties of Persistent Organic Pollutants in Soils

The principal persistent organic pollutants found in soils are PAHs, PCBs, CBs, and PCDD/Fs. Table 1 displays comparative physicochemical data for a number of selected organic chemicals that occur in contaminated soils.

Persistence and Aging Behavior of Organic Chemicals in Soils

Figure 1 illustrates the decay curves of three model nonionic organic compounds after they enter a soil. The persistence and overall loss of organic chemicals from soils following their entry into a soil are dependent on the physicochemical properties of the soil and the chemicals themselves, environmental

Table 1 Physicochemical properties of selected chlorobenzenes, polycyclic aromatic hydrocarbons, polychlorinated biphenyls, and polychlorinated dibenzo-*p*-dioxins and furans

Compound	*Log Kow*	*Log Koc*	*K′ H*	*Log Koa*
1,4-DCM	3.5	3.43	6.46×10^{-2}	4.69
HCB	5.5	5.25	2.72×10^{-2}	7.07
Naphthalene	3.37	3.14	1.74×10^{-2}	5.13
Phenanthrene	4.57	4.4	1.31×10^{-3}	7.45
Benzo(a)pyrene	6.04	6.24	1.86×10^{-5}	10.77
Coronene	6.75	6.4	1.72×10^{-7}	13.51
2,3,7,8-TCDD	6.8	6.2	1.35×10^{-3}	9.67
OCDD	8.2	6.97	2.76×10^{-4}	11.76
PCB-18	5.72	5.15	1.95×10^{-2}	7.43
PCB-52	5.83	4.99	1.75×10^{-2}	7.59
PCB-101	6.3	6.3	8.74×10^{-3}	8.36
PCB-138	6.69	6.24	1.00×10^{-3}	8.69
PCB-206	7.94	6.66	5.59×10^{-3}	10.19

log Kow, octanol/water partition coefficient; log Koc, octanol/carbon coefficient; K′ H, Henry's constant; log Koa, octanol/air coefficient.

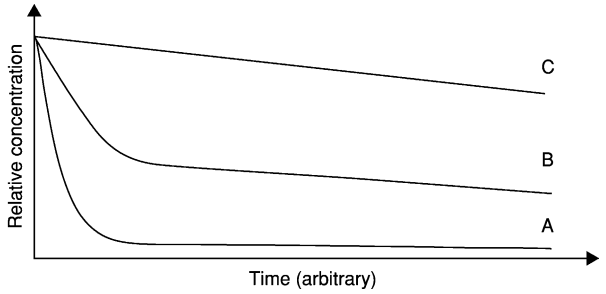


Figure 1 Hypothetical loss profiles for organic chemicals in a soil. A–C are discussed in the text.

factors, including temperature and precipitation, and anthropogenic factors such as cultivation and drainage.

Curve A (Figure 1) is typical of organic contaminants with short half-lives that exhibit relatively rapid dissipation with most of the residues ultimately being removed from the soil. Such decay curves are typical of compounds that are volatile, water-soluble, or easily degraded (or a combination of these). Examples of this chemical type are the lighter chlorinated benzenes and volatile aromatic hydrocarbons. At the other extreme, chemicals that are nonvolatile, relatively water-insoluble and recalcitrant are characterized by persistence profile C (Figure 1). Such chemicals are strongly sorbed and persist in soils over often extremely long periods. Examples include PCDDs and PCDFs. Most organic compounds fall within these two extremes, exhibiting persistence profiles similar to curve B. Curves A and B are essentially similar in shape with an initial period of rapid loss followed by a much slower dissipation of the chemical. This leads to a 'persistent residual fraction' of the compound forming in the soil, the size of which will be determined largely by the physicochemical properties of the compound and soil. As commented earlier, a loss profile of this shape is often referred to as 'biphasic' and is typical of organic chemicals with intermediate sorption potentials.

The 'Aging' of Persistent Organic Pollutants Within Soils

As can be seen in Figure 2, when a nonionic organic compound enters a soil it undergoes a number of loss processes including leaching, volatilization, chemical and biological degradation, and plant uptake. In addition, sorption- and transport-related processes mean that an increasing component of the soil-borne compound becomes 'irreversibly bound' within the soil matrix, i.e., they cannot be extracted by conventional analytical techniques. The result is that, as the residence times of organic compounds increase within a soil, they become increasingly difficult to extract. This phenomenon has become known as the 'aging effect.' There is also an assumption that there is an analogous decline in a compound's bioavailability as soil residence time increases and extraction efficiency declines, but this may be erroneous (see later).

Figure 2 demonstrates that this process of declining extractability of soil-bound compounds is, in part, due to the formation of irreversibly bound residues within the soil matrix. The development of irreversibly bound residues is a major contributing factor in the formation of persistent residual fractions of organic chemicals in soils. Addressing this issue is complicated by the fact that the experimental data are highly dependent on the extraction method used and the knowledge that the efficiency of extraction of an organic chemical from a soil will decline with an increase in contact time between a chemical and the soil, i.e., the 'aging effect.'

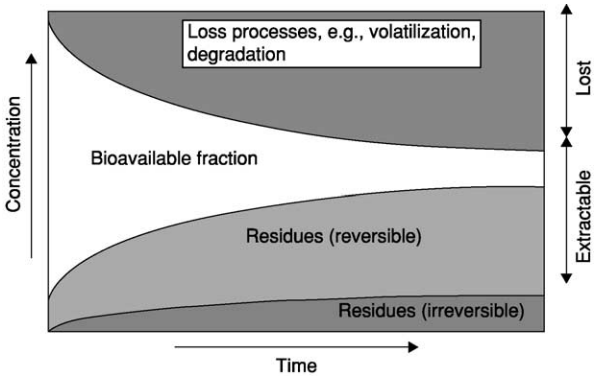


Figure 2 Dynamics of organic chemical persistence, bioavailability and loss in soils.

Mechanisms Contributing to the Persistence of Organic Chemicals in Soils

Several groups of mechanisms are thought to be involved in the 'aging effect'. The processes involved can be split into transport-related and sorption-related mechanisms. Transport-related mechanisms are those where sorption/desorption of organic chemicals is controlled by water flow in soils. Due to their particle size, distribution, and management, soils may develop distinct structural arrangements ranging from simple homogeneous granular sands to pronounced columnar and prismatic clay structures. In the latter soil type, swelling and shrinkage may create a network of cracks that rapidly conduct most water while the remainder of the water moves relatively slowly through the soil matrix. Water flow in sandy soils is more uniform, but it is important to recognize that preferential flow may also occur in such soils. Comprehensive reviews of water movement in structured soils and the implications of this for solute transport and fate will not be discussed in detail here but readers requiring more information can find further reading at the end of this article.

Sorption-Related Mechanisms

The nature and significance of sorption-related mechanisms are less well understood than transport-related mechanisms and much research is still required in this field. The literature still contains many contradictions as to what are the dominant sorption-related mechanisms controlling the persistence of nonionic organic chemicals in soils. Sorption mechanisms can be divided into two distinct classes: diffusive mass transfer and chemisorption. Diffusive mass transfer mechanisms include film diffusion, pore diffusion, retarded intraparticle diffusion, and intraorganic matter diffusion. Film diffusion relates to the resistance encountered by an organic chemical when crossing a thin film of water surrounding soil solids/aggregates, or when moving from the solid phase into the bulk solution or gaseous phases. Film diffusion is generally regarded as insignificant in comparison to other processes discussed here. Pore diffusion is the molecular diffusion of organic chemicals within pores between soil particles inside soil aggregates. The current literature suggests that, similar to film diffusion, pore diffusion rarely contributes to the formation of persistent chemical residual fractions.

Chemisorption occurs when a functional group of an organic chemical interacts with reactive moieties on soil solids through the formation of strong chemical bonds, principally covalent bonds. Such interactions give rise to the formation of a fraction that is irreversibly sorbed, despite exhaustive extraction procedures, and has become known as 'bound residues' in the pesticide science field. However, pesticides are often ionic, in contrast to nonionic compounds such as nonionic persistent organic pollutants, which are less likely to form chemical bonds with soil solids. Thus chemisorption is unlikely to be a major rate-limiting factor governing the persistence and loss of hydrophobic organic chemicals in soils. Where a contaminant has a tendency to ionize or contains polar functional groups, then chemisorption may be important. In addition, biologically mediated degradation processes may alter the structure of nonionic organic chemicals so they are more susceptible to chemisorption. A 1996 study demonstrated that once partial oxidation of a PAH occurs by biologically mediated degradation, the irreversible chemisorption of metabolites can occur and in some cases can be so significant that the process becomes competitive with biodegradation. The significance of chemisorption in the formation of persistent residual fractions has yet to be fully elucidated. Further investigation is clearly necessary because the exact processes involved in persistent residual fraction formation may be important in determining whether the compounds remain chemically and, more importantly, biologically unavailable.

Sorption Retarded Intraparticle Diffusion

Of the mass transfer mechanisms involved in the formation of persistent residual fractions it is generally agreed that sorption retarded intraparticle diffusion (SRPD) and sorption retarded intraorganic matter diffusion (SROMD) are the two primary rate-limiting factors. SRPD has been defined as 'aqueous-phase diffusion of solute within pores of microporous particles (e.g., sand grains) mediated by retardation resulting from instantaneous sorption to pore walls.' However, SRPD is often used as an umbrella term for a number of similar processes that occur by diffusion of organic compounds through pores. The above definition of SRPD refers to the molecular diffusion of chemicals in pore water that is retarded, chromatographic-like, by local sorption on pore walls. This process is more marked in larger mesopores (diameter $>2\,\mu$m), and the sorption may or may not be aided by sorption to and within organic matter lining these pores (see later for description of SROMD). As the pore's diameter decreases, the tortuosity of the pore tends to increase, further hindering diffusion processes.

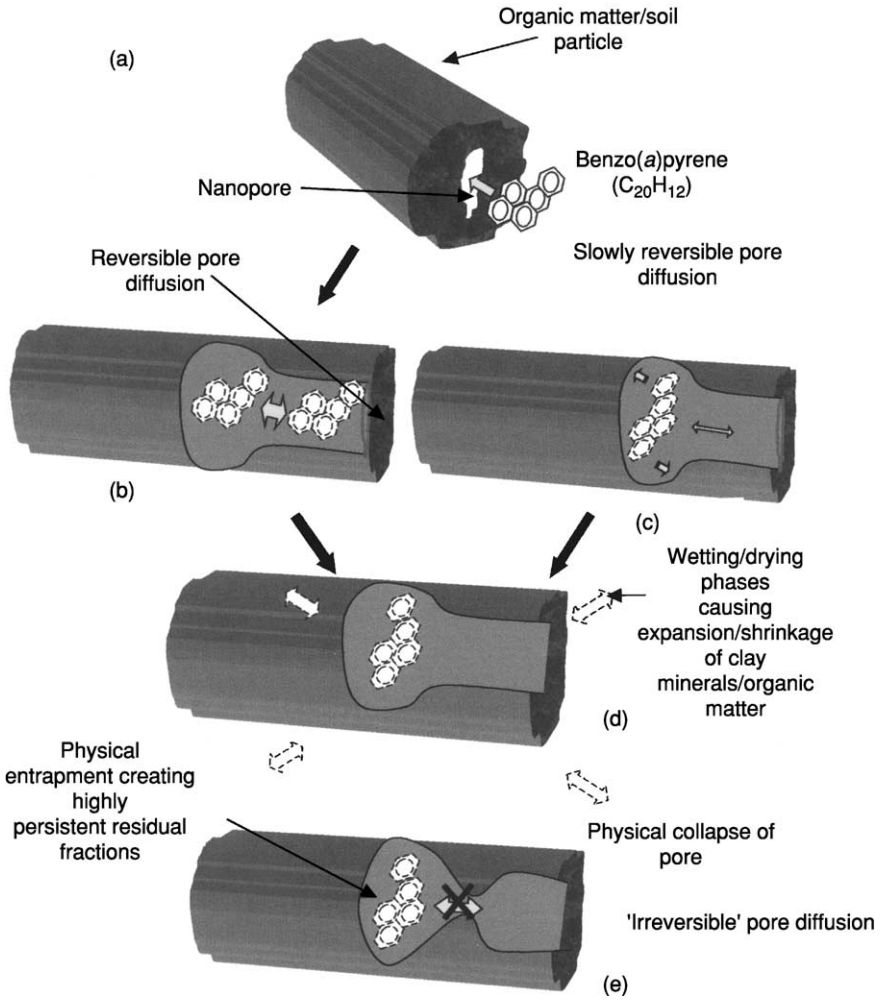


Figure 3 A conceptual diagram showing the mass-transfer limited processes leading to sorption-retarded pore diffusion. (a)–(e) are discussed in the text.

Figure 3 graphically illustrates the various stages of diffusion that occur in exceptionally narrow nanopores traversed by a model PAH, benzo(*a*)pyrene. Compounds tend to enter nanopores as they diffuse along gradually narrowing mesopores (>2 μm) followed by micropores (<2 μm) and eventually end up in nanopores whose diameter may be only a few Angstroms larger than the molecule itself (Figure 3a). In extremely narrow pores, diffusion is often exceptionally slow due to the high viscosity of pore water. Polar minerals have one or more layers of water strongly sorbed on their surfaces. This ordered layering of molecules increases the viscosity of the water and greatly slows molecular diffusion within nanopores (Figure 3b). The water contained in very narrow pores of a few Angstroms may be 'ice-like' and thereby greatly restrict solute diffusion. The tortuosity of pores becomes an increasingly important factor in the retardation of pore diffusion and tortuosity may only influence desorption when the pore diameter is less than 6 nm. In extremely narrow and tortuous pores, steric hindrance may further retard desorption (Figure 3c). When this occurs sorption is only very slowly reversible, contributing to the persistent residual fractions discussed earlier. Natural wetting/drying phases of a field soil will lead to expansion and contraction of clay mineral lattices (Figure 3d), which can induce physical collapse of pores and physical entrapment of solute molecules within 'blind alleys' (Figure 3e).

Figure 3 therefore represents a series of steps that progressively restrict solute diffusion and retard desorption leading to persistent residual fractions of organic compounds in soils. These compounds, although chemically unaltered, are usually unavailable to traditional chemical extraction techniques and the assumption is that they are also not bioavailable. However, it is important to remember that, although such compounds are often referred to as irreversibly bound, this may not always be the case. Physical

destruction of the soil aggregates they are associated with may release the compounds or at least increase their biological and chemical availability. Acidification dissolves the inorganic oxide cements that hold soil aggregates together and hence may similarly increase rates of pore diffusion and sorption.

Intraorganic Matter Diffusion

Intraorganic matter diffusion (IOMD) is the other important rate-limiting process involved in the desorption of persistent organic compounds from soils. Soil organic matter can exist as surface coatings, lining pores, or as discrete particles. Organic matter diffusion may be the primary rate-limiting process in soils with high-organic-matter contents but, since slow sorption also occurs in soils with low to zero organic matter, it clearly is not the only rate-limiting process. IOMD is considered to be diffusion solely within the matrix of organic matter (OM), which is thought to resemble a polyelectrolyte polymer mesh, with retardation resulting from reversible sorption on the internal surfaces of organic matter. OM polymers are said to have glassy (condensed, rigid) or rubbery (expanded, flexible) structures (Figure 4). Diffusion in polymers occurs by either a place change mechanism, in which movement is accomplished by cooperative interchange of position in polymer segments and the penetrating molecule, or by a defect mechanism where the penetrant may jump between lattice defects, voids, and pores. OM polymers consist of condensed, rigid 'glassy' phases through which diffusion is slow (Figure 4b) or more expanded and flexible 'rubbery' structures through which diffusion is faster (Figure 4a).

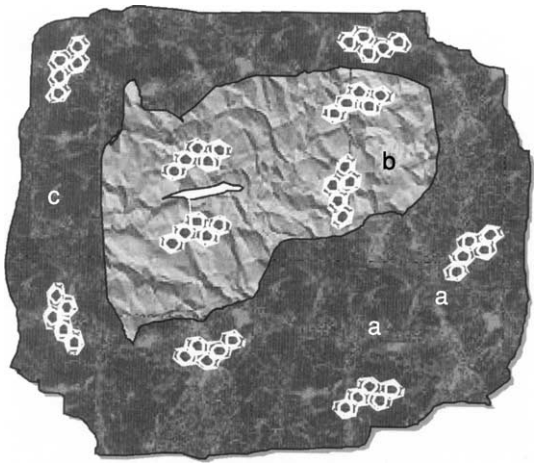


Figure 4 A conceptual diagram describing the slow sorption processes involved in organic matter diffusion (OMD). (a) Diffusion through a rubbery phase; (b) diffusion through a more condensed glassy phase b; and (c) adsorption in a 'Langmuir site' (see text for further details).

In addition, some organic compounds may undergo 'Langmuir-like' adsorption (Figure 4c) in internal voids in the OM matrix, which has been likened to a hole-filling mechanism. These slow diffusion processes lead to the formation of physically protected compounds which, as in the case of SRPD, are often mistakenly labelled as irreversibly bound residues. The OM diffusion and binding processes are in fact transitory since environmental variables may swell or soften the OM matrix, converting it from a glassy to a more rubbery state and accelerating diffusion processes. Increasing temperatures, change in pH, and the presence of cosolvents can all act to increase the diffusivity of tightly bound compounds.

In summary, it is probable that a number of mass transfer mechanisms are responsible for retarded diffusion leading to the formation of persistent residual fractions within most soils. Of these, the two most important are usually SROMD and SRPD. Although much work is needed to determine the exact contribution of each mechanism, it is likely that in most soils it is a synergy of both OMD and SRPD that is primarily responsible for the formation of persistent residual fractions. Residual fractions are often considered both biologically and chemically unavailable. However, as the preceding passages have shown, the mechanisms that lead to their formation are often transitory. This is particularly important when considering biological availability since the very conditions which are known to enhance diffusion rates often occur in the intestines of biological organisms (low pH, mechanical manipulation, and elevated temperatures). Hence, soil ingestion by an organism may have an impact upon a compound's availability and traditional extraction techniques used to determine the risk presented by a persistent residual fraction may be inaccurate.

The (Bio)availability of Organic Chemicals in Soils

The term 'bioavailability' is often discussed in the literature but rarely defined. For the purposes of this article, it may be defined as 'the fraction of a chemical that is absorbed by an organism.' This is a broad definition and encompasses bioavailability to plants and microorganisms via a variety of potential routes. The amount of compound present in an organism will be a function of the rate of absorption versus the rate of excretion.

POPs in soils are potentially bioavailable to many different organisms. In this article, a very brief overview is provided of the bioavailability of POPs to those organisms that are thought to be most important. These are plants, bacteria, and fungi, and

grazing animals. Macroinvertebrates and particularly earthworms are dealt with in more depth as they may represent important routes for POPs to enter the food chain. Organisms such as earthworms that are in constant physical contact with the soil may accumulate significantly higher body burdens of POPs than the surrounding soil. These concentrations may not be great enough to cause either acute or chronic toxicity in the earthworms. However, many species feed on earthworms and will concentrate the lipophilic compounds in the body fat. The result is an increase in POP body burdens as you rise up the trophic level of the food chain. This phenomenon has become known as biomagnification and may lead to symptoms of toxicity in organisms that have little contact with soil POPs.

Availability to Plants

Generally, translocation of nonionic organic compounds from the roots to shoots is negligible due to the low water solubility, high n-octanol/water partition coefficients and strong associations of many compounds to soil particles/organic matter. Plant uptake of nonionic organic chemicals from soils is usually dominated by vegetative uptake of contaminated vapor from the surrounding air. Although high concentrations of soil-borne nonionic contaminants are rarely available for translocation by plant roots, heavily contaminated soils can influence the concentrations of organics in above-ground vegetation by the soil–air–plant route. In a review of crop plant uptake of organic compounds from soils it was concluded that the vast majority of organic compounds in soil are so strongly sorbed by the soil matrix that they exhibit low bioavailability to crop plants and so are accumulated at very low concentrations in the edible portion of food crops.

Availability to Soil Microflora

The 'availability' of organic compounds in soils to microorganisms is critical if they are to be broken down and removed from the environment via either natural attenuation or bioremediation. Microbial biodegradation is a major process affecting the persistence of POPs in soils. The extent of POP bioavailability to bacteria in different soils may differ significantly, even under the same optimum growth conditions. Sorption has an important influence on the bioavailability of hydrophobic organic compounds in soils. A reduction in the degradation rates with increasing soil sorption capacity is apparent. Furthermore, a persistent residual fraction of POPs may remain unavailable to soil bacteria after remediation. This can be attributed to the slow migration of an increasingly large fraction of the POPs into the soil organic matter and mineral matrix with time.

As discussed earlier, POP chemical availability falls with increasing residence time in a soil. This reduced chemical availability is often mirrored by an analogous decline in bioavailability. This 'aging effect' may reduce the efficiency of bioremediation of a contaminated soil as a function of time. However, the reduced bioavailability may also reduce the potential toxicological threat of organic compounds in a soil. There is currently a great deal of research devoted to developing an accurate assessment of the bioavailability of organic compounds in soils. When developed, such an assessment of bioavailability will be able to identify both the suitability of a soil for bioremediation and the potential toxicological threat presented by that soil.

Grazing Animals

Soil ingestion by grazing farm animals has been estimated to be in the range of 1–18% of the dry-matter intake, depending on forage conditions. Thus, ingested soil may represent a potentially major route of uptake for soil-borne organic contaminants by ruminants. It follows that consumption of meat and/or dairy products may be an important exposure pathway to humans. However, the total intake of organic compounds by grazing livestock is relatively low, even when a worst-case scenario is taken into consideration. Furthermore, due to the effect of slow diffusion, discussed earlier, the fraction of a compound 'available' to grazing livestock is likely to be less than that taken up on a weight compound/weight soil basis. It follows that any contemporary risk assessments may overestimate exposure of grazing livestock to soil-borne organic compounds.

Availability to Earthworms

Earthworms are extremely numerous, especially in pasture land where their biomass can be up to $272\,g\,m^{-2}$. Add to that the fact that an individual earthworm can consume up to 400 times its own weight in soil over a 12-month period, and their importance in the dynamics and persistence of organic compounds in soils becomes apparent. This high density and activity mean that up to 25% of the surface layers of soil is turned over annually. A relatively high body lipid content means that earthworms accumulate POPs. Earthworms may facilitate faster release of certain sorbed residues, since a large quantity of soil passes through the gut, and is consequently exposed to digestive enzymes that act on soil organic matter. Hence, the assumption that there is an analogous decline in chemical and biological

availability of organic compounds may break down when an active (in this case gastrointestinal) uptake mechanism is involved as opposed to a more passive mechanism, such as contact between skin and soil.

See also: **Hydrocarbons**; **Organic Residues, Decomposition**; **Pollution:** Groundwater; Industrial; **Remediation of Polluted Soils**; **Sorption:** Organic Chemicals

Further Reading

Beck AJ, Johnson DL, and Jones KC (1996) The form and bioavailability of non-ionic organic chemicals in sewage-sludge amended agricultural soils. *Science of the Total Environment* 185: 125–149.

Brusseau ML and Rao PSC (1990) Modelling solute transport in structured soils – a review. *Geoderma* 46: 169–192.

Edwards NT (1983) Polycyclic aromatic hydrocarbons (PAHs) in the terrestrial environment – a review. *Journal of Environmental Quality* 12: 427–441.

Hatzinger PB and Alexander M (1995) Effect of ageing chemicals in soil upon their biodegradability and extractability. *Environmental Science and Technology* 29: 537–545.

Howsam M and Jones KC (1998) Sources of PAHs in the environment. In: Neilsom AH and Hutzinger O (eds) *The Handbook of Environmental Chemistry 3.1. PAHs and Related Compounds; Chemistry.* Berlin: Springer-Verlag.

Jones KC, Stratford JA, Waterhouse KS *et al.* (1989) Increases in the polynuclear aromatic hydrocarbon content of an agricultural soil over the last century. *Environmental Science and Technology* 23: 95–101.

Jones KC, Alcock RE, Johnson DL, Northcott GL, Semple KT, and Woolgar PJ (1996) Organic chemicals in contaminated land: analysis, significance and research priorities. *Land Contamination and Reclamation* 4: 189–197.

Masclet P, Pistikopoulos P, Beyne S, and Mouvier G (1988) Long-range transport and gas/particle distribution of polycycic aromatic hydrocarbons at a remote site in the Mediterranean. *Atmospheric Environment* 22: 639–650.

Newton I, Wyllie I, and Asher A (1993) Longterm trends in organochlorine and mercury residues in some predatory birds in Britain. *Environmental Pollution* 79: 143–151.

O'Connor GA, Chang RL, and Ryan JA (1991) Bioavailability to plants of sewage sludge borne toxic organics. *Reviews in Environmental Toxicology* 121: 129–165.

Pignatello JJ (1993) Recent advances in sorption kinetics. In: Beck AJ, Jones KC, Hayes MB, and Mingelgrin U (eds) *Organic Substances in Soils and Water*, pp. 128–140. Cambridge: Royal Society of Chemistry.

Weissenfels WD, Klewer HJ, and Langhoff J (1992) Adsorption of polycyclic aromatic-hydrocarbons (PAHs) by soil particles – influence on biodegradability and biotoxicology. *Applied Microbiology and Biotechnology* 36: 689–696.

Wild SR and Jones KC (1995) Polycyclic aromatic hydrocarbons in the UK environment: a preliminary source inventory and budget. *Environmental Pollution* 88: 91.

POLLUTION

Contents
Groundwater
Industrial

Groundwater

H Rubin, Technion, Israel Institute of Technology, Haifa, Israel

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Groundwater contamination presently comprises one of the major global concerns. In developing countries, groundwater is subject to contamination mainly due to poor sanitation associated with high rates of population growth. In rural areas of developed countries, groundwater contamination is mainly attributed to fertilization, use of pesticides and herbicides, and inappropriate habits of irrigation from the point of view of amounts of groundwater pumping and water quantities used for irrigation. In urban areas of developed countries, groundwater contamination is attributed to a variety of species (both chemical pollutants and microorganisms), originating from urban development, and domestic and industrial wastes.

Aquifers are formations of comparatively high permeability capable of storing significant quantities of groundwater. Due to these two properties, groundwater aquifers comprise the major source of fresh

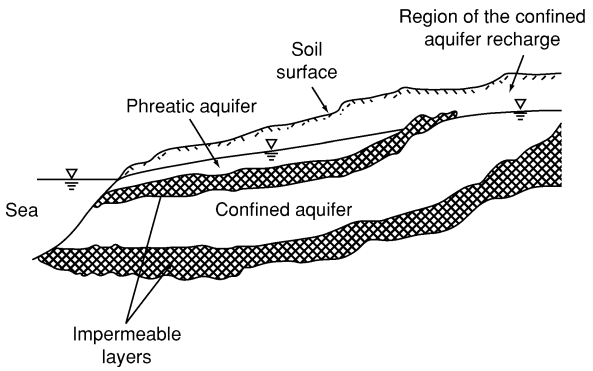


Figure 1 Schematics of phreatic and confined aquifers.

water for humans. Figure 1 provides a schematic of the major two types of aquifer: (1) phreatic aquifers, and (2) confined aquifers. The bottom of the phreatic aquifer, which is also called the 'free surface aquifer,' comprises impermeable formations, whereas the top is unprotected against the penetration of pollutants through the vadose (unsaturated) zone. Such pollutants usually arrive at the phreatic aquifer with infiltrating runoff water (which provides the aquifer recharge) or water from other sources. A confined aquifer is bounded top and bottom by impermeable formations. However, in some areas, where the confined aquifer is subject to recharge, it is basically a phreatic aquifer.

There are several avenues for groundwater contamination. One common avenue of groundwater contamination is associated with direct contact between surface water and groundwater. An example of this type of groundwater contamination is represented by the role of the Arkansas River in the contamination of the aquifers in western Kansas. The Arkansas River is a draining river of surface water contaminated by pollutants of the big farms in Colorado. This type of water flows downstream through Kansas. In western Kansas, due to excessive pumping of groundwater, the Arkansas River supplies water into the aquifer, and thereby pollutes the groundwater with pollutants diverted from Colorado. In the Tel-Aviv area, Israel, the Ayalon River diverts pollutants from its upstream portion into the aquifer of Tel-Aviv, where the groundwater table is often significantly lower than the water level of the Ayalon River.

However, in most cases, groundwater is subject to contamination due to the arrival of seeping contaminants through the vadose zone of the soil that overlies the aquifer. Sometimes groundwater pollution is transferred by various types of mechanisms into the vadose zone, and generally groundwater contamination is intimately connected with the state of the vadose zone.

Association of Pollutant Transport with the Water Cycle

Almost every textbook, guidebook, and manual of hydrology incorporates some descriptions and discussions which refer to the global cycle of water and emphasize its importance for properly maintaining the environment. The global water cycle demonstrates how quantities of water in the global scale are almost unchanged, while water is subject to transfer between different parts of the environment. The dielectric coefficient represents the ratio of the potential (and also the field strength) surrounding an electrostatic charge in a specific environment (medium) to the potential surrounding an identical charge in a vacuum. Due to its large dielectric coefficient, water has an extremely high capability to dissolve salts. Therefore, all chemical processes characterizing the different types of organisms take place in a water environment. By employing soaps and detergents, human society has increased the dissolving capability of water even to materials considered 'immiscible in water.' Society also utilizes water to transport solids as suspended and bed load; where suspended load is the flux of suspended solids and bed load is the flux of solids subject to movement (due to momentum transfer from the flowing water) at the bottom of the conduit. In contemporary urban society, water is mainly used as an agent to dissolve and transport solid material. Therefore, nowadays the involvement of the water cycle stages in pollutant transport in and between different parts of the environment is crucial for the preservation of human health and a balanced environment. It is extremely important to calculate budgets and mass flux balances of contaminants in different areas of the environment. In the bible, in the book of Ecclesiastes, the author expressed his astonishment that all rivers go to the sea, but the sea is never full. However, observations indicate that the salinity of the sea increases, and throughout the history of geology scientists have tried (unsuccessfully) to calculate the age of the earth according to the accumulation of salt in seas and oceans. It is not the appropriate place to evaluate why this calculation has not been successful, but simple mass flux balance shows that fresh water flowing into the sea incorporates small concentrations of various salts, and water leaving the sea in the vapor phase is almost free of salts. Therefore, the salinity of the sea must increase, as it does in lakes such as the Dead Sea (which is saturated with sodium chloride) that are not drained into other big water bodies.

Figure 2 provides a general schematic description showing how water and pollutants are transferred

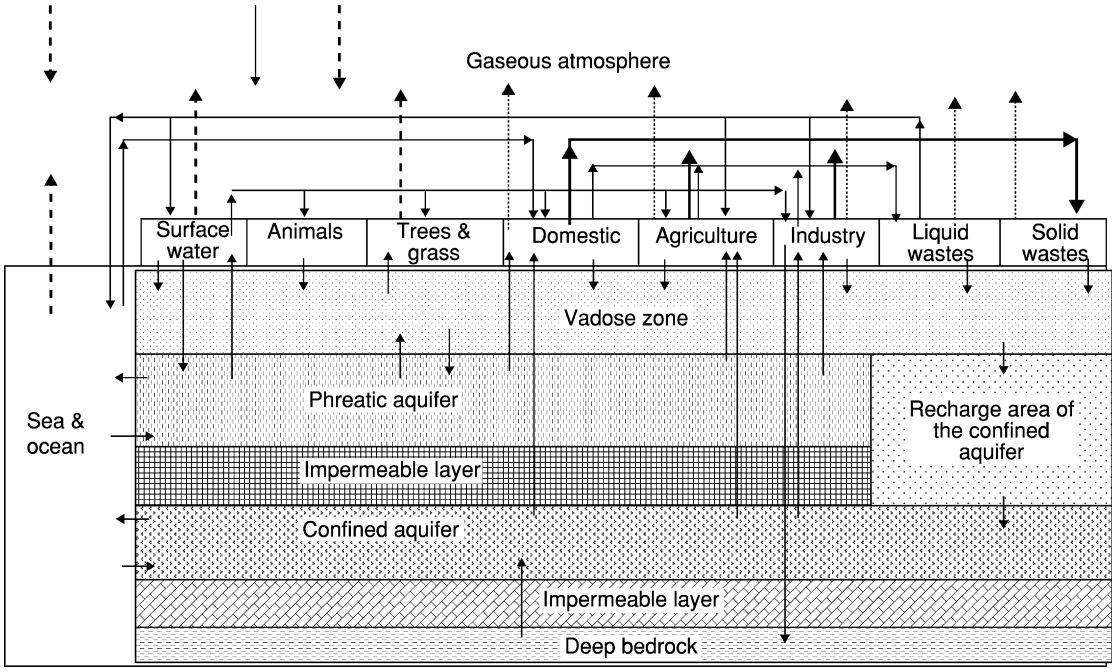


Figure 2 The water cycle associated with pollutant transport.

between different parts of the environment and different water consumers. This figure shows several parts of the environment that can be considered as storage locations for water: (1) seas and oceans, (2) the gaseous atmosphere, and (3) the various geologic formations. Figure 2 also shows some consumers of water. Note that it refers to surface water as one of the consumers, though some professionals may prefer to use another definition for surface water. There are about eight major water consumers (again the exact number of consumers is not as important as is their role in pollution production and transport). Thin solid arrows refer to the direction of transport of water with solutes and pollutants; thick solid arrows refer to solid wastes. Dotted arrows refer to transfer of air pollutants; and thick dashed arrows refer to transfer of pure water (vapors and/or precipitation). Therefore, according to Figure 2, pure water, as precipitation, is transferred from the atmosphere into the sea and the terrestrial part of the environment, but contaminated water is also transferred, e.g., acid rain.

Regarding consumers, domestic consumption also includes municipal uses, transportation, and all activities of an urban society. The major global water reservoir of seas and oceans provides a source of pure water in the atmosphere through evaporation. Through desalination, the atmosphere provides water for domestic use. On the other hand, salinity transport into groundwater might take place due to sea water intrusion into phreatic and confined aquifers, which is not a rare phenomenon, and may be enhanced in locations of excessive groundwater pumping. Phreatic and confined aquifers comprise the major terrestrial water storage of the environment. This storage is connected with the surface water and supplies water to society (domestic, agricultural, and industrial needs). All water consumers discharge water with higher solute concentrations than they originally obtained. Some consumers produce wastewater that should be transferred into liquid-waste collection sites, treated, and then delivered to seas, surface water storage, or agricultural and industrial uses. Some water used by consumers may directly enter the vadose zone, e.g., agricultural irrigation waters. This water may be subject to evapotranspiration (and thereby solutes of this water are accumulated in vegetation and the vadose zone) and/or infiltration through the vadose zone into the phreatic aquifer. (*See* **Evapotranspiration**.) Such infiltration is associated with the accumulation of some additional quantities of solutes. The phreatic aquifer may supply water with contaminants into the vadose zone by capillary movement, induced by water consumption of trees. Volatile contaminants dissolved in groundwater may be transferred as vapors into the vadose zone gaseous phase. The transfer of pollutant between the phreatic aquifer and the vadose zone is greatly enhanced by temporal, seasonal, and annual fluctuations of the groundwater table elevation.

Various industries apply techniques of deep well injection to get rid of their wastes. As an example, pumping of crude oil is associated with pumping of

brines. The latter, in many cases is directly injected into the deep bedrock. In parts of south Florida, practices of deep well injection were used for disposal of industrial wastes.

With regard to the two types of aquifers, the confined aquifer is subject to recharge in restricted areas and discharges water into the sea. Excessive groundwater pumping from the confined aquifer may lead to saltwater intrusion from the sea, as well as brine intrusion from the deep bedrock. The phreatic aquifer is intimately connected with the surface water and through the vadose zone it obtains infiltrating runoff water with solutes accumulated during the infiltration process. The vadose zone is the 'membrane' that separates the phreatic aquifer from the water consumers. The vadose zone changes the solute content (including a variety of solutes, pollutants, suspended and colloidal materials, microbes, parasites, etc.) of the infiltrating water owing to its different soil components, clay minerals, and residual water saturation. During the runoff water infiltration, solutes adsorbed to clay minerals may be released and transferred into the infiltrating water. On the other hand, suspended materials transported with the surface runoff usually cannot penetrate the vadose zone, and they do not arrive at the aquifer. In some cases, the infiltrating water quality may be improved by the adsorption of solutes by clay minerals. However, some of the positive phenomena of water-quality improvement are time-dependent, and therefore they should be carefully evaluated from a quantitative viewpoint.

It is appropriate to assume that every solute whose concentration in water is above its threshold concentration should be considered as a pollutant whose presence is involved in some risks to human health and the environment.

Any part of the environment involved in the water cycle and associated pollutant transport should be preserved in an appropriate risk-free mass budget and flux balance. Figure 2 indicates that special attention should be given to mass balances in the vadose zone and the phreatic aquifer due to the intimate relationships between these two types of strata.

Simplified Mass Budget and Flux Balances in Phreatic Aquifers

This section applies simplified mass budget and flux balance calculations to a variety of environmental development stages typical of different global locations. Figure 3 refers to four different scenarios or stages relevant to groundwater contamination within phreatic aquifers. The first stage, which is not emphasized by Figure 3, considers the natural environment

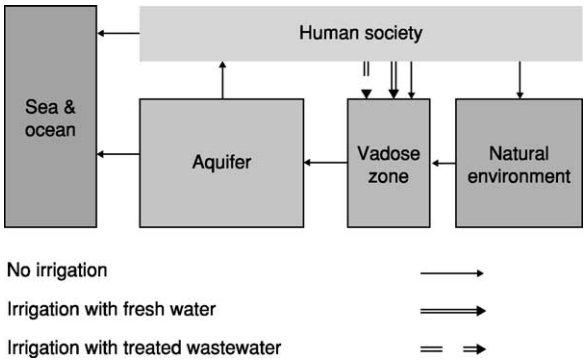


Figure 3 Different stages (scenarios) of pollutant flux development.

with very few effects from society. The second stage refers to involvement of human society with no (or very limited) irrigation practices. The third stage is associated with freshwater irrigation; and the fourth stage is associated with all types of irrigation by fresh water and treated wastewater. The calculations refer to a single contaminant, which is considered as the limiting pollutant of the environment; but similar calculations can be done with regard to any relevant pollutant.

Figure 4 describes the involvement of water and pollutant fluxes with the simplified water cycle. The following types of water fluxes and their relevance to pollutant transport are described by Figure 4:

1. Q_{nr} is the natural recharge of the phreatic aquifer, namely the net recharge (precipitation minus evapotranspiration). Even in the natural environment, the recharging flow rate accumulates some contaminants, ΔC_0 aboveground and ΔC_1 during its infiltration through the vadose zone;
2. Q_{so} is the groundwater flow rate that is emitted into seas and oceans;
3. Q_{hs} is the flow rate consumed by society. Society always supplies some contaminants into the natural recharge and increases its pollutant concentration by ΔC_1. Human society also increases the pollutant concentration by the amount ΔC_4 before delivering the water into seas and oceans;
4. Q_{ir} represents the irrigation flow rate. It is assumed that people apply groundwater for irrigation purposes. The irrigation flow rate comprises a certain portion of Q_{hs}. However, irrigation is associated with evaporation flow rate, Q_{ev1}. Evaporation decreases the flux of irrigation water that infiltrates into the aquifer. However, it is assumed that all pollutants present with Q_{ir} arrive at the phreatic aquifer. Furthermore, the reduced infiltrating discharge ($Q_{ir} - Q_{ev1}$) accumulates additional pollutant concentration ΔC_1 during its infiltration through the vadose zone;

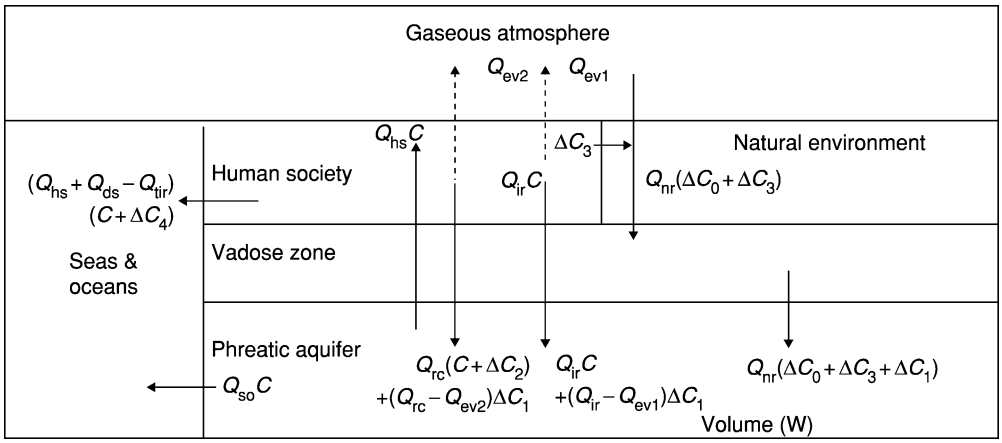


Figure 4 Considered fluxes of water and pollutants in various scenarios.

5. Q_{rc} is the flow rate of recycled wastewater. Wastewater-recycling decreases the discharge from populations that is delivered into seas and oceans. It is assumed that all recycled wastewater is used for irrigation. Probably the pollutant concentration within this type of water is greater by ΔC_2 than pollutant concentration in groundwater. However, again this type of irrigation is associated with evaporation flow rate, Q_{ev2}. Again considerations should be made with regard to the additional pollutant mass flux originating from experiencing wastewater-recycling.

We assume that measures are taken to keep the aquifer total volume W unchanged, and we also assume that initially the pollutant concentration is C_0. Then all pollutant mass budget and flux balance calculations for the different stages, described by Figures 3 and 4, indicate that the pollutant concentration increases exponentially as:

$$C = C_\infty - \Delta C\left[1 - \exp\left(-\frac{Q_{nr} - Q_{ev}}{W}t\right)\right] \qquad [1]$$

where

$$\Delta C = C_\infty - C_0$$

The values of C_∞ and ΔC are given in Table 1, Q_{ev} is the total rate of evaporation. Therefore the time required to obtain steady state conditions, namely when C is smaller than C_∞ by one percent of ΔC is:

$$(t)_{C=C_\infty-0.01\Delta C} = \frac{4.61(Q_{nr} - Q_{ev})}{W} \qquad [2]$$

Table 1 summarizes the results of water flux balance calculations and pollutant mass budget and flux balance calculations. In Table 1, expressions of the two bottom cells of the fourth column have been obtained by assuming that evapotranspiration associated with irrigation is significantly smaller than the natural recharge of the aquifer. Otherwise, these expressions should be replaced by simple deduction of relevant expressions of the third column. However, it should be noted that only the last term of the bottom cell expression in the fourth column represents the real contribution of wastewater recycling to pollutant concentration in the groundwater, provided that irrigation with treated wastewater replaces irrigation with water pumped from the aquifer. Q_{tir} and Q_{ev} in Table 1 are values of total irrigation $(Q_{ir}+Q_{rc})$ and total evaporation $(Q_{ev1}+Q_{ev2})$ respectively.

Table 1 can be useful for initial evaluation of long-term implementation of water resources management practices. This table indicates that irrigation, even with water of drinking quality, increases the pollutant concentration in the aquifer. On the other hand, though the marginal increase in aquifer pollution by wastewater-recycling is greater than irrigation with drinking water, this effect is mitigated if wastewater-recycling is implemented as a substitute for irrigation with drinking water.

Study of Figure 4 and Table 1 may lead to some useful ideas about the preservation of groundwater quality in phreatic aquifers. First of all, such an objective should be achieved by adequate considerations based on the control of all pollutant inputs, which should be minimized, and pollutant outputs, which should be maximized. Figure 4 and the results of calculations in Table 1 indicate that irrigation of all types poses a major threat to groundwater quality, as it circulates pollutants between the groundwater and the human population. Of course, advanced treatment of wastewater to drinking-water quality is appropriate, as it reduces the pollutant input, but such an approach does not necessarily provide sufficient guarantee against the accumulation

Table 1 Water and pollutant mass budget and flux balance results for typical cases of aquifer exploitation scenarios

Stage (scenario)	*Water balance*	*Pollutant concentration at $t\to\infty$*	*Increase of C_∞ (value of ΔC)*
Natural environment	$Q_{so} - Q_{nr} = 0$	$C_\infty = \Delta C_0 + \Delta C_1$	0
No irrigation	$Q_{so} + Q_{hs} - Q_{nr} = 0$	$C_\infty = \Delta C_0 + \Delta C_3 + \Delta C_1$	ΔC_3
Freshwater irrigation	$Q_{so} + Q_{hs} + Q_{ev1} - Q_{nr} - Q_{ir} = 0$	$C_\infty = \dfrac{Q_{nr}(\Delta C_0 + \Delta C_3 + \Delta C_1) + (Q_{ir} - Q_{ev1})\Delta C_1}{Q_{nr} - Q_{ev1}}$	$\dfrac{Q_{ev1}}{Q_{nr}}(\Delta C_0 + \Delta C_3 + \Delta C_1) + \dfrac{Q_{ir} - Q_{ev1}}{Q_{nr} - Q_{ev1}}\Delta C_1$
All types of irrigation	$Q_{so} + Q_{hs} + Q_{ev1} + Q_{ev2} - Q_{nr} - Q_{ir} - Q_{rc} = 0$	$C_\infty = \dfrac{Q_{nr}(\Delta C_0 + \Delta C_3 + \Delta C_1) + (Q_{tir} - Q_{ev})\Delta C_1 + \Delta C_1 + Q_{rc}\Delta C_2}{Q_{nr} - Q_{ev}}$	$\dfrac{Q_{ev2}}{Q_{nr}}(\Delta C_0 + \Delta C_3 + \Delta C_1) + \dfrac{Q_{ev2}}{Q_{nr}}\left(\dfrac{Q_{ir} - Q_{ev1}}{Q_{nr}}\right)\Delta C_1 + \dfrac{Q_{rc} - Q_{ev2}}{Q_{nr} - Q_{ev}}\Delta C_1 + \dfrac{Q_{rc}\Delta C_2}{Q_{nr} - Q_{ev}}$

of high pollutant concentrations in the phreatic aquifers.

As stated, Figure 4 refers to a single pollutant. In an aquifer system subject to a variety of pollutants, whose presence introduces risks to human health and the environment, similar mass budget and flux balance calculations should be made with regard to each one of these pollutants for the evaluation of long-term processes involved with groundwater quality.

Penetration of Solutes into the Phreatic Aquifer

Figure 5 provides a classic description of the distribution of water saturation in the phreatic aquifer and the vadoze zone. This figure does not include the water saturation distribution close to the ground surface, which is subject to significant changes due to precipitation, irrigation, dry seasons, etc.

The region of major interest with regard to solute penetration into the phreatic aquifer is the capillary fringe, which incorporates water subject to pressure below the atmospheric pressure. Due to the capillary forces, water in the capillary fringe is almost stagnant, and does not take part in groundwater horizontal flow. The groundwater table is defined as the level of atmospheric pressure. Therefore, as shown in Figure 5, the water level in a well represents the groundwater table. The contaminant usually migrates as a solute with water that infiltrates, as a wetting front through the vadoze zone. During this process, and mainly in the capillary fringe, where water residual saturation is comparatively high, the infiltrating water is subject to mixing with the residual water saturation and redistribution of contaminant concentrations takes place. This mixing is associated with local reduction of the capillary effect and migration of water with pollutants into the aquifer. However, all these processes may significantly increase the pollutant concentration in the capillary fringe. Here, the effect of groundwater table fluctuation may play an important role in preserving the contamination of the capillary fringe at levels similar to those of the aquifer. However, if such fluctuations are minor, then the capillary fringe may be subject to significantly higher pollutant concentrations than those typical of the aquifer.

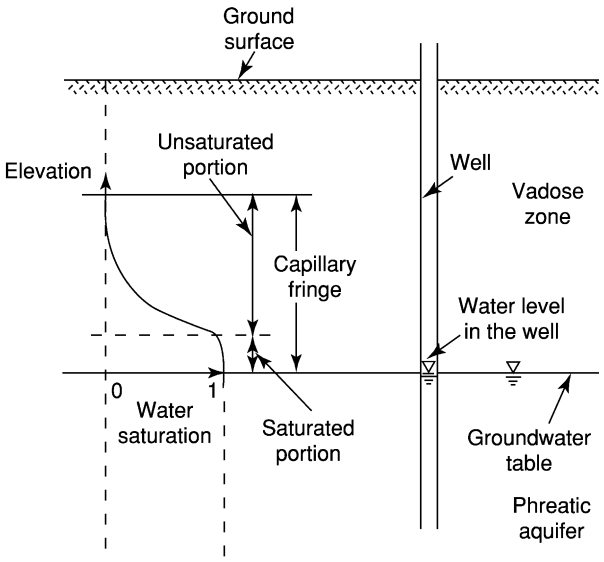


Figure 5 Water saturation distribution in the phreatic aquifer and the vadose zone.

Pollutant Transport in the Aquifer

Pollutants that penetrate through the capillary fringe, due to a wetting-front movement or due to groundwater table fluctuations, are subject to transport in the aquifer. The evaluation of pollutant transport in groundwater incorporates at least two basic differential equations, the equation of groundwater flow, namely Darcy law, and the diffusion–advection equation:

$$\mathbf{q} = \mathbf{K} \cdot \mathbf{J} \qquad [3]$$

$$\frac{\partial C}{\partial t} + \mathbf{V} \cdot \nabla C = \nabla \cdot (\mathbf{D_h} \cdot \nabla C) - R \qquad [4]$$

where $\mathbf{q}$ is the specific discharge vector (Darcy velocity), $\mathbf{K}$ is the hydraulic conductivity tensor (which depends on the permeability and kinematic viscosity of the fluid, namely water), $\mathbf{J}$ is the gradient of the piezometric head, C is the pollutant concentration, $\mathbf{V}$ is the interstitial velocity (the specific discharge divided by the effective porosity), $\mathbf{D_h}$ is the hydrodynamic dispersion tensor (it incorporates effects of diffusion and tortuosity of streamlines), which depends on $\mathbf{V}$ and properties of the permeable domain, and R represents possible chemical reactions (such as adsorption and decay).

In cases of isotropic porous media and fluid of constant density (no density stratification is significant), the hydraulic conductivity is a scalar quantity. Then eqn [3] represents the specific discharge as a vector originating from the gradient of the piezometric head, which is a scalar quantity. Therefore, all approaches and techniques used for the solution of potential flow problems are also applicable to the description of flow phenomena taking place in the aquifer. However, usually in groundwater flow, the streamlines are almost horizontal, and in such cases the Dupuit approximation can be applied.

According to the Dupuit approximation, lines of constant piezometric head (equipotential lines) are vertical. If pollutants have minor effect on the fluid density, the solution of eqns [3] and [4] can be done in a series. This means that after solving the flow with eqn [3] and obtaining a complete description of the streamlines and velocity distribution in the domain, these results are used as an input for the solution of eqn [4]. If the pollutant concentration is comparatively high, then density stratification might be significant, and therefore the velocity and pollutant distributions in the domain should be obtained simultaneously by the simultaneous solution of eqns [3] and [4]. Such cases are typical of saltwater intrusion into aquifers. The solution of eqn [3] is often obtained by analytical means, provided that density stratifications are minor. On the other hand, in most practical cases the solution of eqn [4], under all types of fluid-density distribution in the domain, usually requires the employment of numerical approaches. Only a limited number of analytical solutions to this equation have been developed. However, various methods can be used to obtain approximate first-order descriptions of pollutant transport in the aquifer. **Figure 6** provides a schematic description of the buildup of the contaminant plume at the pollutant penetration site. In some cases, the buildup of this plume can be evaluated by using the approximate approach of boundary layer methods.

It should be noted eqns [3] and [4] refer to an unconsolidated permeable material and a single type of pollutant. If several pollutants are involved, then each pollutant has to be referred to by a particular use of its contaminant transport equation, subject to its particular set of initial and boundary conditions. In some cases also considerations of possible chemical reactions that might occur between the different pollutants have to be taken into account.

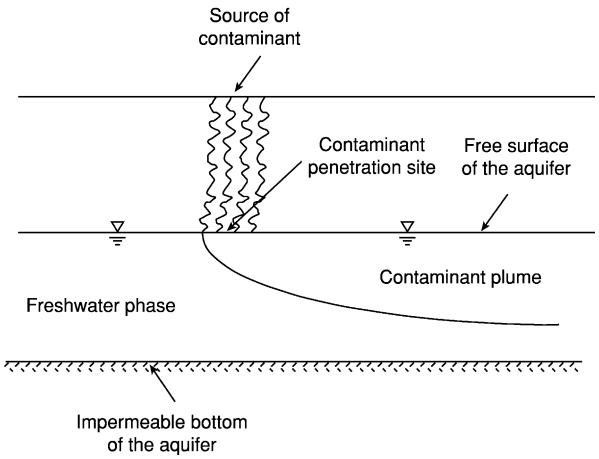


Figure 6 Buildup of a contaminant plume at a contaminant penetration site. Reproduced from Rubin H and Buddemeier RW (2002) Groundwater contamination downstream of a contaminant penetration site. 1. Extension-expansion of the contaminant plume. *Journal of Environmental Science and Health* A37(10): 1781–1812.

Two major issues that are intimately connected should be noted in connection with the appropriate evaluation of pollutant transport in aquifers: (1) upscaling, namely the relevance of phenomena observed on a small scale, such as in a laboratory, to those predicted at the large field scale; and (2) the effect of the aquifer formation heterogeneity on pollutant transport. Physical quantities characterizing groundwater flow through the aquifer are very little affected by the formation heterogeneity, and flow calculations can be done by using the continuum approach, which leads to eqn [3]. With regard to pollutant transport in groundwater, issues of upscaling and aquifer heterogeneity may impose very serious problems on the simple use of the continuum approach, which leads to eqn [4]. Therefore, a variety of approaches, such as double porosity approximation, and statistical presentation of the medium characteristics and transport parameters are very often used by scientists and engineers.

Contamination of Groundwater by NAPLS

Since the mid-1980s, much attention has been directed toward the evaluation of groundwater contamination by nonaqueous-phase liquids (NAPLs). This high level of attention has originated from the realization of how common cases of groundwater contamination by NAPLs are, and how many risks to human health and the environment are associated with such a contamination. The term 'NAPL' has substituted the previous term 'immiscible fluids.' This change indicates that, although the miscibility of these fluids in water is very low (in the order of magnitude of parts per million, ppm), such concentrations in water avoid its safe use by human populations, and they also introduce high risks to the environment. 'NAPL' usually represents an organic liquid-phase material. Such a liquid is very often comprised of a mixture of many organic compounds used for a specific task, e.g., fossil fuels that are employed in driving internal combustion engines, or creosote that is used for treating wood in the wood industry. Various types of chlorinated hydrocarbons are NAPLs used for medical purposes, e.g., chloroform, or for degreasing purposes by the steel industry, e.g., trichloroethylene.

Generally, two types of NAPLs are identified: (1) light NAPLs (LNAPLs), whose density is lower than that of water; and (2) dense NAPLs (DNAPLs),

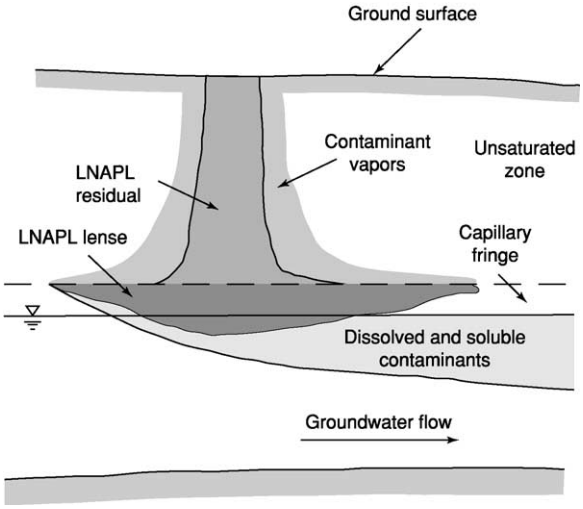


Figure 7 Contamination of the vadose zone and the phreatic aquifer by light nonaqueous-phase liquids (LNAPL). Reproduced with permission from Rubin H, Narkis N, Carberry JB (1998) Overview of NAPL contamination and reclamation. In: Rubin H, Narkis N, Carberry JB (eds) *Soil and Aquifer Pollution – Non-Aqueous Phase Liquids – Contamination and Reclamation*, pp. 3–17. Springer-Verlag: Berlin, Heidelberg, New York. © Springer-Verlag.

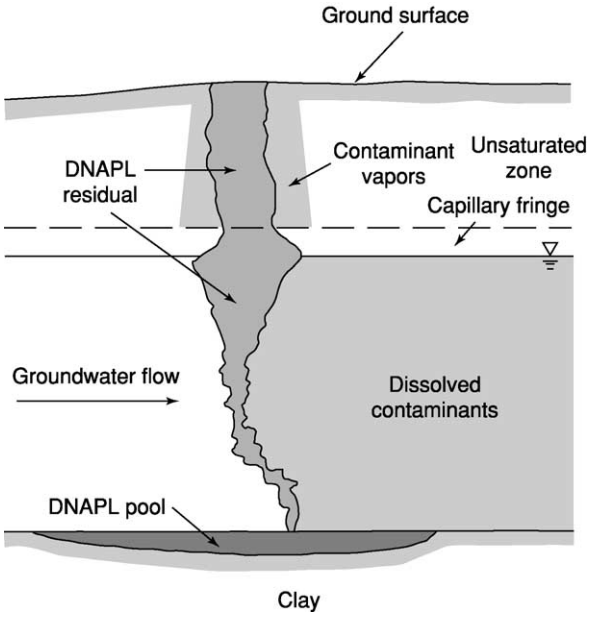


Figure 8 Contamination of the vadose zone and the phreatic aquifer by dense nonaqueous-phase liquids (DNAPL). Reproduced with permission from Rubin H, Narkis N, Carberry JB (1998) Overview of NAPL contamination and reclamation. In: Rubin H, Narkis N, Carberry JB (eds) *Soil and Aquifer Pollution – Non-Aqueous Phase Liquids – Contamination and Reclamation*, pp. 3–17. Springer-Verlag: Berlin, Heidelberg, New York. © Springer-Verlag.

whose density is higher than that of water. This classification is extremely important for the evaluation of NAPL penetration into the aquifer, and its migration with the flowing groundwater. Figures 7 and 8 provide schematic descriptions referring to LNAPL and DNAPL contamination, respectively.

After the release of the NAPL (LNAPL or DNAPL) at the ground surface, it migrates downward within the vadose zone with limited lateral expansion. During the downward migration, the NAPL plume is subject to partial entrapment and the release of some vapors of volatile compounds into the gaseous phase of the vadose zone. Differences between LNAPL and DNAPL contamination take place after the arrival of the contaminant plume at the capillary fringe and the groundwater table.

As shown in Figure 7, at the groundwater table, the LNAPL plume spreads horizontally and creates a lens that floats on top of the groundwater table. If the LNAPL plume originates from the release of a limited and small quantity of NAPL, then possibly the entire mass of NAPL is completely entrapped in the vadose zone without its arrival at the groundwater table. If the released quantity is limited but incorporates a significant volume, then the scenario of migrating LNAPL ends with its complete entrapment within the vadose zone and the floating LNAPL lens on top of the groundwater table.

As shown in Figure 8, after its arrival at the groundwater table, the DNAPL plume penetrates into the aquifer and keeps its downward migration. This process is associated with DNAPL entrapment within the phreatic aquifer. If quantities of released DNAPL are large, then the scenario of DNAPL migration ends with complete entrapment of the released DNAPL in the vadose zone, in the phreatic aquifer, and even possible accumulation of the DNAPL in a pool, at the bottom of the aquifer (or on top of impermeable soil lenses that might be located within the phreatic aquifer).

Scenarios of the NAPL migration in the vadose zone and the phreatic aquifer can be evaluated by using numerical models originating from reference to multiphase flow in porous media. Such models have been used for many years by the oil industry, and have been adapted for the quantification of the environmental issues represented by Figures 7 and 8. In the vadose zone, calculations refer to the solid phase and three fluid phases (gaseous, water, and NAPL). In the aquifer, reference is made to the solid phase and two fluid phases (water and NAPL). The mathematical model of multiphase flow in porous media is generally formulated as:

$$\epsilon_\alpha \rho^\alpha \frac{\partial \omega_{\mathrm{i}}^\alpha}{\partial t} + \rho^\alpha \mathbf{q}^\alpha \nabla \omega_{\mathrm{i}}^\alpha - \nabla \cdot (\epsilon_\alpha \rho^\alpha \mathbf{D}_{\mathrm{h_i}}^\alpha \cdot \nabla \omega_{\mathrm{i}}^\alpha) = \sum_{\beta \neq \alpha} E_{\mathrm{i}}^{\alpha\beta} - \omega_{\mathrm{i}}^\alpha E^\alpha \qquad [5]$$

which is subject to the two following constraints:

$$\sum_{i} \omega_i^{\alpha} = 1; \quad \sum_{\alpha} \epsilon_{\alpha} = 1 \qquad [6]$$

where ϵ_α is the volume fraction of the α phase, ω_i^α is the mass fraction of component i in the α phase, $\mathbf{D}_{h_i}^{\alpha}$ is the hydrodynamic dispersion tensor of the α phase, $\mathbf{q}^\alpha$ is the specific discharge of the α phase, ρ^α is the density of the α phase, and $E_i^{\alpha\beta}$ represents exchange of mass of component i between the α and β phases.

Equation [5] refers to each component i (of the total number of relevant compounds), that may comprise portions of the α and β phases (solid skeleton, water, gaseous phase, and NAPL). Besides eqns [5] and [6], some constitutive relationships should be applied to calculate the propagation of each fluid phase and specific species in the domain.

However, as implied by preceding paragraphs, scenarios of multiphase flow involved in contamination of the vadose zone and the phreatic aquifer usually take place for a limited time, until all quantities of the NAPL are entrapped within the porous medium; then the only mobile fluid phase present in the aquifer is water. Under such conditions the effect of the immobile, entrapped NAPL on flow parameters of groundwater are represented by the decrease in the water saturation and the reduced permeability of the formation. These effects lead to the following modified Darcy law:

$$\mathbf{q} = \nabla \cdot \left[g\left(\mathbf{k} \frac{k_{rw}}{\nu} \right) \cdot \mathbf{J} \right] \qquad [7]$$

where g is the gravitational acceleration, $\mathbf{k}$ is the permeability tensor, k_{rw} is the relative water permeability (which depends on the water saturation), and ν is the kinematic viscosity of the water.

The modified diffusion–advection equation, which is applied to simulation of dissolved constituents transport with the flowing groundwater, is given by:

$$\phi S_w \frac{\partial C}{\partial t} + \mathbf{q} \cdot \nabla C = \nabla \cdot (\phi S_w \mathbf{D}_h \cdot \nabla C) + R + E^{wn} \qquad [8]$$

where ϕ is the porosity, S_w is the water saturation and E^{wn} represents interphase mass transfer, namely mass transfer from the entrapped NAPL into the flowing water phase, due to NAPL dissolution.

The chemical reaction term, R, of eqn [8] is mainly attributed to adsorption and desorption, as well as biological degradation of the organic compound that is used as a substrate for the development of the biomass. The interphase mass transfer term, E^{wn}, of eqn [8] is considered as the major contributor to groundwater contamination due to the presence of the entrapped NAPL in the aquifer. It can also be useful for aquifer cleanup procedures such as pump-and-treat.

Most cases of NAPL contamination in aquifers are attributed to petroleum hydrocarbons, namely fossil fuels. Obviously this phenomenon originates from the widespread use of these types of NAPLs. However, most types of olefins that incorporate the major compounds of fossil fuels do not impose high risks on the environment. Fossil fuel compounds of high risks include various fuel additives and BTEX (benzene, toluene, ethylbenzene, and xylene). Due to the widespread use of fossil fuels, entrapment of LNAPL in the vadose zone and aquifers is a quite common phenomenon. Usually, such an entrapment with comparatively high residual NAPL saturation is limited to the top layers of the aquifer and regions of the capillary fringe. Therefore, the entrapped LNAPL supplies dissolve constituents into the flowing groundwater, as well as volatile organic compounds into the gaseous phase of the vadose zone.

DNAPL contamination of groundwater is less common than LNAPL contamination, and entrapment of DNAPL in aquifers is quite rare. However, the constituents of DNAPLs are usually chlorinated hydrocarbons, which pose a high risk to human health and the environment. Their threshold concentrations are of minute values, of the order of magnitude of several parts per billion. Therefore, the penetration of water solution, which incorporates several parts per million of these compounds, into groundwater may cause a very significant contamination of the aquifer. Furthermore, the chlorinated hydrocarbons are volatile organic compounds (VOCs). The migration of such compounds with groundwater may be associated with the release of poisonous vapors into the vadose zone. Such a phenomenon is somehow mitigated by the immobile water saturation of the capillary fringe. But intensive fluctuations of the groundwater table may significantly reduce the efficiency of this vapor barrier.

Summary

There is an intimate connection between the pollution of groundwater and the vadose zone. Special attention is given here to calculations of overall mass budgets and flux balance of pollutants for the evaluation of phreatic aquifer contamination due to its use as a source of water for human populations. When water shortage becomes a significant barrier for the development of various societies, it is appropriate to evaluate efficient approaches of increasing the number of water resources and avoiding the deterioration of groundwater quality.

Various methods and means are used for the evaluation of the fate of constituents that are dissolved in water and subject to transport with the water cycle, leading to groundwater pollution. The classification of NAPLs to LNAPLs and DNAPLs is extremely important, as these two types of NAPL have different phenomenological effects on the pollutant penetration into the aquifer. There are also significant differences between their potential risk to human health and the environment, their physical properties, rate of biodegradation, and the possible source of organic contaminants in the liquid and vapor phase in the vadose zone.

List of Technical Nomenclature

α, β	indices referring to the particular fluid phase (water, gas, NAPL)
ΔC	difference between final and initial concentration ($kg\,m^{-3}$)
ΔC_0	pollutant accumulated within Q_{nr} in the natural environment ($kg\,m^{-3}$)
ΔC_1	pollutant accumulated within infiltrating water in the vadose zone ($kg\,m^{-3}$)
ΔC_2	pollutant accumulated within complete water-recycling ($kg\,m^{-3}$)
ΔC_3	pollutant accumulated within Q_{nr} by activities of the human population ($kg\,m^{-3}$)
ΔC_4	pollutant accumulated within wastewater released into seas and oceans ($kg\,m^{-3}$)
ϵ	volume fraction
ν	kinematic viscosity ($m^2\,s^{-1}$)
ρ	density ($kg\,m^{-3}$)
φ	porosity
ω	mass fraction
C	pollutant concentration ($kg\,m^{-3}$)
C_0	initial pollutant concentration ($kg\,m^{-3}$)
C_∞	final pollutant concentration ($kg\,m^{-3}$)
E	rate of pollutant interphase transfer ($kg\,m^{-3}\,s^{-1}$)
$\mathbf{D_h}$	hydrodynamic dispersion tensor ($m^2\,s^{-1}$)
$\mathbf{K}$	hydraulic conductivity tensor ($m\,s^{-1}$)
$\mathbf{k}$	permeability tensor (m^2)
k_{rw}	relative soil permeability for water
Q_{ev}	rate of evapotranspiration involved with all types of irrigation ($m^3\,s^{-1}$)
Q_{ev1}	rate of evapotranspiration involved with regular irrigation ($m^3\,s^{-1}$)
Q_{ev2}	rate of evapotranspiration involved with wastewater-recycling irrigation ($m^3\,s^{-1}$)
Q_{hs}	rate of water consumption by the human population ($m^3\,s^{-1}$)
Q_{ir}	rate of water used for regular irrigation ($m^3\,s^{-1}$)
Q_{nr}	rate of water supply into the aquifer by natural recharge ($m^3\,s^{-1}$)
Q_{rc}	rate of water used for irrigation with recycled wastewater ($m^3\,s^{-1}$)
Q_{so}	rate of water supplied by the aquifer into seas and oceans ($m^3\,s^{-1}$)
R	rate of pollutant formation and/or disappearance by chemical reactions ($kg\,m^{-3}\,s^{-1}$)
S_w	water saturation
t	time (s)
W	volume of the aquifer (m^3)

See also: **Evapotranspiration**; **Groundwater and Aquifers**; **Nitrogen in Soils:** Nitrates

Further Reading

Bedient PB, Rifai SH, and Newell CJ (1996) *Groundwater Contamination*. Englewood Cliffs, NJ: Prentice-Hall.

Domenico PA and Schwartz (1997) *Physical and Chemical Hydrology*, 2nd edn. New York: John Wiley & Sons, Inc.

Rubin H and Atkinson JA (2001) *Environmental Fluid Mechanics*. New York: Marcel Dekker.

Rubin H, Narkis N, and Carberry J (eds) (1998) *Soil and Aquifer Pollution (Non-Aqueous-Phase-Liquids – Contamination and Reclamation)*. Heidelberg, Germany: Springer-Verlag.

Rubin H, Nachtnebel HP, Fuerst J, and Shamir U (eds) (2002) *Water Resources Quality – Preserving the Quality of our Water Resources*. Heidelberg, Germany: Springer-Verlag.

Sun NZ (1996) *Mathematical Modeling of Groundwater Pollution*. New York: Springer-Verlag.

Industrial

S P McGrath, Rothamstead Research, Harpenden, UK

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

The term 'pollution' is often used, but seldom defined. A search of the literature reveals that it is often used synonymously with the term 'contamination.' Is there a distinction between pollutants and contaminants? For example, the US Environmental Protection Agency (EPA) defines pollution as "Generally, the presence of a substance in the environment that because of its chemical composition or quantity prevents the functioning of natural processes and produces undesirable environmental and health effects." Under the US Clean Water Act, pollution has been defined as the man-made or man-induced alteration of the physical, biological, chemical, and radiological integrity of water and other media. The word 'integrity' again seems to infer some harm has occurred. However, the US EPA also gives a definition of the term 'contamination' as "Introduction of microorganisms, chemicals, toxic substances, wastes, or wastewater in a concentration that makes the medium unfit for its next intended use." Once more, some damage is inferred, but here microorganisms are specified in addition to chemicals. It would be useful if microorganisms or presence and absence of harm gave a distinction between these terms, but many of the published definitions and recent usage show they are in fact considered synonymous. So although a few authors have tried to distinguish between pollution and contamination, here, the focus will be on the addition of chemical substances to soil that affect its use or affect other parts of the environment through soil.

Time and rate of addition of a substance to soil are important, as substances can be degraded, leached, or volatilized, thereby being removed from soils. Soils can immobilize pollutants, and it is only when the inputs far outweigh these loss pathways that the soil itself will become polluted. Of course, soils may in turn become the source of a pollutant rather than a sink when losses to water or volatilization take place. Another important concept relating to which chemicals become pollutants in soil is the idea of residence time. Because some potential pollutants arriving in soils can be rapidly degraded, leached, or volatilized, they may never become a serious problem. Examples would be some of the easily degraded organic compounds, which disappear quickly, but on the other hand some recalcitrant organic and inorganic pollutants build up in soil because there is little loss by any pathway.

The importance or magnitude of pollution depends on factors such as the rate of addition of the pollutants per unit area. Each route or pathway that pollutants follow to reach a site needs to be summed. Pathways may include wastewater from industry, solid wastes, emission to the air and subsequent deposition, and leaks and spills (Figure 1). Concentrations and total loads in the pathways minus the loss rates due to the processes mentioned above and

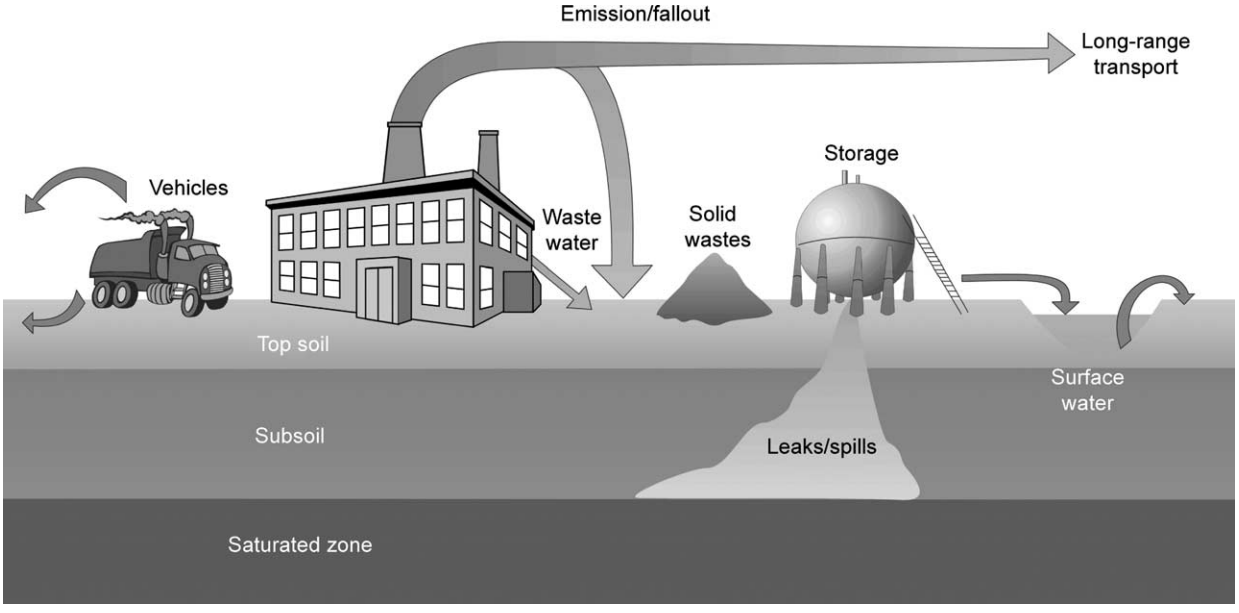


Figure 1 Pathways of industrial pollutants in the soil environment.

Table 1 Estimated number of contaminated sites in Europe and the United States

Country	*Contaminated sites (estimate of total)*	*Contaminated sites (in critical condition)*
Austria	3300	–
Belgium	8300	2000
Denmark	20 000	3600
Germany	>100 000	10 000
Italy	5600	2600
Netherlands	110 000	4000
European Union	1500 000	300 000
Norway	25 000	–
United Kingdom	100 000	–
United States	37 600[a]	>1300[b]

Source: Adriano DC (2001) *Trace Elements in Terrestrial Environments. Biogeochemistry, Bioavailability and Risks of Metals*. New York: Springer; Eijsackers HJP and Hamers T (eds) (1993) *Integrated Soil and Sediment Research*. Dordrecht: Kluwer; Overcash M (1996) European soil remediation research: 1992–1994. *Critical Reviews in Environmental Science and Technology* 26: 337–368; European Environment Agency (2001) *Indicator Fact Sheet Signals 2001 – Soil*.
[a]Reported to the US EPA in 1993.
[b]National priority list ('Superfund sites').

time of exposure all affect the buildup of the pollution and hence its importance.

Because of the difficulty of determining whether sites are polluted, there are only estimates in relatively few countries of the extent of what is normally referred to as 'contaminated land' (Table 1). For example, in the UK it is estimated that 300 000 ha of land are affected to some extent by contamination left by industrial activities. In the European Union 300 000 sites have already been identified as potentially or definitely contaminated, and the best estimate is that there are 1.5 million contaminated areas.

Excluding pesticides and radionuclides, the most common pollutants found at US Superfund sites include: acetone, arsenic, barium, benzene, 2-butanone, cadmium, carbon tetrachloride, chloroform, chromium, cyanide, 1,1-dichloroethene, 1,2-dichloroethane, lead, mercury, methylene chloride, napthalene, nickel, pentachlorophenol, polychlorinated biphenyls (PCBs), polycyclic aromatic hydrocarbons, tetrachloroethylene, toluene, trichloroethylene, vinyl chloride, xylene, and zinc. Common radionuclides at Superfund sites are: americium-241, cesium-137, cobalt-60, iodine, plutonium, radium, radon, strontium, technetium-99, thorium, tritium, and uranium.

Origins of Industrial Pollution

Many different industries and processes emit pollutants into the environment. Use of agricultural chemicals, such as nitrate, and pesticides will not be discussed here. Activities that are nonprimary or 'commercial' also result in emissions and pollution of soil, examples being transport and dredging of waterways, where pollutants are often deposited on the sides and banks of roads or along the banks of rivers. Industrial emissions may also move through air or water to sewage-treatment facilities, where they are often concentrated in the sewage sludge solids that are frequently utilized on land (*See* **Waste Disposal on Land:** Municipal). However, there are also domestic inputs of some pollutants to sewage.

The European Environment Agency considers that most cases of soil contamination arise from the following industries: chemicals, metals, energy, mining, oil, electronics, glass, ceramics, stone, textile, leather, wood, paper, food, trade, and traffic. Some of the activities likely to cause pollution of soils are shown in Table 2. Some of these arise from point sources such as factories, smelters, waste pipes, ditches, or mines. These are stationary sources and may be small or large. One feature of point sources is that the concentrations of pollutants can be high and may build up rapidly. They may also fall off rapidly with distance from the source. Nonpoint sources of pollution are not easily associated with or as regulated as point sources and normally originate from such activities as agricultural practices, urban runoff, and transportation.

Types of Pollutant

Many thousands of pollutants have been dispersed in the environment by human activities, some for a long time. They are broadly classified as inorganic and organic pollutants. However, depending on the source and other local emissions, multiple or 'copollution' of soils may occur. An example would be the land around a factory that used multiple chemicals in its processes or land that has been used by different industries over time.

Table 2 Activities likely to cause pollution in soils

Activity	*Examples*	*Likely inorganic pollutants*	*Likely organic pollutants*
Mineral extraction/processing			
Coal	Deep or open cast mines, crushing	Acidity, pyrites, sulfur	Dust
Metals	Ferrous and nonferrous, deep or open cast, crushing, waste lagoons	All metals and associated anions	
Other minerals	Asbestos, chalk, china clay, gypsum	Sulfides, sulfates	
Waste disposal			
Municipal	Domestic waste, sewage sludge	Many, including leachates	Many, including leachates
Industrial	Unused raw materials, by-products, spent catalysts	Many, including leachates	Many, including leachates
Industries			
Brick works		Fluoride, sulfur dioxide	
Chemical works	Battery, electronics, and pigments manufacture, electroplating	Acidity, alkalinity, As, asbestos, Cd, Cr, Cu, cyanide, Pb, Hg, Ni, sulfate, Zn	Dioxins, halogenated hydrocarbons, pesticides, phenols, PCBs, solvents
Gasworks	Coal gasification	Sulfur, cyanides	Phenols, tar, PAHs
Iron and steel	Blast furnaces	Metal wastes, iron and steel slag	
Nonferrous smelting/ processing	Lead, zinc, copper smelters	Pb, Zn, Cu, Hg, Cd, Ni, acidity	
Oil refineries	Petroleum production	Lead compounds	Tar, oils, phenols, sulfate
Power generation	Coal, gas, nuclear plant	Fuel ash, fly ash, various radionuclides	PCBs, PAHs
Recycling/scrap	Batteries, cars, domestic appliances	Acidity, asbestos, Pb, Ni, Cd	Oils, paint, phenols, PCBs
Sewage treatment	Various processes	Metals, phosphate	Various
Tanneries		Chromium	Dyes
Military bases			
Air/army/navy bases	Munitions, anticorrosion or antifouling chemicals, preservatives	Various	Various, fuels
Defense establishments		Various	Various
Firing ranges		Pb, various	Various
Transportation			
Airports	Runoff, workshops		Fuel, oil, deicing compounds
Canals, waterways, docks	Dredged sediments, dumped waste	Acidity, alkalinity, antifouling compounds, metal wastes	Contaminated sediments
Railways	Railways, borders and sidings/yards	Metal wastes, asbestos	Oil, paint
Roads	Borders, central reservation, runoff areas	Pb, Cd, Zn	PAHs
Energy supply			
Pipelines	Oil pipelines		Oil leaks
Electricity supply network	Cables and towers, substations, transformers	Zn, Cu	PCBs
Oil/petroleum storage		Pb	Tar, petroleum, oils, phenols

PCBs, polychlorinated biphenyls; PAHs, polycyclic aromatic hydrocarbons.

Inorganic Pollutants

Materials that are not based on carbon compounds are termed 'inorganic pollutants.' It is important to realize that some elements are in fact essential nutrients required by organisms, and that they can be deficient in some soils or diets; examples include Zn, B, Cu, Se. A consequence of this is that additions of nutrients will at first be beneficial for the growth and reproduction of organisms. It is only when they become excessive that negative impacts will occur.

This typically occurs because the soil's ability to hold these substances is overwhelmed, along with the ability of the organisms in the ecosystem to cope with the oversupply.

Organic Pollutants

Organic molecules in the environment are usually classified depending on whether they are persistent. For example, solvents may almost completely sterilize the soil if added in large quantities, but they typically volatilize in a relatively short time and the soil may then recover. Some compounds used in industry, such as surfactants and plasticizers, can have disruptive effects on the endocrine system. However, these normally act through the water pathway or directly in packaged and processed food and are not thought to persist in soils. The term 'persistent organic pollutants' (POPs) refers to those which do not degrade, leach out, or volatilize from soils readily. This group can, therefore, build up to concentrations in soil that may inhibit biological processes. Characteristics that make a compound persistent in soil include its low aqueous solubility, low vapor pressure, and a high octanol to water partition coefficient (*See* **Pollutants:** Persistent Organic (POPs)). These characteristics usually mean that only small amounts of POPs are taken up by plant roots because of their low solubility and tendency to bind to organic matter in soil. However, grazing animals and some invertebrates such as earthworms may mobilize POPs from soils in their digestive systems when they ingest soil.

Examples of Pollutants

Many industrial and commercial sources emit pollutants to air and water or direct to soil. Some are short-lived because they are degraded and others leach rapidly in environments with excess rainfall over evapotranspiration. However, others bind to organic matter, clay, and oxides or form insoluble compounds in surface soils and so tend to remain and accumulate in soil with time. Those that have this tendency are metals, metalloids, persistent organic compounds, asbestos, coal tar, fuel ash, and some radionuclides (*See* **Pollutants:** Persistent Organic (POPs)). Some of the industrial sources of these pollutants are shown in Table 3. It is important that any 'new' substances being used in industrial processes or mobilized in the environment are considered also for their persistence and effects in soil.

Legislative Controls on Pollution

The ultimate environmental control method is to prevent the dispersion of pollutants in the environment. This has resulted in complete bans on the production of some compounds perceived to be very harmful. For example, manufacture of PCBs in the USA was banned by the EPA in 1977, though it continued under considerable restrictions in other countries until the mid-1980s. Even after production of any substance ceases, it will continue to be present in products, and thus present a disposal hazard, for a considerable period, e.g., PCBs in the insulation of electrical transformers. Legislation needs to be coordinated such that removing a toxic pollutant from air (e.g., by stopping incineration of certain materials) does not lead to transfer of the pollutants to soil or water. On the other hand, legislation for some classes of pollutants is based on containment or controls on the rate of emission to air or water, or on the purposeful additions of materials to land.

Source–Pathway–Receptor Principle

As mentioned in the Introduction, the term 'pollution' infers that some harm has already been done or will take place. Recent legislation and schemes to determine whether land is polluted or contaminated take a risk assessment approach to evaluate potential harm. The source–pathway–receptor principle requires that all three aspects are established at a site and that there is significant harm. Source is an origin of pollution, which could be wastewater or contaminated soil in the present context. The pathway is a means by which the pollutant can reach and affect a receptor. A receptor is an organism or group of organisms, an ecological system, or a piece of property which is being or could be harmed by a pollutant. A 'pollutant linkage' has to be established, which means the relationship between the pollutant, pathway, and receptor. In some cases there may be more than one pathway (for example, dust and water contamination). Another valuable aspect of the pollutant linkage idea is that the breaking of this linkage can be a very important way of stopping pollution or remediating contaminated sites. Prevention can be brought about by stopping the source or putting a barrier in the pathway to the receptor (*See* **Remediation of Polluted Soils**).

Also, the bioavailability of the pollutant to an organism is an important consideration in the pathway: because a pollutant is present in soil (or another medium) does not necessarily mean that it is likely to transfer and affect an organism. In soils, binding of pollutants by organic matter or soil minerals may render them sparingly or nonbioavailable (*See* **Pollutants:** Persistent Organic (POPs)). An example is chromium; two forms occur in the environment, with valency states III and VI. In soils however,

Table 3 Pollutants likely to occur in soils in contaminated environments

Pollutants	*Industrial sources*
Metals/metalloids	
Antimony	Electronics, metal smelting
Arsenic	Insecticide manufacture and use, wood preservation
Beryllium	Nuclear, munitions, and electronics
Cadmium	Pigments, alloys, electroplating, batteries, mining, and smelting industries
Chromium	Plating, tanning, inks, wood treatment, pigments, and dyes
Copper	Mining and smelting, wood treatment, electronics, plumbing, copper and brass manufacture
Lead	Batteries, tetraethyl lead production, mining, smelting, plumbing
Mercury	Chemical plant, mining and smelting, refining gold and silver, electrolysis industry, chlor-alkali plant, paper mills
Nickel	Plating and metal works, batteries, iron and steel industries, electronics
Selenium	Smelting, paint, pigment, electronics
Silver	Mining, photographic, electroplating, silver manufacture
Thallium	Cement industry, electronics
Zinc	Mining and smelting, plating, galvanizing, runoff from galvanized surfaces, print and paper
Other inorganic pollutants	
Fluoride	Cement and brick production, aluminum smelting, phosphate fertilizer manufacture
Cyanides	Coal gasification, metallurgical processes, solvent/separation of gold
Sulfur	Sulfides, residues from coal processing
Organic pollutants	
Aromatic compounds	
Benzene	Tar, gas, or coke works, petroleum refineries
Ethyl benzene	Tar, gas, or coke works, petroleum refineries
Toluene	Tar, gas, or coke works, petroleum refineries
Xylene	Tar, gas, or coke works, petroleum refineries
Phenols	Tar, gas, or coke works, petroleum refineries
Polycyclic aromatic hydrocarbons (PAHs)	Combustion of coal hydrocarbons and wastes
Chlorinated organic compounds[a]	Pesticide manufacture, deodorants, timber and textile peservatives, transformer insulation
Dioxins and furans	Sewage sludge application, waste disposal, waste incineration
Other pollutants	
Asbestos	Insulation in buildings, boats, other transportation, power stations
Coal tar	Tar, gas, or coke works, petroleum refineries
Fuel ash	Coal-fired power generation
Protons	Oxidation of sulfur dioxide and metal sulfides
Radionuclides	Power generation, weapons production, phosphate manufacture

[a]Aliphatic chlorinated compounds, chlorobenzenes, chlorophenols, polychlorinated biphenyls.

chromium is usually reduced to chromium III, which is insoluble, and as a result little is bioavailable. In some circumstances in aquatic systems, chromium VI may occur, which is bioavailable and very toxic.

Persistent, Bioaccumulative, and Toxic Pollutants

One way of prioritizing pollutants is to choose those that are known to be persistent, bioaccumulative, and toxic (PBT). These are substances that can build up in the soil and the food chain to levels that are harmful to human and ecosystem health. Other features related to their chemical characteristics are that they may travel long distances, and transfer easily between air, water, and soil, and therefore reside for a long time in the environment. In addition to some pesticides, the US EPA's first priority list of PBT pollutants includes the following industrial pollutants: benzoapyrene, hexachlorobenzene, alkyl lead, mercury and its compounds, octachlorostyrene, PCBs, dioxins, and furans. Because of their properties, PBTs have attracted many voluntary or legal controls on their production, use, and entry into the marketplace to decrease their emission in the environment.

Statistics show that the production of some potential pollutants (for example, nonferrous metals) has increased continually over the last decade. However, due to the legislative controls discussed above, and because of controls on emissions of chemicals to the environment, the point and nonpoint (diffuse) sources of pollution are decreasing in the developed countries. However, it should be remembered that, in many rapidly developing countries without strong environmental controls, pollution is actually increasing.

See also: ; **Pollutants:** Persistent Organic (POPs); **Remediation of Polluted Soils**; **Waste Disposal on Land:** Municipal

Further Reading

Adriano DC (2001) *Trace Elements in Terrestrial Environments. Biogeochemistry, Bioavailability and Risks of Metals*. New York: Springer.

Bridges EM (1987) *Surveying Derelict Land*. Oxford: Clarendon Press.

European Environment Agency (2001) *Environmental Signals 2001*. Copenhagen: EEA.

Jones KC (ed.) (1991) *Organic Contaminants in the Environment. Environmental Pathways and Effects*. London: Elsevier Applied Science.

Knox AS, Gamerdinger AP, Adriano DC, Kolka RK, and Kaplan DI (1999) Sources and practices contributing to soil contamination. In: *Bioremediation of Contaminated Soils*, pp. 53–87. Agronomy Monograph No. 37. Madison, Wisconsin: American Society of Agronomy.

UK Department of the Environment, Transport and the Regions (2000) *Environmental Protection Act 1990: part IIA, Contaminated Land*. DETR Circular 02/2000. London: DETR.

POLYMERS AND MICROORGANISMS

M C Rillig, University of Montana, Missoula, MT, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Extracellular polymers, organic macromolecules found outside the cell or hyphal wall, are a common feature in eukaryotic and prokaryotic soil microbes. This summary deals with four main groups of extracellular microbial polymers, believed to be of particular significance in soil biology and the interaction of soil microbes with their environment: exopolysaccharides (EPS), hydrophobins (a group of small fungal proteins), glomalin (a glycoproteinaceous substance produced by arbuscular mycorrhizal fungi, AMF), and extracellular lipid compounds (for example, glycolipids such as rhamnolipids). Extracellular polymers are present in the soil in complex mixtures (spatially and temporally), being produced by a variety of microbes and being subject to modifications after entering the soil environment. They fulfill a range of functions in soil, from biofilm formation and modification of the habitat to increasing solubility of carbon substrates. In addition to the groups discussed here, extracellular enzymes are clearly a crucial component of the soil environment for microbes with absorptive nutrition (such as soil fungi and bacteria), but these are examined in detail elsewhere (*See* **Enzymes in Soils**). Polymeric compounds secreted by microbes in soil are also pertinent in the context of formation and composition of organic matter (*See* **Organic Matter:** Principles and Processes), and in the process of soil aggregation (*See* **Aggregation:** Microbial Aspects).

Extracellular Polymers: Definition and Overview

Commonly microbes in soil are embedded in a matrix of extracellular polymeric substances; the production of these substances is a common feature of archaea, Eubacteria, and eukaryotic microorganisms (e.g., algae, fungi). As a result of the production of these substances, biofilms are formed. Biofilms are mostly discussed in the aquatic microbial ecology literature, because they are uniquely suited to the challenges of inhabiting an aquatic habitat. Clearly, biofilms with their exopolymer matrix occur in soil as well, and can be seen by electron microscopical techniques (see Figure 1). Biofilms are accumulations of microbes (can be mixed communities), extracellular polymeric substances, multivalent cations, and various biogenic particles, as well as dissolved compounds.

Extracellular polymeric substances have been defined as "substances of biological origin that participate in the formation of microbial aggregates." Other definitions highlight the role these substances play in cohesion of microbes and adhesion (binding to the substratum). However, these more structural roles are not the only functions of extracellular polymers (see Table 1).

Extracellular polymers include organic macromolecules (formed by polymerization of similar or identical building blocks) such as polysaccharides, proteins, nucleic acids, and lipids. These polymers may contain nonpolymeric substituents (which can sometimes greatly influence the properties of the polymer). For example, proteins can be glycosylated or there can be acetyl or succinyl groups in polysaccharides.

These polymers are found by definition outside the cell wall, where they may have arrived through active

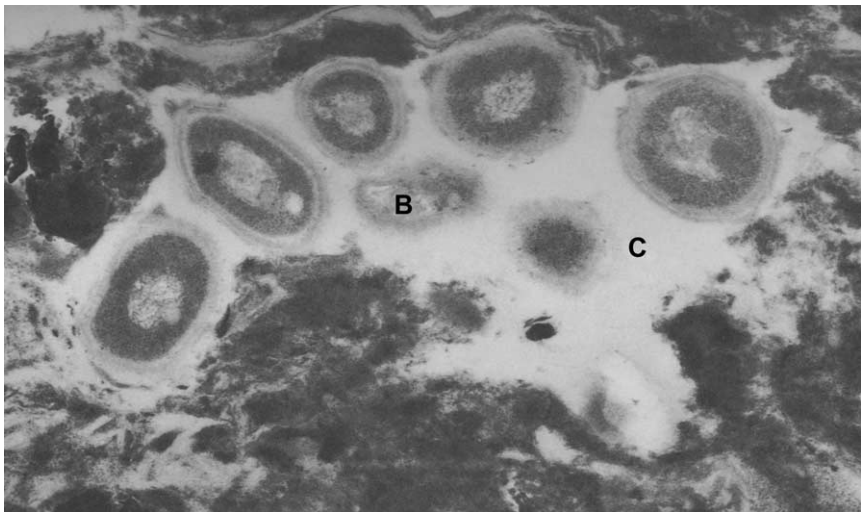


Figure 1 Electron micrograph showing a colony of bacteria (B) surrounded with exopolymer material (C, capsule material) in the rhizosphere of clover. Lanthanum stain, ×30 000. Reproduced with permission from Emerson WW *et al.* (1978) *Modification of Soil Structure*. Copyright 1978. © John Wiley & Sons Limited.

Table 1 Examples of the variety of functions of microbial exopolymers in soil

Exopolymer	*Function*
Exo- and ectoenzymes	Crucial component of absorptive nutrition of microbes
Hydrophobins	Lowering water surface tension, allowing hyphal emergence into air
Hydrophobins, some exopolysaccharides	Microbial attachment/adhesion to surfaces
Glomalin, many other exopolymers	Stabilization/formation of soil aggregates: involvement in the modification of the 'macro'-environment; importance in soil carbon storage via stabilization of labile carbon inside aggregates
Some rhamnolipids	Biocidal (antiviral, antibacterial, antifungal) activity
Exopolysaccharides, many other exopolymers	Biofilm formation: aggregation of microbes (increases potential for communication, synergism, gene transfer); leading to enhanced sorption of nutrients (inorganic, organic); retention of exoenzymes
Exopolysaccharides	Enhanced desiccation resistance
Some rhamnolipids	Surfactant activity (enhanced degradation of hydrocarbons; biosurfactants)
Some exopolysaccharides	Symbiotic signal (e.g., *Rhizobium*); also pathogen recognition
Attached exopolymers	Modification of microbial cell/hyphal wall properties (e.g., hydrophobicity, charge) with potential consequences for microbial transport, nutrient absorption
Exopolysaccharides, other exopolymers	Protection from toxic compounds, e.g., metals, biocides
Attached expolymers	Food web processes: protection from grazers, predation (e.g., amebae)

secretion processes (for example, through the vesicle-mediated pathway at the growing tip of fungal hyphae), sloughing off of cell-surface material, or lysis of microbial cells. Additionally, organic compounds from the environment (e.g., from lysis and hydrolysis) can become part of the polymer matrix.

Once extracellular, polymers are subject to various modifications by enzymatic processes. Various polymer-degrading enzymes such as hydrolases can be present in the matrix. Polymer-breakdown products can sometimes be reutilized by microbes. Due to these processes within a biofilm, the composition of extracellular polymers can be variable both spatially and temporally.

Extracellular polymers can also be broadly distinguished into those that are attached (bound) and nonattached. Bound polymers include capsules and sheaths of bacteria, and loosely bound polymers, while soluble polymers include slime and colloids, as well as other water-soluble macromolecules.

Historically, extracellular polymers were studied by extraction methods. Fractions were defined operationally using the extraction method of choice, e.g., centrifugation, boiling, ultrasonication, filtration, or solubility in various solvents (e.g., NaCl, NaOH, organic solvents). This work was mostly performed using laboratory pure cultures of microbes. In addition, there are now available techniques for the *in situ*, nondestructive study of these polymers, including infrared spectroscopy (e.g., attenuated total reflectance/Fourier transform infrared spectroscopy (ATR/FT-IR), detects signals of surface chemical groups), nuclear magnetic resonance (NMR), and confocal laser scanning microscopy (CLSM). These methods have greatly increased the understanding of the three-dimensional structure and composition of extracellular polymer matrices.

In summary, extracellular polymers occur in complex mixtures of compounds in the environment (to which frequently more than one species or group of microbes can contribute); they can arrive at a certain location by a variety of different mechanisms and may not have been formed where they are found; these mixtures are then subject to processing and modification in the environment as a function of time and space, and the structure of the polymer matrix can also change in response to environmental conditions. It is very important to keep this complexity in mind in the following discussion of selected components of these extracellular polymer mixtures.

Hydrophobins: Secreted Fungal Proteins

Structure and Properties

Hydrophobins are relatively recently discovered, small (ca. 100 amino acids), secreted, moderately hydrophobic proteins produced by fungi in the *Ascomycota* and *Basidiomycota*. Their occurrence in other Eumycotan phyla (*Chytridiomycota*, *Zygomycota*) is not yet clearly established. Based on hydropathy patterns (hydropathy is a measure of the hydrophobic character of an amino acid) and solubility characteristics, hydrophobins have been divided into class I and class II (the latter having been found only in the *Ascomycota* so far). Sequence homology of hydrophobins is very low (especially among class I hydrophobins), but all proteins share eight conserved cysteine residues; these cysteine residues form intramolecular disulfide bridges. Some hydrophobins, like the intensively studied SC3 (isolated from *Schizophyllum commune*, a basidiomycete fungus), are glycosylated. Several fungi have more than one hydrophobin gene. For example, *S. commune* contains at least four hydrophobin genes, which are expressed at different fungal life-history stages (e.g., monokaryotic versus dikaryotic mycelium). Six hydrophobin genes have been identified in *Cladosporium fulvum*. Complementation studies showed that some hydrophobins could only partially substitute for each other, suggesting that hydrophobins may have evolved to fulfill different functions.

All hydrophobins so far studied self-assemble at hydrophilic–hydrophobic interfaces, such as a water–air boundary. This self-assembly is accompanied by conformational changes of hydrophobin monomers that form the amphipathic film or membrane.

Virtually all studies performed with hydrophobins appear to have been conducted under *in vitro* conditions to study molecular and biochemical aspects of their production; hence there is little information about the amount and behavior of hydrophobins found in soils or their persistence in the soil environment.

Functions

Hydrophobins are believed to be essential in several functions in the fungal life-cycle and development. These proteins, by virtue of their ability to self-assemble into amphipathic membranes, allow hyphae to break the water–air barrier (overcoming surface tension), coat air-exposed fungal structures with a hydrophobic film, and allow hyphae to attach to hydrophobic surfaces. Hydrophobins have also been found to alter hyphal wall composition.

Hydrophobins allow fungi to escape their aqueous environment This is a necessary step for the formation of fungal fruiting bodies. However, mycelium growth in soil is also confronted with air–water boundaries as hyphae cross pore spaces between aggregates. The importance of secreted hydrophobins such as SC3 in hyphal growth across a water–air barrier lies in the lowering of water surface tension (Figure 2). A strain of *S. commune* with the *SC3* gene disrupted produces few aerial hyphae, but by adding SC3 protein to the culture medium, water surface tension is decreased to wild-type levels, and formation of aerial hyphae is possible. The time course of decrease in water surface tension of a culture of *S. commune* also correlates with the production and secretion of SC3 into the medium. The mechanistic details of emergence of hyphae as mediated by a hydrophobin-membrane are not known.

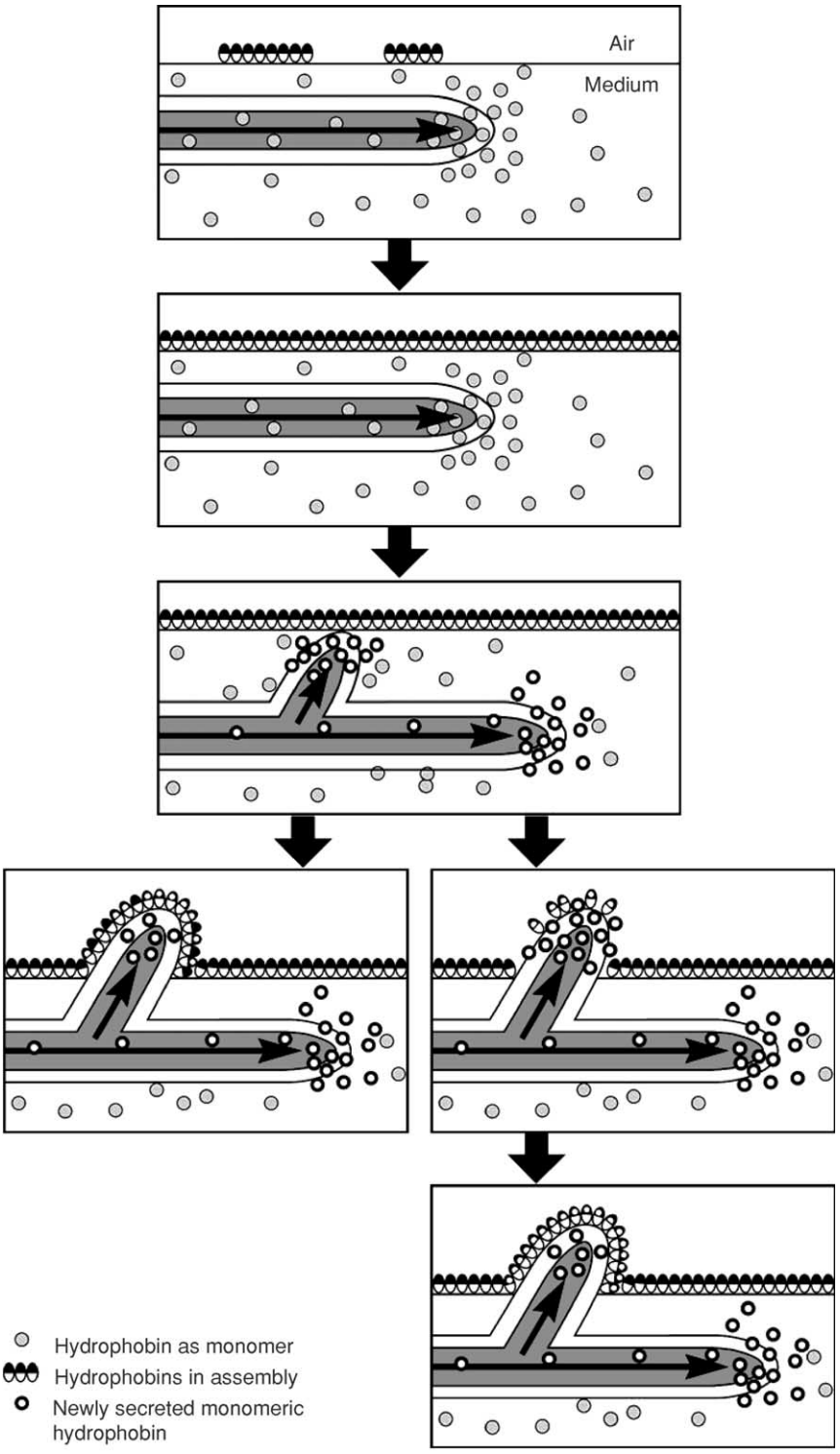


Figure 2 Functioning of fungal hydrophobins at the air–water interface. Monomeric hydrophobins are secreted into the aqueous medium. These spontaneously assemble at the air–water interface, where they greatly decrease the surface tension, allowing hyphae to emerge. In the process of emergence, it is unclear whether hyphae just stretch the hydrophobin film that has formed at the interface (bottom left) or if the film is ruptured as the hypha emerges (bottom right). Hydrophobins coat the hypha exposed to the air. Reproduced from Wösten HAB (2001) Hydrophobins: multipurpose proteins. *Annual Reviews of Microbiology* 55: 625–646. With permission, from the *Annual Review of Microbiology*, Volume 55 © 2001 by Annual Reviews www.annualreviews.org.

Glomalin: An Abundant Glycoproteinaceous Substance Produced by Arbuscular Mycorrhizal Fungi

Properties

Glomalin is a relatively recently discovered glycoproteinaceous substance (of as yet unknown amino acid sequence and structure) produced apparently only by AMF. AMF are obligate biotrophs (in the phylum *Glomeromycota*) living in symbiosis with the roots of ca. two-thirds of higher plants; they are ubiquitous in soils of terrestrial ecosystems (*See* **Mycorrhizal Fungi**).

Glomalin is, like many substances found in soil (e.g., humic acid), operationally defined, and hence definitions are subject to change once structural details become known. Glomalin is the protein (or the proteins, since it is somewhat unlikely that all AMF species produce exactly the same substance) that reacts with a monoclonal antibody (MAb32B11) raised against crushed spores of the AMF species *Glomus intraradices*, and as the protein (as determined with a Bradford assay) that is extracted with 50 mmol l^{-1} sodium citrate (pH 8.0) in successive cycles of autoclaving (121°C) until the supernatant is clear. Several fractions of glomalin are distinguished, based on their immunoreactivity (i.e., if the protein can be detected using an enzyme-linked immunosorbent assay (ELISA) with MAb32B11) and ease of extraction from soil (easily extractable glomalin is extracted in an initial round of autoclaving for only 30 min with 20 mmol l^{-1} sodium citrate, pH 7.0).

Defined in this way, glomalin occurs in very large amounts in soil, typically in the range of several to 10 mg g^{-1} soil (in one case 60 mg cm^{-3} soil was found). Glomalin decomposes slower than the AMF hyphae producing it, and the estimated turnover time for glomalin in soil is on the order of several decades (derived from chronosequence studies and C-14 carbon dating). Glomalin is also produced in sterile *in vitro* cultures of AMF, but most of the work has been done in pot cultures or soil.

Glomalin appears to have N-linked oligosaccharide side chains, and the extract also contains Fe in variable percentages (it is not known in what way Fe is bound in the molecule). In its native state, glomalin is assumed to be insoluble and hydrophobic. Evidence (e.g., from NMR studies) is mounting that glomalin is a compound distinct from humic acid; its ^{1}H NMR spectrum is also distinctly different from tannic acid. The pathway of release of glomalin into soil is not clear; glomalin could be actively secreted at the hyphal tip, or it could be sloughed off hyphae as they grow through soil.

It is not known what factors control glomalin production. However, in an experiment with AMF *in vitro* cultures, glomalin concentrations were significantly higher when AMF hyphal growth was reduced compared with control, perhaps suggesting that glomalin is not constitutively produced but that its production is subject to environmental controls. There is also evidence that AMF species differ in glomalin production rates. Glomalin concentrations in soil are responsive to several disturbances, including various agricultural practices (e.g., crop rotations, tillage, etc.), and factors of global change (e.g., elevated atmospheric CO_2, warming).

Functions

Soil aggregation Possibly the most important function of glomalin from the ecosystem perspective, and the best-documented overall, relates to its role in soil aggregation. While the mechanisms are not clearly understood, there is a strong and consistent correlation of glomalin concentrations in soil (or soil aggregates) with soil-aggregate water stability (a measure of resistance of aggregates to disintegration when sieved in water). The correlation is frequently strongest using the immunoreactive easily extractable glomalin (IR-EEG) fraction of glomalin. The relationship of glomalin concentration with soil aggregate water stability is curvilinear such that at higher glomalin concentrations no additional aggregate stability is achieved (Figure 3). Microscopic observation shows that aggregates are irregularly coated with glomalin (as identified with MAb32B11; see Figure 4), suggesting that a possible mode of action involves sealing off

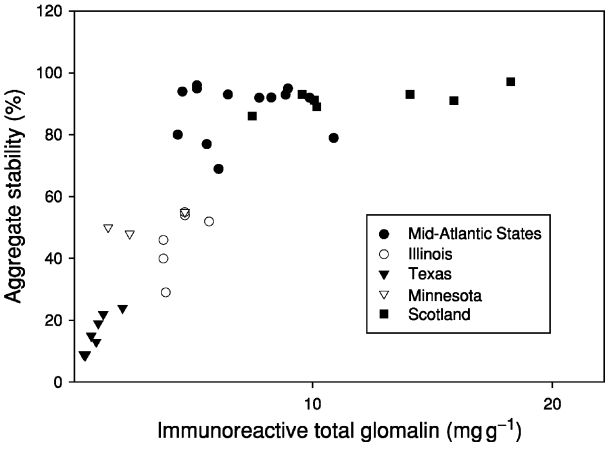


Figure 3 Curvilinear relationship of glomalin concentrations with soil aggregate water stability (1- to 2-mm aggregates) across a variety of soils. Reproduced with permission from Wright SF and Upadhyaya A (1998) A survey of soil for aggregate stability and glomalin, a glycoprotein produced by hyphae of arbuscular mycorrhizal fungi. *Plant and Soil* 198: 97–107 (Figure 3). With kind permission of Kluwer Academic Publishers.

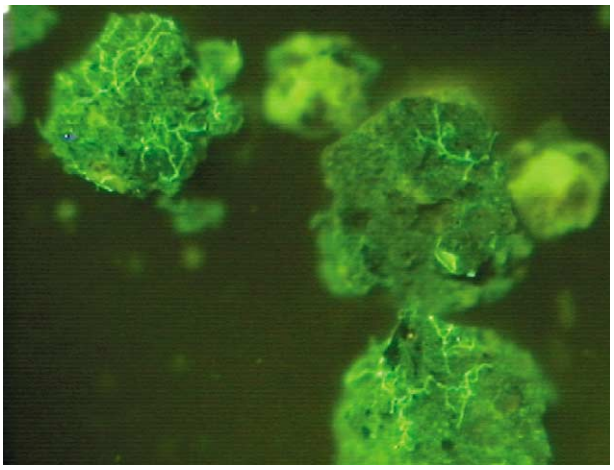


Figure 4 Glomalin (visualized as immunofluorescence using MAb32B11) on the surface of arbuscular mycorrhizal fungal hyphae on soil aggregates (1–2 mm diameter). Reproduced with permission from Wright SF (unpublished).

aggregates to prevent intrusion of water into the aggregate interior. Additionally, glomalin may make AMF hyphae 'sticky' so that the enmeshment mechanism of hyphae could be more efficient. Importantly, the relatively slow turnover of glomalin in soil may explain why aggregates formed by AMF are typically stable for a longer period than aggregates formed by other fungi, for example saprobes.

Role in the biology of AMF and symbiosis It is remarkable that plants and their AMF symbionts invest a significant amount of carbon into glomalin, begging the question what might be physiological roles of the protein. AMF hyphae are less palatable to grazing collembolans than hyphae of other, for example, saprobic fungi; it is hypothesized that glomalin may play a role in the decreased palatability of AMF hyphae in soil. The growth of AMF may be enhanced in aggregated soil compared with soil with no macropore spaces, although this is difficult to show experimentally (due to the presence of confounding factors when comparing aggregated and nonaggregated soil). The investment into glomalin may then serve to increase hyphal growth rate via the effect of the substance on soil aggregation. Results consistent with this idea have been obtained in an artificial bead-culture system. Other hypothesized functions include the adsorption of metals to the protein, and a role in the Fe economy of the rhizosphere.

Exopolysaccharides

EPS are extracellular polysaccharides that occur widely among prokaryotes; bacterial EPS have been studied in greatest detail. Also most microalgae and many filamentous fungi have EPS. EPS are abundant extracellular products that either accumulate on the surface of the cell (for example as a capsule) or that are released into the cell's surroundings. EPS originate from cytoplasmic sugars, and in some cases (e.g., succinoglycan in *Rhizobium*; alginate in *Pseudomonas*) details of the mechanisms of assembly, polymerization, and export are known. In *Rhizobium*, over 20 genes are known to be involved in EPS synthesis.

Structure

EPS are primarily composed of carbohydrates, but, in addition to sugars, there may be organic and inorganic components. Many or most are polyanionic in nature. Organic substituents include various organic acids (acetate, pyruvate, etc.) and amino acids. Inorganic substituents include sulfate, phosphate, and also cations (due to the polyanionic nature). For example, some alginates bind Ca^{2+}. Hetero- and homopolysaccharides are distinguished, depending on whether the monomers in the polymer are different or all the same. EPS can also be further distinguished based on linkage types and branching.

Functions

EPS mediate attachment to roots Work with associative diazotrophs (N-fixing bacteria), for example *Alcaligenes*, has shown that EPS are important in attachment of bacteria to roots. Mutant strains of *Alcaligenes faecalis* that produce either no or reduced levels of EPS, or that strongly overproduce EPS are not able to attach to rice roots as well as wild-type bacteria (the difference being several orders of magnitude of cells per grams of root).

EPS as symbiotic signals: *Rhizobium* EPS are known to play a role in the development of root nodule symbioses between *Rhizobium* and legumes. A mutant of *R. meliloti* that is defective in the production of the EPS succinoglycan (a heteropolysaccharide) is defective in nodule colonization, although (noninfected) pseudonodules are formed. This indicates that succinoglycan is not the initiating signal, but that it is required for successful nodule invasion. Invasion is blocked where the infection thread (which normally carries rhizobia into the nodule cells of the plant) penetrates beyond the epidermal cell layer.

Desiccation resistance Bacteria in soil generally exist within a matrix mainly composed of EPS. This EPS matrix may protect bacteria from drying and from fluctuations in water potential. EPS production may hence serve to alter their microenvironment to

enhance survival of desiccation in soils that experience drought conditions (complementing other mechanisms, such as synthesis of compatible solutes or uptake of inorganic solutes such as K^+). Slowing the rate of drying within the microbe's microenvironment may increase the time for adjustment of metabolism. Additionally, increased water retention and diffusion rates in EPS may help soil bacteria to maintain physiological functions at low water potentials. At various soil water potentials, glucose diffuses faster in EPS or in EPS-amended clay than in pure clay, due to the higher volumetric water content of EPS and their water-saturated porosity.

In experiments using a sand matrix, EPS produced by a soil *Pseudomonas* species has held several times its weight in water at low water potentials (e.g., between 3 and 10 g water per gram of solid EPS at −3 MPa), suggesting that EPS may serve as a water reservoir for cells in the vicinity. EPS-amended sand also holds more water at all water potentials than unamended sand, and the presence of EPS slows the drying rate of the sand. Cultures of the bacteria exposed to desiccation while growing in sand contain more EPS, suggesting that resources are allocated to EPS production during drought. Also, the structure of the bacterial EPS is altered in response to desiccation (as shown by cryoscanning electron microscopy), such that the EPS structures appear less porous and more massive under desiccation.

Soil aggregation Numerous products and activities of microbes in soil contribute to the complex process of soil aggregation (*See* **Aggregation:** Microbial Aspects). Organic binding agents that are involved in stabilizing aggregates can be categorized into transient, temporary, and persistent binding agents based on their persistence in soil. Transient binding agents are decomposed rapidly (and produced rapidly) by microorganisms in soil; the most important of these are polysaccharides, including EPS of microbial origin. Evidence for the involvement of polysaccharides in soil aggregation comes from many sources, spanning the spectrum of soil ecological approaches. Correlative studies have related the concentration of certain fractions of extracted polysaccharide (e.g., NaOH or hot water extraction) with aggregate stability; however, as is the case with many other soil extractions, large amounts of other components of soil organic matter may be coextracted (for example humic acid). Some studies have used the ratio of the monosaccharides mannose to xylose (M/X) to assess the relative origin of carbohydrates in soil (after hydrolysis). The proportion of carbohydrates of microbial origin is assumed to increase with increasing M/X ratios. Other approaches include selective oxidation of polysaccharides (using periodate), electron-microscopical observations of aggregates (showing attachment of clay particles to EPS material), and controlled environment studies with defined microbial EPS and clays.

Alginate: a model system for EPS biology The acidic EPS alginate has been shown to be involved in biofilm formation of the Gram bacterium *Pseudomonas aeruginosa*. Alginate biosynthesis has become a model system in the study of EPS production (perhaps not least due to the biomedical context of cystic fibrosis research). Alginate is composed of β-1,4-linked D-mannuronic acid and α-L-guluronic acid. The monomers can be arranged in a homopolymeric or heteropolymeric fashion. In addition bacterial alginates are frequently O-acetylated to different degrees. The polymeric structure and O-acetylation determine the physicochemical properties of this EPS; for example, increased O-acetylation increases the water-binding capacity of alginate. Bacterial alginate has a high molecular mass and is negatively charged (as a consequence it is hydrated and viscous). Alginate is produced in the environment, in human infections, and industrial systems. It also functions as a virulence factor of plant pathogenic *Pseudomonas* spp., and is involved in the encystment process of *Azotobacter* spp.

A considerable amount of detail is available for the biosynthesis pathway of alginate. More than 20 genes are involved in alginate biosynthesis in *P. aeruginosa* (all located on the bacterial chromosome). Biosynthesis starts with fructose-6-phosphate and proceeds through the phases of activation of precursor, oxidation of precursor (to guanosine diphosphate (GDP)-mannuronate), polymerization, modification (epimerization, O-acetylation), and export. The initial steps in the pathway are common in general carbohydrate metabolism; a critical shunt and kinetic control point in alginate synthesis is GDP-mannose dehydrogenase (*algD*), the *algD* gene being tightly controlled. The export mechanism of alginate is not clearly established yet, but modification of alginate takes place in the periplasmic space: O-acetylation and epimerization. It is not clear how these reactions are controlled, but their rates are important in determining the properties of alginate.

Alginate biosynthesis is triggered by a variety of environmental factors. These include high osmolarity, nutrient limitation, presence of toxic compounds (e.g., copper), and attachment of bacteria to a solid surface. High osmolarity and ethanol (dehydrating agent) enhance transcription of alginate biosynthesis genes (and to a limited extent, alginate production) in *P. aeruginosa* and *P. syringae*. AlgU, a regulatory

element (sigma factor) involved in the regulation of biosynthesis of the EPS alginate in *P. fluorescens*, is essential in providing desiccation tolerance *in vitro* and in soil. ATR/FT-IR and CLSM were used to study the role of alginate and alginate structure in biofilm formation by *P. aeruginosa* on surfaces. Alginate appears to be required for formation of thicker, three-dimensional biofilms.

Extracellular Lipid Compounds

Lipids are typically discussed in the context of membrane composition (e.g., phospholipids) or intracellular storage compounds; however, extracellular lipids are produced by both prokaryotes and eukaryotes. Simple extracellular lipids (e.g., free fatty acids, and various oxidation products of microbes utilizing alkanes) are not discussed here; the focus is on complex extracellular lipids. These contain long-chain acyl moieties attached to a nonlipid component; the latter is frequently a carbohydrate, a protein, or a peptide.

The combination of lipid and nonlipid components in the molecule confers hydrophilic and hydrophobic properties, and hence some of these compounds can be active as surfactants (also referred to as biosurfactants). The emulsifying ability of some complex extracellular lipids explains probably most of their biological role; however, some of these lipid compounds also alter the hydrophobicity of the microbial cell, which also influences (positively or negatively) uptake of, for example, hydrocarbons. Microbially produced surfactants offer the advantage of being potentially less toxic and more biodegradable than some synthetic surfactants, explaining the heightened interest in this compound in biotechnology. The efficiency of biosurfactants in soil remediation has been studied in the case of soil contamination by metals (e.g., Cd^{2+}, Zn^{2+}, Pb^{2+}), polychlorinated biphenyls (PCBs), phenantrene, and polyaromatic hydrocarbons (PAHs).

Glycolipids are lipids with a sugar moiety; the most common groups are sophorolipids (glycolipids containing the disaccharide sophorose; e.g., in *Candida* yeasts), trehaloselipids (containing the disaccharide trehalose; e.g., in *Arthrobacter*, *Rhodococcus*), and rhamnolipids (containing the sugar rhamnose). Rhamnolipids have been extensively studied with respect to their surfactant ability (mostly in the context of bioremediation of metal or hydrocarbon contaminated soils and subsoils), for example in strains of the soil bacterium *P. aeruginosa*.

In addition to having surfactant function (allowing the bacteria to grow on hydrocarbons), rhamnolipids have some bactericidal (against Gram-negative) and antiviral activity. Rhamnolipid from *P. aeruginosa* has zoosporicidal activity against phytophathogenic fungal species such as *Pythium*, *Phytophthora*, and *Plasmopara*. Proposed mechanisms of action include disruption of the fungal plasma membrane by the rhamnolipid. This points to a potential use of glycolipids in biological control of agricultural pests.

See also: **Aggregation:** Microbial Aspects; **Enzymes in Soils**; **Mycorrhizal Fungi**; **Organic Matter:** Principles and Processes

Further Reading

Boulton CA (1989) Extracellular microbial lipids. In: Ratledge C and Wilkinson SG (eds) *Microbial Lipids*, vol. 2, pp. 669–694. London: Academic Press.

Chenu C and Roberson EB (1996) Diffusion of glucose in microbial extracellular polysaccharide as affected by water potential. *Soil Biology and Biochemistry* 28: 877–884.

Foster RC, Rovira AD, and Cock TW (1983) *Ultrastructure of the Root–Soil Interface*. St Paul, MN: The American Phytopathological Society.

Gacesa P (1998) Bacterial alginate biosynthesis – recent progress and future prospects. *Microbiology* 144: 1133–1143.

Leigh JA and Walker GC (1994) Exopolysaccharides of *Rhizobium*: synthesis, regulation and symbiotic function. *Trends in Genetics* 10: 63–67.

Roberson EB and Firestone MK (1992) Relationship between desiccation and exopolysaccharide production in a soil *Pseudomonas* sp. *Applied and Environmental Microbiology* 58: 1284–1291.

Tisdall JM and Oades JM (1982) Organic matter and water-stable aggregates in soils. *Journal of Soil Science* 33: 141–163.

Wingender J, Neu TR, and Flemming HC (1999) *Microbial Extracellular Polymeric Substances*. Berlin: Springer-Verlag.

Wösten HAB (2001) Hydrophobins: multipurpose proteins. *Annual Reviews of Microbiology* 55: 625–646.

Wright SF and Upadhyaya A (1998) A survey of soil for aggregate stability and glomalin, a glycoprotein produced by hyphae of arbuscular mycorrhizal fungi. *Plant and Soil* 198: 97–107.

Poorly Crystalline Alluminosilicates *See* **Amorphous Materials**

POROSITY AND PORE-SIZE DISTRIBUTION

J R Nimmo, US Geological Survey, Menlo Park, CA, USA

Published by Elsevier Ltd.

A soil's porosity and pore-size distribution characterize its pore space, that portion of the soil's volume that is not occupied by solid material. The basic character of the pore space governs critical aspects of almost everything that occurs in the soil: the movement of water, air, and other fluids; the transport and the reaction of chemicals; and the residence of roots and other biota. By convention the definition of pore space excludes fluid pockets that are totally enclosed within solid material – vesicles or vugs, for example, that have no exchange with the pore space that has continuity to the boundaries of the medium. Thus we consider a single, contiguous pore space within the body of soil. In general it has fluid pathways that are tortuous, variably constricted, and usually highly connected among themselves. Figure 1 shows an example of a two-dimensional cross section of soil pore space.

The pore space is often considered in terms of individual pores – an artificial concept that makes possible many quantifications of its essential character. Though many alternatives could serve as a basis for the definition of pores and their sizes, in soil science and hydrology these are best conceptualized, measured, and applied with respect to the fluids that occupy and move within the pore space.

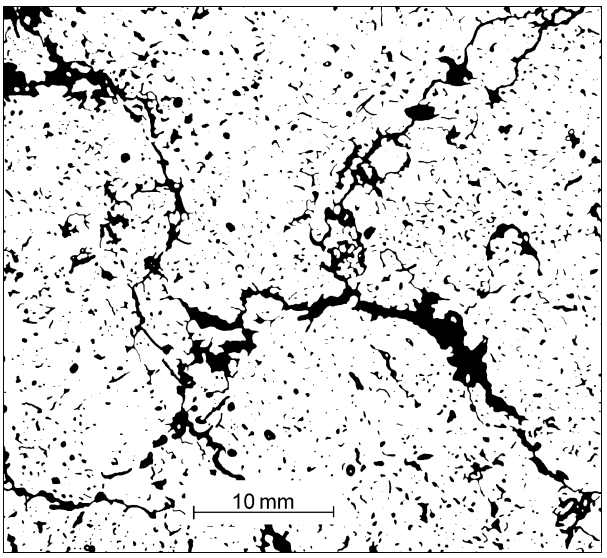


Figure 1 Cross section of a typical soil with pore space in black. Study of this figure would lead to an underestimate of porosity, because pores smaller than about 0.1 mm do not appear. (Adapted from Lafeber D (1965) The graphical representation of planar pore patterns in soils. *Australian Journal of Soil Research* 3: 143–164, with permission.)

Porosity

Porosity, ϕ, is the fraction of the total soil volume that is not occupied by solid material. Thus it is a single-number quantification of the amount of space available to fluid within a specific body of soil. With the assumption that soil is a continuum, porosity can be considered as a function of position. Being simply a fraction of total volume, ϕ can range from 0 to 1, typically falling between 0.3 and 0.7 for soils.

Porosity in Natural Soils

The porosity of a soil depends on several factors, including: (1) packing (bulk) density, (2) the breadth of the particle-size distribution (polydisperse versus monodisperse), (3) the shape of particles, and (4) cementing materials. Mathematically considering an idealized soil of packed uniform spheres, ϕ must fall between 0.26 and 0.48, depending on the packing. Spheres randomly thrown together will have ϕ near the middle of this range, typically 0.30–0.35. A sand with grains nearly uniform in size (monodisperse) packs to about the same porosity as spheres. In a polydisperse sand, the fitting of small grains within the pores between large ones can reduce ϕ, conceivably below the 0.26 uniform-sphere minimum. Figure 2 illustrates this concept. The particular sort of arrangement required to reduce ϕ to 0.26 or less is highly improbable, however, so ϕ also typically falls within 0.30–0.35 for polydisperse sands. Particles more irregular in shape tend to have larger gaps between their nontouching surfaces, thus forming media of greater porosity. In porous rock such as sandstone, cementation or welding of particles not only creates pores that are different in shape from those of particulate media, but also reduces the porosity as solid material takes up space that would otherwise be pore space. Porosity in such a case can easily be less than 0.3, even approaching 0. Cementing material can also have the opposite effect. In many soils, clay and organic substances cement particles together into aggregates. An individual aggregate might have a 0.35 porosity within it, but the medium as a whole has additional pore space in the form of gaps between aggregates, so that ϕ can be 0.5 or greater. Observed porosities can be as great as 0.8–0.9 in a peat soil (extremely high in organic matter).

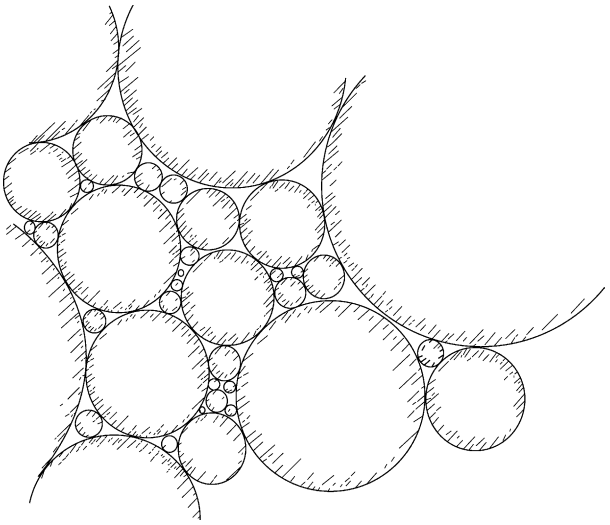


Figure 2 Dense packing of polydisperse spheres. (Adapted from Hillel D (1980) *Fundamentals of Soil Physics*. San Diego: Academic Press, with permission.)

Porosity is often considered to partition into two components, most commonly called textural and structural porosity. The textural component is the value the porosity would be deduced to have if the arrangement of the particles were random, as described above for granular material without cementing. That is, the textural porosity might be approximately 0.3 in a granular medium. (The low end of the 0.30–0.35 range is preferred, because the high end probably reflects nonrandom structural influences in real sands.) The structural component, often considered to represent macropores, is arithmetically the difference between the textural porosity and the total porosity.

The texture of the medium relates in a general way to the pore-size distribution, as large particles give rise to large pores between them, and therefore is a major influence on the retention curve. Additionally, the structure of the medium, especially the degree to which it has aggregation, shrinkage cracks, wormholes, etc., substantially influences the retention curve.

Measurement of Porosity

The technology of thin sections or of tomographic imaging can produce a visualization of pore space and solid material in a cross-sectional plane, as in Figure 1. The summed area of pore space divided by total area gives the areal porosity over that plane. An analogous procedure can be followed along a line through the sample, to yield a linear porosity. If the medium is isotropic, either of these would numerically equal the volumetric porosity as defined above, which is more usually of interest.

The volume of water contained in a saturated sample of known volume can indicate porosity. The mass of saturated material less the oven-dry mass of the solids, divided by the density of water, gives the volume of water. This divided by the original sample volume gives porosity.

An analogous method is to determine the volume of gas in the pore space of a completely dry sample. Of course not only the sampling but also the drying of the soil must be done so as not to compress the soil or otherwise alter its porosity. A pycnometer can measure the air volume in the pore space. A gas-tight chamber encloses the sample so that the internal gas-occupied volume can be perturbed by a known amount while the gas pressure is measured. This is typically done with a small piston attached by a tube connection. The Boyle law indicates the total gas volume from the change in pressure resulting from the volume change. This total gas volume minus the volume within the piston, connectors, gaps at the chamber walls, and any other space not occupied by soil gives the total pore volume to be divided by the sample volume.

To avoid the need to saturate with water or air, measurements of particle density, ρ_p, and bulk density, ρ_b, are used. From the definitions of ρ_b as the solid mass per volume of soil and ρ_p as the solid mass per solid volume, their ratio ρ_b/ρ_p is the complement of ϕ, so that:

$$\phi = 1 - \rho_b/\rho_p \qquad [1]$$

Often the critical source of error is in the determination of soil bulk volume, measured on an uncompacted sample or from the volume of the excavation from which the soil sample originated.

Pores and Pore-Size Distribution

The Nature of a Pore

That soil has pores is obvious, but the precise delineation of an individual pore readily leads to ambiguity and contradiction. Pore space can be partitioned into individual pores only by an arbitrary criterion, which often is not explicitly stated when pores or their sizes are discussed. Because of this inherent arbitrariness, some scientists argue that the concepts of pore and pore size should be avoided. Much valuable theory of the behavior of the soil–water–air system, however, has been built on these concepts, defined using widely, if not universally, accepted criteria.

A particularly useful criterion considers the pore space as a collection of channels through which fluid can flow. The effective width of such a channel varies

along its length. Wide portions can be considered individual pore bodies, and narrow portions can be considered the pore openings that separate them. These are often described with anatomical metaphors, the wide part being the 'belly' or 'waist', and the constrictive part being the 'neck' or 'throat.' In a medium dominated by textural pore space such as sand, pore bodies are the intergranular spaces of dimensions typically slightly less than those of the adjacent particles. At another extreme, a wormhole, if it is essentially uniform in diameter along its length, might be considered a single pore. The boundaries of such a pore are of three types: (1) interface with solid, (2) constriction – a plane through the locally narrowest portion of pore space, or (3) interface with another pore (e.g., a crack or wormhole) or a hydraulically distinct region of space (e.g., the land surface).

Pores conceptualized by this cellular approach are especially relevant to hydraulic behavior, as has been recognized for more than 70 years. The initial application was to Haines jumps, shown in Figure 3, still considered the basic phenomena of capillary hysteresis. The pore openings, which control the matric pressure P at which pores empty, are smaller than the pore bodies, which control the P at which pores fill. As the medium dries and P decreases, water retreats gradually as the air–water interface becomes more curved. At the narrowest part of the pore opening, this interface can no longer increase curvature by gradual amounts, so it retreats suddenly to narrower channels elsewhere. An analogous phenomenon occurs during wetting, when the decreasing interface curvature cannot be supported by the diameter of the pore at its maximum width. The volume that empties and fills in this way is essentially an individual pore. Not all pore space is subject to Haines jumps – water remains in crevices and in films (not seen in Figure 3) coating solid surfaces. Various models and theories treat this space in different ways. By the definition above it is part of a pore in addition to the volume affected by the Haines jump.

Pores can be classified according to their origin, function, or other attributes. A textural/structural

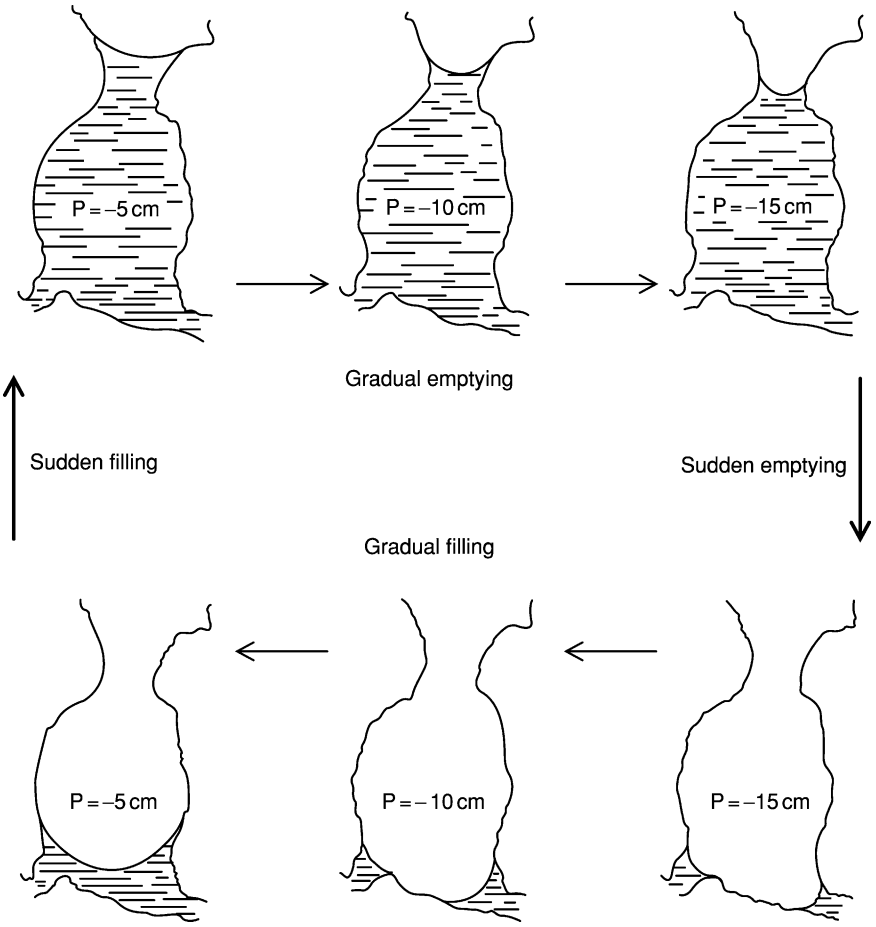


Figure 3 Dynamics of a Haines jump. (Adapted from Miller EE and Miller RD (1956) Physical theory for capillary flow phenomena. *Journal of Applied Physics* 27: 324–332, with permission.)

distinction is possible, analogous to porosity. Hierarchical classifications are also possible. Intragranular or dead-end pores (if not entirely enclosed within solid) might empty or fill with water, but without contributing directly to fluid movement through the medium. Intergranular pores are the major portion of textural porosity, as discussed above. Intraaggregate pores may be essentially equivalent to intergranular pores within an aggregate. Interaggregate pores, including shrink–swell cracks, are common types of macropores. Biogenic pores, for example the channels left by decayed roots and the tunnels made by burrowing animals, are another common type of macropore in soils.

Specific pore sizes are usually specified by an effective radius of the pore body or neck. This relates to the radii of curvature of the air–water interface at which Haines jumps occur, and by capillarity, to the matric pressures at which these occur. Alternative indicators of size include the cross-sectional area or the volume of a pore, and the hydraulic radius, the ratio of the cross-sectional area to circumference, or of pore volume to specific surface.

The pore-size distribution is the relative abundance of each pore size in a representative volume of soil This can be symbolized with a function $f(r)$, proportional to the combined volume of pores of effective radius r. Figure 4 shows examples. As with porosity, the continuum assumption makes valid the concept that $f(r)$ varies from point to point within a body of soil. Also like porosity, $f(r)$ may be taken to comprise textural and structural components.

Measurement

The most obvious and straightforward measurements of pore size are with geometric analysis of images of individual pores. This can be done using various types of microscopy on thin sections or other flat soil surfaces, or tomographs. Dimensions of pore bodies and necks can be measured manually or by computer analysis of digitized images. The lengths of segments of solid or pore along one-dimensional transects can serve similar purposes. As in the case of porosity, isotropy is required for assuming the equality of lineal, areal, and volumetric pore size distribution. For pore size, however, a more important problem is that, when working with fewer than three dimensions, one doesn't know what *part* of pore the selected slice intersects; because it does not in general go through the widest part, it underestimates the pore diameter. Correction techniques are necessary to estimate true pore body and opening sizes.

Three-dimensional analysis is possible with impregnation techniques. In these, the soil pore space is filled with a resin or other liquid that solidifies. After it has done so, the medium is broken up and individual blobs of solid resin, actually casts of the pores, are analyzed as particles.

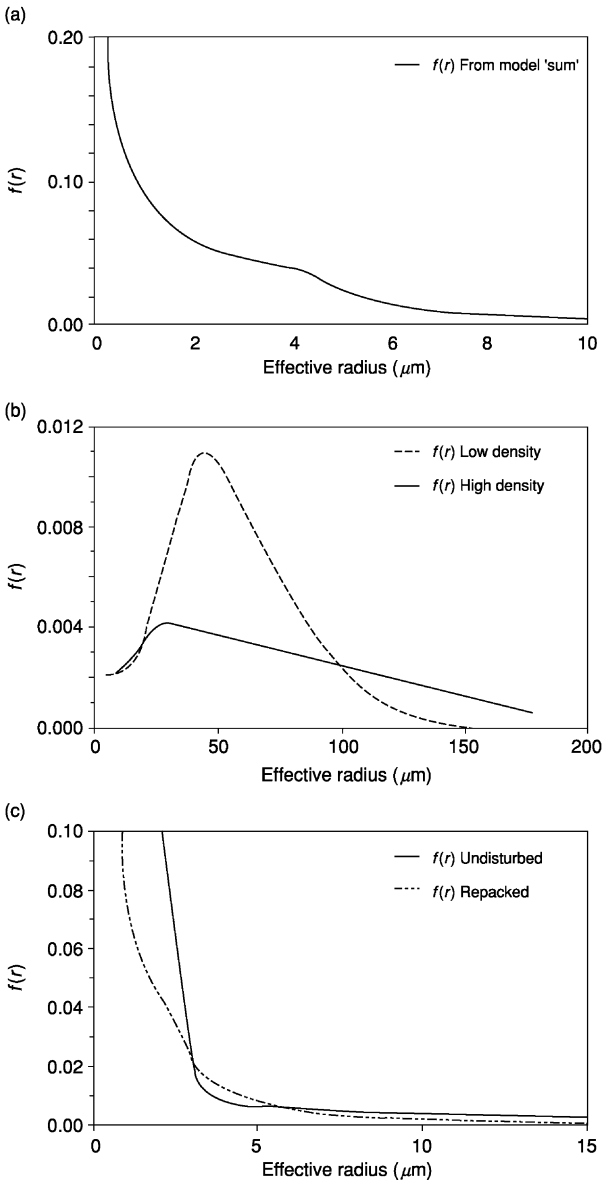


Figure 4 Pore-size distributions based on measured water retention. (a) Loamy soil. (Reproduced with permission from Schofield (1935).) (b) Silty sand at two packing densities. (Data from Croney D and Coleman JD (1954) Soil structure in relation to soil suction (pF). *Journal of Soil Science* 5: 75–84.) (c) Paleosol of sandy loam texture from 42 m depth, as a minimally disturbed core sample, and after disaggregation and repacking to the original density. (Data from Perkins KS (2003) Measurement of sedimentary interbed hydraulic properties and their hydrologic influence near the Idaho Nuclear Technology and Engineering Center at the Idaho National Engineering and Environmental Laboratory. *Water-Resources Investigations Report* 03-4048. Idaho Falls, Idaho: US Geological Survey; and Schofield RK (2002) The pF of the water in soil. *Transactions of the Third International Congress of Soil Science*, pp. 38–48. London: T Murby.)

Image-based techniques can be prohibitively tedious because enough pores must be analyzed to give an adequate statistical representation. They can give a wealth of information, however, on related aspects such as pore shape and connectivity.

More common than imaging methods are those based on effective capillary size. These use data from fluid behavior in an unsaturated medium, usually the emptying or filling of pores during soil drying or wetting (Figure 3), that is, the water-retention curve $\theta(P)$, where θ is the volumetric water content. Because large pores fill or empty at P near 0, a medium that has many large pores will have a retention curve that drops rapidly to low θ at high matric potentials. Conversely, one with very fine pores will retain much water even at highly negative P, thus having a flatter retention curve. By the capillary hypothesis, the P at which a pore empties (or fills) corresponds to the pore opening (or body) size according to:

$$r = \frac{-2\sigma\cos\alpha}{P} \qquad [2]$$

where σ is the surface tension and α is the contact angle. This formula can convert a measured $\theta(P)$ into an equivalent $\theta(r)$ curve. This curve is actually a cumulative pore-size distribution; the water content on a drying $\theta(r)$ indicates the combined volume of all pores with opening radius less than r. Applying the fundamental theorem of calculus, the direct pore-size distribution is simply the derivative:

$$f(r) = \frac{\mathrm{d}\theta}{\mathrm{d}r} \qquad [3]$$

Mercury porosimetry is analogous to the water retention-based method, but uses air as the wetting fluid, corresponding to the water, and mercury as the nonwetting fluid, corresponding to the air in the water–air system. Mercury is forced into the pores of dry soil incrementally, so that the relation between mercury content and mercury pressure is obtained. Applying eqn [2] with the appropriate values of surface tension and contact angle for mercury leads to a pore-size distribution estimate as for the extraction of water from a water–air system.

The different measurement techniques do not give exactly the same results. Emptying and filling depend on more than capillarity – different methods are affected differently by unintended influences. Contact angles, for example, are dynamic, are not likely to be at handbook values in soil, and may deviate quite differently for water and for mercury. The swelling of clays can be a major influence on a water-retention curve, but should have no effect with mercury.

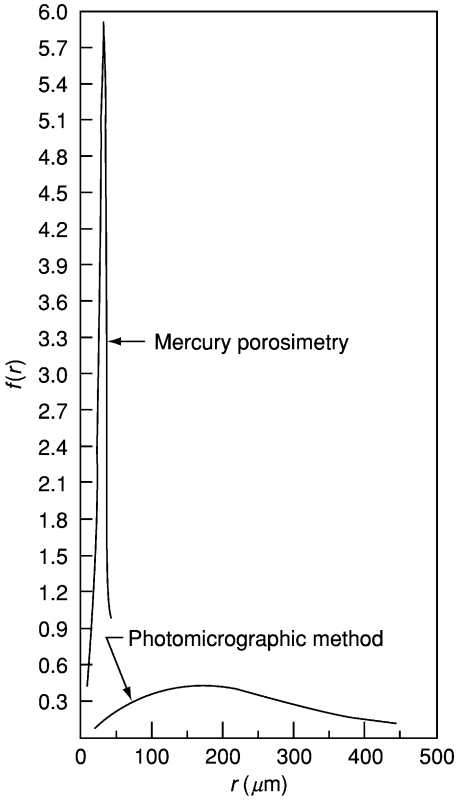


Figure 5 Pore-size distributions obtained for an artificial medium using an imaging method and mercury intrusion. (Reproduced with permission from Dullien FAL and Batra VK (1970) Determination of the structure of porous media. *Industrial Engineering Chemistry* 62: 25–53.)

Imaging techniques are subject to some entirely different influences, such as the chance and subjectivity involved in assessing pore bodies and openings. Figure 5 shows an example of substantially different results from different methods. In soil environmental applications, use of the water-retention curve is most common. Data are more commonly available for this than for any other method. One can also expect a water-based method to give better results for a water-based application (e.g., hydraulic conductivity).

Representation

The basic representation of pore-size distribution is as the function $f(r)$ (Figure 4). Corresponding graphs of the cumulative size distribution can be used equivalently. Often a specific functional form or other representational model of pore-size distribution is useful. A normal or lognormal can be fitted to the data, for example. Structural features may give the soil a bimodal pore-size distribution, leading to several distinctive effects on water flow. Bimodal, trimodal, or other multimodal forms are possible, and can be

represented by superpositions of the normal or lognormal forms. Self-similar or fractal forms of the distribution function are also used, giving a power-law form of the cumulative pore-size distribution.

Typical Features of a Pore-Size Distribution

With the equivalent-capillary concept taken as broadly as possible, the smallest possible pore is limited to the order of a few molecular diameters, approximately 1 nm. Fluid behavior is then likely to be dominated by interaction with the solid material and is not necessarily describable by capillary laws or even standard thermodynamics. The smallest pores measured by water retention or mercury intrusion are typically 50 or 100 times larger than this, though with greater cost and effort, greater magnitudes of pressure and hence smaller measured r can be achieved. As the imputed pore size approaches zero, in effect thin films are associated with effective radii of filled pores. Because r relates inversely to P, the whole dry portion of the retention curve is contained in a very small region near $r = 0$, where the apparent number of pores becomes large. In the extreme case, when θ is held to remain artificially finite at an assumed residual water content as P goes to negative infinity, $f(r)$ at $r = 0$ is a delta function.

There is no real upper limit to pore size, though instability will cause the capillary hypothesis to break down for r of a few millimeters. The experimental limit is typically approximately 0.5 mm.

At either extreme, where capillary phenomena lose dominance, eqn [2] no longer applies, though it still can give r-values corresponding to P. These may be useful for translating one property to another (discussed below) even though they are invalid in terms of the capillary analog.

At intermediate values of r, sometimes $f(r)$ has a pronounced peak, as in **Figures 4b** and 5. Such a peak is assumed to exist in many applications and interpretations of pore-size distributions, for example normal or lognormal representations, though it does not always exist in the measurable range of r. A peak in $f(r)$ corresponds to an inflection point in $\theta(r)$. The inverse relation of r to P means that $\theta(P)$ is much more likely to have an inflection point than $\theta(r)$; if $\theta(r)$ has an inflection point, then $\theta(P)$ is mathematically required to have an inflection point, but the converse is not true. The common case of an S-shaped retention curve (with inflection point in the measured range, and normally a distinct air-entry effect) will have a defined peak in the pore-size distribution only if the middle portion of the retention curve has a slope that differs markedly from that of the two end segments. This is frequently true for monodisperse media and for repacked samples, in which pores larger than a certain r have been destroyed. For many soils, perhaps most, $f(r)$ does not have a peak except at $r = 0$ or so close to 0 that it cannot be measured. This is especially likely for media that are macroporous, polydisperse, or that show an air-entry effect at essentially $P = 0$. Retention curves represented by a fractal or power-law model are also in this category. Lognormal models or other representations that have a defined peak can often still be applied but with the peak outside the measurable range. Some models (e.g., of hysteresis and hydraulic conductivity) may be unaffected by the lack of a measurable peak even though they were derived assuming the pore-size distribution resembles some form of normal distribution.

Significance to Soil and Water Behavior

Significance of Porosity

One obvious significance of porosity is that it is an upper limit of volumetric water content. It is similarly essential to the definition of degree of saturation, θ divided by porosity. Another significance is that, within the pore space, the complement to θ is gas content. That is, volumetric gas content equals the difference between θ and porosity.

The magnitude of porosity roughly indicates complexity of structure, being greater for greater complexity, as noted above for aggregated soil. The spatial variability of porosity is also important, greater variability correlating with greater heterogeneity of the soil. Significant spatial variability on a small scale also implies greater structural complexity, independent of the magnitude of porosity.

Significance of Pore-Size Distribution

A major importance of a soil's pore-size distribution is that it relates to other soil properties in a complex and useful way, as discussed below. It indicates complexity of structure in far more detail than porosity alone. The spatial variation of pore size is an important characteristic of the medium. The pore-size distribution of different parts of soil is the fundamental basis for the concept of aggregates, for example. By some definitions, pore size can permit essential distinctions between micropores and macropores (and mesopores, where that term is used for intermediate-size pores).

The relation of pore-size to particle-size distribution in a randomly structured medium is likely to be monotone: bigger pores are associated with bigger particles. With the nonrandom structure of most soils, the relation to particle size is complex. Bigger pores can be associated with bigger particles, but

also with smaller particles such as clays that promote aggregation and hence the existence of large interaggregate pores.

With more intricate conceptualizations, subdivisions of a pore-size distribution can represent more detailed properties of soils, for example the hysteresis of soil water retention. Figure 6 shows an example in which one graph represents pore-opening and pore-body size distributions. The added information of what fraction of pores with a given opening size have a particular range of body sizes is taken to represent the essential quantitative basis of soil water hysteresis.

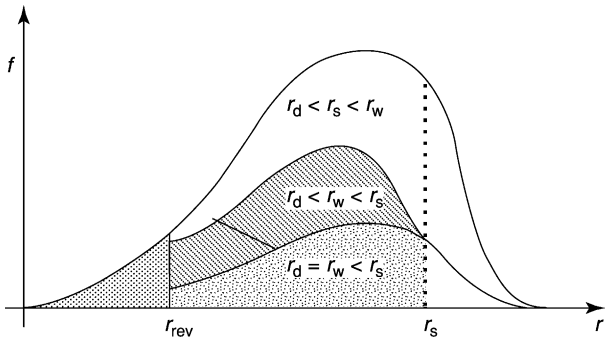


Figure 6 Hypothetical pore-size distribution with subregions distinguished on the basis of pore body and opening radii and drying–wetting history. (Adapted from Nimmo JR (1992) Semiempirical model of soil water hysteresis. *Soil Science Society of America Journal* 56: 1723–1730.)

Pore Dynamics

Natural and artificial soil processes create and destroy pores, and induce changes in their size and other attributes. Table 1 lists some of the common effects on pore-size distribution caused by routine processes. In general, if the soil is compacted by uniform stress, such as a weight imposed at the land surface, it normally loses large pores and gains small pores. Figure 4b shows results of this type. Disturbance from irregular stresses, as during digging or repacking, has a variety of effects on pore size, often with the net effect of a decrease in the number of large pores and an increase in the number of small pores, as shown in Figure 4c. Several types of process can create pores. Though small intergranular pores are seldom completely closed, some processes can close macropores, in effect destroying them.

Applications to Soil Transport Properties

Because the pores it characterizes are conduits and repositories for water, the pore-size distribution relates directly to soil hydraulic properties. In addition to water retention, as described above, because the characterized pores are also fluid conduits, it connects with predictions of hydraulic conductivity

Table 1 Possible effects of routine soil processes on pore-size distribution

Soil process	*Effect*
Shrinkage	Enlarged macropores
	New macropores created
	Within an aggregate, intraaggregate pores decrease in size, or increase in size if clay particles shrink
Swelling	Decrease in the size of macropores
	Closed macropores
	Within an aggregate, intraaggregate pores increase in size, or decrease in size if clay particles expand
Mechanical compression	Decrease in the size of macropores
	Closed macropores
	Aggregates break up, reducing the number of intraaggregate pores, thereby reducing the fraction of the pore space represented by the smallest pores
Disturbance from digging or plowing	Existing macropores destroyed
	Interclod macropores created
	Aggregates break up, reducing the number of intraaggregate pores, also the fraction of the pore space represented by the smallest pores
Biological activity	New macropores created
	Enlarged macropores, as by ongoing traffic of ants or burrowing mammals
	Decrease in the size of macropores, e.g., if they are affected by compression resulting from the expansion of a nearby root
	Increased aggregation, promoting the creation of interaggregate macropores, and possibly smaller intergranular pores within aggregates
	Constricted or obstructed pores, e.g., through growth of microorganisms
Chemical activity	Constricted or obstructed pores through formation of precipitates
	Enlarged pores through dissolution of precipitates
	Increased or decreased interparticle cohesion, with complex effects on pore size and structure

K. The same principles can be applied to the transport of fluids other than water, including gases, though water flow is the most common application.

By analogy to laminar flow in tubes as quantified by the Poiseuille law, the conductance of a single pore can be inferred to be proportional to the fourth power of its effective radius. This makes its hydraulic conductivity proportional to the square of its effective radius. An estimated $f(r)$ distribution tells the relative abundance of each conduit size, thus providing the ingredients needed to predict K.

A K prediction based on the capillary hypothesis assumes pores are filled if smaller than a certain radius determined by θ. The portion of $f(r)$ representing the filled pores is relevant to unsaturated K. The simplest possibility is to integrate $f(r)$ weighted by r^2 over the domain represented by filled pores, giving a number proportional to K for the corresponding θ. Multiplying this integration by a separately determined matching factor gives the actual predicted K.

Many specific models in the published literature are based on these ideas. They differ in how they treat matters such as pore length, connectedness, tortuosity, and the distinction between pore opening and body dimensions. Popular models include those of Mualem and Burdine, which have been analytically combined with widely used empirical formulas for retention curves.

Considering solute transport, the pore-size distribution affects advection as it does hydraulic conductivity. Additionally it is important to dispersion, expected to be greater for a broader pore-size distribution. Adsorptive reactions with other soil constituents are related in a complex way. The smaller pores are associated with longer residence times and greater relative surface area, but most solutes may go quickly through the large pores with minimal opportunity to react. Interchange between fast-transporting and slow-transporting portions of the pore space is a vital and much-investigated aspect of solute transport in soils. Sometimes the terms 'mobile' and 'immobile' are used in this context, but of course the distinction is not as sharp as these terms imply.

For particle transport, many aspects are essentially the same as for K and solute transport. Additionally, the phenomenon of straining depends critically on the proportion of pores smaller than a given particle size. This is the dominant factor in some particle-transport applications.

Pore-size distribution relates indirectly to heat transport. It correlates in some way with the degree of aggregation, the moisture relations, the number of points of particle contact, and other factors affecting thermal conductivity.

Summary

The characterization of pore space is a vital and fruitful aspect of soil investigation. Liquid, solid, and gas constituents of the soil govern the form and development of pores, whose character in turn profoundly influences the nature and behavior of the soil.

Soil porosity is fairly well standardized in definition and measurement techniques. Pore size, however, is not obviously definable, much less measurable. Yet it is central to topics such as macropores, aggregation, fractures, soil matrix, and solute mobility. Pore size plays a key role in various proposed means of quantifying soil structure. It also has a major practical role in the prediction of hydraulic properties. New pore-size concepts, measurement techniques, and relations to transport phenomena are likely to remain a major emphasis in the study of soil.

See also: **Aggregation:** Physical Aspects; **Capillarity**; **Fractal Analysis**; **Hysteresis**; **Isotropy and Anisotropy**; **Macropores and Macropore Flow, Kinematic Wave Approach**; **Spatial Variation, Soil Properties**; **Structure**; **Swelling and Shrinking**; **Texture**

Further Reading

Adamson AW and Gast AP (1997) *Physical Chemistry of Surfaces*, 6th edn. New York: John Wiley.

Baver LD (1938) Soil permeability in relation to non-capillary porosity. *Soil Science Society of America Proceedings* 3: 52–56.

Burdine NT, Gournay LS, and Reichertz PP (1950) Pore size distribution of petroleum reservoir rocks. *Petroleum Transactions – AIME* 189: 195–204.

Childs EC and Collis-George N (1950) The permeability of porous materials. *Proceedings of the Royal Society of London*. In: *Series A: Mathematical and Physical Sciences*, pp. 392–405.

Dullien FAL and Batra VK (1970) Determination of the structure of porous media. *Industrial Engineering Chemistry* 62(10): 25–53.

Haines WB (1930) Studies in the physical properties of soil. V. The hysteresis effect in capillary properties, and the modes of moisture distribution associated therewith. *Journal of Agricultural Science* 20: 97–116.

Kosugi K (1994) Three-parameter lognormal distribution model for soil water retention. *Water Resources Research* 30(4): 891–901.

Marshall TJ and Holmes JW (1988) *Soil Physics*, 2nd edn. Cambridge, UK: Cambridge University Press.

Miller EE and Miller RD (1956) Physical theory for capillary flow phenomena. *Journal of Applied Physics* 27(4): 324–332.

Mualem Y (1974) A conceptual model of hysteresis. *Water Resources Research* 10(3): 514–520.
Mualem Y (1976) A new model for predicting the hydraulic conductivity of unsaturated porous media. *Water Resources Research* 12(3): 513–522.
Nimmo JR (1997) Modeling structural influences on soil water retention. *Soil Science Society of America Journal* 61(3): 712–719.
Nimmo JR and Akstin KC (1988) Hydraulic conductivity of a sandy soil at low water content after compaction by various methods. *Soil Science Society of America Journal* 52(2): 303–310.
Raducu D, Vignozzi N, Pagliai M, and Petcu G (2002) Soil structure of tilled horizons influenced by management practices and implement geometry. *Advances in Geoecology* 35: 149–162.
Ragab R, Feyen J, and Hillel D (1982) Effect of the method for determining pore size distribution on prediction of the hydraulic conductivity function and of infiltration. *Soil Science* 134(6): 141–145.

POTASSIUM IN SOILS

P M Huang, University of Saskatchewan, Saskatoon, Canada
J M Zhou and J C Xie, The Institute of Soil Science, Academia Sinica, Nanjing, People's Republic of China
M K Wang, National Taiwan University, Taipei, Taiwan

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Potassium (K) is a major constituent of soils and plants. The common range of K content in the soil is 0.4–30 g K kg^{-1}. Of the major nutrient elements, K is generally the most abundant in the soil. Of this total K content, 98% is bound in the mineral form, whereas 2% is in soil solution and exchangeable phases.

Among the mineral cations essential for plants, K is the largest in size. Therefore, the number of oxygen ions coordinating K in mineral structure is high. The strength of each K-O is, thus, relatively weak. Potassium has a polarizability of 0.088 nm^3, which is higher than for Ca^{2+}, Li^+, Mg^{2+}, and Na^+, but lower than for Ba^{2+}, Cs^+, NH_4^+, and Rb^+ ions. Ions with higher polarizability are preferentially selected in ion exchange reactions.

The vitality of K in plant growth and yield formation has been known since the work of von Liebig, published in 1840. The availability of K to plants is related in many ways to the structural and surface chemistry of soil minerals. The large amount of K in soil minerals, fixation of applied K, and ion selectivity and dynamics of soil K have led researchers to study the relationships between K uptake by plants and the chemistry of soil K for decades.

Forms of Soil Potassium

Soil K exists in four forms in soils: solution, exchangeable, fixed, and structural (Figure 1). Solution and exchangeable K generally account for relatively small proportions of the total K in soils. Potassium-bearing micas and feldspars are the major K reserves of the soil. Upon reaction with weathered micas, vermiculites, allophane, and zeolites, K is also present in fixed form in soils. In the immediate vicinity of the fertilizer zone, K can react with aluminum hydroxides and acid phosphate solution to form taranakite. However, the nucleation and crystallization of taranakite can be perturbed in soils with high Fe content, especially under reduced and acidic conditions. In acid soils, K could also be coprecipitated with Al and sulfate to form alunite. The order of the soil K forms in terms of bioavailability is: solution > exchangeable > fixed > structural.

Solution K

Plants and microorganisms directly take up soil solution K. This form of K is also most subject to leaching in soils. The concentration of soil solution K varies from 2 to 5 mg K l^- for normal agricultural soils of humid regions and is an order of magnitude higher in arid-region soils. The amount of soil solution K is generally too low to meet the requirement of K by a crop during a growing season. Soil solution K is in dynamic equilibrium with other forms of soil K and is influenced by the soil moisture content, the concentrations of cations in the solution and on the exchanger phase, and the concentration of solution inorganic and organic ligands.

Exchangeable K

Exchangeable K is the portion of soil K that is electrostatically bound as an outer-sphere complex to the surfaces of soil minerals and organic matter. It is readily exchanged with other cations and bioavailable. The replenishment of a K-depleted soil solution is affected by the release of exchangeable K from soils.

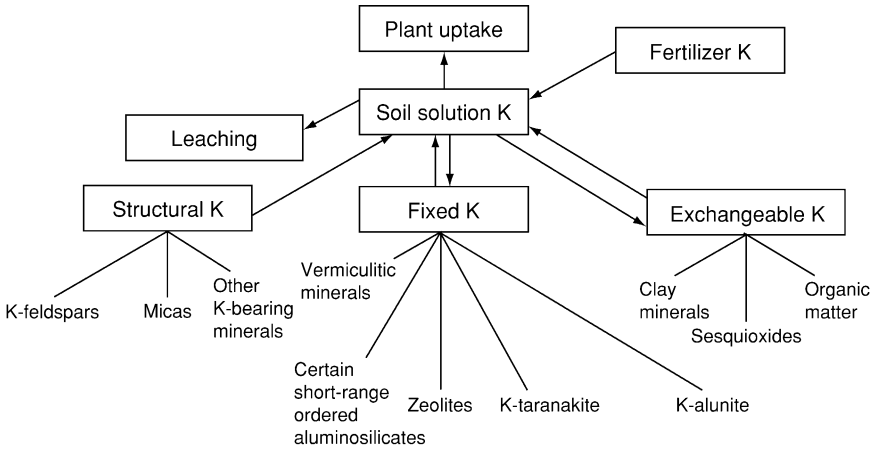


Figure 1 Interrelationship of various forms of soil K. Reproduced with permission from Sparks DL and Huang PM (1985) Physical chemistry of soil potassium. In: Munson RD (ed.) *Potassium in Agriculture*. Madison, WI: American Society of Agronomy, Crop Science Society of America, and Soil Science Society of America.

Fixed K

Fixed K (nonexchangeable K) is different from structural K in that it is not bound within the crystal structures of soil minerals. It is held in the interlayer space of weathered micas, vermiculites, and hydroxy-interlayered vermiculites. Potassium becomes fixed because the binding forces between K and the mineral surfaces are greater than the hydration forces between individual K ions. This results in a partial collapse of the crystal structures and the K ions are physically trapped to varying degrees, making K release a slow, diffusion-controlled process. Fixed K can be found in a series of soil components and is also present in wedge zones of weathered micas and vermiculites. Cations with a size similar to K^+, e.g., NH_4^+ and H_3O^+, can exchange K from wedge zones of interlayers. Large hydrated cations, such as Ca^{2+} and Mg^{2+}, cannot fit into the wedge zones and, thus, cannot release fixed K. Fixed K is moderately to sparingly available to plants. Release of fixed K to the exchangeable form can take place when levels of exchangeable and solution K are decreased by plant removal, microbial activity, and leaching. The ability of plants to take up nonexchangeable K appears to be related to the plant species.

Structural K

Most of the total soil K is in the form of structural K, mainly as K-bearing primary minerals such as feldspars and micas. The total soil K is present in a series of particle-size fractions. The distribution of K in feldspars and micas in a wide range of particle-size fractions of moderately weathered soil profiles is illustrated in Table 1. The K-feldspars and K-micas of soils serve as an important reserve of available K, although structural K is generally assumed to be only slowly available to plants. The availability of structural K depends on the dynamics of K release from micas and K-feldspars, which are, in turn, related to the crystal structure and atomic bonding of these K-bearing primary minerals, the degree of weathering of these minerals, and the level of K in other forms. A substantial amount of K is released from the sand fractions of Delaware soils that are high in K-feldspars. This, along with the large quantities of mineral K in these and other Atlantic coastal plain soils, helps to explain the often observed lack of crop response to K fertilization on these soils.

Ion Selectivity and Fixation of Potassium in Soils

Ion Selectivity

Potassium selectivity of soils is affected by their mineralogical and chemical properties.

Mineralogical properties The layer charge, particle size, configuration of interlayers and wedge zones, position of wedge zones in interlayers, hydroxy interlayers, and variable-charge minerals are important mineralogical properties that affect K ion selectivity of soils.

The site of negative charge and the charge density of minerals may affect K^+ selectivity. The high K^+ selectivity of clay-size muscovite is attributed to the negative charge arising in the tetrahedral sheet. The K^+/Ca^{2+} cation exchange selectivity (CES) values, expressed as the equivalent ratio of exchangeable K^+ to Ca^{2+}, range from 0.1 for montmorillonite to 1.7 for muscovite. This corresponds to the order of

Table 1 Distribution of feldspar K in various size fractions of a Haverhill soil profile (Aridic Haploboroll) in Saskatchewan, Canada

Horizon	Particle size (μm)	Total K ($g\,kg^{-1}$)	Feldspar K ($g\,kg^{-1}$)	Mica K ($g\,kg^{-1}$)	% fraction of K	
					Feldspar	Mica
Ap	<0.2	21.1	–	21.1	–	100
	0.2–2	26.1	0.8	5.3	3	97
	2–5	20.1	6.1	14.0	30	70
	5–20	17.3	8.7	8.6	50	50
	20–50	15.4	9.7	5.7	63	37
	50–500	12.7	7.6	5.1	60	40
	500–2000	17.3	17.3	–	100	–
	>2000[a]	16.9	6.8	10.1	40	60
Bm	<0.2	18.3	–	18.3	–	100
	0.2–2	29.9	0.0	29.9	–	100
	2–5	24.6	3.6	21.0	15	85
	5–20	20.4	7.7	12.7	38	62
	20–50	14.7	8.5	6.2	58	42
	50–500	14.2	8.4	5.8	59	41
	500–2000	18.2	18.2	–	100	–
	<2000	17.6	6.5	11.1	38	63
Ck	<0.2	16.8	–	16.8	–	100
	0.2–2	27.4	2.1	25.3	8	92
	2–5	26.2	8.7	17.5	33	67
	5–20	22.7	9.2	13.5	41	59
	20–50	15.6	8.3	7.3	53	47
	50–500	11.9	9.2	2.7	77	23
	500–2000	20.7	20.7	–	100	–
	<2000	17.7	6.9	10.8	39	61

[a]>Whole mineral fraction.

Reproduced with permission from Somasiri S, Lee SY, and Huang PM (1971) Influence of certain pedogenic factors on potassium reserves of selected Canadian Prairie soils. *Soil Science Society of America Proceedings* 35: 500–505.

increasing layer charge and surface charge density of the minerals.

Potassium selectivity of soils varies with particle size. It increases with particle size in some soils. Potassium selectivity may decrease with particle size at later stages of weathering. Much of the K selectivity apparently occurs in the interlayer regions.

Potassium selectivity is influenced by the configuration of interlayers and the position of the wedge zone in the interlayer. Instead of large hydrated ions such as Ca^{2+} and Mg^{2+}, K^{+} is selected due to the space limitations for diffusion of these large hydrated cations in the wedge zones. If the wedge zone is near the edge of the particle, it can be effective in only a small amount of K selection. If the wedge zone is deep within the particle, much more K can be selected. Wedge zones increase as more interlayers are opened and decrease as the interlayers are opened all the way through the particle.

Potassium selectivity is enhanced by the presence of hydroxy-Al and hydroxy-Fe(III) polymers in the interlayer space. These interlayer materials act as props to hinder the collapse of the layers about the K ions. Potassium ions are able to move more easily in propped-open interlayer spaces than in collapsed interlayers (Figure 2). Besides the propping effect, the preferential occupation by the hydroxy-Al polymers of the interlayer exchange sites that would normally adsorb Ca may also affect K selectivity. Due to their size, the polymers would not be adsorbed at wedge zones but rather in a more expanded interlayer position (Figure 3). Consequently, sites that would normally adsorb Ca^{2+} ions would be occupied by Al polymers, whereas wedge sites would be relatively unaffected by the adsorption of hydroxy-Al

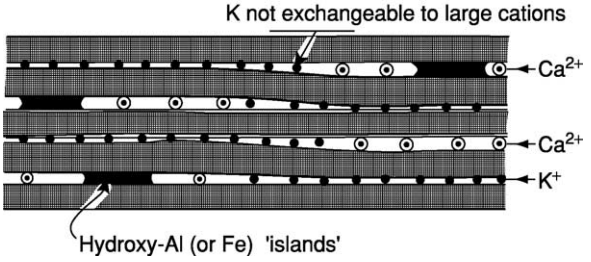


Figure 2 Proposed model of an expandable layer silicate with interlayers indicating effect on K fixation. Reproduced with permission from Rich CI (1968) Mineralogy of soil potassium. In: Kilmer VJ, Younts SE, and Brady NC (eds) *The Role of Potassium in Agriculture*. Madison, WI: American Society of Agronomy, Crop Science Society of America, and Soil Science Society of America.

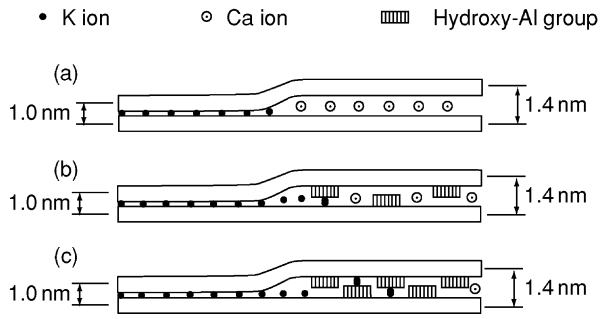


Figure 3 Model of the effects of hydroxy-Al interlayers on K exchange. (a) In the absence of hydroxy-Al interlayers; (b) the propping effect and preferential occupation of Ca-adsorbing sites by hydroxy-Al interlayers; (c) the retarding effect of hydroxy-Al interlayers on the entry of large hydrated Ca ion. Reproduced with permission from Huang PM (1987) Aluminum and the fate of nutrients and toxic substances in terrestrial and freshwater environments. In: Singh M (ed.) *Encyclopedia of Systems and Control*. Oxford: Pergamon Press.

polymers. Hydroxy-Al interlayers may also exert a retarding effect on the entry of the more hydrated Ca^{2+} ions. This would result in the increase of the selectivity toward the less hydrated K ions.

Besides layer silicates, variable-charge minerals also contribute to adsorption and selectivity of K ions. The contribution of these minerals to K selectivity is pH-dependent. Further, both cationic and anionic environments have a profound influence on the adsorption of K^+ and Ca^{2+} in metal oxides such as birnessite (Mn oxide). This is of significance in understanding the chemistry of K in variable-charge soils.

Chemical properties Besides mineralogical properties, chemical properties also affect K selectivity of soils. The hydrated ion size and valence greatly influence K selectivity. In soils containing partially weathered micas and vermiculites where wedge zones exist, NH_4OAc removes more exchangeable K than does $Mg(OAc)_2$, although Mg^{2+} is usually a better replacer of K^+ than NH_4^+. The wedge zones exclude the hydrated Mg^{2+} due to steric hindrance and preferentially adsorb cations that can fit into the wedge sites. $MgCl_2$ can remove more K^+ than can $Mg(OAc)_2$ because of the depression effect of soil pH by $MgCl_2$ and the formation of H_3O^+ by exchange or hydrolysis of exchangeable and polymeric Al. The H_3O^+ ion is similar in size to K and is thus able to fit into the wedge zone and exchange K^+. A combination of NH_4OAc and $MgCl_2$ is more effective in removing exchangeable K^+ than either electrolyte alone. This is attributable to Mg^{2+} opening up the interlayer space and facilitating the exchange of K^+ by NH_4^+.

Ions do not migrate rapidly enough to account for the reaction rate involving H_3O^+ ions. The lifetime of an individual H_3O^+ is exceedingly short, $\sim 10^{-13}$ s, since all of the protons are undergoing rapid exchange. As long as there is a continuous film of H_3O^+ to a K^+ ion, the exchange of H_3O^+ for K could proceed rapidly. Potassium exchange of soils is enhanced by low pH. The effect of pH on ion selectivity varies with the associated cations. pH has little effect for NH_4^+. NH_4^+-saturation collapses the interlayer structure and, thus, excludes H_2O which is the medium for transport of protons to and from the exchange sites. By contrast, the effect of pH for Li^+ and Mg^{2+} is pronounced. The increased efficiency of Li^+ and Mg^{2+} with a decrease in pH is attributed to an expansion of the interlayer structure, which facilitates the entry of water and protons and the exchange of H_3O^+ for K.

Potassium selection in the wedge zones is influenced by the concentration of K^+ in the soil solution. There is K selection at the wedge sites at low K concentration. As the K^+ concentration increases, the silicate layer collapses at the wedge sites, resulting in an entrapment of ions deeper in the interlayer space. A high K^+ concentration is necessary to initiate collapse of the interlayer space of vermiculite and to enhance K selectivity in the absence of wedge zones.

Temperature also affects K selectivity of soils. A rise in temperature decreases K selectivity of Delaware soils (loamy sand, Typic Quartipsamments; sandy loam, Typic Hapludults; silt loam, Typic Hapludults; clay minerals include vermiculite, chloritized vermiculite, mica, kaolinite, chlorite, and quartz). This indicates that K sorption decreases with increasing temperature. The Delaware soils contain wedge zones that would selectively exclude large hydrated cations such as Ca^{2+} and Mg^{2+} and preferentially select smaller cations such as NH_4^+ and K^+. As the temperature increases, the hydration of Ca^{2+} and Mg^{2+} is decreased relative to that of K^+. Therefore, the size is not as important in K^+ selectivity. Furthermore, increased temperature could collapse the interlayer space of hydroxy-interlayered vermiculite and, thus, decrease K^+ selectivity of the soils.

Fixation of Potassium

Potassium fixation in soils has been extensively studied. The extent of K fixation is related to the type of clay minerals and their charge density, the degree of hydroxy interlayering, the concentration of K^+ ions and competing ions such as NH_4^+, the moisture content, and solution pH. The expansible 2:1 layer silicates such as weathered micas, vermiculites, and smectites are the major clay minerals responsible for K fixation in soils. Weathered micas and vermiculites

fix K under moist as well as dry conditions, whereas smectites fix K only under dry conditions. The amount of K fixation by smectites is very small unless their charge density is high. Some soil smectites have a higher charge density and likely have wedge sites near mica-like zones where K selectivity is high and K fixation can take place.

The principal mineral responsible for K fixation in acid soil is probably dioctahedral vermiculite. Counteracting the effect of high-charge density on K fixation in many acid soils is the presence of hydroxy-Al and –Fe(III) polymers in the interlayer space. The polymers act as props between the unit silicate layers and inhibit collapse of the layers about the K ions. The presence of hydroxy-interlayer polymers in vermiculite increases the Gapon coefficient K_G (K/Ca) from 5.7×10^{-2} to 11.1×10^{-2} l(mmol)$^{1/2}$. On the other hand, K fixation is reduced markedly.

Since the pioneer research on the effect of freezing and thawing on K fixation and release, the significance of wetting and drying and freezing and thawing has been recognized by numerous researchers. The degree of K fixation or release on wetting or drying depends on the type of colloids and the level of K^+ in the soil solution.

Potassium fixation in soil and clay systems is influenced by pH. The pH effect on K fixation is related to: (1) the concentration of H_3O^+ and the competing cations and (2) the extent of interlaying and the resultant charge neutralization and propping effect. A marked increase in K fixation in soils was observed when pH was raised to about 9 or 10 with Na_2CO_3. There was no fixation at pH values up to 2.5; the amount of K fixation increased very rapidly between pH 2.5 and 5.5. Fixation of K increased more slowly above pH 5.5. At low pH, the lack of K fixation is attributed to a high concentration of H_3O^+ and its ability to replace K^+. At pH >5.5, Al^{3+} cations precipitate as hydroxy polycations, which increase in the number of OH groups as pH increases until a discrete gibbsite phase is formed. At pH about 8, Al^{3+} does not neutralize the negative charge on the clay and cannot prevent K fixation. The increase in K fixation between pH 5.5 and 7.0 is attributed to the decreased number of hydroxy-interlayer material that decreases K fixation.

Potassium Release from Soils

Primary Minerals

Micas and feldspars are the main K reserves of soils (Figure 1). The release of K from micas and feldspars replenishes the K in the labile pool. Therefore, these K-bearing primary minerals have a significant impact on long-term soil K fertility.

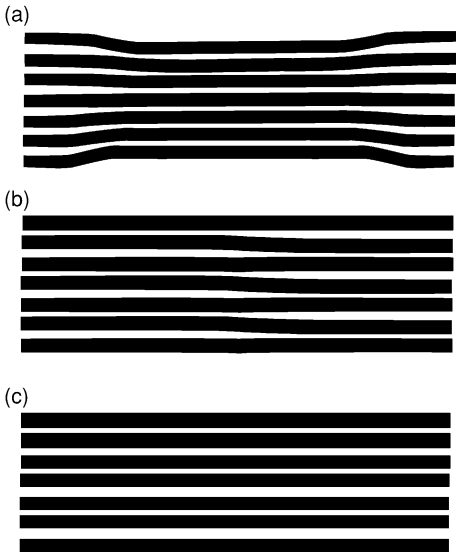


Figure 4 Diagram of a weathered mica particle containing about 50% expanded (vermiculite) layer. (a) 'Frayed edge' and mica core; (b) alternate layers open halfway through interlayer; and (c) regularly interstratified mica-vermiculite (no wedge zone). Reproduced from Rich CI (1972) Potassium in soil minerals. *Proceedings of the 9th Colloquium of the International Potash Institute* 9: 15–31.

Micas Potassium release from micas may proceed by two processes: the transformation of micas to expansible 2:1 layer silicates by exchanging the K^+ with hydrated cations and the dissolution of micas resulting in the formation of weathering products. Both types of reaction occur in soil environments.

The transformation of micas to 2:1 expansible layer silicates through K^+ release proceeds by first, edge weathering and second, layer weathering. These weathering mechanisms are depicted in Figure 4. In models (a) and (b), the expanded interlayers are not continuous but have internal terminations forming 'wedge zones.' In model (c), interlayer spaces are either expanded all the way through the particle or nonexpanded uniformly through each layer. Edge weathering is most common in larger mica particles; in contrast, layer weathering is most common in smaller particles.

The mineralogical factors influence the release of K^+ from micas by both cation exchange and dissolution reactions. These include tetrahedral rotation and cell dimensions, tetrahedral tilting, hydroxyl orientation, chemical composition, particle size, structural imperfections, and layer charge alterations and related reactions. Ionic factors also substantially influence the release of K from micas. The hydronium ion, H_3O^+, is one of the dominant ions in acidic soil solutions. Reaction rates involving the H^+ ion are not restricted by diffusion rates of H_3O^+. Therefore,

proton transfer may be rapid to the wedge zone at the junction of 'open' to 'closed micas.' The rate of H_3O^+–K^+ exchange may be limited by the rate of K^+ diffusion.

The sources of H_3O^+ in soils include water through its dissociation, the dissolution of CO_2 from the air and soil atmosphere, strong mineral acids originated from weathering reactions, and low-molecular-weight organic acids. Hydronium ions are probably the cause of major release of structural K^+ by weathering reactions. This mechanism is especially important in the soil rhizosphere, which is the most important plant feeding zone. Therefore, the extraction of soils with boiling 1 mol l^{-1} HNO_3, dilute HCl solutions, and other acid dissolution techniques have been proposed for measuring the K-supplying power of soils.

The activity of K^+ ions in the soil solution greatly influences the release of K^+ from micas by cation exchange. Therefore, the K release can be greatly driven by K^+ leaching and especially by cropping without K^+ fertilizer application. When the K level is less than the critical value, K^+ is released from the interlayer by other cations from the solution. On the other hand, when the K level is higher than the critical level, the mica-expansible 2:1 phyllosilicates adsorb K^+ from the solution. The critical K level is highly mineral-dependent, being much lower for muscovite than for trioctahedral micas. The nature and concentration of the replacing cations also influence the critical K level in the solution. For example, the critical K levels decrease in the order $Ba^{2+} > Mg^{2+} > Ca^{2+} \cong Sr^{2+}$ for the same concentration of these ions and with a constant mica particle.

Many organic and inorganic ligands in the soil solution also affect K release from minerals. The kinetics of K release from muscovite and biotite is greatly influenced by the nature of organic acids which are common in soil environments, especially in localized zones where biological activity is intense, such as in the rhizosphere and near decomposing plant residues. Besides organic ligands, some strong complexing inorganic ligands such as phosphate can influence the kinetics of K release from micas. In the immediate vicinity of the phosphate fertilizer zone, the concentration of phosphate can be as high as 2.9–4.5 mol l^{-1}. The rates of K release from muscovite, biotite, and illite are significantly enhanced by 1 mol l^{-1} $NH_4H_2PO_4$ (pH 4.0).

Redox potential of soils influences the K^+ release from micas because the interlayer K is more tenaciously held after oxidation of the structural Fe(II). Other factors being equal, the extent of K^+ release from micas appears to be less in oxidized soil environments then in reduced soil environments.

Potassium-bearing feldspars The release of K from feldspars through weathering reactions is an important process in replenishing the labile pool of soil K. The weathering rate of K-bearing feldspars appears to be affected by their structural properties which include the chemical composition, principally the inclusion of Na in the structure, the order of Si and Al in the structure, crystal and particle size, crystallization of more than one mineral within one particle, and the nature of twinning. The rate of weathering reactions of K-bearing feldspars probably increased with Si/A1 disorder and Na content. A perthitic feldspar would likely weather more rapidly than a homogeneous particle. Albite would be weathered more rapidly in perthitic structures than the K-feldspar. Weathering is also likely more rapid at the junction of twins because of disorder at these planes.

Hydronium ions are an important chemical weathering agent of K-bearing feldspars. The initial weathering of K-bearing feldspars proceeds by surface reactions in which K^+ is displaced by H_3O^+. The substitution of K^+ by H_3O^+ weakens the feldspar structure. The kinetics of dissolution of K^+ from K-bearing minerals by 1 mol l^{-1} HNO_3 indicates that, for the temperature range studied, the rates of K^+ release from muscovite, phlogopite, and biotite are about 2, 9–12, and 118 to 190 times, respectively, greater than the rate of K^+ release from microcline (Table 2). The release of K^+ from a tectosilicate is more difficult than from the interlayers of micas. The Arrhenius heat of activation for the release of K^+ from microcline is greater than for the micas. The heat of activation is interpreted as the energy level the structural K must acquire to be able to react. As the heat of activation increases, the rate constant decreases, since the mineral with higher heat of activation for the release of K^+ would release less K per unit time. The Arrhenius heat of activation and apparent rate constant for K^+ release from the K-bearing minerals (Table 2) are fundamental to understanding the availability of K^+ in feldspars and

Table 2 Apparent rate constants and Arrhenius heats of activation for the release of structural K from K minerals

	Rate constant		*Arrhenius heat of activation*
Mineral	*301 K (h^{-1})*	*311 K (h^{-1})*	*(kJ mol^{-1})*
Biotite	1.46×10^{-2}	3.09×10^{-2}	58.66
Phlogopite	9.01×10^{-4}	2.44×10^{-3}	77.81
Muscovite	1.39×10^{-4}	4.15×10^{-4}	85.52
Microcline	7.67×10^{-5}	2.63×10^{-4}	96.24

Reproduced from Huang PM, Crosson LS, and Rennie DA (1968) Chemical dynamics of K-release from potassium minerals common in soils. *Transactions of the 9th International Congress of Soil Science* 2: 705–712.

micas to plants. Although the release of K^+ from K-feldspars is generally more hindered than from micas, the K-feldspars are the largest natural reserves of K in many soils. Therefore, the role of K-feldspars in supplying K to plants cannot be ignored.

Besides H_3O^+ ions in soil solutions, complexing organic and inorganic ligands have significant roles in the release of K from the feldspar structure. Low-molecular-weight organic acids are very effective in releasing K^+ from K-bearing feldspars. Root growth and root length along with organic acids in the soil rhizosphere, thus, merit attention. In the immediate vicinity of the phosphate fertilizer zone, high concentrations of phosphate in the soil solution may also induce K release from K-feldspars.

Climate, topography, and the degree of leaching also influence the stability of K-bearing feldspars in soils. All experimental evidence shows that chemical breakdown of K-feldspars accelerates with raising temperature. Warm climate and high precipitation in tropical soils may have depleted all of their K-feldspars. However, these minerals may persist in temperate soils, especially in sand and silt fractions.

Natural Soils

Chemical methods have been used to extract nonexchangeable K of soils. These include boiling HNO_3, H_2SO_4, hot HCl, electroultrafiltration, Na tetraphenylboron with ethylenediaminetetraacetic acid (EDTA), and ion exchange resins such as H- and Ca-saturated resins.

Soil K released to Ca-resin can be divided into three simultaneous rates of K release from the surface of the soil complex, the weathered periphery, and the micaceous matrix. The diffusion coefficient (3×10^{-19} $cm^2\,s^{-1}$) calculated for the slowest process, which is suggested to be from the mineral matrix, is of the same magnitude as that for illitic clay minerals. Comparing the values of rate constants for K release from the soils with those for K release from minerals, the rate constants of K release from the soils are within the range expected for K release from the mixtures of micas and K-feldspars. The rate constants of K release from six Great Plain soils by successive extraction with Ca-saturated cation exchange resin are highly correlated with mica contents of the soils, cumulative K uptake, and cumulative alfalfa yield.

The release of nonexchangeable K from soils is not only from clay-size K-bearing minerals but also from micas and feldspars in silt and sand fractions. The release of K^+ by the H^+-saturated ion exchanger from these size fractions as a function of time is demonstrated in Figure 5, based on the Elovich equation. The variation in the slope of the Elovich equation correlates significantly with the concentration of micaceous minerals in the sand, silt, and clay fractions, and may have an important impact on K for plant uptake.

Few studies have investigated the effect of anions on the kinetics of K^+ release from soils. The amounts of K^+ released from the Oxisol, Alfisol, and Entisol in $1\,mol\,l^{-1}$ $NH_4H_2PO_4$ (pH 4.0) for 4 weeks are 4.77, 114.0, and $207.8\,mg\,kg^{-1}$, respectively (Figure 6). The results clearly indicate that the K release in $1\,mol\,l^{-1}$ $NH_4H_2PO_4$ at pH 4.0 from the three soils is 2.3–2.9 times higher than that in $1\,mol\,l^{-1}$ NH_4Cl at the same initial pH. Compared with the NH_4Cl systems, the excess K released from the soils in the $NH_4H_2PO_4$ systems is attributed to the phosphate-induced alteration of K-bearing minerals of the soils, since $1\,mol\,l^{-1}$ $NH_4H_2PO_4$ induces alteration of micas and feldspars and the subsequent K release. The amount of the $NH_4H_2PO_4$-induced K^+ release is related to the degree of weathering of the soils, namely, Oxisol < Alfisol < Entisol. The combined effect of phosphate and proton on the alteration of K-bearing minerals and the formation of NH_4-taranakite are the major mechanisms of the $NH_4H_2PO_4$-induced K^+ release from the soils. In the immediate vicinity of the fertilizer zone, the P concentration is very high and the pH is quite low. Therefore, application of $NH_4H_2PO_4$ fertilizer to soils would promote dissolution of K-bearing minerals from the soils, thus enhancing their K-supplying rate.

Availability of Soil Potassium

Quantity/Intensity Relationships

The ratio of the activity of cations such as K^+ and Ca^{2+} is defined by the relation $a_K/(a_{Ca})^{1/2}$ where a is the ion activity. The concepts of quantity (Q) and intensity (I) have been applied to the mineral nutrient status of soils. Based on the ratio law, the I of K in a soil at equilibrium with its solution could best be defined by the ratio $a_k/(a_{Ca} + a_{Mg})^{1/2}$ of the soil solution. This equilibrium activity ratio for K is referred to as AR^K and has often been used as a measure of K^+ availability. However, it has been proposed that the activity of K in the soil solution is of primary importance as it is relevant for K^+ diffusion and K translocation in soils by mass flow to roots.

Soils showing the same value of AR^K may not possess the same capacity for maintaining AR^K while K is removed by plant uptake. Therefore, one must include not only the current potential of K in the labile pool but also the way in which the potential depends on quantity of labile K present. These findings brought about the classic Q/I relationships where the activity ratio AR^K value is related to the change in

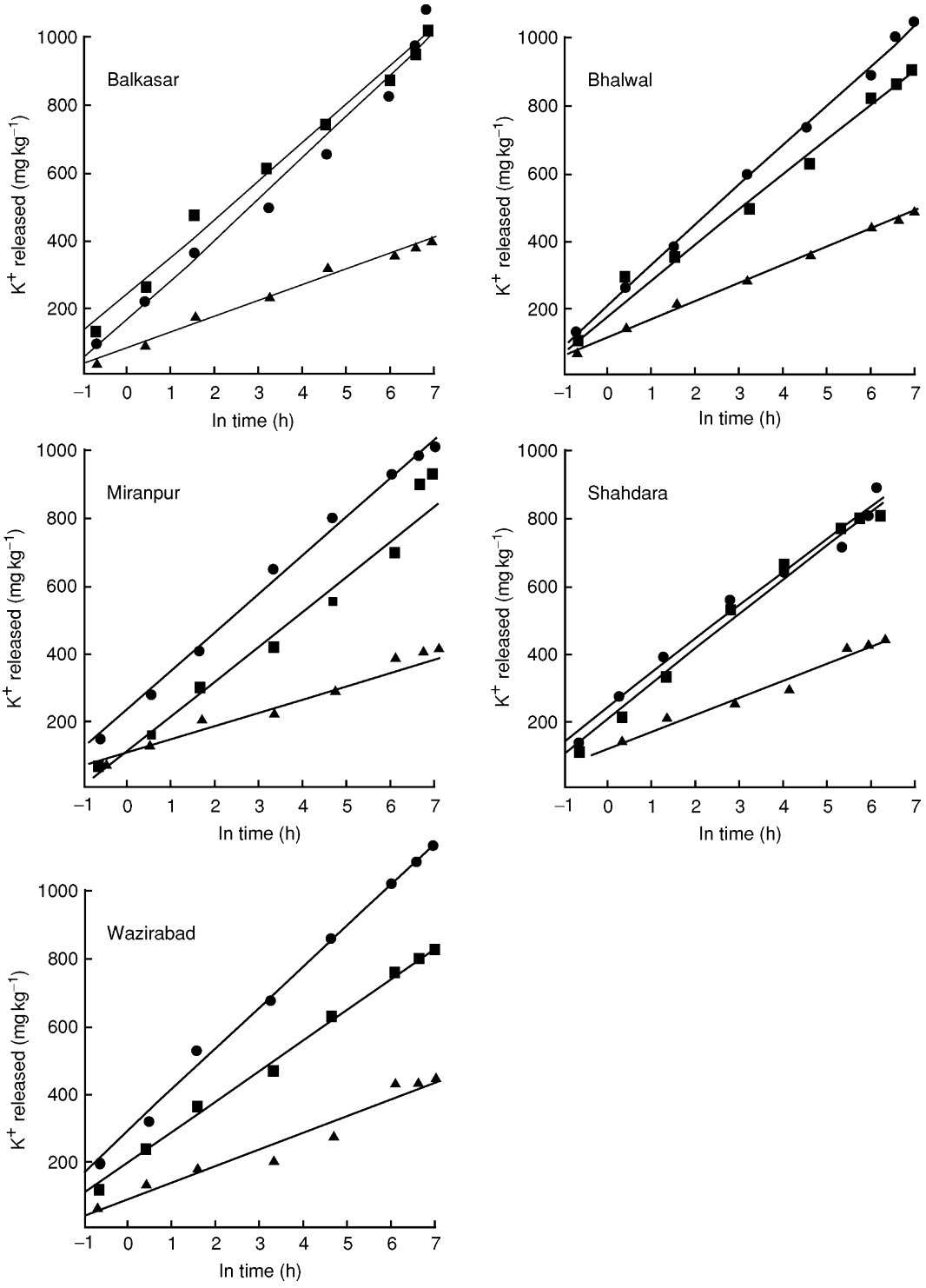


Figure 5 Release of nonexchangeable K^+ from sand (triangles), silt (squares), and clay (circles) fractions. Reproduced with permission from Rahmatullah and Mengel K (2000). Potassium release from mineral structure by H^+ ion resin. *Geoderma* 96: 291–305.

exchangeable K to obtain the effect of quantity of labile K (exchangeable K) on intensity of labile K (AR^K). The *Q/I* concept has been widely disseminated in the soil literature in understanding soil K status.

The method used for constructing a typical *Q/I* curve involves equilibrating a soil with solutions containing a constant amount of $CaCl_2$ and increasing the amount of KCl. The soil gains or loses K to achieve the characteristic AR^K of the soil or remains

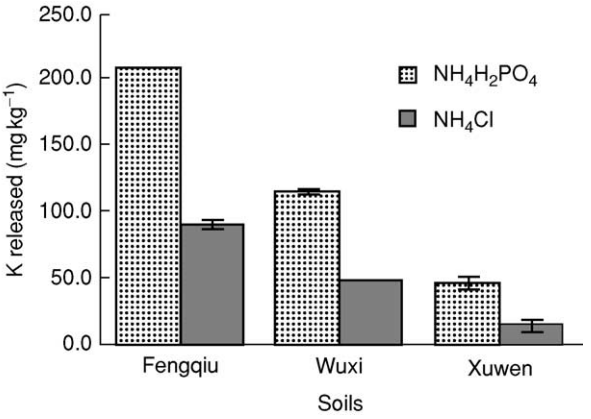


Figure 6 Potassium released from the soil in 1 mol l^{-1} $NH_4H_2PO_4$ and 1 mol l^{-1} NH_4Cl solutions (pH 4.0) at 25°C in 4 weeks. The Fenqiu, Wuxi, and Xuwen soils are Entisol, Alfisol, and Oxisol, respectively. The standard errors of the K release in $NH_4H_2PO_4$-Fengqiu and NH_4Cl-Wuxi soil systems were too small to be shown in the figure. Reproduced with permission from Zhou JM and Huang PM (1995) Kinetics of monoammonium phosphate-induced potassium release from selected soils. *Canadian Journal of Soil Science* 75: 197–203.

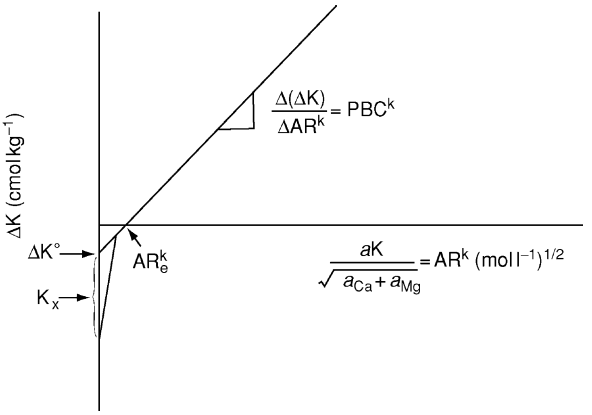


Figure 7 Typical quantity/intensity (*Q/I*) plot. Reproduced with permission from Sparks DL and Huang PM (1985) Physical chemistry of soil potassium. In: Munson RD (ed.) *Potassium in Agriculture*. Madison, WI: American Society of Agronomy, Crop Science Society of America, and Soil Science Society of America.

unchanged if its AR^K is the same as the equilibrating solution. The AR^K values are calculated from the measured concentrations of K^+, Ca^{2+}, and Mg^{2+} converted to their ionic activities by application of the extended Debye–Hückel equation. The AR^K values are then plotted versus the gain or loss of K to form the characteristic *Q/I* curve (Figure 7). Several parameters can be obtained from the *Q/I* plot to characterize soil K status. The value of AR^K when the *Q* factor (ΔK) equals zero is a measure of the degree of K availability at equilibrium or AR^K_e. The value of ΔK when AR^K equals zero is a measure of labile or exchangeable K in soils or $\Delta K°$. The slope of the linear portion of the curve gives the potential buffering capacity of K (PBC^K) and is proportional to the cation exchange capacity of the soil. The number of specific sites for K (K_x) is the difference between the intercept of the curve and the linear portion of the *Q/I* plot at AR^K equals 0. Advances in ion-selective electrode (ISE) technology have allowed for more rapid *Q/I* analysis.

Soils which have the same AR^K_e may not have the same capacity for maintaining AR^K_e while K is being uptaken by plant roots. Soils can have the same AR^K_e but contain different amounts of labile K. K fertilization, liming, and cropping can change AR^K_e values.

The $\Delta K°$ values are a better estimate of soil labile K than normal exchangeable K. Higher values of labile K ($-\Delta K°$) indicate a greater K^+ release into the soil solution, resulting in a larger pool of labile K. The $\Delta K°$ becomes more negative or the labile K pool increases with K fertilization and liming on cropped soil.

The PBC^K indicates the ability of a soil to maintain the intensity of K^+ in the soil solution. A high PBC^K value is indicative of constant availability of K^+ in the soil over a long period. By contrast, a low PBC^K value would indicate the need for frequent K fertilization. However, the lack of response of a Kalmia soil in the Delaware coastal plain in USA, which had a low PCB^K value, to applied K was observed. This was attributed to the presence of the large amount of K-bearing minerals in the soil that with time could release K^+ to exchangeable and to solution form. Cropping and liming can change the PBC^K values of soils.

The Rhizosphere Chemistry and Labile Pool of Soil Potassium

The soil rhizosphere is the narrow zone of soil influenced by the root and exudates. The extent of the rhizosphere may vary with soil type, plant species, age, and many other factors, but it is usually considered to extend from the root surface out into the soil for a few millimeters. Microbial population in the rhizosphere can be 10–100 times larger than the populations in bulk soil. The rhizosphere is, thus, bathed in root exudates and microbial metabolites. Therefore, chemistry and bioavailability of soil K in the rhizosphere should differ from those in bulk soil.

A wide variety and considerable amount of organic compounds are released from plant roots in the rhizosphere. On average, between 30 and 60% of the net photosynthetic C is allocated to the roots in annual species; an appreciable portion (4–70%) of this C is released as organic compounds into the rhizosphere. The major components of rhizodeposition

are sloughed-off cells and their lysates, mucilage, and low-molecular-weight organic compounds. Both the amount and proportion of the various organic compounds of root exudates vary substantially with plant species and cultivars. Moreover, the same plant cultivar grown in different soils varies in the kind and amount of low-molecular-weight organic acids present in the rhizosphere. Biological activity promotes K^+ release from micas. Besides the depletion of the K of the soil solution by plant roots, the overall action of biological activity is more complex when organic acids are produced. Organic acids substantially influence the dynamics of K^+ release from K-bearing minerals. The effect of organic acids in enhancing the release of K is attributed to the dissociated H^+ ions and complexing organic ligands. The weakening of the metal–O bonds by protonation of surface OH groups and the formation of inner-sphere surface complexes with organic ligands certainly enhance mineral dissolution and K release.

In view of the impact of plant species and cultivars and their interactions with soil properties on the formation of the kind and amount of organic acids and the influence of organic acids on the dynamics of K^+ release from the K-bearing minerals common in soils, the role of organic acids in the rhizosphere in affecting the labile pool of soil K merits close attention.

Assessing Potassium Availability by Soil Tests

Most soils have a relatively large quantity of total K but relatively small amounts of plant-available K. Potassium soluble in water plus K adsorbed to exchange sites largely accounts for the labile pool of soil K. The exchangeable K is readily released to solution by the equilibrium process. During the course of a growing season, some initially nonexchangeable K may become water-soluble or exchangeable, and thus be plant-available.

The soil test methods for plant-available K are generally those that measure the soluble and exchangeable K or some parameter that relates to their equilibrium state. Soil test extractants for K are used to measure K easily and rapidly in soils and to estimate K availability. Based on the amounts of extractable K, recommendations based on field test calibrations can then be made on the amount of K required to obtain maximum yields.

Although soil tests for K usually measure the quantity of solution and exchangeable K, acids are usually used as extractants. Therefore, some nonexchangeable and structural K is also extracted. There are a number of soil test methods that are used to evaluate K availability of soils. The suitability of these soil test methods varies with soil types. The $1\,mol\,l^{-1}$ NH_4OAc at pH 7 procedure is probably the most widely used of all soil test methods for measuring so-called plant-available K on a global scale. Water-soluble K^+ or soil solution K^+ is such a small fraction of the total extracted by this method that it is not measured separately but is included along with that extracted (soluble and exchangeable). Both the K^+ in solution and that in the nonexchangeable form are in dynamic equilibrium with the exchangeable form. Therefore, the exchangeable K is sometimes referred to as a moving target.

Several other methods of determining exchangeable K^+ in soils have been developed for various soil conditions or for various methodological expediences. The Mehlich no. 1 (double acid) method ($0.05\,mol\,l^{-1}$ HCl in $0.025\,mol\,l^{-1}$ H_2SO_4) is considered best for sandy, acid, low CEC soils. The Mehlich no. 2 method ($0.2\,mol\,l^{-1}$ HOAc in $0.015\,mol\,l^{-1}$ NH_4F in $0.2\,mol\,l^{-1}$ NH_4Cl in $0.012\,mol\,l^{-1}$ HCl) is considered good for a wide range of soil conditions. The ammonium bicarbonate-diethylenetriamine pentaacetic acid (AB-DTPA) method is considered best for alkaline conditions. The Bray P-1 extractant is also used in some soil-testing laboratories to measure exchangeable K^+.

The soil test methods currently used in the USA are Mehlich no. 1 and Mehlich no. 3 procedures in the north-eastern and south-eastern regions, the Morgan and modified Morgan procedures in parts of the north-eastern region, and the $1\,mol\,l^{-1}$ NH_4OAc at pH 7 procedure in the western region.

For the interpretation of soil tests, it would be useful to know which K^+ quantity originates from the hydrated K^+ fraction and which from the nonhydrated K^+ fraction. The routine method which distinguishes between these two fractions is electroultrafiltration (EUF), which is used in Germany and Austria on a broad scale for K^+ fertilizer recommendation.

Soils which have long-term K-supplying power are attributed to the presence of sufficient K-bearing minerals (micas and K-feldspars). There are soils in the Punjab and Europe which hardly respond to K fertilization in long-term fertilizer trials. These soils release much K^+ from the silt fraction (Figure 5). Such soils differ much from sandy soils in their K fertilizer strategy.

Plant Analysis for Potassium

Plant K analysis is an approach to assess soil K status. This method is based on the principle that if an element such as K is deficient in a plant, growth rates and yields will be decreased. If increasing amounts of K are added, the K concentration in the plant or plant part increases until an optimum level is

Table 3 Critical K concentrations in agronomic crops

Crop	*Time of sampling*	*Plant part*	*Critical concentration*[a] *(g kg^{-1})*
Sugarbeet		Blade	10
Cotton		Leaves	<9–15
Wheat	Jointing (GS6)	Total tops	20–25
	Early boot (GS9)	Total tops	15–20
Rice	Flag leaf to mid-tillering		10–14
Corn	At tassel	Ear leaf	19
	At tassel	Leaves	17–27
	At silk	Sixth leaf from base	13
	At silk	Leaf opposite and just below ear shoot	20
Grain sorghum	Full heading	Second blade below apex	18
Alfalfa		Whole top	8–22
Red clover		Tops	15–22.5
Bermudagrass		Tops	13–15
Orchardgrass		Tops	23–25
Tall fescue		Tops	24–38
Kentucky bluegrass		Tops	16–20

[a]Critical concentration is that nutrient concentration at which plant growth begins to decrease in comparison with plants above the critical concentration. After Westerman RL (1990). Reproduced from Sparks DL (2000) Bioavailability of soil potassium. In: Sumner ME (ed.) *Handbook of Soil Science*. Boca Raton, FL: CRC Press.

attained. This approach is used to express growth and yield as a percentage of the maximum. The zone between the deficient and optimum concentration is referred to as the transition zone. The transition zone is the zone between nutrient concentrations that cause a 20% reduction in growth or yield and those that produce optimum or 100% in the maximum yield. This latter zone is referred to as the critical range where researchers select the yield reduction and nutrient concentrations that are acceptable. The critical K concentrations for various agronomic crops are shown in Table 3. These critical K concentrations are determined by relating yield (percentage of maximum yield or growth rate) to K concentration ($g\,kg^{-1}$) for a specific plant part sampled at a given stage of development.

See also: **Fertigation**; **Fertility**; **Fertilizers and Fertilization**; **Macronutrients**

Further Reading

Huang PM (1987) Aluminum and the fate of nutrients and toxic substances in terrestrial and aquatic environments. In: Singh M (ed.) *Systems and Control Encyclopedia*, pp. 262–268. Oxford: Pergamon Press.

Huang PM (1989) Feldspars, olvines, pyroxenes, and amphiboles. In: Dixon JB and Weed SB (eds) *Minerals in Soil Environments*, 2nd edn, pp. 975–1050. Madison, WI: Soil Science Society of America.

Huang PM, Crosson LS, and Rennie DA (1968) Chemical dynamics of K-release from potassium minerals common in soils. *Transactions of the 9th International Congress of Soil Science* 2: 705–712.

Jackson ML (1963) Interlayering of expansible layer silicates in soils by chemical weathering. *Clays and Clay Minerals* 11: 29–46.

Marschner H (1998) Soil–root interface: biological and biochemical processes. In: Huang PM, Adriano DC, Logan TJ, and Checkai RT (eds) *Soil Chemistry and Ecosystem Health*, pp. 191–231. Madison, WI: Soil Science Society of America.

McClean EO and Watson ME (1985) Soil measurements of plant-available potassium. In: Munson RD (ed.) *Potassium in Agriculture*, pp. 278–308. Madison, WI: American Society of Agronomy, Crop Science Society of America, and Soil Science Society of America.

Mengel K and Kirkby EA (2001) *Principle of Plant Nutrition*, 5th edn. Dordrecht: Kluwer Academic.

Rahmatullah and Mengel K (2000) Potassium release from mineral structures by H^+ ion resin. *Geoderma* 96: 291–305.

Rich CI (1968) Mineralogy of soil potassium. In: Kilmer VJ, Younts SE, and Brady NC (eds) *The Role of Potassium in Agriculture*, pp. 79–108. Madison, WI: American Society of Agronomy, Crop Science Society of America, and Soil Science Society of America.

Rich CI (1972) Potassium in soil minerals. *Proceedings of the 9th Colloquium of the International Potash Institute* 9: 15–31.

Scott AD (1968) Effect of particle size on interlayer potassium exchange in micas. *Transactions of the 9th International Congress of Soil Science* 2: 649–660.

Somasiri S, Lee SY, and Huang PM (1971) Influence of certain pedogenic factors on potassium reserves of selected Canadian Prairie soils. *Soil Science Society of America Proceedings* 35: 500–505.

Song SK and Huang PM (1988) Dynamics of potassium release from potassium-bearing minerals as influenced

by oxalic and citric acids. *Soil Science Society of America Journal* 52: 383–390.
Sparks DL (2000) Bioavailability of soil potassium. In: Sumner ME (ed.) *Handbook of Soil Science*, pp. D38–D53. Boca Raton, FL: CRC Press.
Sparks DL and Huang PM (1985) Physical chemistry of soil potassium. In: Munson RD (ed.) *Potassium in Agriculture*, pp. 201–276. Madison, WI: American Society of Agronomy, Crop Science Society of America, and Soil Science Society of America.
Westerman RL (ed.) (1990) *Soil Testing and Plant Analysis*. Madison, WI: Soil Science Society of America.
Zhou JM and Huang PM (1995) Kinetics of monoammonium phosphate-induced potassium release from selected soils. *Canadian Journal of Soil Science* 75: 197–203.

PRECIPITATION, WATERSHED ANALYSIS

J V Bonta, USDA Agricultural Research Service, Coshocton, OH, USA

Published by Elsevier Ltd.

Introduction

Variability in current and historic weather and long-term climate, acting on spatially variable watershed characteristics, is the driving force that causes observed variability in stream flows, sediment transport, and water quality. Watershed characteristics such as spatial distribution of soil-water storage, hydraulic conductivity, tension characteristics, subsurface flow, and drainage affect flow, erosion, and water-quality sources and transport in a watershed. Weather varies significantly with time at a point on the landscape. Climate and weather also vary spatially over a watershed from one area to another and with elevation, and temporally from season to season.

The most obvious component of weather that affects watershed variability is precipitation, which has both deterministic and random components. Errors in measurement of precipitation, streamflow and other environmental variables, temporal and spatial variability and inadequate quantification of processes of water, sediment, and chemical movement in watersheds cause uncertainties in estimating the impacts of land-management changes on hydrology and water quality. This article summarizes current understanding of watershed processes, precipitation characterization and measurement, and watershed, weather, and precipitation modeling.

Measurement of Precipitation

Precipitation measurements and characterization have inherent inaccuracies that must be recognized before they are used in computer models and in other analyses. Accurate precipitation measurement is complicated because of its significant spatial and temporal variability, and because of its different forms (liquid, solid, mixture). There are three components to precipitation measurement – sensing, representativeness, and resolution. Precipitation sensing involves using a sensor to measure a control volume of atmospheric particles presented to the sensor. Representativeness refers to how accurately the sensed particles represent the volume of precipitation actually reaching the ground, which is important to evaluating a watershed's response to precipitation. Resolution refers to how often (time resolution – minute, hourly, daily) precipitation is measured, and to the depth increment measured (depth resolution – 0.1 mm, 1.0 mm, etc.).

Precipitation measurements must be accurate and representative because watershed model studies have shown that small changes in precipitation can cause large changes in watershed response, highlighting the nonlinearity of watershed systems. Few multidecadal records of representative short-time increment precipitation data with good spatial coverage exist. Synthesized precipitation data are often unrepresentative because precipitation 'data' synthesized by precipitation models are parameterized using the nonrepresentative raw precipitation data. Much has been learned from precipitation-measurement studies that can provide guidance for future precipitation measurements.

Precipitation Sensing and Resolution

A storage gauge is typically a cylindrical container with the orifice about 1 m off the ground. The depth of precipitation is often recorded every 24 h.

A recording weighing gauge (Figure 1a) weighs the precipitation that enters the orifice, and records the weight with a strip chart or data logger. The weighing gauge gives the best measurement of precipitation because the data can be collected at any time or depth resolution (e.g., 1 min and 0.25 mm); there is no limit on high intensities. Also, beginning and ending times of bursts of precipitation can be determined.

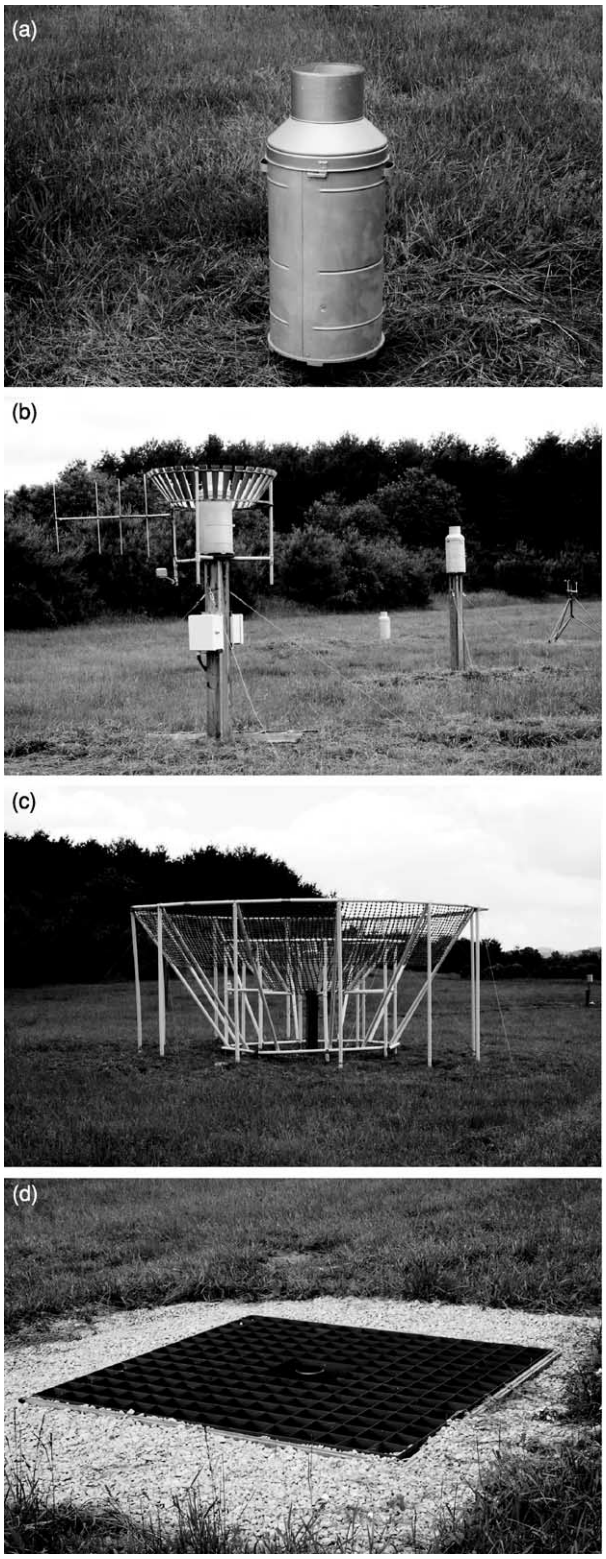


Figure 1 Types of precipitation gauges and shield configurations at the North Appalachian Experimental Watershed at Coshocton, Ohio: (a) standard gauge; (b) dual-gauge system; (c) Wyoming shield; and (d) pit gauge.

In tipping-bucket gauges, water from the orifice is funneled to a small hole where it drops into one side of a small 'bucket' that fills to a known volume before it empties its contents (e.g., 0.1 mm). This device is often used because tips are easily recorded as pulses with a data logger. However, the gauge must be calibrated, because at large rainfall intensities the system becomes overwhelmed and the volume of water from each tip of the bucket is less than the volume of a bucket. Therefore, long-term precipitation records can become truncated at larger intensities, often the ones of interest in watershed studies. Ponding above the buckets under high rainfall rates, and snow measurements can be problems. If heaters are used to melt frozen precipitation, they cause lower measurements because of evaporation and convection currents. Accurate beginning and ending times of storms are unknown with these gauges because water must accumulate in a bucket before it tips.

Precipitation-Measurement Representativeness

Standard precipitation gauges (Figure 1a) typically underestimate true ground-level precipitation, the measurement of interest in watershed modeling and analysis. This is because wind diverts many smaller-diameter rain drops over the top of the gauge orifice. Light-density snow flakes are particularly vulnerable to transport by the wind, bypassing the orifice. Lysimeters (large isolated blocks of soil that act as a ground-level gauge) at the North Appalachian Experimental Watershed (NAEW) at Coshocton, OH, on average measured 9% more annual precipitation and 27% more snow and mixed precipitation than an adjacent recording gauge. Other studies of wind effects on precipitation-gauge catch found that a standard gauge can measure as much as 60% less snowfall than measured at ground level on windy days.

Many studies have shown that shielding gauge orifices increases the precipitation catch. A ring of freely hanging baffles surrounding a gauge orifice (Alter shield) has been shown to be about 75% effective in measuring ground-level precipitation (depending on wind speed). A Wyoming gauge shield (Figure 1c) composed of a standard gauge at the center of two concentric wind shields made of snow fence, and a dual-gauge system (Figure 1b) composed of two elevated standard gauges (one shielded and one unshielded) have been shown to give representative ground-level precipitation measurements.

A pit gauge (Figure 1d) is considered to be the standard measurement system for rainfall. A coniferous forest opening is considered the best place to measure snow. A Tretyakov gauge composed

of dual concentric octagonal snow fences surrounding a gauge is an alternative standard for measuring snow because forest openings are often not available. A round inverted bell-shaped shield (Nipher shield) that surrounds a gauge gives good measurements of snowfall.

Representativeness is also affected by terrain slope. Horizontal orifice measurements can be much different compared with measurements from orifices parallel to the slope. Another factor affecting representativeness is obstacles in the vicinity of precipitation gauges.

Representativeness is also important over large areas as discussed later.

Characterization of Precipitation

Even though precipitation is highly variable across both small and large areas, from season to season, and from storm to storm, hydrologists have developed methods to quantify precipitation for use in models for watershed analysis. Characterization of precipitation involves spatial and temporal components important for watershed evaluations. Recent research has focused on stochastic simulation methods to generate long records of precipitation with similar statistical characteristics for the location of interest. Practical characterization of spatial variability for modeling purposes is a current research topic.

Four elements of a precipitation record require characterization for different purposes: (1) dry time between storms (TBS); (2) storm duration; (3) storm depth; and (4) within-storm intensity variation and corresponding mass curves.

Dry Times Between Storms

In any precipitation record (Figure 2), bursts (B_i) of varying intensities of precipitation are surrounded by dry periods (D_i). These TBS periods can range from a minute to months. A question facing an analyst is: what is the minimum dry-period duration (MDPD) for which $D_i > \text{MDPD}$ separates storms, and $D_i < \text{MDPD}$ is included in a storm. Intuitively, a 5-min period of no rain most likely will be included in the same storm, but a 5-month period would not. Quantification of MDPD is the first step in analyzing a record of precipitation data because storms can be identified, providing a data set for analysis of the other three elements. MDPD can be found by analyzing precipitation data independent of, or dependent on, watershed characteristics.

Arbitrary fixed-storm separation techniques are often used to identify storms. However, a fixed MDPD does not capture the spatial and monthly characterization of dry periods required for some

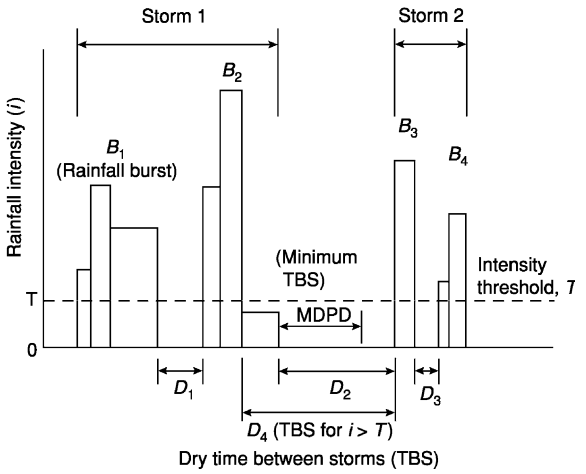


Figure 2 Definition sketch for identifying storms in a precipitation record. See text for details

analyses. Other specific storm-identification methods can be assigned, such as used in determining the 'R factor' (rainfall-runoff erosivity factor) of the Revised Universal Soil Loss Equation (RUSLE), a model used extensively for estimating soil loss on hill slopes (*See* **Erosion:** Water-Induced).

Rank correlation uses equally spaced time intervals of rainfall intensities that are lagged by increasing time intervals until time intervals are not correlated. The rank correlation method uses the rainfall intensity data to determine the dry time between storms instead of using the dry periods. Autocorrelation analysis has also been applied but the procedure does not always give well-defined estimates of MDPD.

Another method of determining MDPD is to assign a fixed time period for rainfall intensities greater than a threshold intensity, such as T in Figure 2. In this case, D_2 would be replaced by D_4. T represents an unknown parameter that is likely to be seasonally and spatially dependent, and interpretation must be qualified.

Another technique is called the 'exponential method.' In this method, MDPD is obtained by approximating groups of bursts of rainfall and dry periods (storms) as a Poisson process. Under this assumption, storms are considered statistically independent 'events' and the interarrival times of events are described by the exponential frequency distribution. In this method, dry periods greater than MDPD form an exponential distribution, and dry periods less than MDPD are included in storms. The exponential distribution is a good method to represent TBS, as it characterizes seasonal variation of TBS. Examples of best-fit exponential frequency distributions for several months at Bridgeport, NE, are shown in Figure 3.

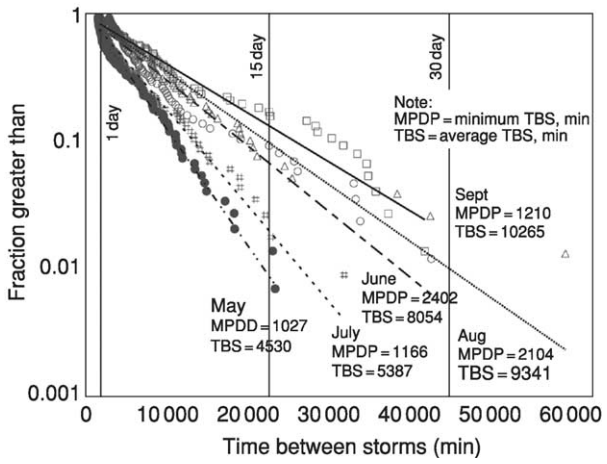


Figure 3 Seasonal variation of exponential frequency distributions of times between storms (TBS) for Bridgeport, Nebraska, USA. —●· May; —○— June; - # - July; ···△··· August; —□— September.

The average TBS is the parameter of the exponential distribution.

Finally, some precipitation analyses have not used the concept of an MDPD value, but used only daily precipitation totals. For example, during the development of the watershed curve number (CN) technique, used worldwide for estimating watershed runoff potential, daily (and not storm) precipitation was used because these were the most common data available from a storage-gauge network. Recent investigations of the CN technique now match actual storms with observed runoff data for curve-number development.

Most methods of determining MDPD assume that storms occur independently from the watershed. However, as in the case of determining precipitation that causes observed runoff hydrographs used in determining a watershed CN, identifying storms that incorporate watershed characteristics will give different results. This is because identifying precipitation that causes runoff events includes watershed storage that attenuates watershed response to rainfall intensities. Increasing watershed size increases storage effects, and spatial variability of precipitation also becomes more important. One or more storms (or partial storms) identified using the exponential method may comprise the causal storm for a watershed hydrograph. The method used for identifying causal precipitation for CN analysis incorporates the threshold T value in Figure 2.

Stochastic simulation of the occurrence of individual storms (alternative to using 24-h total precipitation) is possible by sampling the appropriate monthly exponential frequency distribution of TBS (Figure 3) using the Monte Carlo technique.

Storm Depths and Durations

Storm depths and durations are usually analyzed together. Often daily total precipitation has been used, such as in the development of the CN method, and individual storm durations are ignored. However, times of observation of 24-h totals may be inconsistent, because precipitation may not necessarily have been recorded over the same 24-h period. Yet precipitation analyses often assume 24-h totals at midnight. This inconsistency has led to the approach of identifying individual storms using the exponential method.

One form of depth and duration data characterization is through intensity–duration–frequency curves (IDFs). IDFs summarize the return periods of different intensity rainfalls for a given duration (e.g., 5-year, 30-min intensity). IDFs are widely used in engineering design of culverts and sewerage systems through peak-runoff estimation methods. They are also used in engineering design with synthetic mass curves of rainfall. IDFs represent only the maximum intensities in a storm, with no information on time distribution of rainfall. Maps of isopleths of IDF curves are often available. An assumption made when using IDF curves is that the return periods of precipitation and peak-runoff rate are the same.

Newer precipitation-analysis approaches eliminate the 24-h constraint and use storms as described above using the exponential method. Storm durations can range from a few minutes to several days. Once storms have been identified by monthly MDPD values, storm depth and durations can be analyzed for particular purposes. Storm depths and durations are typically correlated (Figure 4); however, there is much scatter. Storm simulation involves sampling from graphs like Figure 4, using the Monte Carlo method.

Within-Storm Intensities and Mass Curves

There is an infinite number of ways that a storm total can be distributed, as evidenced by the sample of 181 storm mass curves at Coshocton, OH (Figure 5). Historically, the use of only daily precipitation totals has led to simple methods for disaggregating the 24-h or IDF total to determine the time distribution of precipitation. Fixed patterns of rainfall intensities ('design storms') have been used to distribute the total in time (Figure 6). The USDA Natural Resources Conservation Service often uses design storms for determining peak flow rates in design (type I, II, etc. curves). However, fixed patterns do not describe the variability found in nature (Figure 5).

The most satisfactory mass-curve representation has been found to be Huff curves (Figure 6d) because the wide variability of curves is similar to that

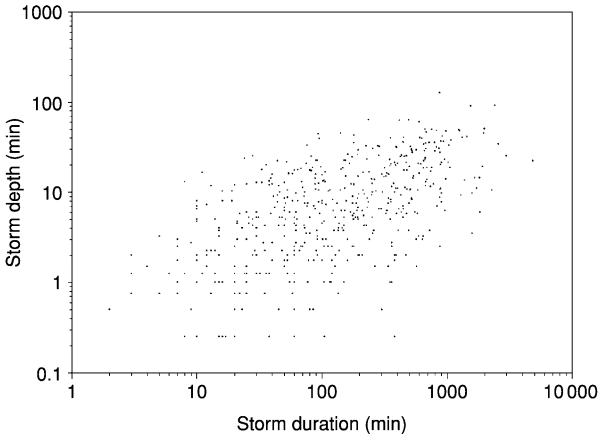


Figure 4 Example of the correlation between storm depths and durations for July at the North Appalachian Experimental Watershed at Coshocton, Ohio, USA.

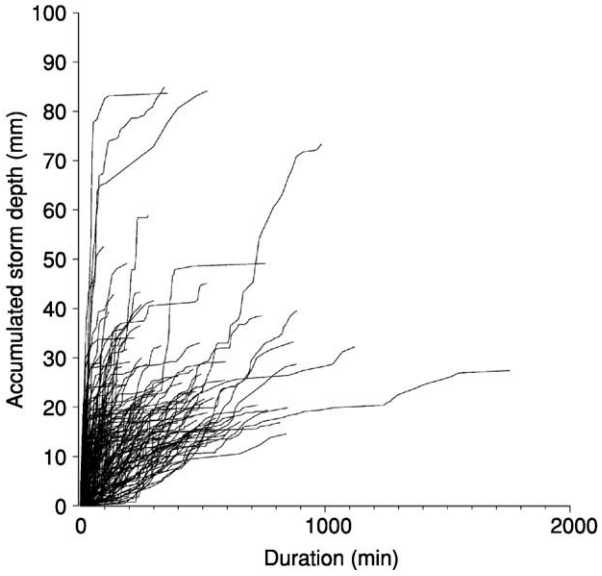


Figure 5 Example of the wide variability in mass curves at the North Appalachian Experimental Watershed at Coshocton, Ohio, USA.

observed in natural precipitation. These are isopleths of the frequency of occurrence of dimensionless accumulated depths for given dimensionless elapsed times in a storm. There are several curves that can be used, compared with a single mass curve of precipitation in a design storm. Documented methods to use the information contained in Huff curves in watershed analysis are an unresolved issue.

Spatial Variability

Spatial variability of storm depth and duration are often linked by snapshots of storm mass curves at the same time by using maps showing isopleths of

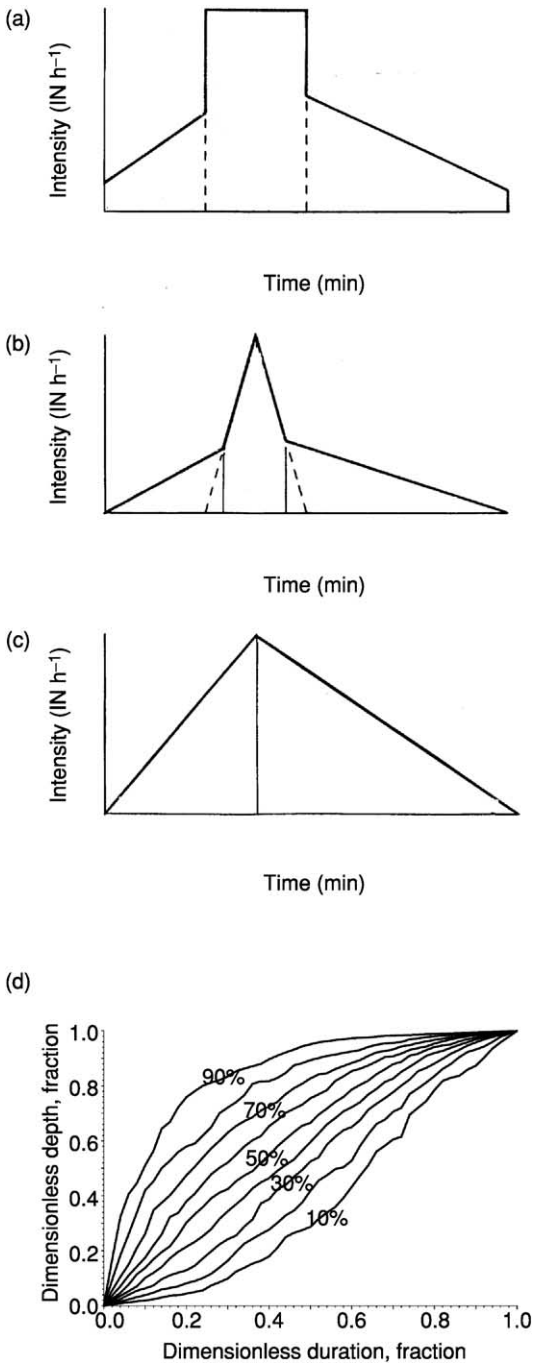


Figure 6 Design storms used for distributing total precipitation (a) mixed trapezoids (Sifalda V (1973) Entwicklung eines Berechnungsregens für die Bemessung von Kanalnetzen. *Wasser Abwasser* 114: 435–440); (b) mixed triangles (Desbordes M and Raous P 1976) Un exemple de l'intérêt des études de sensibilité des modèles hydrologiques. *La Houille Blanche* 1: 37–43); (c) triangular (Yen BC and Chow VT (1980) Design hyetographs for small drainage-structures. *Journal of Hydraulics Division* 106: 1055–1076); and (d) Huff curves at a point (Bonta JV (1997) Proposed use of Huff curves for hyetograph characterization. In: Richardson CW *et al.* (eds). *Proceedings of Workshop on Climate and Weather Research*, pp. 111–124. USDA–Agricultural Research Service 1996–03.

precipitation depth from several rain gauges during the storm (isohyetal maps). Isohyet interpolation can be by linear interpolation between known precipitation amounts, inverse-squared mapping, kriging, and by other methods. Average precipitation is computed by weighting isohyets by the area between isohyets. Thiessen polygons have also been used to capture the spatial variability of precipitation. These are polygons surrounding rain gauges such that the area enclosed by the polygons is closer to the enclosed rain gauge than to adjacent rain gauges. Polygon boundaries are the perpendicular bisectors between adjacent gauges. The polygons provide area-based weights for measured precipitation amounts.

Mapping of precipitation only grossly characterizes the spatial variability of precipitation. This is because precipitation distribution over an area can vary rapidly over time and with elevation in unknown ways over rolling hills and large mountains. Furthermore, precipitation depths can vary on the windward and leeward sides of mountains (i.e., 'rain shadow' effect). The effects of these factors can change monthly. Objective forms of mapping by interpolation do not consider orographic influences, and in mountainous areas, isohyets must often be adjusted manually.

One method of spatial characterization of precipitation that overcomes the problems of spatial variation with elevation is the use of conditional regression. This method of mapping utilizes regression relationships developed from precipitation data, and accounts for elevational, coastal, and other factors known to affect precipitation. Average monthly and annual precipitation have been mapped on a 4-km^2 grid across the USA. These maps are useful for general characterization, exploring relationships between average monthly weather elements and other variables (e.g., TBS), and mapping parameters for weather simulation models.

Watershed Response to Precipitation due to Physical Characteristics

Precipitation and Soil Variability

Watersheds that have similar appearance and land use respond differently to precipitation because of spatial and temporal variability of soil characteristics. Even watersheds having soils mapped in the same soil-map unit from an available soil survey and with the same land management can have significant differences in annual runoff. This is caused by natural spatial variability of soils and complex interactions between precipitation and surface–subsurface water. Studies in experimental watersheds attribute runoff differences to factors such as heavier-textured soil in different parts of the watershed, small seeps on the hillslope due to geologic clay layers that drain perched water to the seeps, and persistently high soil moisture in different areas. Heavier-textured soils and wetter watershed surfaces tend to run off sooner during the higher-intensity parts of storms than coarser-textured and drier areas, and runoff generation during a storm is nonuniform in the watershed. A local soil-survey map cannot necessarily be used to characterize small watersheds adequately because these factors are not typically mapped.

Watersheds exhibit complex interactions between surface and subsurface flow paths. Spatial and temporal variability of precipitation superimposed on watershed variability is an important factor to consider in understanding and modeling watershed hydrology. As watersheds become larger, more diversity in watershed characteristics can be expected.

Watershed Processes

Infiltration at the land surface significantly affects the timing and amounts of high and low watershed flows. Infiltration during rainfall events is affected by soil type, porosity, texture, vegetative cover, land management, preferential flow in soil and bedrock due to roots, cracks, and soil fauna. Macropores in natural systems can contribute significant amounts of infiltrated water and pollutants quickly to shallow and deep waters under high-intensity rainfalls. Advanced infiltration models require short-time increment precipitation to simulate matrix and macropore flow.

Land management can significantly affect infiltration. For example, rainfall on a field plowed with a moldboard plow tends to cause development of a relatively impervious surface due to crusting, whereas surface residue protects the soil surface from crusting with no-till cropping.

Infiltration is also affected by the degree of saturation since the last rainfall event (TBS, described above) as influenced by evapotranspiration (ET) and drainage rates. ET and drainage between storm events vary with soil type, vegetative cover, position on the landscape, aspect, geology, land use, climate, and time and weather since the last rainfall (TBS). The rate of water loss in the soil affects the storage available for water at the beginning of the next rainfall event ('antecedent soil water conditions').

In some watersheds, water moves horizontally near the surface as 'interflow.' The interflow process stores storm water for slow release either during a storm or after a storm has ended. It can also contribute to wetter antecedent conditions. Its effect on runoff, water quality, and erosion depends on the time until rainfall starts

again (TBS) and drainage and ET rates. In addition, interflow can re-emerge at the land surface, causing localized seeps (this is known as 'exfiltration') that immediately generates runoff during a storm.

Interflow, groundwater flow, and spatial variability of soil characteristics affect antecedent water conditions nonuniformly on a watershed surface. Typically, higher soil-water levels can be found in concave parts of the hillslope and along the periphery of stream channels. This is because soil water accumulates by gravity in these low areas during and between storm events. This leads to areas of watersheds that are wetter than others and that can run off sooner in a storm compared with drier areas. However, runoff can be randomly generated from any part of the landscape due to local factors. The nonuniformity of antecedent conditions leads to areas of the watershed that expand as areas become wetter during a storm, leading to a 'variable source area' of runoff generation. Runoff from these varying areas is sensitive to changing rainfall intensities during storms, affecting timing and amounts of runoff, chemicals, and sediment transported.

Watershed and Weather Modeling

Watershed–Precipitation Interactions

Watershed models are one tool for investigating the impacts of alternative land-management practices on hydrology, water quality, and sediment transport. Lack of data and simplifications of watershed processes are necessary. Often some processes are eliminated in modeling because of lack of understanding of these processes.

Few models adequately include the important processes of preferential movement of water as infiltration, subsurface flow, and interflow. The observed effects of scale of watershed on water yield must include a subsurface component to link upland and lowland areas quantitatively. Inadequate modeling of these processes and surface–subsurface precipitation interactions lead to errors in hydrology, erosion, and water-quality modeling.

Few models include spatial variability of infiltration and soil-water storage, except through land-surface discretization methods. These methods include representations of the land surface as square grids, triangles, and irregularly shaped areas. All these methods typically assume uniformity within the discretized element. Even within very small watersheds, however, there is significant variability in the distribution of soil characteristics. When small watersheds are subjected to precipitation, the spatial variability of characteristics can have significant effects on watershed hydrology and water quality. Sometimes a small area of a watershed may contribute a large fraction of sediment or chemicals downstream because it is wetter and tends to runoff sooner under rainfall, or it may receive anthropogenic inputs (e.g., fertilizer, vulnerable land-management practices). Another method for characterizing watershed elements is by using frequency distributions of infiltration characteristics within a 'uniform' element. Knowledge of the spatial variability of infiltration, runoff-producing areas, and antecedent conditions can be useful for developing management practices that minimize the effects of chemicals and eroded sediment.

Watershed Models

Although understanding of natural watershed processes is deficient, there have been many attempts to model watershed hydrology and water quality. The CN model mentioned earlier is widely used for estimating runoff volumes from watersheds, and is often a component of more complex models. Models that measure responses to precipitation have been developed for field-level investigations of surface runoff, sediment, pesticides, subsurface water, leachate and water quality, and erosion on hillslopes.

Weather and Precipitation Modeling

Weather and precipitation provide the deterministic and random driving forces in watersheds, and affect quantities and timing of runoff, sediment, and chemicals. Often measured weather and precipitation data are used as inputs to watershed modeling. Unfortunately, records are often not long, and short time-increment precipitation data exist only at points on a landscape. Furthermore, if several precipitation-gauge records are required, the periods of records may not overlap, the gauges may have had different sampling intervals (15-min vs 24-h), or the gauges inadequately characterize spatial variability of precipitation over watersheds. Consequently, weather and precipitation models have been developed. These models stochastically simulate weather for the long periods needed for analyzing modeled watershed outputs using frequency distributions. Stochastic simulation involves analyzing historic weather data to develop parameters that describe the frequency distributions of measured data. The models are designed so that the long records of weather and precipitation data have the same statistical characteristics as the measured data for a given location.

Two better-known weather models are CLIGEN and GEM (Generation of Weather Elements for Multiple Applications). Both models synthesize daily values of solar radiation, maximum and minimum

air temperature, and precipitation, wind speed, wind direction, and dew point. Both models simulate daily precipitation that depends on whether there was precipitation during the previous day. Both models are 'point' models, meaning that simulated precipitation is modeled at a point and not over an area.

GEM additionally generates weather elements that are dependent on whether a day has precipitation ('wet' or 'dry' days). Intuitively, for example, on a day with rain, the sky is cloudy, and there is less likelihood of a large daily solar radiation value. Consequently, the frequency distribution describing solar radiation is different for dry and wet days. Furthermore, an advanced version of the GEM model incorporates spatially dependent weather elements. Extensive databases for CLIGEN have been created by developing parameters at individual gauges, requiring extrapolation to ungauged areas. GEM databases have been interpolated for many areas in the USA on a 4-km^2 grid.

Spatial Variability of Weather and Precipitation

Under ideal conditions, weather inputs should be spatially available for watershed models at appropriate time steps, particularly for large areas. Often weather elements such as solar radiation and temperature do not vary greatly over large areas (except for variations in elevation), and a daily time resolution is adequate. However, precipitation intensities can vary greatly over small areas and characterization for both spatial and temporal variability for small areas and short time steps are required. Precipitation falling on areas of a watershed with potentially high runoff-generating potential (soil near saturation, feedlots, mined and urban areas, etc.), if not properly characterized in time and space, will result in inadequately modeled volume and timing of runoff. The spatial variability and uncertainties in precipitation, weather, and watershed modeling require watershed data and model outputs to be analyzed in terms of probabilities and frequency distributions (e.g., chemical loads, peak flows, runoff volumes).

Summary

Precipitation characterization for analyzing watersheds depends on data that are representative of ground-level precipitation. Precipitation is characterized in terms of times between storms, storm durations and depths, within-storm intensities, and spatial variation of these storm elements. Spatial and temporal characterization of precipitation is important because watershed landscapes also have significant spatial variability in characteristics that affect runoff generation. Water-movement processes on watershed surfaces are inadequately understood and quantified. However, scientists and engineers have developed many methods and computer models for characterizing natural variability in storms and watersheds to solve practical runoff and water-quality problems.

Much remains to be learned regarding the interactions between precipitation and watershed characterizations. Factors that affect parameterization of weather and precipitation models are weather modification, changing climate, and persistence in weather caused by large-scale atmospheric forcings (e.g., El Niño Southern Oscillation). Atmospheric forcings cause persistent dry or wet periods lasting years in different parts of the world. Unless data are categorized to include these persistent periods, this effect will not be considered. Parameterization to reflect weather modification is lacking. Data sets from which parameters are developed may be changing because of potential changes in climate ('nonstationarity' data). Therefore, long periods of synthesized weather data for watershed modeling may not be representative of future conditions.

See also: **Climate Change Impacts**; **Climate Models, Role of Soil**; **Erosion:** Water-Induced; **Water Cycle**

Further Reading

American Society of Civil Engineers (1996) *Hydrology Handbook,* (Manual No. 28), 2nd edn. Prepared by the Task Committee on Hydrology Handbook of Management Group D of the ASCE.

Amerman CR (1965) The use of unit-source watershed data for runoff prediction. *Water Resources Research* 1(4): 499–507.

Bonta JV (1997) Proposed use of Huff curves for hyetograph characterization. In: Richardson CW *et al.* (eds) *Proceedings of Workshop on Climate and Weather Research*, 17–19 July 1995, pp. 111–124. USDA-Agricultural Research Service 1996–03.

Bonta JV (1998) Spatial variability of runoff and soil properties on small watersheds in similar soil-map units. *Transactions of the ASAE*41(3): 575–585.

Chow VT (1964) *Handbook of Applied Hydrology.* New York, NY: McGraw-Hill.

Daly GH, Gibson WP, Taylor GW, Johnson JL, and Pasteris PA (2002) A knowledge-based approach to the statistical mapping of climate. *Climate Research* 22: 99–113.

Desbordes M and Raous P (1976) Un exemple de l'intérêt des études de sensibilité des modeles hydrologiques. *La Houille Blanche* 1: 37–43.

Hershfield DM (1961) *Rainfall Frequency Atlas of the United States for Durations from 30 Minutes to 24 Hours and Return Periods from 1 to 100 Years.* Technical Paper No. 40. Washington, DC: US Government Printing Office.

Johnson GL, Hanson CL, Hardegree SP, and Ballard EB (1996) Stochastic weather simulation: overview and

analysis of two commonly used models. *Journal of Applied Meteorology* 35: 1878–1896.
Larson LW and Peck EL (1974) Accuracy of precipitation measurements for hydrologic modeling. *Water Resources Research* 10(4): 857–863.
Maidment DR (1992) *Handbook of Hydrology*. New York, NY: McGraw-Hill.
McGuinness JL (1966) A comparison of lysimeter catch and rain gage catch. ARS-41-124.
Nicks AD and Gander GA (1994) CLIGEN: a weather generator for climate inputs to water resource and other models. *5th International Conference on Computers in Agriculture*. ASAE publication 3–94.
Restrepo PJ and Eagleson PS (1982) Identification of independent rainstorms. *Journal of Hydrology* 55: 303–319.
Sifalda V (1973) Entwicklung eines Berechnungsregens für die Bemessung von Kanalnetzen. *Wasser Abwasser* 114: 435–440.
Yen BC and Chow VT (1980) Design hyetographs for small drainage-structures. *Journal of Hydraulics Division* 106: 1055–1076.

PRECIPITATION–DISSOLUTION PROCESSES

W P Robarge, North Carolina State University, Raleigh, NC, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Precipitation and dissolution reactions in soils and the underlying strata have long been of interest to soil scientists and soil chemists, because primary and secondary minerals are inherently part of many soil processes, e.g., chemical weathering, soil genesis, and nutrient availability (which can impact human and animal nutrition). More recently, the importance of precipitation and dissolution in soils has been emphasized because of environmental issues, e.g., mineral dissolution reactions as a buffering mechanism against inputs of acidic deposition, and formation of new solid phases as a means of remediation for trace metal contamination. Indeed, it is a mistake to attempt to limit the importance of precipitation and dissolution reactions in soils to mineral-weathering reactions or formation of secondary precipitates around fertilizer granules. Rather, it is necessary to acknowledge that precipitation–dissolution reactions span a range in physical scales that extends from the atomic and submicron through laboratory, watershed, and even continental processes.

Understanding precipitation and dissolution reactions in soils requires knowledge of the chemical and physical laws assumed to govern these processes. These laws are derived from consideration of either macroscale or microscale reactions. Macroscale processes are based on the science of thermodynamics and allow prediction of precipitation or dissolution under equilibrium conditions. The power of this approach is that it is possible to make predictions of precipitation or dissolution without any knowledge of the actual reactions that occur at mineral surfaces or during the formation of nuclei that lead to the formation of stable precipitates. The disadvantage of this approach is that it cannot easily predict the rate at which reactions will occur. Theories that focus on microscale reactions assume that specific chemical-reaction mechanisms occur at mineral surfaces and are more successful in predicting the rate of mineral dissolution due to changes in the composition of the surrounding soil solution. Unfortunately, equations predicting rates for microscale processes tend to be empirical in nature and require information specific to a given set of conditions that can usually be generated only within the laboratory. Most of the time, laboratory conditions only approximate the actual weathering environment in the soil. A knowledge of both macroscale and microscale processes is important in understanding precipitation and dissolution reactions in soils.

Macroscale Processes

The dissolution of a mineral is a chemical reaction. Thermodynamics predicts that a chemical reaction will proceed when energy can be released. A simple chemical reaction can be written as follows:

$$A + B \Leftrightarrow C + D \qquad [1]$$

where chemical species A and B react to form chemical species C and D. The symbol $\Leftrightarrow$ means that the chemical reaction has reached equilibrium and that there is essentially no net change in the concentrations of A, B, C, or D. Thermodynamics allows eqn [1] to be written in the following format:

$$K = \frac{(C)(D)}{(A)(B)} \quad [2]$$

where K is termed the equilibrium constant and parentheses refer to the activity of the chemical species in solution. The activity of a chemical species is a function of its concentration in solution. The equilibrium constant is related to the energy of the chemical reaction through the expression:

$$\Delta G_r = -RT \ln K \quad [3]$$

where in thermodynamics ΔG_r is termed the standard Gibbs energy change, R is the molar gas constant, T is absolute temperature, and K is the equilibrium constant when all the reactants and products are in their standard states. For sparingly soluble compounds, K is written as K_{so} and defined as the solubility product when the chemical reaction is written as a dissolution reaction. Most secondary minerals found in soils can be considered sparingly soluble in the presence of water: dissolution does not noticeably change the total mass of the solid phase present. Standard state assumptions concerning the solid phase are thus applicable, and the solubility of the mineral can be described by a solubility product. The solubility product of a mineral, therefore, is a relative index of its potential stability (whether it will dissolve, remain stable, or conditions are favorable for its formation) in the soil environment.

Representative values of $\log K_{so}$ calculated from standard Gibbs free energies are listed in Table 1. Negative values of $\log K_{so}$ indicate low solubility in the presence of water. Positive values of $\log K_{so}$ suggest that the corresponding mineral phases are not stable and would not attain equilibrium in the typical aqueous soil environment. This comparison is only qualitative, however, as the dissolution reaction for most silicates and phyllosilicates requires the presence of H^+ as well as water, thus the dissolution reaction (and therefore mineral stability) is also a function of pH. Calculated values of K_{so} are also dependent on the degree of uncertainty associated with the standard Gibbs free energies for each of the constituents in the reaction and variations in calculated $\log K_{so}$ values of several whole numbers or more are not uncommon.

Ion Activity Products

The dissolution of a mineral can be written as the transition from one solid phase to another:

$$\text{mineral A} + \text{water} \Leftrightarrow \text{mineral B} + \text{solution species C} \quad [4]$$

or the destruction of the original mineral with the stoichiometric (in proportion to the atomic structure of the mineral) release of the constituent ions to solution. The dissolution reaction for the secondary mineral kaolinite can be written as:

$$Al_2Si_2O_5(OH)_4(s) + 6H^+(aq) \Leftrightarrow 2Al^{3+}(aq) + 2H_4SiO_4^0(aq) + H_2O(l) \quad [5]$$

where the subscript '(s)' refers to the solid phase, '(aq)' to the ion in solution, and '(l)' for the bulk aqueous phase. The calculated $\log K_{so}$ for this reaction is 5.4 (Table 1). Solving for K_{so} yields the following equation:

$$K_{so} = \frac{(Al^{3+}(aq))(H_4SiO_4^0(aq))^2(H^+(aq))^{-6}(H_2O(l))}{(Al_2Si_2O_5(OH)_4(s))} \quad [6]$$

Under standard conditions, the expressions for the solid-phase kaolinite and the bulk liquid-phase water are defined as being equal to a value of 1, leaving the numerator as an ion activity product (IAP) for the constituent ions in solution:

$$IAP = (Al^{3+}(aq))(H_4SiO_4^0(aq))^2(H^+(aq))^{-6} \quad [7]$$

When the IAP $> K_{so}$, the solution is said to be supersaturated with respect to the solid phase (conditions favor the precipitation of kaolinite). When the IAP $< K_{so}$, the solution is undersaturated with respect to kaolinite (conditions favor the dissolution of kaolinite). When IAP $= K_{so}$, the solution is considered saturated and at equilibrium (kaolinite should neither dissolve nor precipitate). With tabulated values of solubility products (Table 1) and measurements of the dissolved concentrations of ions in soil solutions or streams, soil scientists can predict the possible absence or presence of minerals by calculating IAPs.

Activity Diagrams

The IAP provides a numerical way to estimate the stability of a mineral in a given soil environment. A visual representation of the relative stability of a mineral can be obtained using an activity diagram. Activity diagrams vary in complexity, depending on the number of mineral phases involved. They are all, however, based on rearrangement of dissolution reactions (e.g., eqn [5]) into a linear form where the y- and x-axes are represented by constituent ions common to all the mineral phases being addressed. This rearrangement often involves the introduction of H^+ into the dissolution reaction by adding the hydrolysis reaction of water. The resulting K values are sometimes designated as K_{diss} (where 'diss' is 'dissolution') in contrast to K_{so} to reflect the mathematical convenience of incorporating H^+ into the dissolution reaction.

Table 1 Calculated solubility products for selected minerals using standard Gibbs free energies of formation[a]

Name	*Chemical formula*	*Dissolution reaction*	*log K_{SO}*
Carbonates			
Calcite	$CaCO_3$	$Ca^{2+} + CO_3^{2-}$	−8.3
Dolomite	$MgCa(CO_3)_2$	$Mg^{2+} + Ca^{2+} + 2CO_3^{2-}$	−17.6
Siderite	$FeCO_3$	$Fe^{2+} + CO_3^{2-}$	−10.1
Rhodochrosite	$MnCO_3$	$Mn^{2+} + CO_3^{2-}$	−9.9
Sulfates			
Gypsum	$CaSO_4 \cdot 2H_2O$	$Ca^{2+} + SO_4^{2-} + 2H_2O$	−4.6
Jarosite	$KFe_3(SO_4)_2(OH)_6$	$K^+ + 3Fe^{3+} + 2SO_4^{2-} + 6OH^-$	−95.3
Alunite	$KAl_3(SO_4)_2(OH)_6$	$K^+ + 3Al^{3+} + 2SO_4^{2-} + 6OH^-$	−79.9
Phosphates			
Variscite	$AlPO_4 \cdot 2H_2O$	$Al^{3+} + PO_4^{3-} + 2H_2O$	−21.8
Strengite	$FePO_4 \cdot 2H_2O$	$Fe^{3+} + PO_4^{3-} + 2H_2O$	−26.1
Vivianite	$Fe_3(PO_4)_2 \cdot 8H_2O$	$3Fe^{2+} + 2PO_4^{3-} + 8H_2O$	−35.5
K-Taranakite	$H_6K_3Al_5(PO_4)_8 \cdot 18H_2O$	$6H^+ + 3K^+ + 5Al^{3+} + 8PO_4^{3-} + 18H_2O$	−176.5
Octacalcium phosphate	$Ca_8H_2(PO_4)_6 \cdot 5H_2O$	$8Ca^{2+} + 2H^+ + 6PO_4^{3-} + 5H_2O$	−92.6
Hydroxyapatite	$Ca_{10}(OH)_2(PO_4)_6$	$10Ca^{2+} + 2OH^- + 6PO_4^{3-}$	−114.9
Fluorapatite	$Ca_{10}F_2(PO_4)_6$	$10Ca^{2+} + 2F^- + 6PO_4^{3-}$	−314.3
Sulfides			
Alabanite	MnS	$Mn^{2+} + S^{2-}$	−12.7
Sphalerite	α-ZnS	$Zn^{2+} + S^{2-}$	−24.4
Pyrite	FeS_2	$Fe^{2+} + S_2^{2-}$	−26.6
Covellite	CuS	$Cu^{2+} + S^{2-}$	−35.6
Black-metacinnibar	β-HgS	$Hg^{2+} + S^{2-}$	−51.0
Oxides/hydrous oxides			
Quartz	α-SiO_2[a]	$H_4SiO_4^0$	−4.0
Quartz (amorphous phase)	SiO_2[a]	$H_4SiO_4^0$	−2.7
Gibbsite	γ-$Al(OH)_3$	$Al^{3+} + 3OH^-$	−33.5
Gibbsite (amorphous phase)	$Al(OH)_3$	$Al^{3+} + 3OH^-$	−31.9
Boehmite	γ-AlOOH[b]	$Al^{3+} + 3OH^-$	−33.4
Amorphous iron hydroxide	$Fe(OH)_3$	$Fe^{3+} + 3OH^-$	−37.9
Hematite	α-Fe_2O_3[b]	$Fe^{3+} + 3OH^-$	−41.4
Goethite	α-FeOOH[b]	$Fe^{3+} + 3OH^-$	−41.5
Silicates[c]			
Olivine	$Mg_{1.6}Fe_{0.4}SiO_4$	$1.6Mg^{2+} + 0.4Fe^{2+} + H_4SiO_4^0$	25.8
Diopside	$CaMg(SiO_3)_2$	$Mg^{2+} + Ca^{2+} + 2H_4SiO_4^0$	20.9
Pyroxene	$CaAl_2SiO_6$	$Ca^{2+} + 2Al^{3+} + H_4SiO_4^0 + 2H_2O$	34.8
Albite	$NaAl_2Si_3O_8$	$Na^+ + Al^{2+} + 3H_4SiO_4^0$	3.6
Microcline	$KAlSi_3O_8$	$K^+ + 2Al^{3+} + 3H_4SiO_4^0$	1.0
Anorthite	$CaAl_2Si_2O_8$	$Ca^{2+} + 2Al^{3+} + 2H_4SiO_4^0$	23.0
Phyllosilicates[c]			
Muscovite	$KAl_2(AlSi_3O_{10})(OH)_2$	$K^+ + 3Al^{3+} + 3H_4SiO_4^0$	13.3
Vermiculite	(see below)[d]	$2.71Mg^{2+} + 0.02Fe^{2+} + 0.46Fe^{3+} + 0.06Ca^{2+} + 0.1K^+ + 1.14Al^{3+} + 2.91H_4SiO_4 + 0.36H_2O$	38.1
Montmorillonite	(see below)[e]	$0.49Mg^{2+} + 1.71Al^{3+} + 0.22Fe^{3+} + 3.81H_4SiO_4^0$	2.7
Kaolinite	$Al_2Si_2O_5(OH)_4$	$2Al^{3+} + 2H_4SiO_4^0 + H_2O$	5.4

[a]Standard Gibbs free energies of formation taken from Lindsay WL (1979) *Chemical Equilibria in Soils*. New York: John Wiley; Stumm W and Morgan JJ (1996) *Aquatic Chemistry: Chemical Equilibria and Rates in Natural Waters*, 3rd edn. New York: John Wiley.
[b]Water added as a reactant (e.g., $SiO_2 + 2H_2O = H_4SiO_4^0$).
[c]Water and H^+ added as reactants.
[d]$Mg_{2.71}Fe(II)_{0.02}Fe(III)_{0.46}Ca_{0.06}K_{0.1}(Si_{2.91}Al_{1.14})O_{10}(OH)_2$.
[e]$Mg_{0.2}(Si_{3.81}Al_{1.71}Fe(III)_{0.22}Mg_{0.29})O_{10}(OH)_2$.

Simple activity diagrams usually include a series of linear lines, reflecting differences in the solid phase under consideration (Figure 1). These differences are often associated with variations in solubility due to assumed crystallinity or size of the solid phase. Experimental data points plotted on such graphs which fall along the linear lines are taken to reflect equilibrium with the solid phase. Data points falling above or below the line indicate undersaturation or supersaturation of the soil solution with respect to the mineral phase (Figure 1). More detailed activity diagrams can also be used to delineate regions of stability among several mineral phases (Figure 2). Experimental data can be plotted on such plots, but the intersection of the

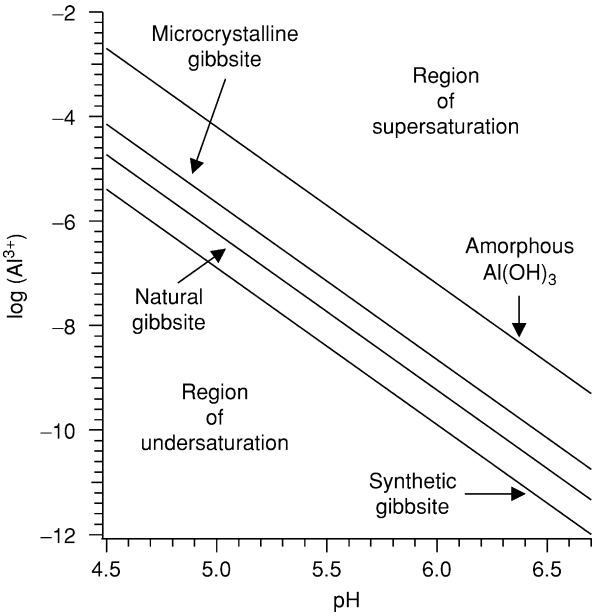


Figure 1 Activity–ratio diagram for synthetic gibbsite, natural gibbsite, microcrystalline gibbsite, and amorphous aluminum hydroxide. Reproduced with permission from Johnson NM, Driscoll CT, Eaton JS, Likens GE, and McDowell WH (1981) 'Acid Rain', Dissolved Aluminum and Chemical Weathering at the Hubbard Brook Experimental Forest, New Hampshire. *Geochimica et Cosmochimica Acta* 45(9): 1421–1437.

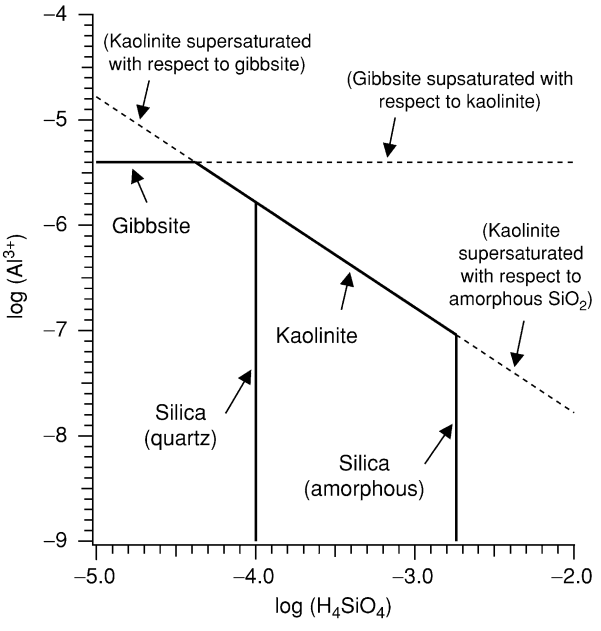


Figure 2 Activity–ratio diagram for the gibbsite, kaolinite, silica mineral assemblage for pH 4.5. Dashed line representing supersaturation of kaolinite with respect to quartz is not shown. Adapted from Robarge WP (1999) Precipitation/dissolution reactions in soils. In: Sparks DL (ed.) *Soil Physical Chemistry*, 2nd edn. Boca Raton, FL: CRC Press, with permission.

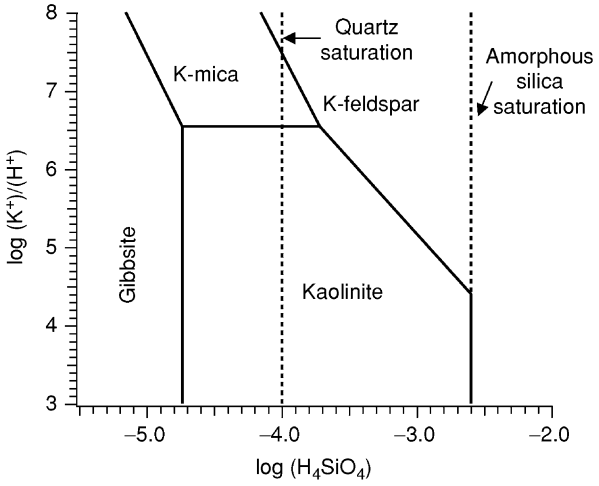


Figure 3 Phase diagram for gibbsite, kaolinite, K-feldspar, K-mica, silica mineral assemblage. Adapted from Garrels RM and Christ CL (1965) *Solutions, Minerals and Equilibria*. New York: Harper and Row, with permission.

various linear representations of mineral solubility also provide predictive information concerning the stability of minerals in soil systems. For example, kaolinite is the most stable phase compared with gibbsite in soil environments where the log ($H_4SiO_4^0$) in soil solution is more than −4.5. The region of stability is increased if the activity of $H_4SiO_4^0$ in solution is controlled by amorphous silica in contrast to quartz (Figure 2).

Dissolution reactions as depicted in eqn [4] can also be displayed visually through a variation in activity plots known as a phase diagram (also known as an activity–activity diagram) (Figure 3). Linear lines in such plots represent equilibrium between two solid phases. Departure away from the linear lines indicates that only one of the mineral phases in question would be stable for the values of the constituent ions plotted along the axes.

Understanding the dissolution of minerals as a macroscale process relies on the assumptions inherent in thermodynamics and congruent dissolution of the solid phase. Congruent dissolution requires that the rates of release of the chemical species represented by the formula weight for a mineral are equal. In other words, it is assumed that the various chemical species that comprise a mineral enter the surrounding soil solution at the same time during the dissolution reaction. For many carbonate, sulfate, phosphate, sulfide, and even silicate-based minerals, this assumption is probably valid. With increasing complexity in mineral structure, however, the assumption of congruent dissolution becomes questionable. For example, the congruent dissolution of montmorillonite or muscovite (Table 1) involves the concurrent

release to soil solution of up to five different ionic species. Given the differences in chemical bonding between these constituent species within the mineral structure, it is highly unlikely that dissolution of these minerals proceeds in a congruent fashion. It is more likely that the dissolution reaction is incongruent: the ratio of ions released to solution is not the same as that found in the bulk phase of the mineral. The extent to which incongruent dissolution invalidates the macroscale approach to understanding mineral solubility in soil systems is not easily determined. In general the macroscale approach is valid for predicting long-term, steady-state conditions or delineating boundaries for possible reactions.

There are also limits to using the condition of $IAP > K_{so}$ (supersaturation) to predict precipitation reactions in the soil environment. Minerals (including soil clay minerals) represent the selective accumulation of two or more constituent ions into an organized solid matrix. This can occur on the bulk soil solution (homogeneous precipitation) or on the surfaces of other soil solids that act as a catalyst to promote the growth of a new solid phase (heterogeneous precipitation). It is possible for many solids with relatively simple structures (carbonates, sulfates, sulfides, and some phosphates and oxides) to form in soils over relatively short time periods. For example, soil horizons that experience periodic wetting and drying due to a variable water table lose and gain color due to the dissolution and then subsequent precipitation of iron oxides. In more arid climates, it is not uncommon to find layers of carbonates and even sulfates that precipitate from the soil solution as the wetting front from a rain event slows and finally stops as it moves through the soil. Formation of more complex solids, especially silicates, phyllosilicates (which include most soil clay minerals) and apatites, however, should not be assumed to occur readily in the bulk soil environment. These solids generally require unique conditions within the soil that favor their formation, or occur as heterogeneous precipitation reactions on the surfaces of or near other minerals that are undergoing dissolution. Even when such solids form, they are often amorphous in nature, meaning that they have the expected chemical composition for the solid phase, but require more time to mature and develop the internal crystalline, well-ordered structure characteristic of a given mineral. Amorphous phases of a mineral are more soluble than its crystalline counterpart and may not survive long enough to develop into a stable separate phase within the soil. As with dissolution reactions, predicting the rates of precipitation in specific soil environments is best viewed as a microscale process.

Microscale Processes

Precipitation

Formation of a new solid phase when viewed as a microscale process requires a series of steps that must occur before a new solid can survive and be characterized as having a specific K_{so} value. The first requirement is that supersaturation ($IAP > K_{so}$) conditions must exist whether the reaction results in homogeneous or heterogeneous precipitation. Supersaturation conditions are necessary because they promote the formation of nuclei as a result of the collisions in solution (or on a foreign surface) of the constituent ions that eventually could become the new solid phase. Once formed, the nuclei must resist the tendency to dissolve and must strive to grow in size by accumulating other constituent ions and transforming into crystallites. The formation of crystallites is a critical step, because they are large enough in size to begin to lower the supersaturation of the solution and lower the chances of other nuclei forming. This effectively promotes their own survival, because it lowers the competition for the remaining constituent ions in solution. The final step is the growth of the crystallites into crystals through the process called Ostwald ripening: an increase in the size of a selected number of crystallites to form crystals by the net transfer of material from smaller nuclei and crystallites via dissolution and precipitation. The larger crystals being formed in fact can be defined by an IAP and continue to grow until ideally $IAP = K_{so}$ (saturation).

Which crystallites survive the transformation into crystals and the eventual formation of a new solid phase can be predicted by consideration of two sources of energy which influence the growth of crystals in solution or on the surface of a foreign particle. The first source of energy is the formation of the chemical bond between the constituent ions that will compose the solid (here designated as ΔG_{bulk}). The second energy term is related to the amount of work for the nuclei actually to be present within the solution (here designated as ΔG_s). In other words, the act of forming nuclei must result in the displacement of the surrounding water molecules to create a volume in solution where the nuclei can physically exist. How well a nucleus or crystallite can do this is related to its size and its ability to interact with the surrounding water molecules. This latter ability is referred to as interfacial tension. The summation of the two energy terms equals to amount of energy available for the formation of a stable nuclei in solution (or on a foreign surface):

$$\Delta G^0_{r(nucleus)} = \Delta G_{bulk} + \Delta G_s \qquad [8]$$

For any nucleus, ΔG_{bulk} is related to the number of molecular units in the nuclei or crystallites and the degree of supersaturation in solution. As long as IAP $> K_{so}$, ΔG_{bulk} is negative, which promotes nuclei or crystallite growth. ΔG_s is a function of the interfacial tension between the surrounding water molecules and the surface area of the nuclei or crystallites. This term is often positive in value, meaning that it requires energy to make room for the particles to grow in solution.

Mathematical expressions for the two energy terms in eqn [8] can be derived but are a function of the geometry selected for the nuclei and will not be discussed here. Equation [8] does, however, provide a qualitative explanation for the results of precipitation and dissolution reactions often observed in the soil environment. For example, inclusion of ΔG_s illustrates why dissolution rates are often a function of particle size, i.e., smaller particles dissolve faster than larger particles. Small particles, typically less than 1 μm in size, have significant positive ΔG_s values because of the combination of their relatively large surface area to volume ratios and interfacial tension. Small particles therefore exhibit a different K_{so} value than their larger counterparts, such that a solution may appear supersaturated with respect to a solid phase when in fact that solid may actually be present but as very small particles. The presence of the ΔG_s in eqn [8] also explains why amorphous phases may be present in soils even though they are more soluble than their more crystalline counterparts (e.g., compare K_{so} values for amorphous silica with quartz in **Table 1**). For two solids having the same chemical composition, the solid that is more crystalline in nature should have a larger negative ΔG_{bulk} value. However, if the amorphous-phase nuclei have a lower interfacial tension than the crystalline phase, ΔG_s will have smaller positive value such that the sum of the two terms actually results in a $\Delta G^0_{r(nucleus)}$ value that favors formation of the amorphous phase. Experimental evidence suggests that formation of amorphous phases is almost always preferred over crystalline phases under most conditions that favor precipitation. More crystalline phases usually develop later either through internal rearrangement of the amorphous-phase solid, or through the dissolution of the amorphous phase and subsequent growth of the crystalline phase. In either case, it should not always be assumed that the crystalline phases of minerals will dominate soil environments simply because of differences in K_{so} values.

Lastly, the role of interfacial tension also offers an explanation of why heterogeneous precipitation occurs in the presence of foreign surfaces, often at supersaturated conditions below those necessary to promote homogeneous precipitation. If the interfacial tension between the constituent ions of a nuclei or crystallite is lower when interacting with a surface than with bulk water, the formation of a nucleus on the surface of a foreign particle would reduce the surface area exposed to solution, in effect lowering the value of ΔG_s in eqn [8]. This yields a $\Delta G^0_{r(nucleus)}$ value that favors nuclei attempting to grow on the surface of a foreign particle in contrast to those attempting to survive in the bulk solution. Thus solid phases that nucleate and grow on a foreign surface perform less work in order to survive. Depending on the surface of the foreign particle (i.e., the interfacial tension between the foreign particle and the nuclei of the new solid phase), the growth of the nuclei into crystallites may actually proceed along the surface of the foreign particle before developing as a separate, three-dimensional phase extending away from the foreign surface. This affords one explanation for the development of coatings, such as iron oxide coatings on clay minerals, which are often observed in soils.

Dissolution

The dissolution of a mineral can be defined as the detachment of constituent ions from the surface of a mineral and their subsequent transport to the bulk solution. It is now generally accepted that the rate of dissolution of minerals in soils is controlled by chemical reactions along the mineral surfaces at distinct surface features that can include dislocations, etch pits, and micropores. Dislocations represent areas in the mineral structure where the normal growth of the mineral has been disrupted or caused to alter in some fashion. Etch pits and micropores form in the surface of minerals and are visible as holes in the mineral surface when viewed using devices such as a scanning electron microscope. The presence of surface features such as dislocations, etch pits, and micropores is taken as an indication of a population of surface reactive sites where dissolution reactions are favored above those possible along the total surface area of a mineral.

The dissolution reaction at the surface reactive sites is considered a two-step process. The first step involves the relatively rapid formation of a surface species as the result of the reaction with entities in soil solution such H^+, OH^-, H_2O molecules, and inorganic and organic ligands. The second step, and the one that is considered rate-limiting, involves the transformation of the surface species into an activated complex with the eventual release of constituent ions to the bulk solution. Dissolution at the microscale therefore can be considered a function of the reactive surface area of the mineral, the nature of the inherent chemical bonding structure that comprises the

mineral, and the presence of various chemical species in the soil solution. Thus the dissolution rate of a mineral is not a fixed quantity but dependent upon a number of variables, which may change with time due to changes in the physical size of the mineral itself, as well as changes in the surrounding soil solution.

Rate law A general form of a rate law for chemical dissolution at heterogeneous mineral surfaces has been formulated to account for the various factors important in the solubility of minerals. A simplified version of this rate law can be expressed as:

$$\text{Rate} = k_o \cdot S_{min} \cdot \exp^{(-E_a/RT)} \cdot \int\left(\text{pH} \cdot a_i^{N_i}\right) \cdot \int(\text{IAP}) \quad [9]$$

where k_o is the intensive rate constant (moles per meter squared per second), S_{min} is the reactive surface area of the mineral, and E_a is the overall activation energy for the reaction. The possible effects of pH, ionic strength, and other ligands, reductants, oxidants, or sorbed ions ($\int(\text{pH} \cdot a_i^{N_i})$) are written as products to emphasize that these effects can be studied separately or in concert with one another, but that one cannot ignore the influence of any of these variables on the rate of dissolution. The final term, $\int(\text{IAP})$, is added to account for the fact that rate of dissolution is also a function of the degree of supersaturation or undersaturation of the soil solution. Technically eqn [9] can represent the rate of dissolution or precipitation, with rate = 0 at saturation (IAP = K_{so}).

The range in dissolution rates possible for various minerals, primarily from laboratory dissolution experiments, is shown in Table 2. Among the silicates and the phyllosilicates, those minerals referred to as framework silicates (quartz, albite, microcline, anorthite) and the sheet silicates (kaolinite, montmorillonite, muscovite) are more insoluble than the chain silicates (diopside) or the ring and orthosilicates (olivine). Assuming a mineral particle with an initial radius of 1 mm, minerals with log rate values of less than −13 could have lifetimes of more than 1 000 000 years, while log rate values of −8 to −5 would have lifetimes of less than one hundred to several thousands of years. Such predictions are generally valid as compared to actual field observations across many different types of geologic and hydrological systems. However, dissolution rates and mineral stability within specific local environments are dictated by the variables included in eqn [9]. Laboratory dissolution rates for calcite and dolomite (Table 2) suggest that these minerals are not stable across geologic time spans, yet under proper conditions these minerals are found naturally in soils.

Table 2 Rates of dissolution for selected minerals

Mineral	*Log rate* ($mol\,m^{-2}\,s^{-1}$)	*Conditions*
Silicates		
Quartz	−13.2	40°C, $I = 0$[a]
	−11.2	25°C, pH 5
	−10.7	70°C, pH 5.4, 0.25 mol l^{-1} Na^+
	−9.9	70°C, pH 6, $I = 0.35$
Microcline	−12.5	25°C, pH 5
	−11.8	25°C, pH 3, $I = 0$
	−11.1	25°C, pH 3.1, 1 mmol l^{-1} oxalic acid
Diopside	−11.3	25°C, pH 3.6
	−10.2	25°C, pH 5
Albite	−12.3	25°C, pH 5
	−11.3	25°C, pH 5, oxalic acid
	−11.6	25°C, pH 3, oxalic acid
Anorthite	−12.0	25°C, pH 6.2, $I = 0$
	−8.6	25°C, pH 5
Olivine	−5.7	25°C, pH 3–5
Phyllosilicates		
Kaolinite	−13.8	25°C, pH 5
	−13.0	25°C, pH 8
Montmorillonite	−14.1	20°C, pH 8
	−13.4	25°C, pH 5
	−12.5	25°C, pH 3
Muscovite	−13.0	25°C, pH 5
	−12.5	25°C, pH 3
Other		
Gibbsite	−12.5	25°C, pH 5
	−11.3	25°C, pH 3.1, $I = 0.0001$
Calcite	−5.6	pH 5, low P_{CO_2}
	−4.5	pH 3, low P_{CO_2}
Dolomite	−6.5	pH 5
	−5.0	pH 3

Data taken from Lasaga AC (1998) *Kinetic Theory in the Earth Sciences*. Princeton, NJ: Princeton University Press; Stumm W (1990) *Aquatic Chemical Kinetics: Reaction Rates of Processes in Natural Waters*. New York: John Wiley; and White AF and Brantley SL (eds) (1995) *Reviews in Mineralogy*, vol. 31. Washington, DC: Mineralogical Society of America.

[a] I, ionic strength.

Summary

The dissolution of minerals is a vital geochemical process that has led to the development of life on Earth as we know it. The presence of minerals in soils and underlying soil parent material is critical to the sustainability of soils, especially forest soils, where fertilization to replace mineral elements such as calcium or magnesium lost by timber harvesting or enhanced leaching due to acidic deposition is currently cost-prohibitive. Intensive investigation into the solubility of minerals continues using both the macroscale and microscale approaches. New efforts are now being directed at incorporating *ab initio* quantum-mechanical calculations of mineral surface reactions to gain a better understanding of heterogeneous kinetics. This work is made possible by the

rapid changes in computer technology, hardware, and software that have occurred since the 1980s. Just how molecular modeling will impact our understanding of the solubility of minerals and the precipitation of new solid phases, especially for larger and more complex chemical systems, is uncertain, but the merger of computational and experimental science has begun and will continue for the foreseeable future.

See also: **Acid Rain and Soil Acidification**; **Agroforestry**; **Chemical Equilibria**; **Clay Minerals**; **Factors of Soil Formation:** Biota; Climate; Human Impacts; Parent Material; Time; **Fertility**; **Metal Oxides**; **Micronutrients**; **Minerals, Primary**; **Productivity**

Further Reading

Bragg L, Claringbull GF, and Taylor WH (1965) *Crystal Structures of Minerals: The Crystalline State,* vol. IV. Ithaca, NY: Cornell University Press.

Cygan RT (2001) Molecular modeling in mineralogy and geochemistry. In: Cygan RT and Kubicki JD (eds) *Molecular Modeling Theory: Applications in the Geosciences*. Reviews in Mineralogy and Geochemistry,vol. 42, pp. 1–35. Washington, DC: Mineralogical Society of America.

Dixon JB and Weed SB (eds) (1989) *Minerals in Soil Environments,* 2nd edn. Soil Science Society of America Book Series No. 1. Madison, WI: Soil Science Society of America.

Garrels RM and Christ CL (1965) *Solutions, Minerals and Equilibria*. New York: Harper and Row.

Lasaga AC (1995) Fundamental approaches in describing mineral dissolution and precipitation rates. In: White AF and Brantley SL (eds) *Chemical Weathering Rates of Silicate Minerals*. Reviews in Mineralogy, vol. 31, pp. 23–86. Washington, DC: The Mineralogical Society of America.

Lasaga AC (1998) *Kinetic Theory in the Earth Sciences*. Princeton, NJ: Princeton University Press.

Lasaga AC and Gibbs GV (1990) *Ab initio* quantum-mechanical calculations of surface reactions – a new era? In: Stumm W (ed.) *Aquatic Chemical Kinetics: Reaction Rates of Processes in Natural Waters.*New York: John Wiley.

Lindsay WL (1979) *Chemical Equilibria in Soils*. New York: John Wiley.

Robarge WP (1999) Precipitation dissolution reactions in soils. In: Sparks DL (ed.) *Soil Physical Chemistry*, 2nd edn. Boca Raton, FL: CRC Press.

Sposito G (1981) *The Thermodynamics of Soil Solutions*. London, UK: Oxford University Press.

Stumm W (ed.) (1990) *Aquatic Chemical Kinetics: Reaction Rates of Processes in Natural Waters*. New York: John Wiley.

Stumm W and Morgan JJ (1996) *Aquatic Chemistry: Chemical Equilibria and Rates in Natural Waters*, 3rd edn. New York: John Wiley.

White AF and Brantley SL (1995) Chemical weathering rates of silicate minerals: an overview. In: White AF and Brantley SL (eds) *Chemical Weathering Rates of Silicate Minerals*. Reviews in Mineralogy, vol. 31, pp. 1–22. Washington, DC: The Mineralogical Society of America.

Precision Agriculture *See* **Site-Specific Soil Management**

Preferential Flow *See* **Unstable Flow**; **Macropores and Macropore Flow, Kinematic Wave Approach**

PRODUCTIVITY

D L Karlen, USDA Agricultural Research Service, Ames, IA, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Soil productivity is defined as the capacity of a soil to produce a certain yield of agricultural crops or other plants using a defined set of management practices. It encompasses soil fertility plus all of the other factors affecting plant growth and development, even though it is impossible to explain all of the potential interactions. Soil productivity is most accurately defined in terms of a specific kind of soil, for a specific crop or crop sequence, and with regard to a specific set of management practices. As a result of all these constraints, the concept is perhaps more accurately defined as an economic relationship than one of soil science.

Soil productivity is generally measured by comparing inputs with outputs. For agronomic situations, 'output' generally means crop yield, while 'inputs' usually refer to water, plant nutrients, genetic resources, pest control, or labor. Therefore, for a specific input (e.g., water or plant nutrients) productivity is sometimes calculated as efficiency (i.e., water-use efficiency or nutrient-use efficiency). With regard to sustainability, unintended off-site effects such as runoff, erosion, or leaching losses are also recognized as cropping system outputs that must be considered when evaluating productivity with regard to the social acceptability. The relationships between productivity and economic return are very evident at the farm level, while, from a global perspective, productivity is strongly associated with food sufficiency and sustainable food, feed, and fiber production systems. These multiple scales at which productivity can be evaluated mean that economically beneficial inputs at the field scale may or may not be environmentally and socially acceptable when evaluated at the whole-farm, watershed, regional, national, or international scale.

Productivity is also dependent upon the physical, chemical, and biological characteristics that for each soil reflect numerous interactions among the five basic soil-forming factors (parent material, climate – especially moisture and temperature effects – macro- and microorganism species and populations, topography, and time). Therefore, even though soil productivity *per se* is not an inherent soil characteristic, it will be different for every soil and is therefore one of the many characteristics used to classify more than 18 000 soil series in the USA alone.

Productivity is also affected by what some refer to as a sixth soil-forming factor – humankind. Others prefer to classify this response as 'dynamic soil properties,' since the changes are often controllable and reflect the soil-management decisions or practices that are imposed by humankind on the soil resource. Processes affecting the dynamic soil properties include erosion, nutrient mining (where removal is greater than the inputs), loss of organic matter, acidification, salinization, water logging, and compaction. Productive soils generally exhibit resistance or resilience, two properties that enable the resource to withstand degradation by natural (e.g., drought or flooding) or human-induced (e.g., tillage) stress without losing their ability to provide critical functions.

Productivity Effects of the Inherent Soil-Forming Factors

The specific effect of each soil-forming factor on productivity will differ among soils, but for high agronomic productivity the soil must provide physical support, plant-available water, air for respiration, and the essential nutrients (C, H, O, N, P, K, Ca, Mg, S, B, Cu, Fe, Mn, and Zn). Parent material influences the availability of essential plant nutrients (those required to complete the life cycle) through its mineral composition and predominant particle size (i.e., sand, silt, or clay). Particle size is important because it influences soil texture as well as the type and rate of chemical reactions that occur within a specific soil. The process through which parent material was derived (e.g., volcanic activity, sedimentation, residual weathering of rocks) or transported to various sites (e.g., water, wind, or glaciers) affects productivity, primarily through soil texture and landscape topography. Climate, especially precipitation and temperature, influences inherent soil productivity through its effect on weathering of the primary minerals, erosion by wind or water, and the native vegetation (i.e., forest versus prairie). Chemical reaction rates for processes such as solution, hydration, and leaching are also regulated by water and temperature regimes. The primary type and quantity of vegetation found at various landscape positions, the type and amount of microbial growth, the type of enzymes produced, the number and type of soil animals (e.g., grazers versus burrowers), the predominant soil biota, and amount of organic material added to the

soil from above- and belowground plant parts each year are also influenced by climate.

Topography influences soil productivity primarily through its effects on soil water, temperature, and erosion. Soils on sloping land, especially those with a medium or fine texture, will almost always capture less precipitation, irrigation, or runoff water and are more likely to be degraded through erosion than soils with similar vegetation on more level areas. Soils in low-lying or depression areas can be highly productive if rainfall or irrigation is evenly distributed, but they can be nonproductive if water accumulates, creating aeration problems or increasing leaching. Slope orientation and elevation affect productivity primarily through their effects on soil-forming processes as related to soil temperature and water content. Time of formation exerts its influence through the degree of soil development by processes of eluviation (loss of material) and illuviation (accumulation of material) at different positions within the soil profile.

Vegetation influences productivity through photosynthesis, which determines the quantity, quality, and spatial distribution (above- or belowground) of carbon input each year. This annual carbon input as soil organic matter affects nutrient cycling, soil aggregation, and soil structure. In this context, vegetation quality generally refers to carbon-to-nitrogen (C:N) or lignin-to-nitrogen ratios. These quality factors affect the food source for soil biota and hence nutrient cycling. Aggregation, the process through which primary soil particles (sand, silt, and clay) are bound together by natural forces and organic compounds derived from root exudates and microbial processes is important because, if soil aggregates are stable, rainfall or irrigation water are more likely to enter (infiltrate) the soil than to runoff. Aggregates also provide the building blocks for soil structure which influences productivity through its effect on root proliferation, aeration, and the volume of soil available for water retention. Vegetation also provides the fuel for naturally occurring and intentional fires, a natural ecosystem process that plays an exceptionally important role in nutrient cycling and reseeding of grasslands and coniferous forests. Periodic fires also prevent encroachment of trees into grassland areas, and also, in arid areas, they prevent shrubs or other woody species such as honey mesquite (*Prosopis glandulosa*) from becoming established.

Humankind's Impact on Productivity

Humankind can have a tremendous impact on soil productivity, because with external inputs distinctly different soils can have very similar outputs. For example, two soils that respond similarly with low input (e.g., simply tilling, planting, and harvesting) may respond in very different ways if the inputs include fertilizer, lime, animal manure, or rotations with deeply rooted legumes. These differences illustrate the impact of humankind on dynamic soil properties through their land use and other decision-making processes.

For comparative purposes, the US Department of Agriculture (USDA)–Natural Resources Conservation Service (NRCS) soil survey program defines soil productivity assuming some specified level of management. A productive soil is generally defined as one that will produce good yields of adapted crops in relation to the level of labor, nutrient, and energy inputs. This interpretation of inherent soil productivity provides a basis for comparing crop yield per unit area among soil map units throughout the USA. The caveat for each comparison is that the reported yields are those that can be expected under a high level of management, but the specific management practices are not defined. Also, when a yield is not given for a specific soil map unit within a survey report, it implies that either the soil is not suited for that particular crop or the crop is generally not grown on that particular soil.

Long-term soil management through its effect on the dynamic soil properties has created some general expectations regarding soil productivity when evaluated at regional or major land resource area (MLRA) scales. Soils formed under grassland in humid regions often have higher organic matter content, an increased percentage of water-stable aggregates, and a structure that accommodates rapid plant growth, unlimited root proliferation, and thus increased productivity (i.e., crop yield). These soil characteristics generally occur because grasses in these regions replace most of their roots and top growth each year. As the aboveground plant dies, it falls on the soil surface, where it is decomposed or mixed into the upper part of the soil. Earthworms and other soil organisms incorporate a large portion of this material into the soil. The dead plant material serves as a food source for fungal and bacterial decomposition processes, leading to the formation of soil aggregates, and modifies effects of soil texture with regard to water and air relationships and root penetration. The increased water-stable aggregation, annual proliferation of plant roots, and associated wetting and drying cycles gradually develop a soil structure that is stable and, with good management, generally resists degradation (Figure 1).

Grassland soils are often less acid than those developed under forest vegetation. This occurs because grassland soils were generally formed with less annual precipitation which can cause greater leaching

Figure 1 A well-structured grassland soil showing characteristics associated with productivity.

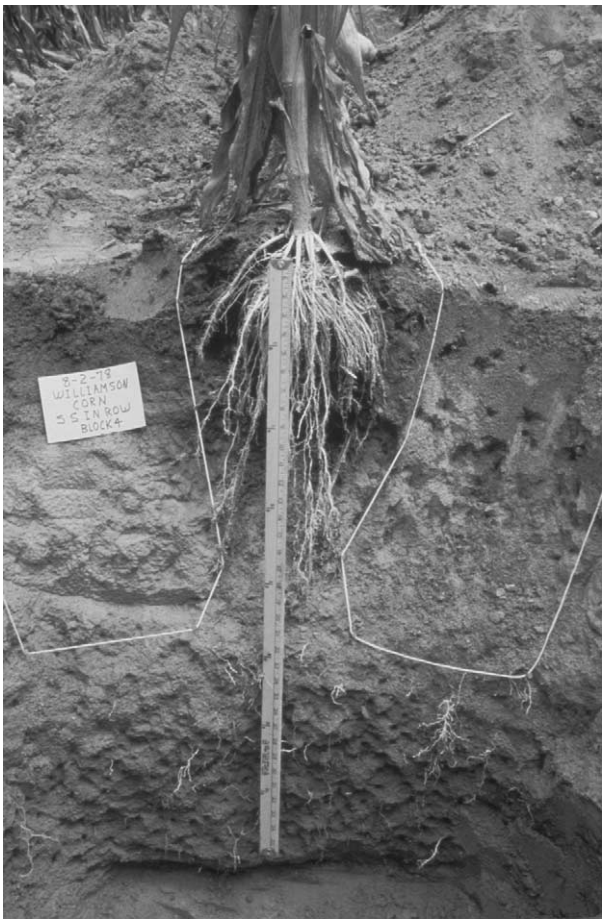


Figure 2 Corn root proliferation restricted to the area where in-row subsoiling (deep tillage) disrupted the eluviated layer in a southeastern Coastal Plain soil.

of soluble elements such as K, Ca, Mg, and NO_3-N. Grassland soils also tend to have a darker color and organic matter accumulation to a greater depth than soils developed under forest. This occurs because trees have larger and longer-living roots than grasses, thus reducing the amount of new organic matter added to the soil each year. Furthermore, most organic matter under forest vegetation is concentrated in the litter layer and remains in the upper few inches of the surface or A horizon. Below that relatively shallow zone in forest-derived soils, the soil organic matter is often more eluviated (leached) than under grassland. The increased amount of organic matter and prolific plant root systems in grassland soils also enable them to be more effective at recycling leached nutrients than forest-derived soils. Nutrient recycling occurs when plant roots absorb ions from the soil solution and transport these materials to aboveground to leaves and other plant parts, which then die and return the nutrients to the soil through litter and root decomposition. As a result of these soil-formation processes, grassland soils (Mollisols) are often considered to have a higher inherent productivity for commonly grown crops than forest soils (Alfisols or Ultisols). The increased aggregation and desirable soil structure enhance air exchange, water entry and retention, and nutrient cycling for subsequent crops.

The eluviated soil layers typically found within forest-derived soils are often more dense and subject to compaction when tilled. This subsurface compaction reduces inherent productivity by decreasing the volume of soil that is available for water retention, air exchange, and plant root development. In the US southeastern Coastal Plain, eluviated horizons are dominated by sand-sized particles (0.05–2.0 mm) and generally have less than 10% by weight of highly weathered kaolinite or similar clay minerals. This results in low plant-available water content, nutrient retention, and organic matter, and thus reduces the absorption of water and nutrients by plants. The most common soil-management practice to enhance the productivity of these soils would be to use some type of deep tillage or mechanical disruption such as in-row subsoiling to loosen the very dense, root-restrictive eluviated soil layers (Figure 2). The short-term benefit of deep tillage on several of the dynamic soil properties is often positive. However, tillage also enhances oxidation of the already low levels of soil organic matter and, in the long term, results in even greater compaction due to the wheel traffic.

Soil Productivity Comparisons

To illustrate soil productivity, selected chemical and physical characteristics, as well as the resultant corn and soybean productivity ratings, for a midwestern Mollisol representing the northern US Corn and Soybean Belt (Nicollet) and an Ultisol from the southeastern Coastal Plain (Norfolk) can be compared (Table 1). The soils differ in their productivity ratings for soybean and corn by 290 kg ha^{-1} and 375 kg ha^{-1} when given a high degree of management. Several factors contribute to crop yield differences, but a comparison of the soil-forming factors, dynamic soil properties resulting from management decisions, and the subtle interactions provide a better understanding of soil productivity.

The cation exchange capacity (CEC) for the two soils reflects inherent differences in clay mineralogy and the amount of organic matter. This impacts productivity by influencing the total amount of cations (primarily Al^{3+}, Ca^{2+}, H^{+}, Mg^{2+}, K^{+}, and Na^{+}) that the soil can retain. However, the specific effect of these differences on nutrient availability (*See* **Nutrient Availability**) and crop yield will depend on the total nutrient management program (*See* **Nutrient Management**). The lower pH within the Norfolk soil (i.e., more acid), coupled with the lower CEC, means that smaller but more frequent lime applications may be needed to optimize productivity. The lower amount of organic matter in the Norfolk soil will affect both soil structure and water-retention characteristics. The lower plant-available water capacity (AWC) means that plants growing on the Norfolk soil must receive rainfall or irrigation water more frequently during the growing season than those growing on Nicollet soil. Fortunately, annual rainfall in the southeastern Coastal Plain is greater, but so is the average temperature. Also, because of the restricted rooting associated with compaction, the volume of soil available for each plant to obtain water and nutrients effectively is generally less for the Norfolk soil than for the Nicollet. The effects of compaction and the lower nutrient and water retention can be partially compensated for by including deep tillage as part of the high-level management plan for Norfolk soils. However, use of a subsoiler/ripper requires approximately 16 l ha^{-1} of diesel fuel,

Table 1 The soil-forming factors, resultant chemical and physical properties, and productivity ratings for a US Midwestern Mollisol and a southeastern Coastal Plain Ultisol

	Nicollet	*Norfolk*
Soil-forming factor		
Parent material	Loamy glacial till, calcareous	Loamy marine sediments, acid
Time	Deposited during late Pleistocene	Deposited during Pliocene/Miocene
Climate	Mean annual temperature 8°C (mean winter −7°C, summer 22°C); annual precipitation 864 mm	Mean annual temperature 17°C (mean winter 7°C, summer 26°C); annual precipitation 1162 mm
Topography	Nearly level to gently sloping glacial till plains and moraines, 1–3% slope	Nearly level to sloping marine terrace uplands, 0–2% slope
Natural vegetation	Prairie vegetation characterized by Indian grass, little bluestem, switchgrass, needlegrass, blue grama, sideoats grama, and scattered trees and shrubs of bur oak, juniper, and sumac	Mixed pine and oak forest vegetation characterized by loblolly, longleaf, slash, and shortleaf pines; yellow poplar; sweetgum; and red and white oaks as overstory, and dogwood and farkleberry as understory
Resultant properties		
Mineralogy	A mixture of minerals in the sand, silt, and clay fractions	Dominated by resistant sand grains and low-activity clays
Drainage	Moderately permeable but somewhat poorly drained, seasonal high-water-table 60–100 cm	Moderately permeable but well-drained, seasonal high-water-table 100–180 cm
Texture	Loam, clay loam, silty clay loam	Loamy sand, sandy loam
Organic matter (g kg^{-1})	40–80	5–20
CEC (cmol kg^{-1})	25–40	1–5
pH	5.6–7.8	4.0–5.5
AWC (cm cm^{-1})	0.17–0.22	0.06–0.11
Productivity ratings		
Corn	7275 kg ha^{-1}	6900 kg ha^{-1}
Soybean	2960 kg ha^{-1}	2690 kg ha^{-1}

CEC, cation exchange capacity. A measure of the ability of the soil to retain positively charged ions such as potassium (K), calcium (Ca), or magnesium (Mg) that are essential for plant growth and development.
AWC, available water capacity. The amount of plant-available water that a soil can retain.

increasing the annual costs of production on that soil and further decreasing economic productivity when computed using output per unit input.

Productivity Ratings

As illustrated by the comparison of Norfolk and Nicollet soils under natural or nonirrigated conditions, the critical functions that strongly influence the relative productivity rating of a soil for crop production are those needed to provide a nonlimiting rooting environment. For maximum productivity, this environment must be capable of meeting the plant's water and nutrient requirements throughout the growing season. The suitability of a soil for plant root proliferation thus provides a scientific basis for several soil productivity indices. One of the most common is:

$$\mathrm{PI} = \sum_{\mathrm{i}=1}^{r} (A_{\mathrm{i}} * C_{\mathrm{i}} * D_{\mathrm{i}} * \mathrm{WF})$$

where PI is the relative productivity index (0–1), A_{i} is the sufficiency of AWC, C_{i} is the sufficiency of bulk density, D_{i} is the sufficiency of pH, WF is a weighting factor representing an idealized plant root distribution, and r is the number of horizons in the rooting depth. Within this index, sufficiency is defined as the pH, D_{i}, and C_{i} levels that do not restrict plant growth or development throughout the life cycle of the species for which the assessment is being made.

This approach for evaluating productivity uses readily available information, assumes a high level of management, assesses only those changes associated with nonreplaceable inputs, and assesses long-term productive potential without annual fluctuations in crop yield. It was used extensively during the 1980s to project the potential effects of soil erosion and demonstrate the usefulness of soils databases, especially the SOILS-5 and the National Resources Inventory (NRI). One conclusion drawn from a study designed to predict the effects of soil erosion in the US Corn and Soybean Belt was that the nonreplaceable productivity would decline by an average of less than 8% during the next 100 years. The rationale for such a small decline was that the soils in that region are typically deep and fertile, and the cropland tends to be on level-to-gently-rolling terrain. However, for soils in landscape positions where the slope exceeds 6%, the decline in long-term productivity could be much greater. Also, for soils with unfavorable subsoil material (e.g., bedrock or gravel), the loss of productivity could be much greater if erosion was allowed to continue.

Processes Reducing Soil Productivity and Practices for Restoring It

Soil erosion, nutrient mining (removal or output greater than input), loss of organic matter, acidification, salinization, water logging, and compaction are among the most common causes for reduced soil productivity. As shown by the productivity index, reductions in plant-available water, nutrient retention, and the creation of a less-favorable environment for plant growth are the primary reasons for loss of soil productivity via erosion.

Erosion primarily removes surface soil, which often has the highest biological activity and greatest amount of soil organic matter. This results in an economic loss of essential plant nutrients and a potential environmental loss if those nutrients are transported to water resources, where they can create problems such as algal blooms and eutrophication. Visual indicators that suggest soil erosion may be reducing productivity include changes in soil color, decreasing soil horizon depth, deposition at field boundaries, or, in desert soils, the presence of moss and algal (crypotogam) crusts. The erosion processes occur not only through the effects of wind and water, but also through mechanical processes such as the removal of soil through the harvest of root crops or tillage and cultivation practices that move soil downslope. Management practices that can help minimize erosion and the subsequent productivity loss include modifying the landscape to control runoff amounts and rates (e.g., terracing, shortening slope length and steepness, strip-cropping, planting shelterbelts) or maintaining a protective cover on the soil (e.g., protecting the surface with crop residue with no-tillage, growing cover crops, improving aggregate stability, increasing water infiltration).

Loss of organic matter decreases productivity by reducing air and water movement within the soil, decreasing cation and anion exchange sites, increasing the potential for crusting and subsequent runoff, and reducing the availability of N, P, and S. Soil organic matter can be lost through erosion or decomposition processes. When soils are tilled, organic matter decomposes faster than under no-till conditions because of changes in water, aeration, and temperature. Although the actual amount of organic matter lost after clearing a wooded area or tilling native grassland will vary, a large amount (e.g., 30–50%) will often be lost within the first 10 years.

Practices that hasten the loss of soil organic matter and potentially decrease productivity include: (1) decreasing the production of plant materials (e.g., replacing perennial vegetation with short-season vegetation, replacing mixed vegetation with

monoculture cropping, using cultivars with high harvest indices, increasing the use of bare fallow, or introducing more-aggressive but less-productive plant species); (2) decreasing the supply of organic materials by burning forest, range, or crop residue or through excessive grazing; and (3) increasing the rate of organic matter decomposition through tillage, drainage, or excessive N fertilization. Strategies to enhance soil productivity by increasing soil organic matter include: (1) increasing plant material production with irrigation, optimum fertilization, use of cover crops, reforestation, or restoration of grassland; (2) increasing organic matter supply by protecting from fire (provided the area is not degraded grassland, where fire actually improves productivity by preventing encroachment of shrubs and other woody plant species), applying animal manure or other carbon-rich materials, and controlling insects and rodents; or (3) decreasing rates of decomposition by reducing or eliminating tillage, preventing excessive drainage, or reducing soil temperature by preserving surface vegetative cover.

Acidification or decreases in soil pH often occur because of the excessive application of nitrogen fertilizers. Acidity reduces productivity by decreasing the availability of essential plant nutrients such as Ca, Mg, or P; decreasing CEC through accelerated mineral weathering; and increasing the solubility of Al, which can interfere with plant root development. Acidity also affects the activity of many microorganisms responsible for organic matter decomposition and chemical transformations within the soil. A soil pH between 6.6 and 7.3 is favorable for microbial activities that contribute to the availability of N, S, and P in soils.

Salinization decreases soil productivity by limiting the ability of plants to take up water. It occurs in response to any practice that affects the soil–water balance and allows water-soluble salts to accumulate in the soil. Irrigation, drainage, changes in plant community composition that alter soil cover or rooting characteristics, and summer fallow are some agricultural management practices that can result in salinization if the climate and hydrology conditions are favorable. The process can occur on the soil surface whenever the following conditions occur together: soluble salts, such as sulfates (SO_4^{-2}) of Na, Ca, or Mg are present in the soil, the water table is near the soil surface, evaporation rates are high, or annual rainfall is low. In semiarid areas, this process often results in saline seeps on the rims of depressions and edges of drainage-ways, at the base of hill-slopes, and in flat, low-lying areas surrounding sloughs or shallow bodies of water.

Early signs of salinization include: (1) increased soil wetness in semiarid and arid areas to the point that the soil does not support agricultural equipment, (2) growth of salt-tolerant weeds, and (3) irregular patterns of crop growth and lack of plant vigor. Advanced signs include a white crusting on the surface, a broken ring pattern of salts adjacent to a body of water, white spots and streaks in the soil (even where no surface crusting is visible), or the presence of weeds or other naturally growing, salt-tolerant vegetation.

Restoring soil productivity in saline areas is done primarily through water management. Sufficient irrigation water should be applied to leach salts below the root zone, but care should be taken not to overirrigate, especially in areas that do not have adequate natural drainage. It is very important to provide suitable drainage water outlets that do not degrade water bodies or other natural resources into which the drainage water flows. The use of cropping and tillage systems that promote adequate infiltration and permeability can also help restore productivity. This includes building soil organic matter for better soil aggregation and avoiding compaction. Planting salt-tolerant crops that use available soil moisture, removing excess water from recharge areas with actively growing, deep-rooted plant species such as alfalfa, and reducing summer fallow by continuous cropping are among the most effective cropping practices for controlling salinization, restoring productivity, and minimizing negative impacts on the environment.

Soil compaction reduces productivity by restricting rooting depth and the uptake of water and nutrients. Compaction occurs when soil particles are pressed together, increasing the weight of solids per unit volume of soil (bulk density). This reduces pore size and volume, increasing the proportion of water-filled to air-filled pore space. Although in some soils compaction can increase thermal conductivity, causing the soil temperature to be higher than under noncompacted conditions, soils in humid regions with a high water content often have a lower soil temperature, which suppresses the activity of many thermophilic and mesophilic microorganisms that drive soil organic matter decomposition and the subsequent release of nutrients. Compaction also decreases infiltration, increasing the potential for runoff and erosion.

Tilling, harvesting, or grazing when soils are too wet are some of the primary causes for soil compaction. The depth at which compaction occurs, the shrink–swell potential of soil (which is determined by the clay mineralogy), and climate are critical factors determining the persistence of problems and overall impact on productivity. Subsoiling or deep tillage may be necessary to break up compacted layers in some soils. Applying animal manure, compost,

newspaper, or woodchips with an appropriate C:N ratio, or municipal sludge that does not contain toxic compounds can help restore productivity by improving soil structure and biological activity. The negative effects of compaction on soil productivity can also be minimized by reducing wheel traffic, the pressure or weight of equipment, and maintaining or increasing soil organic matter.

Summary

Soil productivity encompasses soil fertility plus the inherent and management-related factors affecting plant growth and development. It is generally measured in terms of inputs versus outputs, which for agronomic situations generally refers to water and/or nutrient input versus crop yield. The critical soil functions influencing productivity within any soil are those that provide physical support, a rooting medium with plant-available water, air for respiration, and essential nutrients. Humankind can also have a tremendous impact on soil productivity through its effects on the dynamic soil properties. Agricultural management decisions regarding tillage, fertilization, crop rotation, irrigation, and drainage are among the practices that can significantly affect soil productivity.

See also: **Nutrient Availability**; **Nutrient Management**

Further Reading

American Society of Agricultural Engineers (1985) *Erosion and Soil Productivity*. Publication 8-85. St. Joseph, MI: ASAE.

Foth HD and Ellis BG (1997) *Soil Fertility*, 2nd edn. Boca Raton, FL: CRC Press.

Francis CA, Flora CB, and King LD (eds) (1990) *Sustainable Agriculture in Temperate Zones*. New York: John Wiley.

Hatfield JL and Karlen DL (eds) (1994) *Sustainable Agriculture Systems*. Boca Raton, FL: Lewis Publishers/CRC Press.

Hatfield JL and Stewart BA (eds) (1992) *Limitations to Plant Root Growth, Advances in Soil Science,* vol. 19. NY: Springer-Verlag.

Hillel D (1991) *Out of the Earth: Civilization and the Life of the Soil*. Berkley, CA: University of California Press.

Lal R (ed.) (1999) *Soil Quality and Soil Erosion.* Ankeny, IA: Soil Water Conservation Society.

Larson WE, Foster GR, Allmaras RR, and Smith CM (eds) (1990) *Proceedings of Soil Erosion and Producitivity Workshop*. St. Paul, MN: University of Minnesota.

Magdoff F and van Es H (2000) *Building Soils for Better Crops*. Burlington, VT: Sustainable Agriculture Publications.

National Academy of Sciences (1994) *Soil and Water Quality: An Agenda for Agriculture*. Washington, DC: National Academy Press.

Profile *See* **Morphology**

PROTOZOA

W Foissner, Universität Salzburg, Salzburg, Austria

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Protozoa are unicellular, heterotrophic, eukaryotic organisms comprising four organization types: amebae, flagellates, ciliates, and parasitic sporozoans. About 1600 species, of which some are restricted to certain geographic regions, are known to live in terrestrial habitats; however, at least the same number is still undiscovered. Small body size and the ability to produce protective resting cysts are the main adaptations of protozoa to the peculiarities of the soil environment. Many soil protozoa feed, more or less selectively, on bacteria, while others are omnivorous or highly specialized fungal feeders. Protozoa (active and cystic) inhabit the soil in great numbers, that is, some 10 000–1 000 000 individuals per gram dry mass, and produce many generations annually. They significantly enhance the flow of nutrients and growth of plants and earthworms. Accordingly, they are important soil inhabitants, and studies on their dynamics and community structures thus provide a powerful means for assessing and monitoring changes in biotic and abiotic soil conditions. Unfortunately, methodological and taxonomical problems still limit the use of protozoa as bioindicators in terrestrial environments.

This article covers the diversity, ecology, and bioindicative value of soil protozoa. The knowledge on autotrophic soil protists, which mainly live on the soil surface because they depend on light, is still in its infancy. However, they play an important role, especially as a symbiotic partner of lichens in the crust soils of extreme regions, such as desert and high mountain areas.

A Brief History of Soil Protozoology

Although there were predecessors, such as Ehrenberg, Goodey, and Russel and Hutchinson, soil protozoology as a science was established by the Austrian–Hungarian naturalist H.R. Francé and the English microbiologist D.W. Cutler in the 1920s. Their monographs showed that unicellular organisms inhabitate the soil in great number and diversity, and thus contribute significantly to soil fertility and sustainability. Francé's booklet saw many editions, the latest in the 1960s.

This first bloom of soil protozoology followed a long period where a few specialists added significantly to the knowledge available, mainly J.F. Darbyshire in the UK, L. Bonnet in France, J.D. Stout in New Zealand, S.S. Bamforth in the USA, and V.F. Nikolyuk in Russia. It was only in the last two decades of the twentieth century that soil protozoology was rejuvenated by the studies and reviews of W. Schönborn in Germany, F. Ekelund and R. Rønn in Denmark, W. Foissner in Austria, and J.F. Darbyshire, who edited the first book devoted entirely to the biology and ecology of soil protozoa. Today, the field is well established but still a domain of a limited number of specialists, while general soil ecologists have problems to incorporate the new knowledge appropriately, in contrast to marine and limnetic plankton researchers, where the 'microbial loop,' that is, the intimate connection of dissolved organic matter, bacteria, and protists, greatly changed views and research strategies.

Diversity and Biology

Diversity and Geographic Distribution

About 23 000 free-living, extant protozoa are known. Of these, roughly 1600 were originally described or later reported from terrestrial habitats (Table 1). For a long time, the soil protozoan community was considered as an ubiquitous part of limnetic species and some even suggested that nearly the same species of protozoa occur in soil, sewage, and activated sludge. However, more detailed research disproved these findings and showed that they were based on misidentification of species. Today it is widely accepted that the main part (>70%) of the soil protozoan species evolved in and are restricted to terrestrial environments. Refined morphological methods and gene sequence data indicate that a considerable portion of soil and freshwater protists have a restricted geographical distribution (Figures 1 and 2), disproving the old hypothesis that, in microorganisms, 'everything is everywhere, the environment selects.' This matter, however, is still under discussion.

Table 1 Known and proposed diversity of soil protozoa

Taxa	*Known species*	*Proposed species*
Naked amebae[a]	60	600
Testate amebae	300	500
Flagellates	270	600
Ciliates	1000	2000
Microsporidia and sporozoans	3	?

[a]Without several hundreds of slime molds.

Table 1 shows the known and proposed soil protozoan diversity. While numbers of testate amebae and ciliates are supported by detailed data, those of the other groups are very speculative. Likewise, little is known of protozoan parasites. But even the known soil protozoan diversity is high and a major problem in basic and applied soil biology because comprehensive identification literature is lacking and few taxonomic specialists are available. The decreasing interest of the public and science in classical taxonomy, will soon dramatically increase our ignorance. Although molecular methods likely can solve some of the problems, nobody will be available to add names to the sequences. Thus, great efforts are necessary to educate young scientists in classical alpha-taxonomy.

Morphological and Physiological Adaptations

The soil is a very special environment for protozoa. Three factors are most important for their existence and activity: (1) the structure of the porous space; (2) the astatic water supply; and (3) intraspecific and interspecific interactions. These factors caused two basic adaptations: small body size and the universal capability to produce protective resting cysts. These and some other important adaptations will be briefly described.

Morphological adaptations Mean body length, mean body width, and mean biomass are significantly smaller in soil ciliates and testate amebae than in ciliates and testaceans from fresh water. The small size, and volume, is achieved either by an overall reduction in size or by reduction of body width. Many soil ciliates are thigmotactic creepers with a reduced ciliature and a flattened and/or worm-like

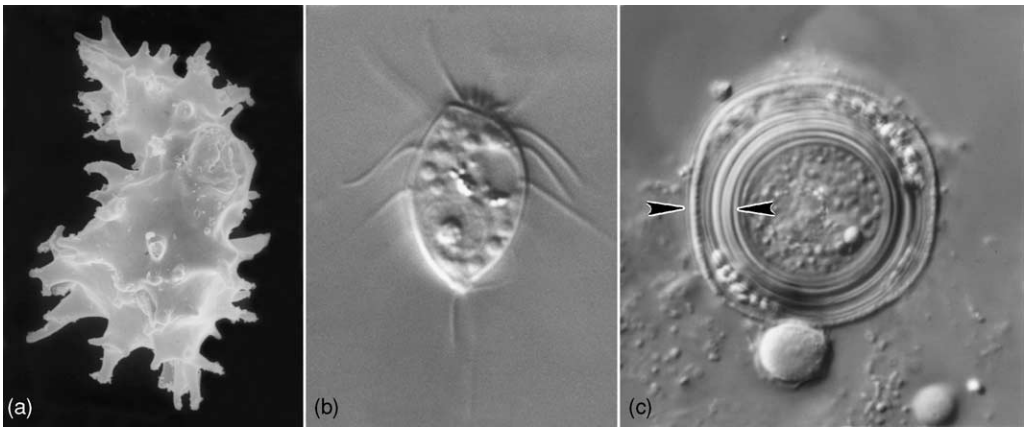


Figure 1 Small-sized and cystic soil protozoa. (a) A naked ameba about 30 μm long with many tiny pseudopodia which can exploit even minute soil pores for food. (b) A 15 μm long soil flagellate, *Hemimastix amphikineta*, which only occurs in southern terrestrial environments, that is, in Gondwana. Note the two rows of flagella which make the organism look like a ciliate. (c) Resting cyst (diameter about 35 μm) of the ciliate *Exocolpoda augustini*. This ciliate occurs mainly in hot and dry soils, e.g., the Namib desert, and thus has a very thick cyst wall (arrowheads) composed of many membranes. The organism can remain viable in this condition for years.

body (Figure 3b). These features and several physiological peculiarities produced a soil ciliate community of unique structure, characterized by a high proportion of colpodids (Figure 3a and 3b) and hypotrichs and a strong underrepresentation of sessile peritrichs and suctorians compared with fresh water ciliate communities. The soil testacean community is characterized by having a high proportion of species with reduced pseudostome size (opening of the shell), a flattened ventral side, and a globular shape (Figure 2). Soil naked amebae and flagellates have similar adaptations as ciliates and testate amebae (Figure 1a and 1b): most are very small and the amebae, for instance, can exploit with their tiny and flexible pseudopodia micropores which have a diameter of only 1 μm.

Physiological adaptations Few physiological adaptations have been investigated in detail; it is probable that several are still to be discovered. Some physiological adaptations cause conspicuous morphological specializations.

Food and feeding Many soil protozoa, especially most naked amebae and flagellates, entirely or partially feed on bacteria, skimming off the dense bacterial standing crop provided by many soils. However, under natural conditions soil protozoa cannot significantly reduce bacterial abundance, and the absolute amount of organic matter consumed by the testacean community is small – 234 mg m^{-2} during a 3-month period.

In soil ciliates, 39% feed mainly on bacteria, 34% are mainly predaceous, and 20% are omnivorous. Some are strictly mycophagous and the most characteristic species of the soil ciliate community. They have reduced the classical ciliate mouth almost completely and evolved a minute feeding tube producing discrete holes in the hyphae and spores of fungi and yeasts (Figure 3b–e). This type of feeding has been termed 'perforation lysis' and is also found in several naked amebae. There is evidence from pot experiments that such activities can reduce the inoculum level of plant pathogenic fungi in soils and can reduce the severity of take-all disease of wheat by the fungus *Gaeumannomyces graminis* var. *tritici*. Some flagellates are also strictly mycophagous, but take up the fungi by an ordinary cytostome. Many soil testaceans seem to feed on humus particles and/or fungal hyphae and spores, but other materials such as algae, protozoa, and bacteria are also ingested; some are polyphagous. A Gondwanan species, *Apodera vas*, can even feed on nematodes (Figure 2b). In some other species, *Difflugia lucida* and *Schoenbornia humicola*, a remarkable specialization has been described: they collect and store humus particles around the pseudostome during optimal periods and transport them into the cytoplasm during suboptimal environmental conditions (Figure 2c).

r/K-*selection* Field data and laboratory experiments show that colpodid ciliates are more *r*- than *K*-selected. Thus, they are particularly diverse and frequent in unpredictable and extreme habitats, such as leaf surfaces, lichens, and alpine and desert soils. In contrast, polyhymenophoran ciliates (mainly hypotrichs) are more *K*-selected and thus particularly diverse and frequent in predictable habitats (Figure 4). The ratio of these ciliate groups proved to be a sensitive indicator for biotope stability and can thus be used in studies on human-influenced soils.

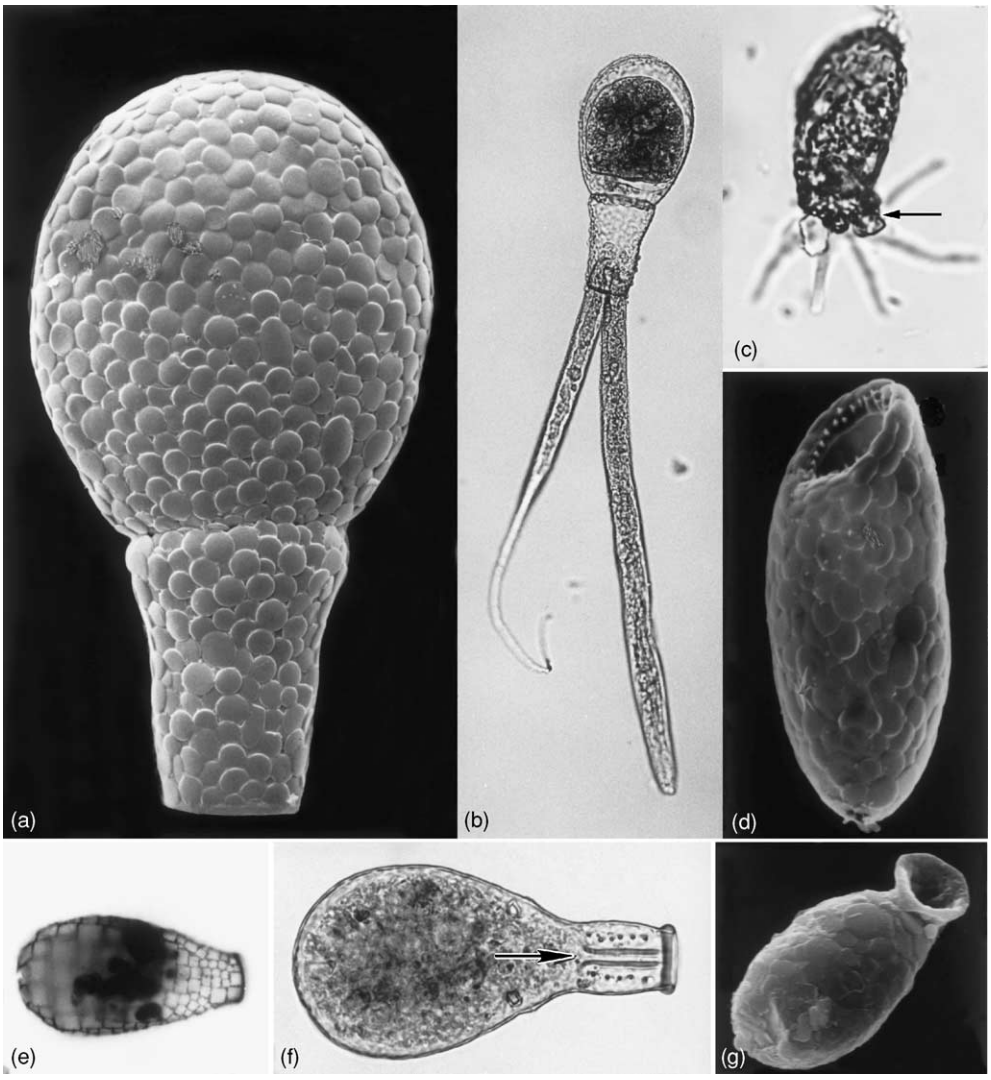


Figure 2 Soil testate amebae. (a, b) *Apodera vas* is about 170 μm long and only occurs in the southern hemisphere, i.e., in Gondwana. It can feed on nematodes (b) and build its shell of platelets taken from the shell of another prey organism, the testate ameba *Trinema lineare* (d). (c) *Difflugia lucida* is about 60 μm long and produces food-bundles (arrow) around the shell opening under optimal environmental conditions. These food accumulations are ingested and digested when conditions are less favorable. (d) *Trinema lineare* is about 35 μm long and one of the most common soil testate amebae globally. Its shell is made of silicium platelets, which *A. vas* uses for building its own shell (a). (e) *Quadrulella* makes the test of quadrangular silicium platelets. (f) *Nebela certesi* is about 150 μm long and only occurs in the southern hemisphere (Gondwana). Note the oral tube (arrow) surrounded by minute teeth. (g) *Edaphonobiotus campascoides*, a soil testate ameba about 35 μm long has, as yet, been found only in central Europe (Laurasia). Its shell is highly similar to that of some marine sand testaceans.

Testate amebae are, in general, on the *r* side within the *r/K*-continuum, when compared with ciliates, naked amebae, and flagellates. However, a more fine-scaled rating suggests that autochthones are more *K*-selected than allochthones.

Cysts All soil protozoa survive detrimental environmental conditions in a dormant stage, termed 'protective resting cyst.' When encysting, the cell rounds up, loses water, and secretes a thick wall (Figure 1c). In encysted condition, protozoa can withstand, for instance, decades of drought and food depletion, liquid nitrogen, and strong acids. When conditions become favorable again, the cell excysts and commences feeding and reproduction. The receptors indicating favorable conditions to the encysted cell are still not known. Likewise, few reliable data are available on the factors stimulating excystment, but substances secreted by potential food organisms and CO_2 might be important. A slight increase in the CO_2 level to 0.5–0.8% enhanced excystment of soil amebae, but a further increase inhibited it.

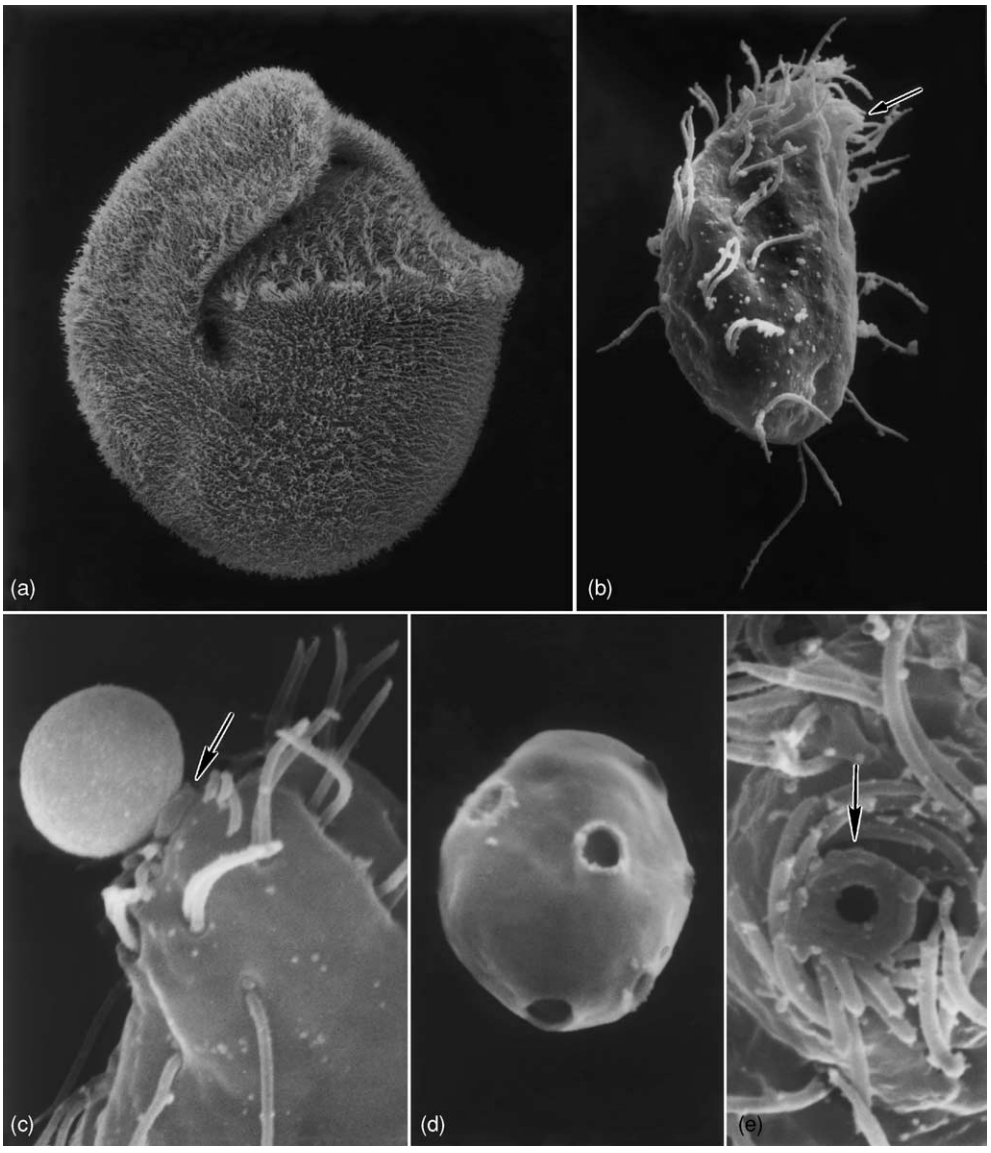


Figure 3 (a) *Bresslauides discoideus* is a colpodid ciliate up to 600 μm large and likely restricted to the northern hemisphere. It is a predator with a large mouth and one of the largest ciliates occurring in soil. (b–e) *Pseudoplatyophrya nana* is a minute species about 25 μm long that feeds only on yeasts and fungi. It feeds by perforation lysis, that is, has a minute oral tube (arrows) that enzymatically produces small holes in the yeast's cell wall (c, d).

There is evidence that cysts are adapted to the overall environmental conditions in which the trophic cells live. Cysts formed, for instance, in arid environments can survive desiccation for decades, while those produced in rainforests survive drought for only a few weeks or months. The same is true for temperature. In experimental microcosms, diversity of soil ciliates and reproduction are highest at or near the mean annual temperature of the habitat, that is, around 20°C for soils from the region of Bonn, Germany, and 25–30°C for those from an African savannah. These data show the existence of many local physiological races adapted to the specific conditions of the habitat.

There are further cyst variations in soil protozoa. Testate amebae can produce a so-called 'precystic stage' for bridging short unfavorable periods, an adaptation to the often quickly changing soil environment. When forming this stage, the organism produces a thin membrane (epiphragm) closing the shell opening (pseudostome), and rounds up, but does not form a thick cyst wall. During excystment, the epiphragm is lysed or digested. Another spectacular cyst is produced by the epiphytic, colpodid ciliate genus *Sorogena*. Under favorable conditions, it multiplies rapidly, aggregates, encysts, and produces noncellular stalked aerial sori which look highly similar to those

of cellular slime moulds. The sorogenesis requires alternate light and dark conditions, just as in the pink slime mold *Acrasia rosea*.

Anabiosis Anabiosis (cryptobiosis, anhydrobiosis) is the property of organisms to resist desiccation without marked morphological changes, such as the production of a special cyst or membrane. It is common in some groups of soil- and moss-inhabiting animals, such as rotifers, tardigrades, and nematodes, but is very restricted among soil protozoa, where it likely occurs in a few naked amebae and in a suctorian ciliate, as I myself have observed.

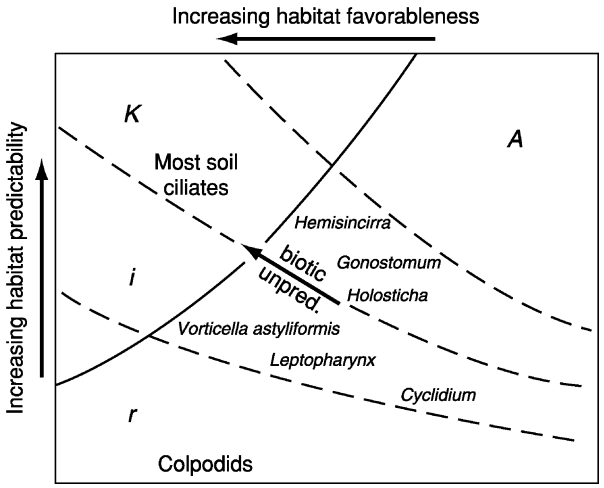


Figure 4 Ordination of soil ciliate taxa on a two-dimensional template. Colonizing *r*-organisms, colpodids, characterize ephemeral habitats of low resources (lower left to lower right of diagram). A few intermediate *K*-selected species (e.g., *Cyclidium* and *Holosticha*) can adapt to harsh conditions. Species diversity increases towards the upper left corner of the diagram where the abiotic factors of the habitat improve, enabling more biotically unpredictable interactions due to competition and predation. *A*, adversity selection; *i*, intermediates. Reproduced with permission from Bamforth SS (2001) Proportions of active ciliate taxa in soils. *Biology and Fertility of Soils* 33: 197–203.

Number and Biomass

In soil, protozoa are present as active (trophic) and inactive (encysted, dormant) cells. It is likely that many soil protozoan species are encysted most of their life. Thus, the encysted portion of the total number is usually much larger than the active. There is currently no reliable method available for separating active and dormant specimens. Furthermore, most counts rely on culture techniques followed by a most probable number (MPN) calculation of protozoan abundance. However, only few of the many protozoan species present excyst and grow to detectable numbers under the limited conditions of such cultures. All these methodological shortcomings mean that many of the quantitative data available in the literature are doubtful, especially as concerns the active and thus ecologically relevant portion. Only numbers obtained by direct microscopic counts in soil suspensions are reliable (Table 2). Unfortunately, naked amebae and flagellates can hardly be counted with direct methods because they strongly adhere to the soil particles and thus cannot be seen. Accordingly, all numbers on active naked amebae and flagellates are highly questionable.

In spite of these problems, it is clear that protozoa inhabit soil in great numbers, that is, some 10 000–1 000 000 individuals per gram dry mass (Tables 2–4, Figure 5). Biomass tends to correlate with abundances, and values between 0.1 g and 5 g m^{-2} are frequent. Naked amebae and flagellates (active plus cystic) usually show numbers between 2000 and 1 000 000 g $^{-1}$ dry mass of soil. Standing crop numbers of active testate amebae are moderately high i.e. 100–1000 individuals per gram dry mass in arable land and mineral soils, 1000–10 000 in meadow and grassland topsoils, and 10 000–100 000 in forest leaf litter and upper humus horizons. However, as many testate amebae are rather voluminous, their standing crop and production biomass often surpasses that

Table 2 Mean numbers of active ciliates per gram dry meadow soil (0–5 cm), spruce litter (0–2 cm), and beech litter (0–2 cm)

	Direct counts[a]		*MPN simplified*[b]		*MPN complete*[c]	
Sample	*Mean*[d]	*CV (%)*[e]	*Mean*[d]	*CV (%)*[e]	*Mean*[d]	*CV (%)*[e]
Meadow ($n = 7$)	60*	80	102 000*	177	88 000*	207
Spruce ($n = 3$)	400+	57	79 000*	70	71 000*	93
Beech ($n = 3$)	8100++	13	136 000*	56	127 000*	65

[a]Microscopic counts in soil suspensions.
[b]Dilution culture method by Singh and most probable number (MPN) equation by Fischer.
[c]Dilution culture method by Singh and refined MPN equation as available in computer programs.
[d]Values followed by the same symbol are not significantly different ($P \leq 0.05$).
[e]Coefficient of variation.
Reproduced with permission from Foissner W (1999) Soil protozoa as bioindicators: pros and cons, methods, diversity, representative examples. *Agriculture, Ecosystems and Environment* 74: 95–112.

Table 3 Effects of irrigation on the microfauna of a spruce forest soil

Parameters [a]	*Soil depth (cm)*	*Irrigated* [b]	*Control*	n	*Statistics*
Soil moisture (% wet mass of air-dried soil)	0–3	$50.0^{*} \pm 9.9$	$41.3^{\dagger} \pm 12.6$	15	ANOVA ($P \leq 0.05$)
	3–9	$48.8^{*} \pm 4.3$	$44.7^{*} \pm 5.9$	8	ANOVA ($0.2 \geq P > 0.1$)
Ciliates					
Abundance	0–3	$311^{*} \pm 141$	$489^{\dagger} \pm 258$	15	U-test ($0.05 < P \leq 0.1$)
	3–9	$10^{*} \pm 10$	$14^{*} \pm 24$	8	U-test ($P \geq 0.2$)
Species number	0–3	$12.8^{*} \pm 4.5$	$8.4^{\dagger} \pm 3.3$	15	ANOVA ($P \leq 0.005$)
	3–9	$2.4^{*} \pm 2.1$	$1.6^{*} \pm 1.9$	8	U-test ($P \geq 0.2$)
Testate amebae					
Abundance					
Living	0–3	$22\,203^{*} \pm 7040$	$17\,908^{*} \pm 3175$	4	ANOVA ($P \geq 0.2$)
Empty tests	0–3	$361\,720^{*} \pm 17\,204$	$319\,081^{*} \pm 58\,390$	4	ANOVA ($P \geq 0.2$)
Species number					
Living	0–3	$11.5^{*} \pm 1.3$	$9.8^{\dagger} \pm 1.0$	4	ANOVA ($0.05 < P \leq 0.1$)
Total	0–3	$18.0^{*} \pm 0$	$19.8^{*} \pm 2.2$	4	U-test ($P \geq 0.2$)
Nematodes					
Abundance	0–3	$1197^{*} \pm 349$	$824^{\dagger} \pm 349$	15	ANOVA ($P \leq 0.01$)
	3–9	$308^{*} \pm 80$	$391^{*} \pm 173$	8	ANOVA ($P \geq 0.2$)
Rotifers					
Abundance	0–3	$227^{*} \pm 88$	$181^{*} \pm 122$	15	ANOVA ($P \geq 0.2$)
	3–9	$38^{*} \pm 25$	$31^{*} \pm 18$	8	U-test ($P \geq 0.2$)

[a]Abundances (individuals g^{-1} dry mass of soil; arithmetic mean $\pm$ standard deviation) were estimated with a direct counting method in soil suspensions. Values followed by the same symbol are not significantly different.
[b]Irrigated plot (15 m^2) received 25 l m^{-2} water every fourth day.
ANOVA, analysis of variance.
Reproduced with permission from Foissner W (1999) Soil protozoa as bioindicators: pros and cons, methods, diversity, representative examples. *Agriculture, Ecosystems and Environment* 74: 95–112.

Table 4 Comparison of community parameters for soil testaceans from different sites of the world

Parameters	*Moss under beech forest (moder)*	*Ash-maple forest (mull)*	*Alpine rendzina (moder)*	*Alpine mat (moder)*	*Aspen woodland (mor)*	*Beech forest (mull)*
Annual mean density ($\times 10^6\ m^{-2}$)	1.7	33.6	39.6	20.1	261	84
Annual mean biomass (mg m^{-2})	15.5	1033	2209	1165	723	1715
Number of generations ($year^{-1}$)	16.0	12.5	ND	8	90	ND
Mortality rate (% day^{-1})	3.0	8.5	ND	ND	10.8	ND
Production numbers ($\times 10^6\ m^{-2}$)	145	940	29	118	90 930	358 000
Production biomass (g m^{-2})	0.11	25.9	1.4	5.4	206	73
Biomass turnover (P_B/B)	8.1	43.9	0.7	5	285	43

ND, not determined.
Reproduced with permission from Foissner W (1987) Soil protozoa: fundamental problems, ecological significance, adaptations in ciliates and testaceans, bioindicators, and guide to the literature. *Progress in Protistology* 2: 69–212.

of all other protozoans (Table 4). Thus, their omission in many recent studies is an unfortunate mistake. Abundances of active ciliates are low in evolved soils (humus and mineral horizons), usually <100 cells g^{-1} (Tables 2 and 3). However, up to 10 000 individuals per gram dry mass can be reached in developing soils and leaf litter (Table 2), showing that ciliates are important during the early stages of soil formation.

Vertical and Horizontal Distribution

Like as many soil organisms, protozoa are usually most abundant and diverse in the upper 0–10 (30) cm soil layer, where food is concentrated in the form of leaf litter and the rhizosphere. However, there is a distinct, not yet fully understood, niche separation of ciliates and testate amebae. While active ciliates occur mainly in the uppermost (fresh) leaf litter layer, where abundances of up to 10 000 individuals per gram dry mass are reached, testate amebae peak underneath, that is, in the evolving humus horizon ('fermentation layer'; Table 3). Thus, habitats such as meadow topsoils and arable lands contain very few active ciliates, usually $<100\ g^{-1}$ dry mass, although a lot of viable resting cysts are present. The inability to excyst (germ) in evolved soils has been termed ciliatostasis,

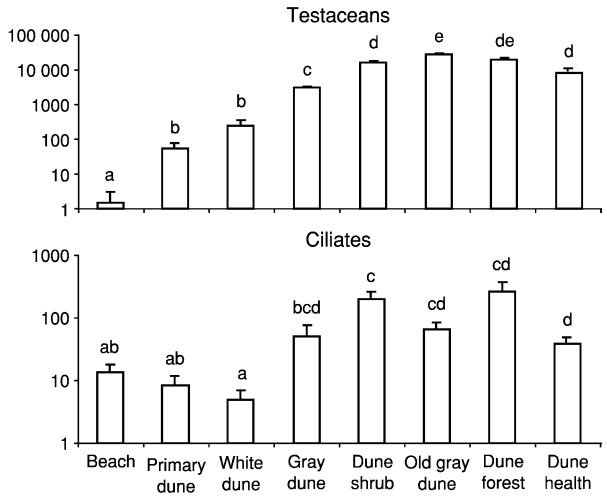


Figure 5 Abundance (individuals g^{-1} dry mass of soil) of active testate amebae and ciliates in dune soils along a series of soil (dune) development. Significant differences ($P < 0.05$) are indicated by different letters. Reproduced from Verhoeven R (2002) The structure of the microtrophic system in a development series of dune soils. *Pedobiologia* 46: 75–89.

in analogy to fungistasis. The reasons for ciliatostasis are not yet clear, but enrichment of certain detrimental substances during humification, density-dependent processes, and general food depletion are discussed.

Little is known about the horizontal distribution of soil protozoa. However, a very clumped distribution is evident from several faunistic and 'hot spot' investigations. Recent data suggest that spatial organization of protozoan communities occurs at scales below 10 cm.

The 'Water Paradox'

Protozoa are aquatic organisms, and free water is thus indispensable for their life. Accordingly, it appears reasonable to assume that water is the key factor for soil protozoan abundance and diversity. Surprisingly, carefully designed field and laboratory studies do not support this assumption, but over the years have provided many conflicting results. Two examples from recent studies should demonstrate the 'water paradox.'

In one study, aspen woodland plots were irrigated during the seasonal dry period every 3 days with 3.78 l m^{-2} and testate amebae were sampled on three occasions. The general effect of adding water to the soil was to increase significantly the number of active testaceans and to decrease the number of encysted specimens. Numbers of active testaceans and of total living testaceans (active plus encysted) showed a significant, positive correlation with soil moisture content. At higher soil moisture levels, the larger (>60 μm) species present tended to have a larger proportion of living individuals as cysts than did the smaller forms which numerically dominated the population. Generation times were shorter and secondary production higher in the watered plots.

In the other study, spruce forest plots were irrigated during the seasonal dry period every 4 days with 25 l m^{-2} (Table 3). This caused a marked decrease ($P \leq 0.05$) in the abundance of the active ciliates in the litter layer; species richness, in contrast, increased in both ciliates and testate amebae, while testacean abundance did not increase significantly. The nematode numbers increased dramatically by about 45%, while rotifer abundance did not change significantly.

As yet, there is no convincing explanation for these and many other contrasting results. However, most soil protozoologists agree on the basic assumption that water is a key factor for protozoan life, not only because they basically depend on water but also due to several morphological peculiarities most meaningfully explained as adaptations to the astatic water regime of the soil environment in general. I suggest, very briefly, two hypotheses for explaining the 'water paradox':

1. Water correlates with other, not yet recognized factors, especially the presence or growth of appropriate food organisms. This hypothesis is supported by the observation that laboratory microcosms, where food has been tested to be edible for the protozoans under investigation, show more consistent results, that is, increasing protozoan abundance and biomass with increasing soil moisture.
2. Protozoa can tolerate a wide range of soil moisture; only under extreme conditions do communities react. This hypothesis is supported by the occurrence of local, physiological races and the ability of many protozoan species to live in very thin films of water covering the soil pores and soil particles. In the spruce forest experiment described above, the litter of the control plot was hand-dry; most of the water present was a constituent of the needles (Table 3).

Ecological Significance

Standing Crop, Respiration, and Annual Production

Several ecosystem studies show consistently that the microflora (bacteria and fungi) accounts for about 90% of the heterotrophic soil respiration. Only about 10% of the total comes from the 'animals,' and about 70% of the 'animal' respiration is due to the protozoa, although their contribution to the standing crop is 'only' approximately 30% (Figure 6). This is due to their small size and fast reproduction. However, in earthworm-rich soils the proportion of

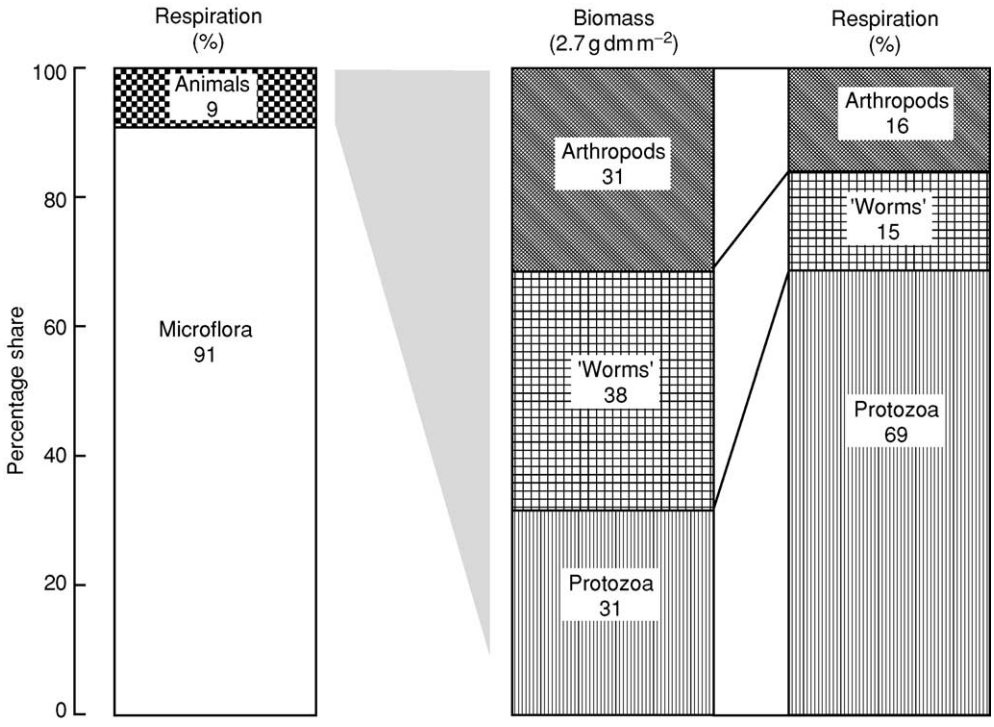


Figure 6 Contribution of protozoa to dry biomass (dm) and respiration of soil organisms. The graphs show the mean of 14 ecosystem studies from various sites of the world. Reproduced with permission from Foissner W (1987) Soil protozoa: fundamental problems, ecological significance, adaptations in ciliates and testaceans, bioindicators, and guide to the literature. *Progress in Protistology* 2: 69–212.

protozoa to the total biomass is considerably lower than stated above, while it is higher in extreme regions, where earthworms are few or lacking, for instance, in coniferous forests, deserts, and high mountains.

Field studies of production of soil protozoa are mostly restricted to testate amebae, the shells of which can be used to estimate their mortality. The annual production is in the same order of magnitude as that of the earthworms, especially when productions of naked amebae and flagellates are added (Table 4). The high contribution of the protozoa to the total annual animal production is due to the many generations produced and shows that protozoa play an important role in the energy flux of the soil ecosystem.

Mineralization and Plant Growth

Considering the fast growth of protozoa and their high annual production (Table 4), it is not surprising that they enhance flows of nutrients in soil to the benefit of plants and microorganisms. Experiments suggest that a complex 'microbial food-web' is operating rather than a simple 'microbial loop' during decomposition and growth. Convincing evidence is available that protozoa enhance plant growth significantly ($\geq$20%) either directly by increasing the availability of nitrogen in soil and leaching water and/or indirectly by nonnutritional effects, such as selective grazing on microbial populations, modification in the concentrations of plant hormones in the rhizosphere, or suppression of pathogenic bacteria. This matter is still under discussion. About 40% of the ingested microbial biomass N will be excreted by protozoa, leading to a constant release of nutrients available for plant uptake. Protozoa are not only responsible for respiring about 10% of the total carbon input but also for 20–40% of net N mineralization in the field. Furthermore, protozoa enhance growth of the plant root system and thus nutrient uptake (Figure 7).

Protozoa also cause a strong decrease in phosphate leaching. Phosphate is one of the most immobile nutrient ions in soil and continuous microdisturbance by protozoa and other microfaunal organisms (e.g., nematodes) is likely particularly important for microbial phosphate availability.

Earthworm Growth

Early reports that earthworms feed on protozoa and are an essential constituent in the diet of the compost worm *Eisenia fetida* to reach sexual maturity were recently confirmed and extended by carefully designed experiments. These showed that *Aporrectodea caliginosa* actively searchs for places with high

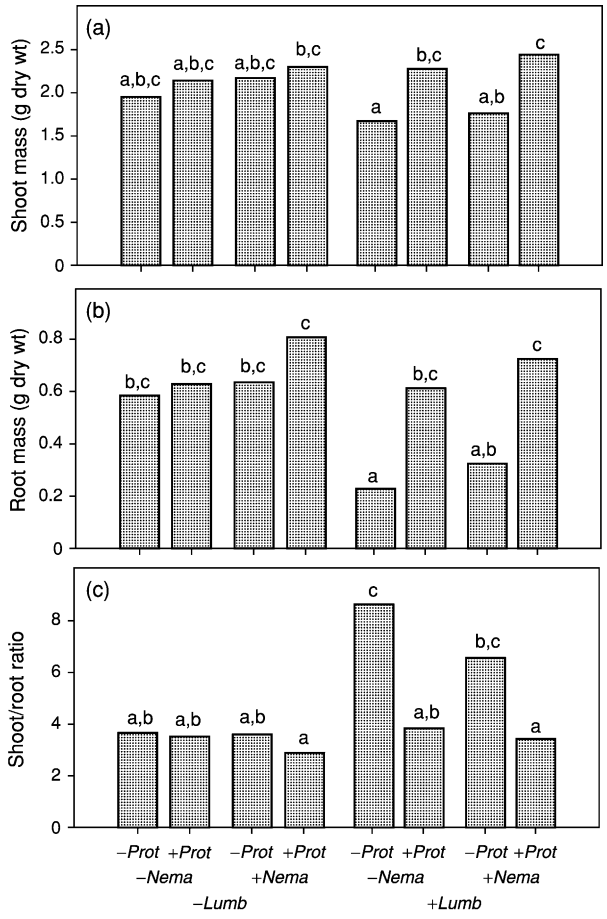


Figure 7 Effects of protozoa (−*Prot* without protozoa, +*Prot* with protozoa), nematodes (−*Nema* without nematodes, +*Nema* with nematodes) and earthworms (−*Lumb* without earthworms, +*Lumb* with earthworms) on (a) shoot biomass; (b) root biomass; and (c) shoot/root ratio of *Hordelymus europaeus*, a grass, in experimental chambers at the end of the experiment (112 days). Bars sharing the same letter are not significantly different. $P > 0.05$ (three-way ANOVA). Reproduced with permission from Alphei J, Bonkowski M, and Scheu S (1996) Protozoa, Nematoda and Lumbricidae in the rhizosphere of *Hordelymus europaeus* (Poaceae): faunal interactions, response of microorganisms and effects on plant growth. *Oecologia* 106: 111–126.

protozoan densities and feeds on them. The weight gain of young *A. caliginosa* in soil with naked amebae was twice that in soil without amebae.

Protozoa as Bioindicators in Terrestrial Environments

There is considerable literature on the use of protozoa as bioindicators in terrestrial environments, and recently appropriate soil protozoan bioassays became available. Previous reviews summarized data obtained until 1998 and discuss general problems and methodology of bioindication. It has been shown that the full power of bioindication needs identification of species and/or functional groups because the information content of individual numbers and biomasses is often rather limited. Here, some recent studies are reviewed to demonstrate the great potential of protozoa in the investigation of natural and human-influenced soil ecosystems.

Compared to the soil meso- and macrofauna, protozoa have several specific features making them especially useful for bioindication:

1. Protozoa are an essential component of terrestrial ecosystems, because of their large standing crop and production and their high food consumption and respiration per mass unit (Figure 6, Table 4). Thus, changes in their numbers and diversity influence the rate and kind of soil formation and soil fertility.
2. Protozoa, with their rapid growth, short lifetime, and delicate external membranes, can react more quickly to environmental changes than any other eukaryotic organism and can thus serve as an early warning system and excellent tool in bioassays. Results are obtained within a few days, faster than with any other eukaryotic test system.
3. The genome of the protozoa is similar to that of the metazoa. Their reactions to environmental changes can thus be related to higher organisms more convincingly than those of the prokaryotes.
4. Morphologic and genetic differentiation among globally distributed protozoan species is rather low, suggesting that many indicator species can be used worldwide.
5. Protozoa inhabit and are particularly abundant in those soil ecosystems that almost or entirely lack higher organisms, especially earthworms, due to extreme environmental conditions.
6. Protozoa are not readily dislodged in soil. The difficult problem of horizontal migration, especially with the epigaeon, does not affect the investigations.

Protozoa also pose problems when used as bioindicators; specifically, counting methods are time-consuming and species identification is troublesome due to the lack of comprehensive identification literature. There are indications that molecular methods and taxonomic reviews can partially solve these drawbacks.

Soil Protozoa as Indicators in Natural Ecosystems

The main humus types, mull, moder, and mor, can be finely discriminated by certain protozoan species, especially testate amebae. Further, some ciliates are excellent indicators for periodic or sporadic oxygen depletion, which is often difficult to recognize and

measure with physicochemical methods. Of course, protozoa can also be used to characterize certain habitats and soil development, for instance, in coastal dunes of Germany (Figure 5). This study shows, *inter alia*, that the content of organic matter is a key factor for the abundance and diversity of the microfauna, and testate amebae are the most important group, reaching up to 50 000 individuals per gram dry mass. Further, active ciliates are comparatively abundant. These data are in accordance with niche separation of testate amebae and ciliates and the concept of ciliatostasis reported in the section on vertical distribution.

Protozoa as Indicators in Reclaimed Open-Cast Coal Mining Areas

Topsoil removal, as occurs in open-cast mining areas and during the preparation and use of high-mountain ski slopes, heavily impoverishes protozoa and soil life in general (Figure 8). However, colonization occurs within a few months. Abundances and biomasses in 2- up to 46-year-old afforested mine soils are in the same order as in undisturbed forest soils, but typical humus-inhabiting, large-sized testacean species are lacking or occur rarely, showing that reclamation was only partially successful. Six ubiquists out of 48 taxa contributed 61–87% to the overall abundance. Generally, all test sites had a distinct testacean community, whose structure depended on age, substrate, and stocking.

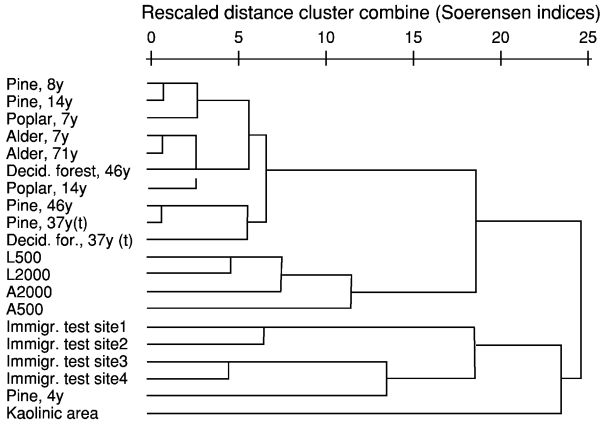


Figure 8 Cluster based on similarity (Soerensen index, UPGMA-linkage) of numbers and species of testate amoebae in various immigration test sites, minicontainers (sites L, A), and recultivated forests. The distances are rescaled to fall in the range of 1–25. Quality and development of the substrates exposed for different time intervals were more important for immigration and colonization success than the adjacent source habitats. Reproduced with permission from Wanner M and Dunger W (2001) Biological activity of soils from reclaimed open-cast coal mining areas in Upper Lusatia using testate amoebae (protists) as indicators. *Ecological Engineering* 17: 323–330.

Protozoa as Indicators of Soil Decontamination

Decontamination of polluted soils is a current challenge for scientists from various fields. Few data are available as concerns the soil fauna, and only one study included protozoa (Figure 9). Immediately after decontamination, active protozoa, nematodes, and collembola were not detected in treated soil. However, protozoan cysts survived treatment (5700 ± 4200 cysts per g dry soil). Following 1 week of exposure, no active protozoa were found. From the second sampling date onwards (after 3 weeks), protozoa were observed with significantly increasing biomass until the end of the experiment 12 weeks after exposition.

Effects of Biocides

A considerable amount of literature is available on the effects of biocides on soil protozoa. Several reviews showed that: (1) the general reaction of soil protozoa to biocide stress matches that of other organisms; (2) many protozoan species seem to be just as sensitive to pesticides as other more commonly used test organisms; (3) insecticides are usually more toxic than herbicides; (4) insecticides disturb soil protozoa critically, i.e., populations often do not fully recover within 60 days; and (5) investigations should be performed at species and/or functional group level because the simple measure 'total protozoan abundance' is often too insensitive. This is also obvious from a recent study, where the effects of the fungicide fenpropimorph have been investigated in laboratory microcosms and field experiments. While the microcosms showed that bacterivorous protozoa were affected by fenpropimorph at concentrations

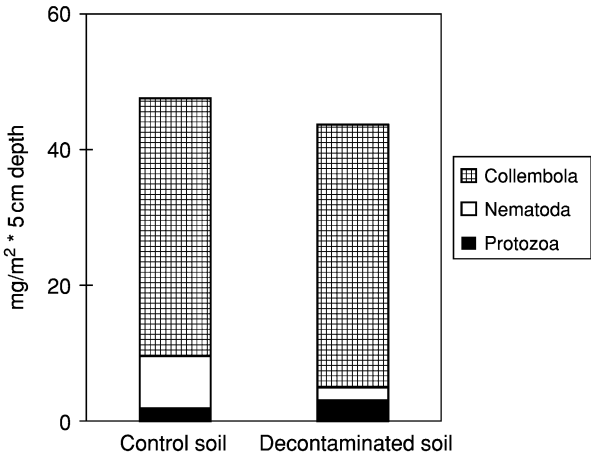


Figure 9 Biomass of Collembola, Nematoda and Protozoa in control and decontaminated soil after 12 weeks of exposure in an arable field. Reproduced with permission from Böckl M, Blay K, Fischer K, Mommertz S, and Filser J (1998) Colonisation of a copper-decontaminated soil by micro- and mesofauna. *Applied Soil Ecology* 9: 489–494.

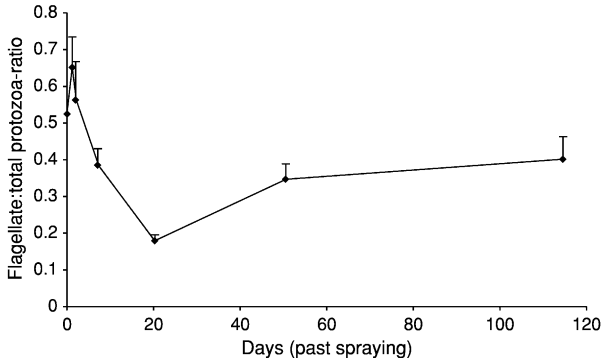


Figure 10 A significant decrease in the flagellate:total protozoan ratio was recognizable when fenpropimorphic acid, a degradation product of fenpropimorph, peaked 20 days after pesticide application. Total protozoan abundance was, however, not significantly changed. Reproduced with permission from Bjørnlund L, Ekelund F, Christensen S *et al.* (2000) Interactions between saprotrophic fungi, bacteria and protozoa on decomposing wheat roots in soil influenced by the fungicide fenpropimorph (Corbel®): a field study. *Soil Biology and Biochemistry* 32: 967–975.

lower than those expected in soil after normal field application, the field experiments showed only a slight decrease in the flagellate-to-total protozoan ratio after 20 days (Figure 10).

See also: **Bacteria:** Plant Growth-Promoting; Soil; **Fauna**; **Fungi**; **Microbial Processes:** Environmental Factors; **Mycorrhizal Fungi**; **Nematodes**; **Pollutants:** Biodegradation; **Root Exudates and Microorganisms**

Further Reading

Alphei J, Bonkowski M, and Scheu S (1996) Protozoa, Nematoda and Lumbricidae in the rhizosphere of *Hordelymus europaeus* (Poaceae): faunal interactions, response of microorganisms and effects on plant growth. *Oecologia* 106: 111–126.

Bamforth SS (2001) Proportions of active ciliate taxa in soils. *Biology and Fertility of Soils* 33: 197–203.

Bjørnlund L, Ekelund F, Christensen S *et al.* (2000) Interactions between saprotrophic fungi, bacteria and protozoa on decomposing wheat roots in soil influenced by the fungicide fenpropimorph (Corbel®): a field study. *Soil Biology and Biochemistry* 32: 967–975.

Böckl M, Blay K, Fischer K, Mommertz S, and Filser J (1998) Colonisation of a copper-decontaminated soil by micro- and mesofauna. *Applied Soil Ecology* 9: 489–494.

Bonkowski M and Schaefer M (1997) Interactions between earthworms and soil protozoa: a trophic component in the soil food web. *Soil Biology and Biochemistry* 29: 499–502.

Corliss JO (2000) Biodiversity, classification, and numbers of species of protists. In: Raven PH and Williams T (eds) *Nature and Human Society*, pp. 130–155. Washington, DC: National Academy Press.

Darbyshire JF (ed.) (1994) *Soil Protozoa*. Oxon: CAB International.

Foissner W (1987) Soil protozoa: fundamental problems, ecological significance, adaptations in ciliates and testaceans, bioindicators, and guide to the literature. *Progress in Protistology* 2: 69–212.

Foissner W (1994) Soil protozoa as bioindicators in ecosystems under human influence. In: Darbyshire JF (ed.) *Soil Protozoa*, pp. 147–193. Oxon: CAB International.

Foissner W (1999) Soil protozoa as bioindicators: pros and cons, methods, diversity, representative examples. *Agriculture, Ecosystems and Environment* 74: 95–112.

Verhoeven R (2002) The structure of the microtrophic system in a development series of dune soils. *Pedobiologia* 46: 75–89.

Wanner M and Dunger W (2001) Biological activity of soils from reclaimed open-cast coal mining areas in Upper Lusatia using testate amoebae (protists) as indicators. *Ecological Engineering* 17: 323–330.

Q

QUALITY OF SOIL

B J Wienhold, G E Varvel, and J W Doran, USDA Agricultural Research Service, Lincoln, NE, USA

Published by Elsevier Ltd.

Introduction

Soil forms a thin layer over the surface of the earth that performs many functions essential to life. It serves as a substrate to support plant growth, a reservoir for many nutrients essential for plant growth, a filter to maintain air quality through interactions with the atmosphere, and a storage and purification medium for water. Soil also supports biological activity involved in the decomposition and recycling of animal and plant products; and it has been identified as having a major role in carbon sequestration and mitigation of climate change associated with anthropogenic emissions of carbon dioxide and other greenhouse gases. Maintaining these soil functions in natural and managed systems is essential for environmental quality, system productivity and biodiversity. Human activity has affected soil functions in many natural and managed systems. Management practices have been developed to manipulate the soil and its functions, to facilitate planting and harvesting of food, forage, and fiber, and to improve productivity. Commonly used management practices include grazing systems, prescribed fire, forest management, tillage, fertilization, crop rotation, water management, liming, and cover crops.

'Soil quality' refers to the capacity of a soil to sustain biological productivity, maintain environmental quality, and promote plant and animal health. Land-management practices influence soil quality to varying degrees, and this influence can be positive or negative. Any management practice misapplied can result in soil degradation, and this has been the case with many traditional practices. As our understanding of how management practices affect soil quality improves, new practices are developed that maintain or improve soil quality. Monitoring soil quality allows those managing the land to assess the impact their decisions are having on the soil and the functioning of that resource. Hence, soil quality serves as a tool for monitoring soil function and system sustainability. Soils are not managed for soil quality per se; rather, soils are managed to optimize inputs to maintain soil functions and generate an acceptable rate of return on investment. In managed systems, soils are managed to produce food, forage, and fiber needed by society. In addition, producers and other land managers may use soil-quality indicators to focus on the intended result of soil management, i.e., improvement of the fitness of a soil for a particular use, not merely the characterization of the soil material itself. Soil quality is a tool that managers can use to identify problems in the short term and to develop management practices that are sustainable over the long term.

Management Systems

Soil-quality and land management have the dual goals of meeting the needs of society and conserving essential soil, water, and air resources for future generations. Rangelands are managed for forage production, water harvesting, recreation, and wildlife production. On rangelands, management is heavily focused on vegetation. In these systems, animal stocking rates are managed to maintain vegetation composition and productivity, while water-source and mineral-supplement locations are used to improve animal distribution and avoid overgrazing and compaction of the soil. Avoiding compaction and maintaining vegetative cover reduces the potential for erosion and maintains water quality in rangeland watersheds. Prescribed fire, fertilization, and seeding are other management practices that can be used to improve degraded rangelands and maintain critical soil functions.

In forests, soils are managed for wood and fiber production and other multiple uses such as water harvesting, recreation, and wildlife. During timber-harvesting operations, soil quality is maintained by using practices that minimize the potential for erosion and

restore vegetative cover as quickly as possible. Prescribed fire, fertilization, and seeding (or seedling transplants) can be used to improve site productivity and maintain critical soil functions.

In agriculture, soils are managed primarily for the production of food and fiber. Historically, tillage has been used to prepare a seedbed, incorporate residue, control weeds, and distribute agricultural chemicals. Inorganic fertilizer, animal manure, or other organic by-products are applied to provide nutrients needed by the crop but not supplied in sufficient quantities by the soil or to replace nutrients removed by the harvested crop. Lime is applied to increase soil pH. Crop sequences and crop rotations are used to break disease cycles, diversify crop production, and provide additional marketing options. Soil quality monitoring allows agricultural managers to make assessments of the effects of various combinations of these management factors in the overall system.

Soil Quality Assessment

Soil quality is an effective tool for monitoring soil function. Assessing soil quality involves measuring physical, chemical, and biological soil properties and using these measured values to identify properties of the soil that may be inhibiting soil function or to monitor how changes in management are affecting soil functions.

The initial steps in this assessment are to identify the management goals (e.g., food or fiber production, environmental protection) and to determine the soil functions that contribute toward accomplishing them. For a soil managed for food production, functions such as providing a suitable plant substrate and serving as a nutrient reservoir are important. For a soil managed for environmental protection, key functions are storage and purification of water, and interactions with the atmosphere.

Once soil functions important to a management goal are identified, soil properties influencing those functions can be selected for measurement (Table 1), depending on location and soil type. For example, soil properties important for a soil being managed for irrigated crop production in an arid environment (e.g., sodicity, depth to water table) will be different from those for a soil in a rain-fed environment (e.g., infiltration rate, potentially mineralizable N, extractable P). Selection of soil properties for soil quality assessment also needs to take into account the expertise and resources available to the person doing the assessment. Some soil properties are better indicators of soil quality than others, but special facilities or equipment may be required to quantify the property; in which case they are often of little practical use.

A one-time measure of these soil properties may identify properties that fall outside an acceptable range of values for that soil. Once soil properties exhibiting an unacceptable value are identified, management practices can be implemented to improve them and restore the associated soil function. If multiple management practices are being used on a single soil type, a one-time comparison of soil properties among the practices can be used to determine the effect the various practices have on soil quality. Such one-time assessments are termed 'comparative soil quality assessments.' A more effective approach is to sample the soil repeatedly (e.g., two or three times each year or at the same time of year for several years) and assess management effects on soil properties over time (i.e., during the year or from year to year). Repeated measurement of soil properties is termed

Table 1 Potential soil properties for assessing soil functions

Soil function	*Physical properties*	*Chemical properties*	*Biological properties*
Substrate for plant growth	Bulk density Aggregate stability Soil depth	Electrical conductivity Sodium adsorption ratio	
Nutrient reservoir		Inorganic N content Extractable P pH	Potential mineralization
Atmospheric interactions	Water-filled pore space	CO_2 flux NO_x flux Methane flux NH_4 flux	
Storage and purification of water	Infiltration rate Water-holding capacity	NO_3-N concentration Extractable P	
Biological activity	Water-filled pore space		Microbial biomass Soil respiration Potential mineralization Earthworm population

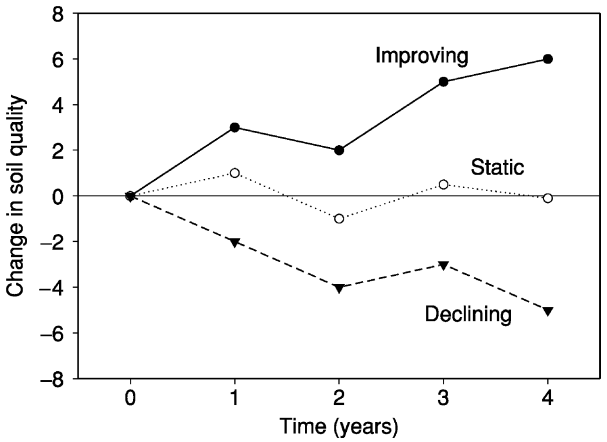


Figure 1 Hypothetical outcomes for a dynamic soil quality assessment.

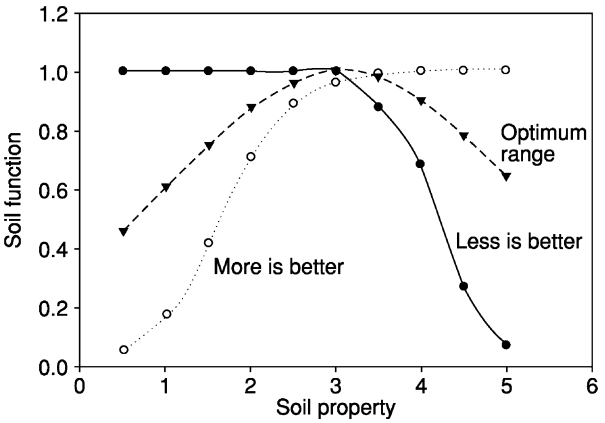


Figure 2 Hypothetical indicator curves describing the relationship between a soil property and soil function. (Adapted from unpublished data from Andrews SS, Karlen DL, and Seybold C.)

'dynamic soil-quality assessment.' By repeatedly measuring a soil property, the direction and magnitude of change can be assessed and management practices can be adjusted (Figure 1).

Soil Function–Soil Indicator Relationships

Relationships exist between soil properties and soil functions, which may take the form of: (1) lower or decreasing in value is better, (2) higher or increasing in value is better, or (3) an optimum range (Figure 2). As soil functions are identified, appropriate soil properties are selected to assess those functions. Identifying critical soil functions and selecting soil properties to be measured for assessing soil-quality management goals are decisions that are best made on a case-by-case basis.

Evaluation of soils functioning as a substrate for plant growth includes measurement of soil properties such as bulk density, soil depth, and electrical conductivity. These properties affect the rooting activity of plants and influence the ability of the plant to acquire water and nutrients. The relationship between bulk density and rooting activity takes the form of 'less is better', with rooting activity being depressed when the bulk density exceeds a certain value. The value at which bulk density restricts rooting activity varies with soil texture and ranges from $1.4\,g\,cm^{-3}$ for a clay soil to $1.7\,g\,cm^{-3}$ for a sandy soil. The relationship between soil depth and rooting activity takes the form of 'more is better', with the relationship leveling off when soil depths exceed the rooting depth of the crop. The relationship between rooting activity and electrical conductivity takes the form of 'less is better'. As electrical conductivity increases, the effect on rooting activity and subsequent plant growth is dependent on plant species. Crops such as bean (*Phaseolus vulgaris* L.), with a threshold of $1.0\,dS\,m^{-1}$, are much more sensitive to electrical conductivity (salinity) than others such as barley (*Hordeum vulgare* L.), with a threshold of $8.0\,dS\,m^{-1}$.

Evaluation of soils functioning as nutrient reservoirs includes measurement of soil properties such as the concentration of nutrients essential to plant growth and pH. These properties are measures of the soil's ability to provide nutrients needed by the plant. The relationship between nutrient concentration and ability to meet plant demand takes the form of 'more is better' for most nutrients. For some nutrients the relationship takes the form of 'more is better' at low concentrations and 'less is better' at high concentrations, due to negative effects on crop quality. The range of concentrations needed to meet crop needs varies among crops and soils. Values needed to develop these relationships are available in the soil nutrient-testing literature. The relationship between plant growth and pH is of the form of an optimum range. The availability of many nutrients essential to plant growth are highest at pH values between 5.5 and 7.5. At higher pH values, nutrient availability is reduced; at lower pH values, nutrient availability is reduced and availability of toxic elements such as aluminum increases.

Evaluation of soil interactions with the atmosphere includes measurement of soil properties such as respiration and flux of greenhouse gases. These properties provide a measure of whether the soil is serving as a source or sink for greenhouse gases – carbon dioxide, (CO_2), methane (CH_4), and nitrogen oxides (NO_x). The relationship between respiration and atmospheric interactions takes the form of an optimum range with lower values than optimum suggesting that beneficial

microbial activity is too low (e.g., suboptimal release of nutrients from organic matter, low N-fixation rates, poor development of soil structure) and high values suggesting that microbial activity is too high (e.g., degradation of soil organic matter and release of nutrients in amounts exceeding what plants can utilize). Optimum value is determined by comparing organic matter inputs (crop residue and animal manure) with net respiration. The relationship between greenhouse gas fluxes and the atmosphere takes the form of 'less is better'. Low gas fluxes are a consequence of the activity of microorganisms naturally present in the soil. High gas fluxes represent the presence of labile C and nutrient substrates (e.g., nitrogenous fertilizers) under conditions (e.g., water-filled pore space of ~80%) that stimulate microorganisms that produce greenhouse gases as products of their metabolism (Figure 3).

Evaluation of soils as a medium for the storage and purification of water includes measurement of properties such as infiltration rate and nutrient concentrations. These properties provide a measure of the potential for contamination of surface-water and groundwater resources. The relationship between infiltration rate and storage and purification of water takes the form of 'more is better' and differs among soils as a function of soil texture. Low infiltration rates contribute to runoff and erosion losses of nutrients to surface water. The relationship between soil nutrient concentrations and storage and purification of water takes the form of 'less is better'. High nutrient concentrations in surface soils can contribute to contamination of surface waters when they are transported in runoff and erosion sediments. High nutrient concentrations in subsurface soils can contribute to groundwater contamination when they are leached from the root zone.

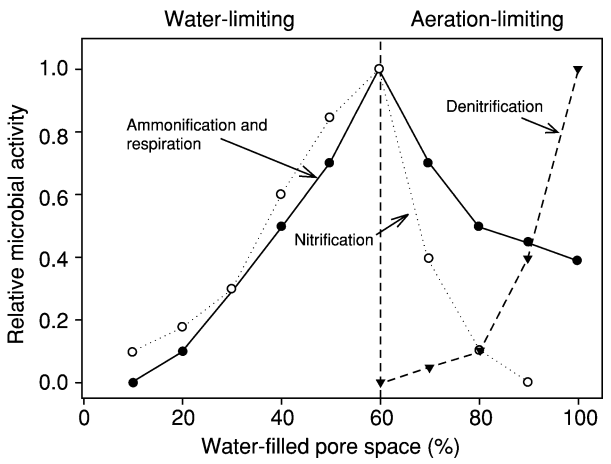


Figure 3 Hypothetical relationship between soil water-filled pore space and microbial activity associated with respiration, nitrification, and denitrification. (Source: Linn DM and Doran JW (1984) Effects of water-filled pore space on carbon dioxide and nitrous oxide production in tilled and non-tilled soils. *Soil Science Society of America Journal* 48: 1267.)

Evaluation of the biological activity of a soil includes measurements of soil properties such as water-filled pore space and microbial biomass. These properties provide a measure of the soil's ability to decompose added material and to recycle nutrients contained in that material. Including physical soil properties in the evaluation of biological activity provides information about whether the products resulting from biological activity are from aerobic metabolic pathways (e.g., carbon dioxide, nitrate, sulfate) or from anaerobic metabolic pathways (e.g., methane, oxides of nitrogen, hydrogen sulfide). The relationship between water-filled pore space and biological activity takes the form of an optimum range. The optimum range corresponds to a moist, well-aerated soil. Microbial activity is greatest at 60% of water-filled pore space and declines linearly as the soil dries (Figure 3). When water-filled pore space reaches 80%, anaerobic processes such as denitrification increase. Microbial biomass responds more quickly to management than does soil organic carbon. Increases in microbial biomass are usually associated with improved management and an increase in soil quality.

For each soil function, soil quality is the integrated total of the selected set of properties i.e., the integration of soil physical, chemical, and biological processes specific to a particular soil and environment. Because it is location-specific, care must be taken when making soil-quality comparisons between soils. A soil that has been poorly managed for a long period of time may have a relatively low soil quality as compared to a similar, yet well-managed soil. Improved management is likely to have a much greater effect on the soil quality of the poorly managed soil.

Summary

Since soil quality is assessed in terms of the fitness of a soil for its intended use, conflicts can arise when there are multiple uses and management goals for a soil. The quality of an agricultural soil may be quite different if the management goal being assessed is crop production than if the management goal being assessed is environmental services. A manager may incorporate residue using tillage to produce a seedbed and promote mineralization of nutrients. Fertilizers may also be used to provide adequate amounts of nutrients needed by the crop. Assessing the crop-production management goal for a given field may reveal that soil properties such as nutrient concentration, potentially mineralizable

N, and soil respiration are within the optimum range. Assessing the environmental-services goal for the same field may reveal that nutrient concentrations may be near levels with potential for leaching losses to groundwater, that denitrification rates are high following precipitation or irrigation, or that decomposition of the incorporated residue results in high carbon dioxide fluxes. Conflicting assessment outcomes are not a failure of the soil-quality concept. Knowledge of the soil-quality status for multiple management goals provides the potential for making adjustments in management practices to minimize negative aspects while maintaining positive aspects. The goal is that such assessment of management goals and adjustment of management practices will improve the sustainability of the system.

Concerns regarding the soil quality concept include the need to test underlying concepts in more soils and production systems and to develop procedures for addressing the potential for conflicting outcomes. Soil quality assessment is a management tool that is being improved to include more management goals, soil functions, and soil properties. The concept is being used in a wider range of soils and production systems. The utility of the soil quality concept improves as our understanding of the soil resource improves. Proponents and sceptics of the soil quality concept share a common goal: to improve management of the soil resources to maintain critical soil functions.

List of Technical Nomenclature

Ammonification	Biological process leading to ammonium formation from nitrogen-containing organic compounds
Denitrification	Biological reduction of nitrate and nitrite to nitrogen and nitrogen oxides
Labile carbon	Carbon-containing compounds that can be readily metabolized by microorganisms
Nitrification	Biological oxidation of ammonium to nitrate and nitrate
Respiration	Oxidative reactions providing energy to an organism

See also: **Acidity**; **Carbon Cycle in Soils:** Dynamics and Management; **Compaction**; **Conservation Tillage**; **Crop Rotations**; **Crop-Residue Management**; **Cultivation and Tillage**; **Degradation**; **Erosion:** Water-Induced; Wind-Induced; **Fertility**; **Manure Management**; **Nutrient Management**; **Organic Matter:** Principles and Processes

Further Reading

Acton DF and Gregorich LJ (1995) *The Health of Our Soils: Toward Sustainable Agriculture in Canada*. Ottawa, ON, Canada: Agriculture and Agri-Food Canada.

Andrews SS, Karlen DL, and Mitchell JP (2002) A comparison of soil quality indexing methods for vegetable production systems in Northern California. *Agriculture, Ecosystems and Environment* 90: 25–45.

Doran JW and Jones AJ (1996) *Methods for Assessing Soil Quality*. Soil Science Society of America Special Publication No. 49. Madison, WI: Soil Science Society of America.

Karlen DL, Mausbach MJ, Doran JW, Cline RG, Harris RF, and Schuman GE (1997) Soil quality: a concept, definition, and framework for evaluation. *Soil Science Society of America Journal* 61: 4–10.

Schjonning P, Emholt S, and Christensen BT (eds) (2004) *Managing Soil Quality – Challenges in Modern Agriculture*. Wallingford, UK: CABI Publishing.

Sojka RE and Upchurch DR (1999) Reservations regarding the soil quality concept. *Soil Science Society of America Journal* 63: 1039–1054.

R

RADIATION BALANCE

J L Hatfield, T J Sauer, and J H Prueger,
National Soil Tilth Laboratory, Ames, IA, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Life on Earth depends upon the flow of energy. Living objects balance their energy inputs with outputs to create a comfortable environment. Humans have an advantage because they create clothing and shelter to help modify their environment. Plants and soil also have inputs of energy, but have limited controls on the dissipation of energy and even less control on the energy inputs into the system. When we begin to think about the balance of energy in the soil or plant system we can begin to understand some of the remarkable balances that exist within the natural environment. Understanding the surface radiation and energy balances provides an opportunity to explore the role of different components, e.g., crop residue, tillage, cover crops, on soils and the environment.

The balance of radiation is similar to a checkbook or savings account. How much money is in it at any one time depends upon how much has been put into the account and how much has been taken out. This balance can be expressed very simply in word form. Net radiation is the difference between the gains (incoming solar radiation and absorbed longwave radiation) and losses (reflected and transmitted solar radiation and emitted longwave radiation). This 'budget' can be expressed mathematically as:

$$R_n = (1 - \alpha)S_t + L_d - \epsilon\sigma T_s^4 \qquad [1]$$

where R_n is the net radiation at the surface expressed as the amount of energy (joules) per unit area (meters squared) per unit time (seconds) or watts per meter squared. All of the radiation terms (S_t, L_d, and $\epsilon\sigma T_s^4$) in eqn 1 must have the same units to avoid problems. As we expand these terms, S_t is the solar radiation input, α the albedo or reflectance of the surface, L_d the input of longwave radiation, $\epsilon\sigma T_s^4$ is the energy lost as longwave radiation from the surface. Shortwave radiation represents the energy in the 0.3- to 3-μm range, while longwave radiation is derived from wavelengths of more than 3 μm. Another way to consider this difference is that shortwave originates from the sun, with a surface temperature near 6000 K, while longwave is derived from objects with temperatures near 300 K. A representation of the radiation balance for two different days is shown in Figure 1. Radiation balance values throughout the day do not always produce a smooth curve because of the presence of clouds. Nighttime values are zero or below because the Earth's surface tends to lose more energy than it receives. Energy available as net radiation is further divided into other components. We express the energy balance as:

$$R_n = \mathrm{LE} + H + G \qquad [2]$$

where LE represents the energy used to evaporate water, H the energy to heat the air, and G the amount stored in the soil volume. The more energy input to the surface the greater the potential magnitude of these components. (*See* **Energy Balance**.)

Incoming Solar Radiation

Solar radiation originates at the Sun and travels through space and the atmosphere before it reaches the Earth's surface. Solar radiation consists of two

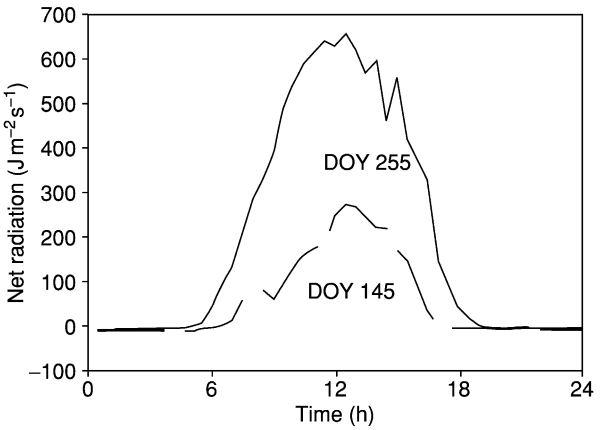


Figure 1 Net radiation balance for 2 days (days 145 and 255) during the midsummer growing season in central Iowa.

components, the direct component (S_b) and the diffuse (S_d) component. Diffuse radiation comes from all angles, while the direct beam follows the path of the Sun throughout the day. Diffuse radiation is formed by the scattering of the direct beam by particles in the atmosphere. Physically, we can observe the effect of direct and diffuse radiation by sharpness of shadows. On a clear day, the outlines of the shadows are clearly defined, while on a cloudy day there are no shadows yet we still can see. Mathematically, we represent incoming solar radiation as:

$$S_t = S_b \sin\beta + S_d \quad [3]$$

The angle of the Sun's beam above a horizontal surface, β, distributes the energy in the beam over a different area than if the surface were perpendicular to the beam. The magnitude of direct and diffuse radiation depends upon the solar elevation, the angle of the Sun's beam above the horizon, the amount of scattering and adsorption by atmospheric gases and aerosols, and the amount of scattering and adsorption by water droplets and ice particles in the clouds. The effect of clouds is easy to understand; however, particles in the atmosphere have a large impact, because they scatter radiation back to space and some toward the Earth's surface. The portion of diffuse radiation increases as the number of particles in the atmosphere increases.

There is large variation in the incident solar radiation at many locations throughout the year because of the changing angle of inclination of the Earth with respect to the Sun. At the equator there is little difference; however, at the midlatitudes there are large differences between summer and winter. Clear-sky solar radiation incident on the Earth at midday is approximately $26\,J\,m^{-2}$ per day during summer, but approximately $9\,J\,m^{-2}$ per day during winter at 42° N latitude. Clouds affect these totals and need to be considered. There are tables of clear-sky solar radiation incident on the Earth's surface for any location: these tables have been assembled through measurements and empirical calculations.

The slope of the surface is a critical factor in the radiation balance. If there is a south-facing slope in the northern hemisphere then more energy is going to be absorbed than if there is a north-facing slope. More subtle differences exist if the slope of the soil surface is changed with tillage implements. The creation of a furrow with a tillage tool that runs east to west across the field creates a surface in which one part will receive more direct solar radiation than the other. This will cause the soil surface and soil temperatures to be warmer on the south side and cooler on the north side in the northern hemisphere. Temperatures for bare soil at the surface under these conditions may be 10–20°C different at midday. This same furrow shape placed in a north–south direction will have differences in surface temperatures from morning to afternoon. The geometry of the surface plays a critical role in the radiation balance through the effect of direct-beam solar radiation. Manipulation of the soil surface through tillage implements can create a specific microclimate at the soil surface that influences temperature-dependent processes within the soil.

Albedo

Albedo of a surface is the ratio of amount of solar radiation reflected back to the atmosphere to the amount of incoming solar radiation (S_r:S_t). We see a highly reflective surface as white, and a low-reflectance surface as black. A dark surface will absorb more radiation than a light-colored surface. The range of albedo values for different surfaces is shown in Table 1.

Albedo is more complex than assuming a single constant value for a surface. Albedo values for a soil surface change with soil water content, type of soil, tillage, amount and type of crop-residue cover, and age of residue. Starting with a smooth soil surface with an albedo of 0.15, the addition of water to wet the soil surface decreases the albedo to 0.10, and then there is a slow return to the original value as the surface dries. If the same surface is roughened, there is a further decrease in the albedo of a dry surface to near 0.10. This decrease is due to the change in the soil surface that traps more direct- and diffuse-beam solar radiation. Wetting this surface further decreases the albedo. Adding a fresh crop residue increases the albedo because of the highly reflective nature of the crop residue; however, as the residue ages, the albedo decreases and may be only slightly different from the soil surface in the spring. Soil surfaces are continually undergoing changes that affect their albedo. This will impact the radiation balance and ultimately the evaporation of water from the surface, heating the air, and warming the soil.

Table 1 Albedo values for different natural surfaces

Surface	*Albedo*
Bare soil, dry	0.20
Bare soil, wet	0.15
Bare soil, fresh corn residue	0.35
Bare soil, weathered corn residue	0.22
Corn canopy	0.22
Wheat canopy	0.26
Rice canopy	0.24
Deciduous forest with leaves	0.18
Coniferous forest	0.16

The albedo values shown in Table 1 represent an integrated value over the shortwave wavelengths and are determined by measured the incoming solar radiation with a pyranometer and reflected radiation with an inverted pyranometer of the same type. Within the solar radiation spectrum, there are differences in reflectance, as shown in Figure 2 for a corn leaf. There is a dramatic change in the reflectance between the visible and near-infrared wavelengths from 0.6 to 0.7 μm. The difference in reflectance spectra across the wavelengths in plant material shows low reflectance values in the blue and red portions of the spectrum, an increase in the green wavelengths, and then a sharp increase in the near-infrared wavelengths. Reflectance values for a bare soil across the solar radiation wavelengths show a different pattern from a leaf and tend to increase linearly with wavelength, not showing the large differences among portions of the spectrum (Figure 3). Tilling the soil to change the roughness and increase the water content changes the reflectance of the soil. There is a decrease in the reflectance values for all wavelengths (Figure 3). Adding crop residue to the soil surface also changes the reflectance of the surface and changes the energy-balance components.

When solar radiation impinges on a leaf or any other object, the radiation may be reflected, absorbed, or transmitted. If we imagine ourselves underneath a thin layer of white cloth, from the upper side we would see the white color owing to reflection of all wavelengths; however, underneath the cloth we would still see owing to light being passed through (transmitted). Some of the light would be absorbed by the cloth. Under a 10-cm layer of soil, there would be no transmitted light and there would be no shortwave radiation, therefore no visibility. This same principle applies when layers of material are added to the soil surface. A layer with a high light transmittance allows solar radiation to reach the soil surface and warm the soil, evaporate water, or warm the air next to the soil surface. When a highly reflective and low-transmittance material is placed on the soil surface, there is little change in the temperature of the soil surface with varying solar radiation; hence snow affects the radiation balance. Transmittance values for different materials vary widely, as shown in Table 2 for both the shortwave and longwave regions of the spectrum. Transmittance of solar radiation through materials, especially plant material, has a large effect on the radiation balance at the soil surface. Very dense canopies transmit some light through all of the layers of leaves to the soil surface. Humans are able to see in this region; however, a leaf receives little energy. Plants grown under other canopy layers often have very large leaves with less chlorophyll (more yellow) than the same plant grown in full sunlight. The radiation balance of these plants and the soil under dense canopies is much different from the environment without this canopy.

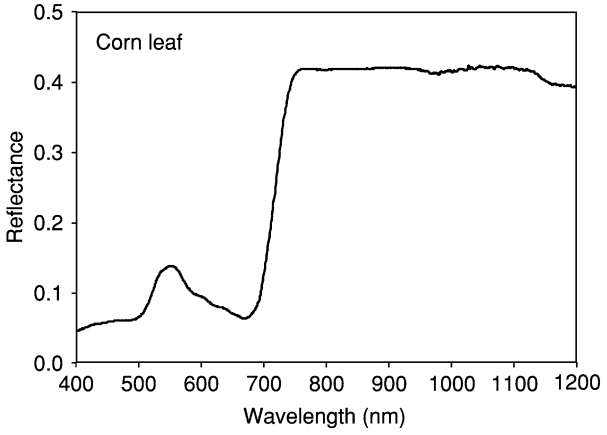


Figure 2 Reflectance spectra for a corn leaf in the visible and near-infrared wavelengths.

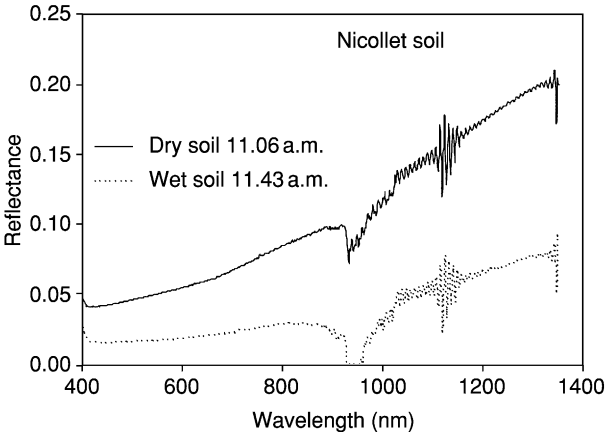


Figure 3 Reflectance of solar radiation across visible and near-infrared wavelengths for a smooth, tilled, Nicollet soil when wet and dry.

Table 2 Transmittance values for different natural surfaces

Surface	*Transmittance*
Corn leaves, visible	0.08
Corn canopy, visible	0.05
Wheat leaves, visible	0.07
Wheat canopy, visible	0.06
Fresh corn residue, visible	0.05
Weathered corn residue, visible	0.04
White plastic, visible	0.95
Black plastic, visible	0.0

Incoming Longwave Radiation

All objects with an absolute temperature of more than 0 K emit energy. Objects in our atmosphere such as gases, particles, ice crystals, or water droplets have temperatures above absolute zero and emit energy. The energy emitted is equivalent to the 4th power of their temperature as described by the Stefan law, $E = es\ T^4$. The Stefan constant, s, is $5.67 \times 10^{-8}\,\mathrm{W\,m^{-2}\,K^{-1}}$. The amount of energy emitted by an object greatly increases as the temperature increases. The emissivity term can be considered an efficiency factor. The higher the emissivity of an object the more effective the emission of longwave radiation away from the object. Longwave radiation is measured with radiometers designed to capture only these wavelengths; however, there are relatively few of those instruments located around the world. Incoming longwave radiation is generally estimated from mathematical equations in which the emissivity of the atmosphere, e_a, is estimated through a combination of water vapor or temperature measurements to modify the energy relationship to $e_a s T_a^4$. In one of the original relationships for this estimation, it was proposed that $e_a = 0.51 + 0.66 e_0^{1/2}$, where e_0 is the water vapor pressure of the atmosphere. There have been several attempts to refine this relationship, for example, where e_a is estimated by $0.575 e_0^{1/7}$ and e_a is best represented by $0.70 + 5.95 \times 10^{-5} e_0 \exp(1500/T_0)$, with T_0 the air temperature in degrees Kelvin. Both e_0 and T_0 are obtained from meteorological station data at a height of 2 m above the surface and are available every hour throughout the day. A comparison of all of these different methods with data collected across the USA has found that there is little difference among the methods and all predicted longwave radiation to within 5% of the measured values. Emissivity values for clear skies can range from 0.6 to 0.75, and these equations apply to the clear-sky condition. When there are clouds, the e_a term can be considered to be near 1. Consider what happens to air temperature on a clear versus a cloudy night. On the clear night there is a rapid drop in air temperature because the atmosphere is contributing little energy to the Earth's surface. On a cloudy night or an extremely humid night there is little change in air temperature because the longwave radiation lost from the Earth is captured by the clouds or water vapor and reemitted back to Earth's surface.

Terrestrial Longwave Radiation

Objects on the Earth's surface have a temperature and emit energy in relationship to their temperature as discussed for longwave energy emitted from the atmosphere. The same principle applies to the soil surface or objects on the Earth's surface. The last term in eqn 1, $e_s s T_s^4$, represents the loss of energy away from the surface into the atmosphere. Emissivity values for objects on Earth are close to 1; however, there is variation, as shown in Table 3. The variation is small and often the assumption is that 0.95 or 0.97 are adequate mean values that can be used. The T_s term in the relationship represents the surface temperature of the object. This would be the temperature at the outer surface of an object or at the very top layer of the soil surface. There are large gradients of temperature in the soil, and for a dry soil the surface temperature at midday may exceed 50°C while at 10 mm below the surface the temperature is less than 40°C. In the estimation of the longwave energy lost from the surface, a longwave radiometer can be inverted over the surface to measure the amount of energy escaping or estimated using measurements of surface temperature obtained from infrared thermometers combined with an estimate of emissivity for the surface. The latter approach works quite well for smooth or uniform surfaces; however, for rough or undulating surfaces or surfaces comprised of a combination of plants and soil, there is a large variation in surface temperature and measurements would have to be made that would represent the average value for the entire surface area. The magnitude of the energy lost from surfaces with different temperatures is shown by Figure 4. As the surface temperature increases, the energy emitted increases rapidly, because emission is a function of the 4th power of surface temperature.

Table 3 Emissivity values for different natural surfaces

Surface	*Emissivity*
Bare soil, dry	0.97
Bare soil, wet	0.99
Bare soil, fresh corn residue	0.97
Bare soil, weathered corn residue	0.98
Corn canopy	0.94
Wheat canopy	0.96
Deciduous forest, foliage present	0.99
Coniferous forest	0.99

Radiation Balance

Each of the terms in eqn [1] has been discussed separately in order to develop an appreciation for the variation within each component and the factors that affect the magnitude of the energy within each term. As these components are reassembled into the radiation balance for a surface it is possible to see how soil-management factors might affect the

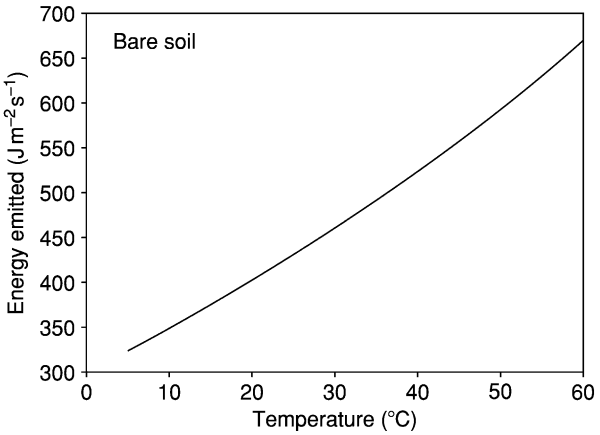


Figure 4 Energy emitted from a soil surface under different surface temperatures.

radiation balance. There are changes throughout the day as shown in Figure 1. If the radiation balance is expanded into the components for each hour throughout the day, the differences that each component adds to the balance can be seen. The largest factor is the solar radiation term, and the difference between the two days is a direct result of the incoming solar radiation. The least amount of variation among days is introduced by the incoming longwave radiation, L_d, since this factor is primarily affected by air temperature. Longwave loss is actually highest during the day when the surface temperatures are highest. As these relationships are expanded into annual radiation balances, surface temperature and air temperature variations among days across the seasons have a large impact. If the annual radiation balance over the year is computed, the sum of the inputs and losses is nearly zero.

On a daily interval, changing the soil surface through tillage or crop-residue management has a large impact on the radiation balance. If a smooth, dry soil is tilled, there are several changes in the radiation balance components. Incoming solar and longwave radiation do not change. The albedo decreases so there is a greater adsorption of solar radiation leading to a greater amount of energy available at the soil surface. However, evaporation of soil water from the wet, rough surface decreases the surface temperature, creating a situation in which the surface emits less energy. The end result is a surface that has an increased amount of energy available than did the surface before tillage. This effect lasts only as long as there is soil water to evaporate, but there is some effect of the rough surface on reducing the albedo of a dry soil; however, this effect is small compared with the initial period with a wet soil surface.

The final example is the soil surface to which crop residue has been added. In the case of fresh crop residue there is an increase in the albedo of the surface and a reduction in net radiation. Weathered residue decreases the albedo and allows the surface to have an increased net radiation loading. Wetting and drying of the crop residue has some effect on the radiation balance because of the effect of water on both the albedo and emissivity of the surface. Soil-management practices have a large impact on the radiation balance, so understanding which factor is affected by management will help provide an improved understanding of the consequences of management decisions.

Implications

The radiation balance of a surface represents the balance of shortwave gains and losses and longwave gains and losses. The radiation balance changes throughout the day and year owing to incoming solar radiation, which has the largest effect on the radiation balance, followed in importance by the cloud cover. The next major factor is the slope of the surface, which affects the amount of solar radiation absorbed. The slope factor can be both a large-scale (slope of hills and mountains) and microscale factor (creation of furrows and ridges with tillage implements). Humans have a large impact on the radiation balance and can manipulate the energy balance through all of the components in the radiation balance.

See also: **Climate Change Impacts**; **Crop Water Requirements**; **Crusts:** Structural; **Cultivation and Tillage**; **Energy Balance**; **Evaporation of Water from Bare Soil**; **Evapotranspiration**; **Remote Sensing:** Organic Matter; **Water-Use Efficiency**

Further Reading

Hatfield JL, Reginato RJ, and Idso SB (1983) Comparison of long-wave radiation calculation methods over the United States. *Water Resources Research* 19: 285–288.

Monteith JL (1973) *Principles of Environmental Physics*. New York: Elsevier.

Rosenberg NJ (1974) *Microclimate: The Biological Environment*. New York: John Wiley.

Radionuclides *See* **Isotopes in Soil and Plant Investigations**

Rainfed Farming *See* **Dryland Farming**

RANGE MANAGEMENT

G L Anderson, USDA Agricultural Research Service, Sidney, MT, USA

Published by Elsevier Ltd.

Introduction

Proper rangeland management includes good stewardship of the rangeland biological components to preserve soil health and soil conservation. Certain management practices can also contribute to a host of soil-related problems (e.g., excessive wind and water erosion, nutrient loading, and increased sedimentation in reservoirs), ultimately leading to a general reduction in the land's ability to sustain healthy rangeland systems. Many of these problems can be avoided by the proper management of the aboveground or biological component of the range. While the concept of managing a healthy rangeland may seem simple and straightforward, different definitions of what constitutes a healthy rangeland result in divergent management approaches.

What are Rangelands?

The natural evolution of different types of rangeland occurs under varying conditions and disturbances. Environmental factors such as drought, wind, and temperature affect the different types of plants that make up the rangeland plant communities found around the world. Other disturbances that influence the development of specific range types include: grazing (native animals and livestock), fire, weeds, and diseases. These disturbances have always been an important part of rangeland development and health.

In the USA a general example of these varying rangeland types includes: the presence of tall-grass prairies in the eastern Central Plains, mid-grass prairies in the Central Plains and Midwest, and short-grass prairies in the Midwest and West. Environmental conditions also result in varying amounts of shrubs, forbs, and trees. Rangeland in south Texas is dominated by trees, shrubs, and forbs; sagebrush dominates prairies in the west; and grasses dominate the northern and middle portions of the country.

Internationally, rangelands are divided into six broad categories: nonwoody grasslands, open shrublands, closed shrublands, savannas, woody savannas, and tundra. Africa is dominated by savannas with some open shrublands and closed shrublands, Australia is largely savannas along the coast and open shrublands in the interior, and Eurasia is dominated by nonwoody grasslands and open shrublands. All categories of rangeland are found in South America, except tundra, with the most diverse mixture of the different classes occurring along the lengthy western coast. Tundra is primarily limited to the extreme northern latitudes.

How Much Area is Covered by Rangelands?

Worldwide, approximately 40.5% of terrestrial area is grasslands. This means that rangelands comprise more than 52.5 million km^2. This area can be further broken down into the different types of rangeland. Approximately 13.8% of terrestrial rangelands are woody savanna, 12.7% are open and closed shrublands, 8.3% are nonweedy grasslands, and 5.7% are tundra.

What Determines How Rangelands are Managed?

Widely different environmental conditions and impacts of multiple disturbances make rangeland health difficult to define. One of the most influential factors behind range-management decisions is the manager's understanding of what comprises a healthy

rangeland. There are essentially three major rangeland models or concepts that influence a rangeland manager's understanding and actions.

The climax model evaluates range condition and trend based on vegetation composition, plant production, ground cover, and soil erosion. By measuring and monitoring existing conditions, proponents of this approach hope to develop a meaningful index of rangeland health. The underlying assumption of this model is that a definable mixture of plant species composes the 'climax' or 'healthiest' vegetation type for a given ecosystem. Plant communities, over long periods of time, come to an equilibrium of predictable species based on the impact of human use, soil condition, geographic position, plant competition, animal use, microbial activity, burning, and climatic factors. Areas with minimal or no discernible impact by humans are often referred to as 'pristine' rangelands or rangelands that exemplify the purest, uncorrupted mixture of soils, microbes, plants, and naturally occurring animals, diseases, and pests.

While the climax model considers human use as a component associated with rangeland health, many people now view human activities (e.g., livestock grazing) as the primary cause of rangeland degradation. Therefore, any departure from the climax state is often attributed to misuse of the land by humans. An example of this type of thought underlies the beliefs of those proposing the development of 'buffalo commons,' which involve returning vast areas of US rangelands to a more 'natural' state, devoid of domestic grazing and other human uses. A synopsis of the climax model is: (1) We know what the composition of native plant and animal species should be, and (2) by limiting human influence the climax state can be maintained, and, given enough time, degraded ecosystems will ultimately return to a predetermined state of equilibrium.

A similar ecologic model of a healthy rangeland ecosystem also presumes a climax state exists, but that this state is rarely, if ever, achieved, because of natural and human impacts. The basic difference between this model and the previous one is the belief that nature often has as great, or greater, an impact on rangelands as do humans. In this model, plant community equilibrium is prevented by discrete occurrences of drought, fire, disease, pests, weeds, or human influences. But the basic idea that an ultimate climax state exists, remains. A synopsis of the ecologic model disturbance is: We know what the composition of native plant and animal species should be, but due to natural and human disturbance, this ideal state is rarely, if ever, achieved. However, the closer the system is to the expected equilibrium the healthier it is perceived to be.

The above two models presuppose that we know the basic combination of biotic and abiotic conditions that should exist in a given area. In reality we rarely, if ever, can know with any degree of certainty what a climax community should look like or even if climax in a grassland ever existed. This uncertainty has led to the development of the state and transition model, which presupposes that an ecologic system can transition (change) into several states (multiple types of vegetation communities), dependent on changing conditions that occur at critical junctures in a rangelands development. According to this model, the biological composition of rangelands can transition over time into a variety of different plant communities. Each transition can happen relatively quickly (e.g., from a grassland to a pine–juniper savanna) if the right environmental conditions occur, but once the change occurs it is very difficult to reverse. This means that some changes are ultimately permanent. Thus, this model recognizes that rangeland ecosystems are dynamic and changing, and that there is no 'one' combination of plants that is ecologically the 'right or ultimate' state.

An illustration can be found in an article written by Dr. Donald Jameson, who found that the combination of overgrazing followed by at least three wet years could result in the establishment of pines on rangelands previously devoid of the tree and that this change could ultimately be permanent and progressive without further human intervention. A synopsis of the state and disturbance model is: Many types of rangeland plant communities can develop in an area over time, and each can be 'healthy' and stable for long periods of time. However, the model takes into account that rangeland ecosystems can, and do, change when conditions are right.

In our brief discussion of what constitutes a healthy rangeland ecosystem, we have traversed from an absolute model (climax) to a more relative model of multiple steady states. How range managers approach their job is directly influenced by which concepts they use. As one progresses from one end of the ecologic spectrum to the other, a divergence of management approaches occurs. In the first instance (climax), managers view an ecologically defined set of specific plants as a healthy rangeland ecosystem, while at the other end (the transition model), managers expect to see any number of ecologic states where rangeland 'health' is not dependent on the existence of a specific type of plant community. In the former case, climax theory leaves the manager striving to achieve something that probably never existed on rangelands, while in the latter case the transition model incorporates change as a component of ecologic health. Under this conceptual model a

range manager strives to achieve or maintain rangeland health; however, the ecologic definition of health depends on relatively 'recent' environmental conditions and the current mix of plants in the community.

Biological Diversity

While divergent opinions on how best to manage rangelands often result in different impacts to the rangeland ecosystem, agreement generally exists on one key component: a healthy rangeland must maintain a mix of desirable (preferably native) plants that are self-sustaining. Biological diversity (or the number of different types of plants and organisms in the system) is an important consideration, because it provides food and habitat needs for different native and domestic grazers, birds, and the living microbial component of the soil.

The health of the belowground component of the plant and soil interface also depends on diversity (e.g., different plants using and influencing different portions of the ecosystem), a sustained rate of returning dead plant and animal material back into usable nutrients, maintenance of soil porosity (the rate at which water enters the soil), and water-holding capacity (the amount of water that the soil will retain for future use). For example, in the Northern Great Plains of the USA, most native rangeland plant species have fairly shallow roots, capable of capturing the limited amount of precipitation that occurs during the growing season. This shallow root mass is one of the primary zones of competition between plants, and it is the part of the biological component that minimizes soil loss. To a large degree, it is the soil–root interface that ultimately determines the stability and sustainability of the rangeland ecosystem. Therefore, it is here that changes in the aboveground plant community can have the greatest impact.

A common idiom in ecology is that 'nature abhors a vacuum.' When applied to this discussion, it simply means there are many environmental and soil conditions that exist across a rangeland, with different plants living and utilizing the resources in each of these areas or niches. If a specific plant group is missing (the vacuum), something will move in to take its place, and it is the range manager's job to ensure the niche is filled with the proper plant type. Unfortunately, more often than not, invading undesirable plants (weeds) exploit these areas and establish before management steps can successfully close the empty niches. Thus, the introduction of invasive weeds into the system can have a devastating impact on the soil–root interface, plant biodiversity, soil health, and finally the economic worth of rangelands.

Introduction of Undesirable Plant Species

The impact of weeds on rangelands and range ecology is dramatic: First, native plants disappear, wind and water erosion are accelerated, and eventually the seedbank is altered to the point where there is no natural way for the system to recuperate. Second, weeds can also alter the soil chemistry so that little else is capable of growing in that location, and, finally, they often use excessive amounts of 'deep' water that give them a jumpstart in the early spring before native species begin their annual growth cycle.

The preponderance of weeds (undesirable plant species) that are proliferating across all types of rangelands is perhaps the single greatest biological threat to the maintenance of rangeland health today. Many of these undesirable plants have been introduced through several pathways: as contaminated seeds or plants used in landscaping, uncertified hay and landfill, and as unwelcome residents of improperly disposed waste material. While there are natural checks and balances for many of the native weeds present on rangelands, newly introduced species often propagate and disperse like a highly virulent disease, because existing natural control mechanisms are lacking. Consequently, many of these highly invasive species can establish in even the healthiest rangelands and quickly dominate the landscape.

One such plant is leafy spurge. Leafy spurge was introduced into the USA in 1827. The weed out-competes native vegetation for water in early spring and readily establishes in a wide range of climatic types. Figure 1 shows leafy spurge stands dominating the riparian habitat in and around Devils Tower, Wyoming. The inset shows leafy spurge's ability to survive even in the harshest of areas (in this case growing out of a crack in a sandstone rock). Most invasive weeds are not used by domestic or native grazers. Limited grazing pressure increases a weed's ability to out-compete and native plants can result in monocultures that exist for decades.

Weeds impact most ecosystems; however, their impact on rangelands is usually much more pronounced for several reasons. First, the large area covered by rangelands, and their remoteness, may mean that invasive weeds are not discovered until they are well-established. Second, the individual cost to landowners, for combating weed infestations, is usually several times the economic return of the land. Finally, a failure to respond rapidly and aggressively to a weed infestation (either using herbicides or other control methods) leads to extensive problems that even most state and federal governments are financially unable to combat.

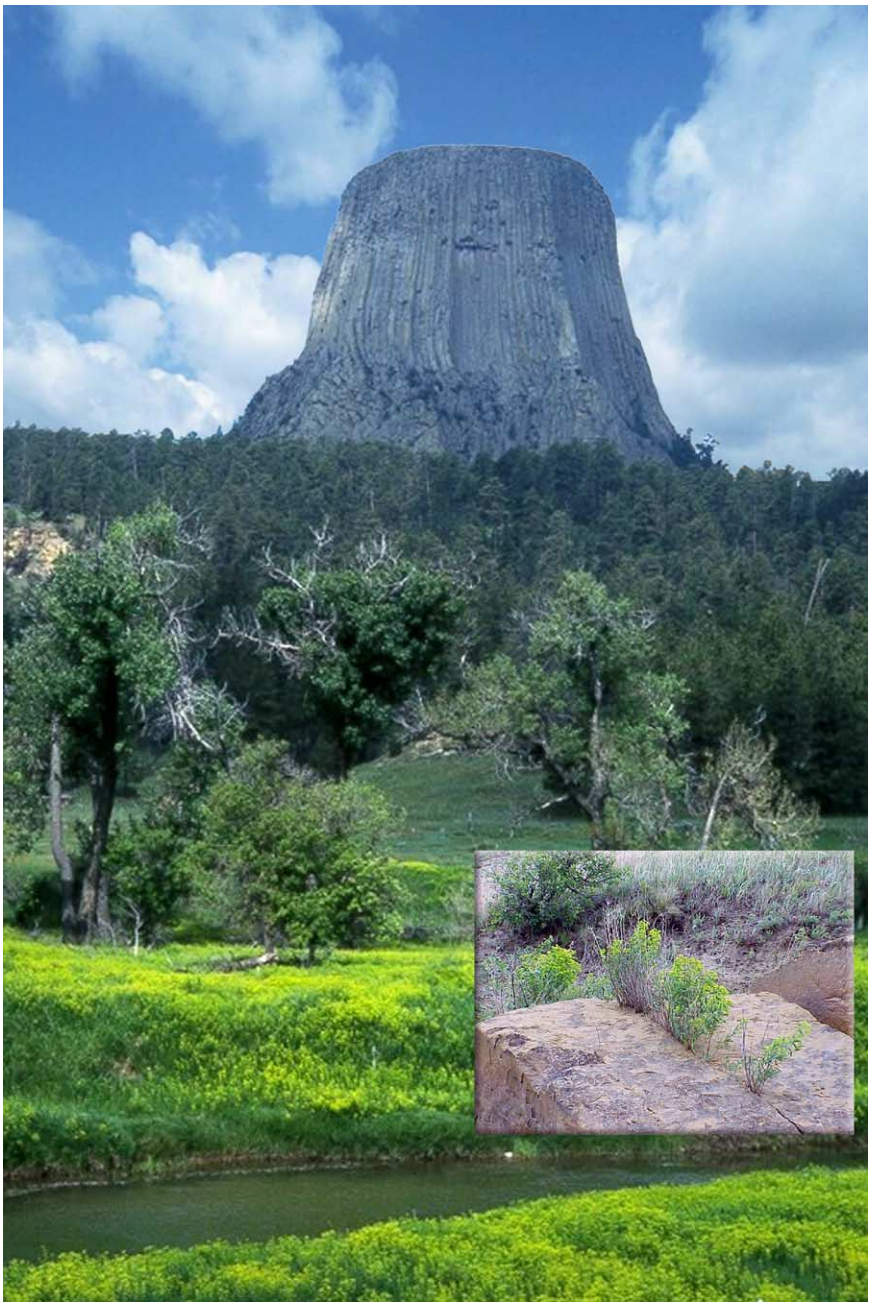


Figure 1 Leafy spurge invasion near Devils Tower, Wyoming, USA. The yellow-green weed dominates the riparian flood zone and can even establish in the harshest of areas (inset, leafy spurge growing out of a sandstone rock).

Invasive and noxious weeds are expected to infest 57 million hectares of rangeland in the USA by the year 2010. When combined with the impact of other invasive species (not restricted to rangeland), the cost to the USA is estimated to be $137 billion per year. Given the rangeland's economic disadvantages and remoteness, range managers must prioritize the set of weeds they will attempt to manage and develop new approaches that are economically and ecologically sustainable. One innovative control method, refined since the mid-1990s, is to find natural enemies that occur where the weed originated (e.g., Europe, Asia, or the Middle East), ensure the natural enemy affects only the target plant, and establish and spread these populations across areas in North America infested by the weed. Secondly, managers need to look at additional control approaches (e.g., cultural, mechanical, and chemical) to aid the natural enemy and diminish further weed establishment. The scientific term for this type of control is 'biologically based

integrated pest management,' because it combines natural enemies with other control options to provide affordable, ecologically sound, and sustainable weed control.

The Role of Disturbance in Rangeland and Soil Health

Rangelands develop under conditions of periodic disturbance (e.g., drought, fire, grazing by domesticated animals and native wildlife, weed incursions, outbreaks of insect pests, disease, and human activities). These disturbances often help maintain rangeland ecosystem biological diversity and soil health over long periods of time. If the magnitude of the disturbance(s) is too great, the results can be devastating to the resource and result in permanent ecologic changes that are undesirable and frequently destabilizing to the point that the system may not recover.

Given that excessive disturbance can have an adverse effect on the rangeland ecosystem, managers generally accept that healthy rangelands depend on a certain amount of disturbance. However, the beneficial effects of certain disturbances are subject to debate, with livestock grazing being one of the most contentious. Many in the general public and some range managers believe rangeland health has been substantially degraded by livestock grazing. The idea is that healthier rangeland ecosystems will result from the elimination of livestock grazing. This was the rationale used to eliminate livestock grazing on many hectares of rangeland in the 1990s. But, the question remains: Is this policy effective in improving rangeland and soil health?

Livestock Grazing

No Domestic Grazing

Since some individuals attribute changes in rangelands to human influences, such as grazing, one solution often proposed is to do away with domestic grazing. While this may seem reasonable to those unfamiliar with the complexities and history of the rangeland ecosystem, the results are typically more detrimental to the resource than livestock grazing.

Grazing, even overgrazing for limited periods, has always been a part of rangeland ecosystems. The vast herds of buffalo and other native grazers migrating across the landscape no longer exist and the reestablishment of the dynamic interactions of native species that existed 100–200 years ago is difficult, if not impossible. While different native grazers may be introduced, historical movement of the herds throughout the landscape (typically memorized by the oldest member of the herd) has been lost. Physical boundaries, such as roads and fences, also preclude movement of larger native grazers and would concentrate native grazing much the same as they do for domestic livestock. Therefore, the reestablishment of native grazers is seldom effective over large areas and the cost–benefit ratio often precludes introduction altogether. The net result is the elimination or reduction of rangeland grazing pressure that was typical 200 years ago.

Given the limitations of reestablishing a natural symbiotic plant–animal interaction, several things happen to the rangeland resource if livestock grazing is entirely removed. One is the reduction in the turnover rate of nutrients from aboveground plants to the soil. This often stresses existing plants that grow in the nutrient-poor soil. As plants become stressed, they begin to lose root biomass. Soil compaction typically follows (reducing the ability of the soil to take up water) and dead, aboveground plant tissue can block incoming sunlight from living plants. There also tends to be an overall loss of biodiversity. The ultimate result is usually the dieback of a portion of the root system, an increased susceptibility to disease, and an increased amount of bare soil that results in unprotected areas (gaps) where weeds can readily establish and insect populations (e.g., grasshoppers that feed on the remaining plants) can reach high levels. While this situation may occur when the range is utilized for domestic grazing, it is nearly inevitable when rangelands are not managed at all.

Therefore, the removal of grazing or human management of native rangeland does not always result in a more natural or healthier rangeland. In fact, reduced human management often results in declining range health and a rapidly changing plant component that requires an inordinate amount of resources to maintain.

Domestic Grazing

Grazing livestock on rangelands can have a negative impact when not managed properly. The existing body of scientific literature indicates that proper grazing is an important part of maintaining healthy rangelands. When rangelands are grazed using a proper grazing system (individual pastures are grazed at different times of the year to prevent overutilization of plants that prosper at different times of the year) and by only removing aboveground plant material to acceptable levels (leaving approximately 25% of the produced plant material), there is a positive impact on the soil–root interface. Roots continue to receive the photosynthetic materials needed to flourish and maintain a healthy mixture, many of the nutrients needed for plant growth are recycled

quickly by the cattle and soil microbes back to the soil, and the stress of aboveground plant removal does not favor any single plant component (warm-season or cool-season plants). Rangelands flourish under proper grazing management. Rooting systems and species composition are maintained, soil structure and composition are protected from wind and water erosion, and the health of the system is more resilient to weed, insect, and disease impacts.

Overgrazing

In this instance, the overutilization of the range due to excessive use or during periods of extended drought can worsen and ultimately cause increased soil erosion, soil compaction, and the permanent alteration of the soil seedbank and the rangeland plant community.

In any management application, there is always the possibility that the rangeland resource can be overutilized. However, the overwhelming majority of ranchers are good stewards of the range, more so because they choose to ranch as a lifestyle rather than for the earning potential of the job. To do that on an often limited income, they must maintain the quality of the rangeland needed for their livestock.

There are, however, times when even the best management plans can be taxed. Balancing grazing pressure during periods of drought, disease, or pest infestation is difficult. It is hard to predict how long adverse situations will last or how extensive problems will become. Therefore, there is often a lag-time between management changes and negative environmental impact. The result is, during periods of extended adverse conditions, management responses may not come quickly enough to prevent some damage to the rangeland resource.

What happens to rangelands when they are overgrazed for long periods or when improper management strategies are applied? Perhaps the most predominant impact of long-term overgrazing (or drought coupled with grazing) is the removal of more plant material than is recommended. For example, in the US Great Plains, the removal of more than 75% of aboveground production is considered detrimental to rangeland health. At this point there is not enough photosynthetically active plant material to provide the roots with the nutrients needed to maintain a healthy balance. The result is host dieback, and associated susceptibility to disease, repeated weeds and insect infestations.

Summary

Maintaining or improving rangeland health is the primary concern of most range managers regardless of the ecologic concept they espouse. But the approaches they use to achieve that goal often diverge, particularly when a single rangeland concept or model is overemphasized. This is particularly true when the impact of periodic disturbances is considered.

Rangelands are dynamic and complex. Therefore our understanding and management strategies must also be dynamic to account for new knowledge. This is important for ensuring that rangeland health and soil health can be maintained for generations to come.

See also: **Grassland Soils**

Further Reading

Devine R (1998) *Alien Invasion: America's Battle with Non-native Animals and Plants*. Washington, DC: National Geographic Society.

Halbach DW, Runge CF, and Larson WE (eds) (1987) *Making Soil and Water Conservation Work: Scientific and Policy Perspective*. Ankeny, IA: Soil Conservation Society of America.

Satterlund DR (1972) *Wildland Watershed Management*. New York: John Wiley.

Sheley RL and Petroff JK (eds) (1999) *Biology and Management of Noxious Rangeland Weeds*. Corvallis, OR: Oregon State University Press.

Stoddart LA, Smith AD, and Box TW (1955) *Range Management*. New York: McGraw-Hill.

Vallentine JF (1980) *Range Development and Improvement*. Provo, UT: Brigham Young University Press.

Recycling of Organic Wastes *See* **Pollutants:** Biodegradation

Redistribution *See* **Water Cycle**

REDOX POTENTIAL

R D DeLaune, Louisiana State University, Baton Rouge, LA, USA
K R Reddy, University of Florida, Gainesville, FL, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Geochemists, soil scientists, and limnologists have used redox potential (E_h) measurements to characterize oxidation–reduction status of surface environments. The redox potential of soil, water, and marine systems is a measure of electrochemical potential or electron availability within these systems. Electrons are essential to all inorganic and organic chemical reactions. Redox potential measurements allow for rapid characterization of the degree of reduction and for predicting stability of various compounds that regulate nutrients and metal availability in soil and sediment. Redox potential is diagnostic for determining whether an area is functioning as wetland or nonwetland.

Oxidation and reduction reactions regulate many of the biogeochemical reactions in surface environments. Redox potential (E_h) is determined from the concentration of oxidants and reductants in the environment. The inorganic oxidants include oxygen, nitrate, nitrite, manganese, iron, sulfate, and CO_2, while the reductants include various organic substrates and reduced inorganic compounds.

Theoretical Relationships

Oxidation and reduction reactions involve transfer of electrons from one compound to another and play a major role in regulating many reactions in biological systems. Oxidation–reduction reactions are coupled reactions.

Oxidation is defined as removal of electrons from a compound. Such a compound is usually referred to as 'electron donor' or 'reductant.' During this process it is oxidized and increases its oxidation number:

$$\text{Reductant} = \text{Oxidant} + e^-$$

Reduction is defined as the addition of electrons to a compound.

A compound that accepts electrons is usually referred to as 'electron acceptor' or 'oxidant.' During this process the compound is reduced and its oxidation number is decreased:

$$\text{Oxidant} = \text{Reductant} + e^-$$

The tendency of compounds to accept or donate electrons is expressed as 'reduction potential' or 'redox potential.' The redox potential of a substance depends upon:

- Affinity of molecules for electrons;
- Concentration of reductants and oxidants (referred to as redox pair);

Reductants (electron donors) in wetland soils are:

- Organic matter and various organic compounds;
- Reduced inorganic compounds such as NH_4^+, Fe^{2+}, Mn^{2+}, S^{2-}, CH_4, and H_2.

Oxidants (electron acceptors) are inorganic compounds such as O_2, NO_3^-, MnO_2, FeOOH, SO_4^{2-}, and HCO_3^-.

Anaerobic environments such as wetland soils are usually limited by electron acceptors and have an abundant supply of electron donors. Aerated soils are usually limited by electron donors and have an abundant supply of electron acceptors (primarily O_2).

Redox–pH Relationships

Redox reaction in surface environments can be represented by the following half-cell reduction equation:

$$\text{Ox} + m\text{H}^+ + ne^- \rightarrow \text{Rd} \qquad [1]$$

where Ox is the oxidized component or electron acceptor, Rd is the reduced component or electron donor, m is the number of hydrogen ions involved in the reaction, and n is the number of electrons involved in the reaction. The electrons used in eqn [1] must be supplied by an accompanying half-cell oxidation reaction. In soils, organic matter (CH_2O) is the primary source of electrons. The oxidation reaction balances the reduction reaction. This is illustrated below with the reduction of $Fe(OH)_3$:

$$4Fe(OH)_3 + 12H^+ + 4e^- \rightarrow 4Fe^{2+} + 12H_2O \text{(reduction)}$$
$$CH_2O + H_2O \rightarrow CO_2 + 4H^+ + 4e^- \text{(oxidation)}$$
$$4Fe(OH)_3 + CH_2O + 8H^+ \rightarrow 4Fe^{2+} + CO_2 + 11H_2O$$

The reduction reaction can be defined quantitatively through the change in Gibbs free energy (ΔG):

$$\Delta G = \Delta G^0 + RT \ln\left(\frac{\text{Rd}}{(\text{Ox})(\text{H}^+)^m}\right) \qquad [2]$$

where ΔG^0 is the standard free energy change, R is the gas constant, and T is absolute temperature. The Nernst equation expresses the reduction reaction in terms of electrochemical energy (millivolts) using the relationship $\Delta G = -nEF$

$$E_h = E^0 - \frac{RT}{NF}\ln\left(\frac{(Rd)}{(Ox)}\right) - \frac{mRT}{nF}\ln(H^+) \qquad [3]$$

where E_h is the electrode potential, E^0 is the standard half-cell potential, F is the Faraday constant, n is the number of electrons exchanged in the half-cell reaction, m is the number of protons exchanged, and the activities of the various oxidized and reduced components are shown in parentheses. Substituting values of $8.31\,J\,K^{-1}\,mol^{-1}$ for R, $9.65 \times 10^4\,C\,mol^{-1}$ for F, and 298 K for T, and using the relationship $\ln(x) = 2.303\log(x)$, eqn [3] simplifies to:

$$E_h(mV) = E^0 - \frac{59}{n}\log\left(\frac{(Rd)}{(Ox)}\right) + 59\frac{m}{n}pH \qquad [4]$$

Inspection of eqns [3] and [4] reveals that redox potential (E_h) increases with increasing activity of the oxidized component, decreases with increasing activity of the reduced component, and increases with an increase in H^+ activity (or a decrease in pH).

Equation 4 shows the importance of pH in redox reactions. When the ratio of protons to electrons (m/n) is equal, there is a predicted 59 mV change in E_h per pH unit. The E_h/pH slope can vary depending on the oxidants and reductants involved.

The E_h/pH slope predicted from the Nernst equation assumes that the redox couple controls the pH of the system. This is only true for pure chemical solutions. Measured slopes in natural systems due to mixed potential or various redox couples deviate from the predicted values. Applying the 59 mV correction factor cannot be used under those conditions.

The reduction reaction in eqn [1] can also be defined by an equilibrium constant:

$$K = \frac{(Rd)}{(Ox)(e^-)^n(H^+)^m} \qquad [5]$$

which can be rewritten in log form:

$$\log K = \log(Rd) - \log(Ox) - n\log(e^-) - m\log(H^+) \qquad [6]$$

The $-\log(e^-)$ term in eqn [6] can be defined as pE (relative electron activity) in the same way pH is defined as $-\log(H^+)$. Rearranging eqn [5] yields:

$$pE + pH = \log K - \log(Rd) + \log(Ox) \qquad [7]$$

Equation [7] can be used for plotting equilibrium solubility diagrams, because the pE + pH sum can be used as one axis and species activity as the other axis. E_h is the more-common expression of soil redox potential, since it is readily determined with a platinum electrode. Soil pE can be calculated from E_h using eqn [8]:

$$pE = \frac{E_h(mV)}{59} \qquad [8]$$

Normal Limits of pH and Redox Potential (E_h) in the Environment

Carbon dioxide from the atmosphere and organic acids formed by decaying organic matter are the two sources of acidity in most near-surface environments, resulting in acid pHs in the range 5–6. Lower pHs are found in acid sulfate soils, where values as low as 4.0 are sometimes recorded. As a result, a pH of 4 can be selected as the usual lower limit of pHs in natural environments (**Figure 1**).

The upper end of the pH limits is associated with CO_2-free water in contact with carbonate rocks and can acquire a pH of 10 and, in contact with some silicates, a pH up to nearly 11. But since most surface waters emit CO_2 to the atmosphere, such alkalinities are not attained. A reasonable upper limit of pH in most near-surface environments is near 9.0.

Oxygen is the strongest oxidizing agent commonly found in nature. Stronger agents than this cannot persist, because they react with water to liberate oxygen. Thus the upper limit of redox potentials is defined by the reaction:

$$H_2O \Leftrightarrow \frac{1}{2}O_2 + 2H^+ + 2e^- \qquad E^0 = +1.23\,V$$

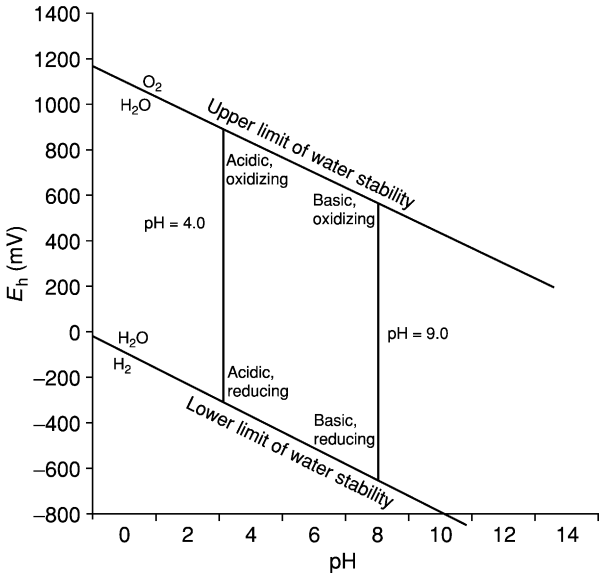


Figure 1 The E_h–pH range found in soil surface environments. Adapted from Krauskoft KB (1967) *Introduction to Geochemistry*. New York, NY: McGraw Hill.

The potential of this half-reaction, which reflects the upper limitations of water stability, clearly depends on the pH, as shown by the equation:

$$E = +1.23 + 0.03\log[O_2]^{1/2}[H^+]^2$$

The lower limit of redox potential is that of the hydrogen reaction,

$$H_2 \Leftrightarrow 2H^+ + 2e^- \qquad E^0 = 0.00\,V$$

For which

$$E = 0.00 + 0.03\log[H^+]^2 - 0.03\log[H_2]$$
$$= -0.059\,pH - 0.03\log[H_2]$$

Since the pressure of hydrogen in surface environments could not exceed 101 kPa, the maximum possible reducing potential in the presence of water would be:

$$E_h = -0.059\,pH - 0.03\log(1) = -0.059\,pH$$

Aerobic and/or Anaerobic Redox Potential Range in the Environment

The E_h scale used to quantify reduction intensity (or oxidation intensity) in chemical and biological systems is shown in Figure 2. The scale is in electromotive force units or volts (or millivolts) and can be divided into zones ranging from oxidized (where the aerobes function) to highly reduced (where the methane producers and sulfate reducers function).

Typically in wetland soils, E_h values range from −300 to 700 mV, with a total range of approximately 1000 mV. In aerobic soils where the O_2/H_2O redox couple functions, the E_h range is between 300 and 700 mV.

Redox Couples in Wetlands

Although several redox couples function in wetlands soils, the following are the most common reduction reactions involving a specific redox couple:

$$O_2 + 4e^- + 4H^+ \rightarrow 2H_2O$$
$$2NO_3^- + 10e^- + 12H^+ \rightarrow N_2 + 6H_2O$$
$$MnO_2 + 2e^- + 4H^+ \rightarrow Mn^{2+} + 2H_2O$$
$$Fe(OH)_3 + e^- + 3H^+ \rightarrow Fe^{+2} + 3H_2O$$
$$SO_4^{2-} + 8e^- + 8H^+ \rightarrow S^{2-} + 4H_2O$$
$$CO_2 + 8e^- + 8H^+ \rightarrow CH_4 + 2H_2O$$

The various inorganic redox systems found in soil and sediment become unstable at critical redox potentials (Figure 3). Sequentially following flooding, oxygen is reduced first, followed by nitrate and oxidized manganese compounds, and then ferric iron compounds. After the reduction of ferric iron, the next redox compound to become unstable is sulfate, followed by the reduction of carbon dioxide to methane.

Soil redox potential represents an indication of the oxidation–reduction status of the various redox couples. For example, a redox potential of 0 mV indicates that oxygen and nitrate are not likely to be present and that the bioreducible iron and manganese compounds are in a reduced state. At this same potential, however, sulfate is stable in the soil with no production of sulfide, which is toxic to plants. A redox potential of +400 mV indicates that oxygen may be present even though there may be excess water.

Redox potential measurements in pure systems can be used to predict the ionic distribution between chemical species which may interact with the transfer of electrons, such as ferrous and ferric iron or nitrite and nitrate nitrogen. However, in natural systems, there are many redox couples present, and not all redox couples are chemically interactive with others. Unless the concentration of a given redox couple is relatively high, inert electrodes (generally platinum) used for redox measurements are not specific for a specific redox couple. The redox electrode responds to the collective electrochemical potentials of all redox couples present.

The measured redox potential in soil is generally a mixed potential which reflects a weighted average of the potentials contributed by each of the redox couples present. As a result of the continuous

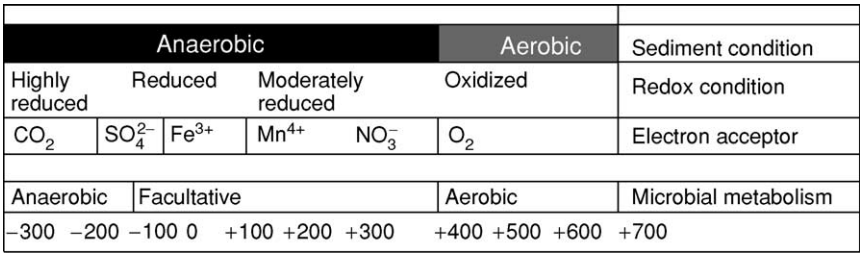


Figure 2 Redox potential range in soil and sediment, showing the microbial metabolism process and electron acceptor.

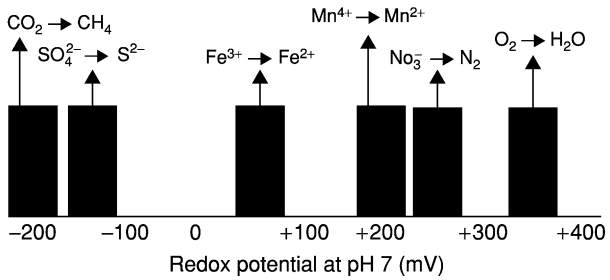


Figure 3 Critical redox potential for the transformation of redox couples.

addition of organic matter, which oxidizes and serves as an electron donor, a redox equilibrium is almost never attained in a natural system.

With the theoretical limitations involved in the use of redox potentials to quantitatively describe a specific ionic distribution in a mixed system, redox measurements can be successfully applied to characterize the oxidation–reduction transformations of many elements, including heavy metals and plant nutrients in soil.

Predicting Mineral Stability

E_h and pH can also be used for plotting stability fields for specific minerals. For example, the E_h–pH stability field for iron in Figure 4 can be interpreted that hematite (Fe_2O_3) is a stable iron mineral under oxidizing environments. Under reducing environments, a stable mineral may be pyrite. If dissolved sulfur is low, siderite ($FeCO_3$) may be the predominant mineral under reducing conditions.

Intensity and Capacity of Reduction

Redox potential measurement is only one measure of intensity of reduction in soils and sediment, as described below.

Intensity

The ease of reduction is represented by the free energy of reduction or equivalent electromotive forces of the reaction (oxidation–reduction potential or redox potential).

The reduction processes that take place in a soil as a result of the oxidation of organic matter utilizing alternate electron acceptors of nitrate, manganic manganese, ferric iron, sulfate, and carbon dioxide do not proceed together. As indicated previously, nitrate is utilized first after oxygen disappears, followed by manganic manganese, ferric iron, sulfate, and finally carbon dioxide. For this reason the soil must be considerably more reduced for sulfate to be

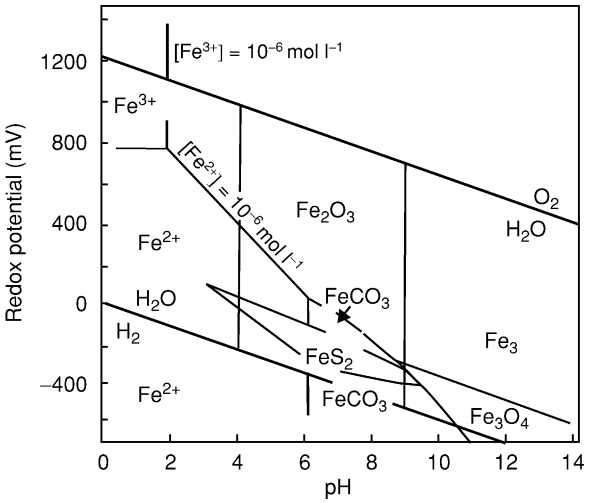


Figure 4 Redox–pH effect on stability of iron minerals. Adapted from Krauskoft KB (1967) *Introduction to Geochemistry*. New York, NY: McGraw Hill.

reduced than for nitrate to be reduced. If nitrate and sulfate and a suitable organic substrate are all present in a flooded, oxygen-deficient soil at the same time, the denitrifiers will reduce all of the nitrate before any of the sulfate is reduced by the sulfate reducers. The intensity of reduction is different for the two reduction processes. The soil is not very reduced (low reduction intensity or low electron pressure) for the reduction of nitrate, while it is very reduced during the reduction of sulfate (high reduction intensity or high electron pressure).

This intensity of reduction can be thought of in terms of the pressure of the electrons that the microorganisms need to dispense with as they carry out oxidation of the energy source. It does not require much electron pressure for the nitrate to accept electrons (and even less electron pressure for oxygen, if it is present, to acceptor electrons). This is why a large energy yield can be obtained from the oxidation of the energy source, since not so much of the possible energy from the oxidation process is utilized in forcing the electron acceptor to accept the electrons from the organic matter oxidation.

Electron acceptors that are difficult to reduce require more electron pressure to force the electrons to move to the electron acceptor. For example, sulfate reducers in soil and sediment can oxidize organic substrates, but a great deal of the energy derived from the oxidation is required to force the sulfate ion to accept the electrons and be converted to sulfide. This leaves less energy than the microorganism derives from the reaction that it can use for its energy needs. A great intensity of reduction is required for

sulfate to accept the electrons produced during the oxidation of the energy source.

This approach to determine the intensity of reduction is called the oxidation–reduction or redox potential of the soil. Measurement of redox potential is conducted by placing a platinum electrode in the soil, connecting it to a suitable circuit that balances and measures the electron pressure from the reduction processes.

Reduction capacity described below differs from intensity of reduction.

Capacity

The total amount of electrons accepted by oxidants or number and redox system undergoing reduction is the 'capacity.' The capacity can be best described in terms of its O_2 equivalent (not redox potential measurement).

The capacity factor of the various redox systems varies from one soil to another. The amount of oxygen present in the soil at the time of flooding is usually low. The amount of NO_3^- present in the soil is dependent on soil organic matter and can be variable. Biologically reducible Mn can be present in soils at varying levels (less than 100 to 1000 mg kg^{-1}), while reducible Fe can be in the range of 500–3000 mg kg^{-1} of soil.

Although the critical redox potential at which the inorganic redox systems become unstable provides valuable information, it does not provide any indication of the total capacity of the system to accept electrons or support microbial respiration. Capacity factor is equivalent to the total amount of electrons accepted by the oxidants in support of microbial respiratory activity. The redox couple present in lowest quantity has the smallest capacity for supporting microbial respiration. In mineral-flooded soils, reducible Mn and Fe primarily support the organic matter decomposition. In estuarine areas that receive sea water containing large amounts of SO_4^{2-}, the reduction of SO_4^{2-} to S^{2-} supports the microbial respiratory activities.

Redox Condition and Metal/Metalloid Chemistry

Redox conditions of soil and sediment deposits also govern heavy metal chemistry (Figure 5). For example, chromium can exist in several oxidation states from Cr(0), the metallic form, to Cr^{6+}. The most stable oxidation states of chromium in the environment are Cr^{3+} and Cr^{6+}. Besides the elemental metallic form, which is extensively used in alloys,

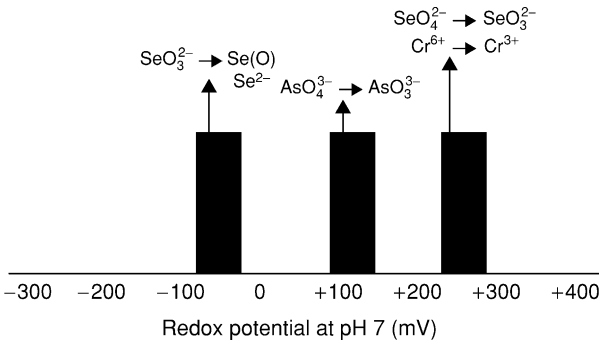


Figure 5 Redox potential as related to metal and/or metalloid chemistry or transformation.

chromium has three important valence forms: Cr^{2+}, Cr^{3+}, and Cr^{6+}.

The trivalent chromic Cr^{3+} and the tetravalent dichromate Cr^{6+} are the most important forms in the environmental chemistry of waters and sediments. The presence of chromium Cr^{6+} is of particular importance, because in this oxidation state Cr is water-soluble and extremely toxic. They are also the only forms that undergo valence changes in the E_h–pH ranges encountered in natural systems. The solubility and potential toxicity of the chromium that enters into aquatic or sediment columns are governed to a large extent by oxidation–reduction reactions. Cr^{6+} reduces to Cr^{3+} at approximately +300 mV (Figure 5).

Oxidation–reduction processes also play a role in arsenic (As) and selenium (Se) chemistry (Figure 5). For the pH and redox conditions in most soils and sediments, As exists as an oxyanion in the arsenate ($H_2AsO_4^-$, and $HAsO_4^{2-}$, As^{5+}) or arsenite [H_3AsO_3; As^{3+}] form. Monomethylarsenic acid (MMAA) and dimethylarsenic acid (DMAA) are important organoarsenic chemical species. Se occurs as selenate (SeO_4^{2-}, Se^{6+}), selenite ($HSeO_3^-$ and SeO_3^{2-}; Se^{6+}), and elemental selenium (Se(0)), but selenide (DMSe) is the most important. At higher soil redox levels (+200 to +500 mV), As^{5+} is the predominant As species present. The reduction of As^{5+} to As^{3+} occurs at redox levels corresponding within the nitrate-reducing zone of soils characterized by a soil redox level of approximately +300 mV. When Fe reduction starts and soil redox levels drop below +50 mV, selenite is reduced to elemental Se or metal selenides. Reoxidation of reduced As and Se species occurs at similar soil redox levels to those identified for reduction reactions. Both microbial-mediated and chemical-oxidation processes are thought to be involved in the oxidation of reduced As and Se species.

Measurement of Redox Potential

Inert metal electrodes (usually platinum) are widely used to measure the oxidation–reduction potentials of chemical systems. The inert metal electrode is used in conjunction with a reference electrode to form a complete cell. The most widely used reference electrodes for redox potential measurement are the calomel (mercury–mercurous chloride) and the silver/chloride electrode. The redox potential (E_h) is obtained from the potential of the cell (E) by adding the appropriate value for the reference electrode potential corrected for liquid junction effects. The potential measured with the platinum electrode and the secondary reference electrode must be corrected to give true E_h or redox potential.

The platinum electrode referenced to a saturated calomel electrode will give E_h or millivolt readings, provided corrections are made for the potential of the calomel electrode.

Performance of the platinum electrodes is checked by immersing the platinum electrode along with the reference calomel electrode in pH 4.00 and pH 7.00 buffer to which a small amount of quinhydrone has been added. The reading on the E_h meter in millivolts should agree with that shown in Table 1. The meter reading is corrected relative to hydrogen by adding the calomel potential (Table 1). This value is the quinhydrone E_h in millivolts relative to hydrogen. For example, a meter reading of 218 mV at 25°C for quinhydrone in pH 4.00 buffer gives an E_h of 462.3 mV:

$$E_h = \text{meter reading} + \text{calomel potential}$$
$$E_h = 218.0 + 244.3 = 462.3\ \text{mV}$$

An electrode is considered satisfactory if it does not deviate from the above values by more than 5 mV.

When making soil E_h measurements, the calomel potential must be added to the meter reading (millivolts). The procedure is as follows:

The potential is measured by inserting the electrodes in the substance to be measured and reading the millivolt scale on the pH meter. Since the calomel electrode is used as a reference, +244.3 mV is added to the recorded potential for reference to the hydrogen electrode.

Table 1 Effect of temperature on the standard potential of a calomel half-cell and platinum electrode

Temperature (°C)	Calomel potential (mV)	Meter reading[a]	
		pH 4 (mV)	pH 7 (mV)
5	257.0	236.0	66.5
10	253.9	231.6	60.1
15	250.8	227.3	53.7
20	247.3	222.7	47.3
25	244.3	218.0	40.8
30	240.9	213.4	34.3
35	237.5	208.5	27.7
40	234.1	203.5	21.0
45	230.6	198.4	14.2

[a]Meter reading with platinum and calomel electrode immersed in buffer saturated with quinhydrone.

Further Reading

Bass Becking LGM, Kaplan IR, and Moore D (1960) Limits of the natural environment in terms of pH and oxidation–reduction potentials. *Journal of Geology* 68: 243–286.

Faulkner SP, Patrick WH Jr, and Gambrell RP (1989) Field techniques for measuring wetland soil parameters. *Soil Science Society of America Journal* 53: 883–890.

Garrels RM and Christ CM (1985) *Solutions, Minerals, and Equilibria*. New York: Freeman and Cooper.

Grenthe I, Stumm W, Lacksuharju M, Nilsson AC, and Wikbers P (1992) Redox potentials and redox reactions in deep groundwater systems. *Chemical Geology* 98: 131–150.

Patrick WH, Gambrell RP, and Faulkner SP (1996) Redox and measurement of Soils. In: Sparks DL *et al.* (eds) *Methods of Soil Analysis*, part 3, *Chemical Methods*, pp. 1255–1273. Soil Science Society of America Box Series No. 5. Madison, WI: Soil Science Society of America.

Ponnamperuma FN (1972) The chemistry of submerged soils. *Advances in Agronomy* 24: 29–96.

Reddy KR and Delaune R (2004) *Biogeochemistry of Wetlands*. CRC (in press).

REDOX REACTIONS, KINETICS

P S Nico, California State University–Stanislaus, Turlock, CA, USA
S Fendorf, Stanford University, Stanford, CA USA

© 2005, Elsevier Ltd. All Rights Reserved.

Oxidation and reduction (redox) reactions constitute a diverse, complex, and extremely important class of reactions in soils. Redox reactions influence properties, such as solubility and reactivity, of both organic and inorganic species/phases. For example, the two environmentally important oxidation states of Cr, Cr(III) and Cr(VI), have dramatically different properties. Cr(III) is sparingly soluble, tending to precipitate as Cr or Fe/Cr oxyhydroxides, while Cr(VI) exists as the soluble and quite mobile CrO_4^{2-} oxyanion. Similarly, properties of fulvic and humic acids, including their contrasting solubilities and sorption behavior, are fundamentally a consequence of their oxidation state. Moreover, all biological life is dependent on redox reactions proceeding at disequilibrium.

Because of the dramatic effect that oxidation state can have on the properties of important compounds, soil chemists are often interested in being able to understand, model, and predict the redox behavior of a certain species under various conditions. We therefore need two kinds of information: thermodynamic and kinetic. Thermodynamics describes the overall path and time-independent energetics of a reaction. Kinetics describes the time-dependent behavior of a reaction. Thermodynamic data are generally more accessible; they answer unequivocally whether a reaction is 'favored' (i.e., possible) under certain conditions. However, reaction kinetics dictate the rate at which a reaction will proceed; reactions that do not proceed 'instantaneously' are often termed 'kinetically controlled.' For example, the oxidation of carbon as diamond to CO_2 by atmospheric O_2 is thermodynamically favored (by $\sim$400 kJ/mol^{-1}) but proceeds appreciably only on geologic time scales. As we study kinetics, it is crucial to keep in mind that thermodynamic driving forces do not dictate reaction kinetics.

The vast majority of the redox reactions of interest to the soil chemist are kinetically controlled, and all life depends on kinetic constraints maintaining disequilibrium. Because of this, the study of redox kinetics is of central importance to understanding almost any soils system. Kinetic data are usually combined into a mathematical description of the reaction kinetics known as the kinetic rate law. A complete rate law contains terms for all the variables in a system that affect the time-course of the reaction. These factors include the chemical species involved in the reaction as well as other factors such as temperature, pressure, or ionic strength. The rate law also contains a constant that is an indication of the reaction's relative speed, known as the rate constant. The initial form and some of the important terms in a rate law can occasionally be deduced from equilibrium data or from a theoretical description of the reaction, but complete rate laws, other than the simplest gas phase reactions, are experimentally determined. Rate laws are essential for a variety of purposes, including modeling experimental data, understanding the detailed molecular steps involved in the reaction (the reaction mechanism) or, most importantly, to model and predict actual field behavior.

Kinetic Rate Laws

We mathematically define the rate of a reaction in terms of how quickly a given reactant is consumed or product is formed. In order to do this we must already have good equilibrium data on the reaction that tells us the reactants, the products, and the stoichiometric relationships between these species. For example, a reaction like that described in eqn [1], where the reactants A and B combine to form products C and D, and the stoichiomteric coefficients of each reactant or product, are given by a, b, c, or d respectively.

$$aA + bB \rightleftarrows cC + dD \qquad [1]$$

The rate can be defined based on any of the reactants or products as follows, where square brackets denote the concentration (and not activities) of a species.

$$\text{Rate} = -\frac{1}{a}\frac{d[A]}{dt} = -\frac{1}{b}\frac{d[B]}{dt} = \frac{1}{c}\frac{d[C]}{dt} = \frac{1}{d}\frac{d[D]}{dt} \qquad [2]$$

Eqn [2] is valid as long as the stoichiometric coefficients in eqn [1] are accurate under the given conditions (i.e., there are no side reactions that consume or produce species within the reaction). For cases involving side reactions, the proportionality of the reaction/products would not depend solely on the coefficients in eqn [1].

The rate as defined in eqn [2] is the difference between the rate of the forward reaction (eqn [1] read from left to right, which will be a function of, among other things, the concentration of A and B) and the reverse reaction (eqn [1] read from right to left, which will be a function of the concentrations of C and D).

$$\text{Rate}_{net} = \text{Rate}_{forward}([A],[B]) - \text{Rate}_{reverse}([C],[D]) \qquad [3]$$

Equilibrium is the point where the concentrations of reactions and products remain constant, or, in

other words, that Rate_{net} is zero. We can see from eqn [3] that equilibrium does not mean that eqn [1] has ceased to occur but simply that the $\text{Rate}_{\text{forward}}$ is equal to the $\text{Rate}_{\text{reverse}}$. As long as eqn [1] is far from equilibrium, we can neglect the rate of the reverse reaction and concentrate solely on the forward reaction. As stated already, the $\text{Rate}_{\text{forward}}$ will be a function of the concentrations of A and B, known as the rate law. The exact nature of this function must be determined experimentally, but for many reactions the rate law takes the form shown in eqn [4], where $[A]$ and $[B]$ are the concentrations of reactants A and B; concentration units of molarity (M or mol l^{-1}) are usually used for dissolved species. Solid species can also be reported in terms of molarity or in mass (mg l^{-1}) or surface area (m^2l^{-1}) concentrations. The constant k is an experimentally determined proportionality factor known as the rate constant. The constants α and β are also experimentally determined and are denoted as the reaction order. If α is equal to 1 and β is equal to 2, then the reaction is said to be first-order in A, second-order in B, and third-order overall. The units of k are dictated by the reaction order such that the units of the rate will come out appropriately; usually rates are reported in units of concentration per time, such as $\text{mol l}^{-1}\,\text{s}^{-1}$.

$$\text{Rate}_{\text{forward}} = k[A]^{\alpha}[B]^{\beta} \qquad [4]$$

If eqn [1] was an elementary reaction, the coefficients in eqn [1] (a and b) would be equal to the α and β of eqn [4]. Unfortunately, most soil systems are sufficiently complex that we must deal in overall reactions, meaning reactions whose reactants, products, and stoichiometry have been determined macroscopically. In this case, there is no theoretical link between the coefficients in eqn [1] and the exponents in eqn [4]. In fact, the exponents in eqn [4] need not even be whole numbers.

Integrated Rate Laws

Remember that eqn [4] is in fact a differential equation of the form shown in eqn [5]. If we were able to integrate eqn [5] we would be able to calculate the concentrations of A or B at any given time.

$$\text{Rate}_{\text{forward}} = -\frac{1}{a}\frac{d[A]}{dt} = -\frac{1}{b}\frac{d[B]}{dt} = k[A]^{\alpha}[B]^{\beta} \qquad [5]$$

The integration of a general equation like eqn [5] is not trivial. For this reason, let us first consider a simpler system. Eqn [6] shows the rate law for a reaction that depends only on the concentration of A and is first-order in A.

$$\text{Rate} = -\frac{d[A]}{dT} = k[A] \qquad [6]$$

Eqn [6] is easily integrated to yield eqn [7], where $[A]_0$ is the initial concentration of A, $[A]_t$ is the concentration of A at any time t, and k is the first-order rate constant.

$$[A]_t = [A]_0 e^{-kt} \qquad [7]$$

The integrated rate laws for zero- and second-order, single-species rate laws are shown in **Table 1**.

Table 1 Integrated rate law expressions

Reaction order	*Rate law*	*Integrated rate law*	*Units of* k	*Linear plot*
0	$-\frac{d[A]}{dt} = k$ [8]	$[A]_t = [A]_0 - kt$ [9]	$\text{mol l}^{-1}\,\text{s}^{-1}$	$[A]_t$ versus t
1	$-\frac{d[A]}{dt} = k[A]$ [10]	$[A]_t = [A]_0 e^{-kt}$ [11]	s^{-1}	$\ln[A]_t$ versus t
2	$-\frac{d[A]}{dt} = k[A]^2$ [12]	$\frac{1}{[A]_t} = \frac{1}{[A]_0} + kt$ [13]	$\text{mol}^{-1}\,\text{l}\,\text{s}^{-1}$	$\frac{1}{[A]_t}$ versus t
2	$-\frac{d[A]}{dt} = k[A][B]$ [14]	$\frac{1}{[B]_0 - [A]_0}\ln\frac{[A]_0[B]_t}{[B]_0[A]_t} = kt$ [15] $\frac{1}{[A]_t} = \frac{1}{[A]_0} + bkt$ [16]	$\text{mol}^{-1}\,\text{l}\,\text{s}^{-1}$	
$\alpha+\beta+\chi$	$-\frac{1}{a}\frac{d[A]}{dt} = k[A]^{\alpha}[B]^{\beta}[C]^{\chi}$ [17]	None	$\text{mol}^{-(\alpha+\beta+\chi-1)}\text{l}^{(\alpha+\beta+\chi-1)}\text{S}^{-1}$	

Reactions whose rates are controlled solely by one species and are exactly zero; first-, or second-order do exist, and sometimes the order of a reaction can be identified solely by collecting concentration data over time for a single reactant and comparing to one of the simple rate laws. As seen in **Table 1**, if the reaction is first-order in A then a plot of the $\ln[A]_t$ versus t should be a straight line. Similarly, if the reaction is second-order in $[A]$, then a plot of $[A]_t^{-1}$ versus t should be a straight line.

Unfortunately, the majority of reactions of interest do not fall neatly into one of the easily integrateable rate laws, and it is more common that we must use mathematics or experimental design to simplify a more complex rate law, like eqn [17]. The most common of these techniques are discussed below, but a clever experimentalist or mathematician can always devise new approaches.

Initial Rates

The technique of initial rates can be used to determine the order of an individual reactant within a complex rate expression. The initial rate is the linear reaction rate at the beginning of a reaction before a significant amount (<10%) of any of the reactants has been consumed and before there is any appreciable reverse reaction. This is expressed mathematically in eqn [18].

$$\text{Rate}_{\text{initial}} = -\frac{1}{a}\frac{d[A]}{dt} \approx -\frac{1}{a}\frac{\Delta[A]}{\Delta t} \approx k[A]^{\alpha}[B]^{\beta}[C]^{\chi} \quad [18]$$

For example, if we wanted to determine α, the order of A in eqn [17], we need only to measure the initial rate at two different A concentrations. We can then take the ratio of these two rates as shown in eqn [19].

$$\frac{\text{Rate}_1}{\text{Rate}_2} = \frac{k[A_1]^{\alpha}[B]^{\beta}[C]^{\chi}}{k[A_2]^{\alpha}[B]^{\beta}[C]^{\chi}} \quad [19]$$

Since the rate constant, k, and the concentration of B and C should be equal in both equations, eqn [19] reduces to eqn [20].

$$\frac{\text{Rate}_1}{\text{Rate}_2} = \frac{[A_1]^{\alpha}}{[A_2]^{\alpha}} = \left(\frac{[A_1]}{[A_2]}\right)^{\alpha} \quad [20]$$

α can be determined by taking the log of both sides and rearranging, to give eqn [21].

$$\alpha = \frac{\text{Log}\left(\frac{[\text{Rate}_1]}{[\text{Rate}_2]}\right)}{\text{Log}\left(\frac{[A_1]}{[A_2]}\right)} \quad [21]$$

This process can be repeated to determine the order of any number of reactants. Because these are reaction orders for a complex overall reaction, α, β, and χ need not be whole numbers.

Although it is true that in theory one needs only to determine the initial rate at two different concentrations of A, it is good practice to employ several different concentrations to make sure that the orders calculated with any pair of A concentrations are the same (i.e., the reaction rate is determined at varying concentrations of A while all other concentrations are maintained at the same value). In fact, one can rearrange eqn [18] to give eqn [22]. Using this equation a plot of $\log(\text{Rate}_{\text{initial}})$ versus $\log[A_{\text{initial}}]$ should give a straight line with slope α. The analogous equation can be used to determine the remaining reaction orders.

$$\text{Log}(\text{Rate}_{\text{initial}}) = \alpha\log[A_{\text{initial}}] + \log(k[B]^{\beta}[C]^{\chi}) \quad [22]$$

Once the orders of all the various reactants have been determined, the rate constant can be calculated simply by rearranging any of the expressions of the form of eqn [17] into eqn [23].

$$k = \frac{\text{Rate}_{\text{initial}}}{[A]^{\alpha}[B]^{\beta}[C]^{\chi}} \quad [23]$$

The rate constant should be the same, or very similar. The rate constants should also be equal (within experimental error) or the reaction is incorrectly modeled and the calculated orders are not valid.

The initial rate approach was used to determine the orders for the reactants involved in the reduction of birnessite ($MnO_{1.7}$) to Mn(II) by catechol ($C_6H_6(OH)_2$). According to the overall reaction equation (eqn [24]), the general rate expression should contain three factors, birnessite, catechol, and H^+, according to eqn [25].

$$MnO_{1.7(s)} + 0.7C_6H_6(OH)_{2(aq)} + 2H^+ \rightarrow Mn(II)_{(aq)} + 0.7C_6H_4O_{2(aq)} + 1.7H_2O \quad [24]$$

$$\frac{d[Mn(II)]}{dt} = k[MnO_{1.7}]^{\alpha}[C_6H_6(OH)_2]^{\beta}[H^+]^{\chi} \quad [25]$$

Each reactant was varied independently and the initial rates were measured. Plots of initial rates versus $\log[C_6H_6(OH)_2]$ or $\log[MnO_{1.7}]$ were linear, with slopes equal to one indicating that the reaction was first-order in both catechol and birnessite. On the other hand, plots of initial rates versus $\log[H^+]$ were flat, i.e., slope of zero, indicating a zero-order dependence on H^+ concentration. This information allowed the rate law for this reaction to be simplified to eqn [26].

$$\frac{d[Mn(II)]}{dt} = k[MnO_{1.7}][C_6H_6(OH)_2] \quad [26]$$

Pseudo *n*th-Order Reactions

Some reactions are so rapid as to preclude the collection of initial rate data. In these situations, we can reduce higher-order reactions to 'pseudo' lower-order reactions by adjusting the concentrations of the various reactants; this is often referred to as the method of isolation. Returning to eqn [17], if we set the concentration of B and C such that they are much greater than the concentration of A, we can assume, after taking into account appropriate stoichiometries, that A can be entirely consumed without significantly altering the concentration of B or C. Thus, eqn [17] reduces to eqn [27], where k' is the 'pseudo α order rate constant.' (k' is also frequently denoted k_{obs} because it is the observed (measured) rate constant for the specific experimental conditions employed.)

$$-\frac{1}{a}\frac{d[A]}{dt} = k'[A]^{\alpha} \quad [27]$$

Using this approach, we can reduce a rate expression of any order into one with only one or two reactants of concern and determine if it conforms to one of the simple rate expressions in **Table 1**, in other words that α is $\cong 0$, 1, or 2. This was the case for the reduction of carbamate pesticides by Fe(II). The initially postulated rate expression for this process was eqn [28], but by monitoring the reduction of the carbamate pesticide in the presence of large excesses of Fe(II), eqn [28] simplifies to eqn [29].

$$-\frac{d[\text{pesticide}]}{dt} = k[Fe(II)][\text{pesticide}] \quad [28]$$

$$-\frac{d[[\text{pesticide}]}{dt} = k_{obs}[\text{pesticide}] \quad [29]$$

Plots of ln[pesticide] versus time were found to be linear, indicating that eqns [29] and [28] are indeed first-order with respect to pesticide concentration.

Second-Order, Two-Reactant Rate Laws

If we suspect that we a have a rate law in the form of eqn [14], we can model the data and determine the rate constant, k, in one of two ways. If it is relatively simple to measure both A and B during the course of reaction, it is possible to fit the data to eqn [15] using k as a fitting parameter. Alternatively, if it is difficult to measure A and B simultaneously, but we know the stoichiometric relationship between A and B, we can use that relationship to simplify eqn [15]. We can substitute an expression of the form $[B]_t = [B]_0 - ab^{-1}[A]_t$ into eqn [15], where a and b are the stoichiometric coefficients for A and B analogous to those in eqn [1]. Or, more elegantly, we can design our reaction such that A and B are initially present in their stoichiometric ratios according to eqn [30].

$$\frac{[A_0]}{[B_0]} = \frac{a}{b} \quad [30]$$

When eqn [30] is valid, then the integrated form of eqn [14] is simply eqn [16]. This expression is similar to eqn [13] except for the stoichimetric factor b.

This approach was employed to determine the rate constant for the reduction of Cr(VI) by Fe(II). The rate expression for this reaction was postulated to be first-order in each reactant and second-order overall (eqn [31]). Three Fe(II) ions are necessary to provide the three electrons needed to reduce one Cr(VI) to Cr(III). This leads to an Fe(II) to Cr(VI) stoichiometry of 1:3, respectively. By controlling the initial reaction conditions such that $[Fe(II)_0] = 3[Cr(VI)_0]$, the researchers were able to use eqn [32] to fit their data and obtain a value for k.

$$-\frac{d[Cr(VI)]}{dt} = k[Fe(II)][Cr(VI)] \quad [31]$$

$$\frac{1}{[Cr(VI)]_t} = \frac{1}{[Cr(VI)]_0} + 3kt \quad [32]$$

Reaction Diagrams and Activation Energies

The progress of a reaction can be represented as in **Figure 1**, in which the y-axis represents energy and the x-axis is what is known as the reaction coordinate, and is simply a way of showing the progress of the reaction from reactants to products. **Figure 1** represents what is known as a reaction diagram; this can be a very effective and concise way of summarizing the energetics of a reaction. For a reaction to be thermodynamically favored, the total energy of the products must be less than the total energy of the reactants, or, in other words, the products must be more stable than the reactants. The energy barrier shown between the reactants and the products is known as the activation energy (E_a). The magnitude of this barrier (as represented by its height in **Figure 1**) determines the rate of the reaction. Reactions with smaller activation

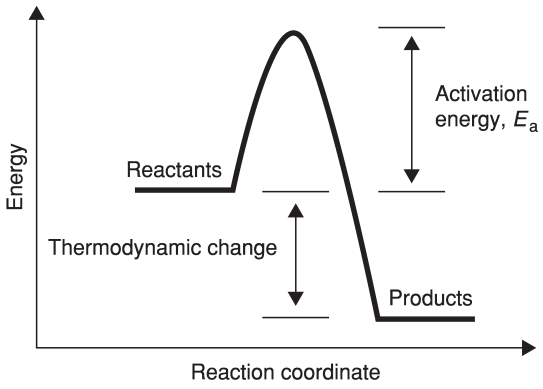


Figure 1 Reaction diagram for generalized thermodynamically favored but activation energy-limited reaction.

energies will proceed rapidly from reactants to products as compared with those with larger activation energies. This is the reason why reactions can be thermodynamically favored but kinetically unimportant (owing to huge activation energies).

Reactants gain the energy necessary to overcome the activation energy by colliding with each other. For reactions with very small activation energies, reactants will form products virtually every time there is a collision. These reactions are known as diffusion- or transport-controlled reactions because the rate of reaction is only limited by how quickly the reactants can approach one another through diffusion. At normal temperatures, homogeneous aqueous diffusion-controlled reactions are too rapid to be of much interest to soil chemists. On the other hand, for heterogeneous reactions, in which reactants must diffuse toward a surface or into and out of pore spaces, rates of diffusion are much slower. Under these conditions it is possible to have a relatively slow, but diffusion-controlled reaction. Diffusion-controlled reactions generally obey rate laws of a different form than the ones discussed here.

For reactions with larger activation energies, only a small percentage of collisions will have enough energy to push the reactants over the activation energy barrier. For these reactions, it is the activation energy that determines the magnitude of the rate constant, k. The mathematical relationship between the activation energy and the rate constant was developed by Savante Arrhenius and is known as the Arrhenius equation (eqn [33]), where E_a is the molar activation energy in $\mathrm{J\,mol^{-1}}$, R is the universal gas constant ($8.3145\,\mathrm{J\,K^{-1}\,mol^{-1}}$), T is temperature in K, and A is a proportionality constant known as the preexponential factor.

$$k = Ae^{\frac{-E_a}{RT}} \qquad [33]$$

Many people are qualitatively familiar with the observation that most reactions tend to proceed more rapidly at higher temperatures, but it is the Arrhenius equation that allows this temperature dependence to be quantified.

The activation energy of a reaction can be determined by measuring k at several different temperatures and using the Arrhenius equation. A plot of $\ln(k)$ versus T^{-1} should yield a straight line with slope, $-E_a R^{-1}$ according to eqn [34], which is obtained simply by taking the natural log of both sides of eqn [33] and rearranging.

$$\mathrm{Ln}(k) = \mathrm{Ln}(A) - \frac{E_a}{\mathrm{RT}} \qquad [34]$$

Similarly, two temperature points can be used to calculate E_a according to eqn [35].

$$E_a = \frac{\mathrm{Ln}(k_1) - \mathrm{Ln}(k_2)}{\left(\frac{1}{T_2} - \frac{1}{T_1}\right)} \qquad [35]$$

Because both of these approaches are simple rearrangements of the Arrhenius equation, they are theoretically equivalent. Once the activation energy has been determined we can model and predict the temperature dependence of a reaction. The magnitude of E_a can help determine whether the reaction is diffusion-controlled or activation energy-controlled. Diffusion-controlled reactions tend to have an E_a less than $\sim 40\,\mathrm{kJ\,mol^{-1}}$.

Reaction Mechanism (Rate-Limiting Step and Preequilibria)

Understanding reaction kinetics has additional benefits beyond simply being able to model the time-dependent progress of a given reaction. The kinetics of a reaction can often be used to make conclusions regarding the mechanism of that reaction, although this must be done with caution. The mechanism of a reaction is the series of elementary reactions that combine to give the complete reaction. With complex soil chemical reactions, it is rarely, if ever, possible to break a reaction mechanism into all of its elementary reactions. This being said, all metal redox reaction must progress through three stages: (1) formation of a prereaction complex; (2) electron transfer; and (3) separation of postreaction complex. In organic redox reactions, the second step often involves the transfer of an entire atom or group of atoms instead of simply an electron, but the fundamental process is similar.

The individual steps that make up a reaction mechanism usually have different activation energies and

therefore different rates. It is unlikely that for a given redox reaction each of the three stages outlined above will have the same rate constant. In fact, it is often the case that the activation energy of one of the steps will be much greater than that of the others and therefore will be much slower. The slowest step is known as the rate-limiting or rate-determining step (it is the 'bottle-neck' in the reaction progression). If the rate-determining step is the first step in a reaction mechanism, none of the following reactions or intermediate species will appear in the rate law. This effect can be seen in the reduction of carbon tetrachloride (CCl_4) by surface-absorbed Fe(II); the overall reaction is as seen in eqn [36].

$$CCl_4 + 2Fe(II) + H^+ \rightarrow CHCl_3 + Cl^- + 2Fe(III) \quad [36]$$

$$\frac{d[CHCl_3]}{dt} = k[CCl_4][Fe(II)] \quad [37]$$

The experimentally determined rate expression for this reaction was found to be first-order in carbon tetrachloride, second-order in Fe(II), and zero-order in H^+ (eqn [37]). The absence of H^+ in the rate law can be explained by postulating the following mechanism, in which eqn [38] is assumed to be the rate-limiting step.

$$CCl_{4(sorbed)} + 2Fe(II)_{(sorbed)} \rightarrow CCl^-_{3:(aq)} + Cl^- + 2Fe(III)_{(sorbed)} \quad [38]$$

$$CCl^-_{3:(aq)} + H^+_{(aq)} \rightarrow CHCl_{3(aq)} \quad [39]$$

However, if there are fast reactions before the rate-determining step, then these reactions will appear in the rate law. Because these early reactions are so much faster – both forward and reverse reactions – than the rate-limiting step, we can assume that they always remain at equilibrium throughout the course of the reaction. Acid–base reactions and metal speciation reactions are generally rapid and are frequently treated as preequilibria. Similarly, surface absorption can sometimes be treated as a preequilibrium, as was done in the case of the oxidation of fulvic acid by an Mn-oxide. The proposed reaction mechanism involved rapid absorption of the fulvic acid on to the Mn-oxide followed by a slow electron transfer step (eqns [40] and [41]), where Mn≡ denotes a Mn-oxide surface site.

$$Fulvic_{(aq)} + Mn{\equiv}OH \rightleftarrows Mn{\equiv}Fulvic (rapid) \quad [40]$$

$$Mn{\equiv}Fulvic \rightarrow Mn(II) + oxidized\ Fulvic\ (slow) \quad [41]$$

If eqn [41] is the rate-limiting step then the rate equation for this process should be like that seen in eqns [42], and [40] can be treated as an equilibrium expression (eqn [43]).

$$\frac{d[Mn(II)]}{dt} = k_{41}[Mn{\equiv}Fulvic] \quad [42]$$

$$K_{40} = \frac{[Mn{\equiv}Fulvic]}{[Fulvic][Mn{\equiv}OH]} \quad [43]$$

Combining eqns [42] and [43] and the equality that the total oxide surface sites must be equal to the free sites plus the fulvic occupied sites, $[Mn{\equiv}_{tot}] = [Mn{\equiv}OH] + [Mn{\equiv}fulvic]$, it was possible to obtain a rate expression that matched the data (eqn [44]).

$$\frac{d[Mn(II)]}{dt} = k_{41}K_{40}[Mn{\equiv}_{tot}]\frac{[Fulvic]}{1 + K_{abc}[Fulvic]} \quad [44]$$

Although this equation may initially seem more complicated, it is practically speaking much simpler because it is based on total reactant concentrations that are much easier to measure.

Biologically Controlled Redox Reactions

Many important redox transformations in soils are controlled by enzymes produced by microorganisms. Enzymatically controlled reactions are frequently modeled using either Michaelis–Menten kinetics or a derivative of Michaelis–Menten kinetics known as Monod kinetics. The proposed enzyme-controlled reaction mechanism used to develop Michaelis–Menten kinetics is shown in eqns [45] and [46], where E is the enzyme, S is the metal or organic molecule of interest known as the substrate, ES is the enzyme substrate complex, and P is the product.

$$E + S \rightleftarrows ES \quad [45]$$

$$ES \rightarrow P \text{ and } E \quad [46]$$

To convert this qualitative mechanism into a quantitative rate law, we must make an assumption known as the steady-state assumption. This assumption is similar but not identical to the preequilibrium situation described above. In the steady-state assumption, we assume that certain species are present at relatively constant values throughout the course of the reaction, although not necessarily their equilibrium values. To develop a steady-state situation we must assume that the substrate binds rapidly and

reversibly to the enzyme – eqn [45] is rapid in both the forward and reverse directions. Next it is assumed that the transformation of the enzyme substrate complex into the product (eqn [46]) is a first-order process and is the rate-limiting step in the overall reaction (eqn [48]). These assumptions allow us to assume that ES is present in a steady state or, in other words, that the rate of ES formation is equal to the rate of ES destruction. Eqn [47] shows this mathematically where the equation for ES formation is shown on the left and is set equal to the sum of the equations for ES destruction.

$$k_{45}[E][S] = k_{-45}[ES] + k_{46}[ES] \quad [47]$$

$$\text{Rate} = \frac{d[P]}{dt} = k_{46}[ES] \quad [48]$$

If we mathematically combine these two assumptions by substituting in for ES, we derive the basic Michaelis–Menten equation (eqn [49]).

$$\text{Rate} = \frac{k_{222}[E]_{tot}[S]_{tot}}{K_m + [S]_{tot}} \quad [49]$$

K_m is called the Michaelis–Menten or the Michaelis constant and $[E]_{tot}$ and $[S]_{tot}$ are the total enzyme and substrate concentrations, respectively.

The development of Monod kinetics is similar to that for Michaelis–Menten, with the insertion of X as the total cell concentration and K_s as the half saturation constant.

$$\text{Rate} = \frac{k[X][S]_{tot}}{K_s + [S]_{tot}} \quad [50]$$

Because it is frequently more feasible to control/measure cell concentration than enzyme concentration, Monod kinetics often have a practical advantage over Michaelis–Menten kinetics.

Conclusion

It is important to remind ourselves of primary assumptions made throughout this chapter. First, we have almost exclusively dealt with reactions that are far from equilibrium and were therefore able to ignore reverse reactions. For systems closer to equilibrium these reactions must be included. Although this complicates the mathematics, the general approach for developing the rate laws remains the same. Second, because these rate laws are empirically derived relationships, they are only rigorously applicable to systems with reactant concentrations, temperatures, ionic strengths, and other variables within the ranges used experimentally to develop the rate laws. They can sometimes be applied to slightly different systems, but this must always be done with caution. Lastly, there are many more detailed and complex rate laws that have been developed for both abiotic and biotic redox reactions, but they are beyond the scope of this chapter. The interested reader is referred to the further reading as a place to start.

List of Technical Nomenclature

X	Concentration of X ($mol\,L^{-1}$)

See also: **Anaerobic Soils**; **Redox Potential**

Further Reading

Amonette JE, Workmann DJ, Kennedy DW, Fruchter JS, and Gorby YA (2000) Dechlorination of carbon tetrachloride by Fe(II) associated with goethite. *Environmental Science Technology* 34: 4606–4613.

Buerge IJ and Hug SJ (1999) Influence of mineral surfaces on chromium(VI) reduction by iron(II). *Environmental Science Technology* 33: 4285–4291.

Lasaga AC and Kirkpatrick RJ (eds) (1981) Kinetics of geochemical processes. In: *Reviews in Mineralogy*, vol. 8. Washington, DC: Mineralogical Society of America.

Matocha CJ, Sparks DL, Amonette JE, and Kukkadapu RK (2001) Kinetics and mechanism of birnessite reduction by catechol. *Soil Science Society of America Journal* 65: 58–66.

Mortimer RG (1993) *Physical Chemistry*. Redwood City, CA: Benjamin/Cummings.

Sparks DL (1989) *Kinetics of Soil Chemical Processes*. San Diego, CA: Academic Press.

Sparks DL (1995) *Environmental Soil Chemistry*. San Diego, CA: Academic Press.

Sposito G (1994) *Chemical Equilibria and Kinetics on Soils*. New York: Oxford University Press.

Strathmann TJ and Stone AT (2002) Reduction of the pesticide oxamyl and methomyl by Fe(II): effect of pH and inorganic ligands. *Environmental Science Technology* 36: 653–661.

Waite TE, Wrigley IC, and Szymczak R (1988) Photoassisted dissolution of a collodial manganese oxide in the presence of fulvic acid. *Environmental Science Technology* 22: 778–785.

REMEDIATION OF POLLUTED SOILS

E Lombi and R E Hamon, CSIRO Land and Water, Glen Osmond, SA, Australia

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Public concern about environmental pollution has increased substantially since the beginning of the twentieth century with growing awareness of the stresses being placed on the environment due to activities arising from both industrial activity and intensive agriculture, and as a consequence of high-profile incidents that have highlighted the risks of pollution for human health. Examples of such incidents include industrial pollution at Love Canal, USA, the chemical factory explosion at Bhopal, India, and the nuclear near-accident at Three Mile Island, USA. Pollutants tend to have longer residence times in soil than in water or air. The soil acts as a sink or a filter in which pollutants are accumulated rapidly but are depleted slowly. The decontamination of polluted soils is becoming an issue of societal and political interest, because of its importance for environmental protection and human health, and its economical relevance. New technologies based on physical, chemical, and biological processes to reclaim polluted soils have been developed during the last decades. This article briefly describes the available technologies for the remediation of organic and inorganic contaminants in soils. A synopsis of these technologies is depicted in Figure 1.

Containment Versus Clean-Up Strategies

In terms of managing soil pollution, there are two types of strategy. The first, which at the beginning of the twenty-first century is the preferred option for many regulatory agencies, has the remediation goal expressed as a maximum allowable total concentration of pollutant in the soil. To comply with this approach, the total concentration of pollutant in soil must be lowered and this can be achieved using the group of technologies termed 'clean-up' technologies (Figure 1). The second strategy does not require a reduction in the total concentration of a pollutant in soil, but aims to manage exposure to the pollutant, and hence environmental and health risks, through technologies that lower pollutant mobility and bioavailability. These are termed 'containment' technologies (Figure 1).

In Situ Versus *Ex Situ* Strategies

In situ remediation refers to a soil-treatment regime in which the soil remains confined to its original location throughout the remediation process. Some earth-moving activities may be required in order to ensure complete exposure of the soil to the remedial treatment. Protocols that minimize off-site transport of dusts, vapors, or contaminated liquors are critical. *In situ* strategies have advantages in that costs, and environmental and health risks associated with excavation and off-site transport of polluted soil are reduced. However *ex situ* strategies (transport of

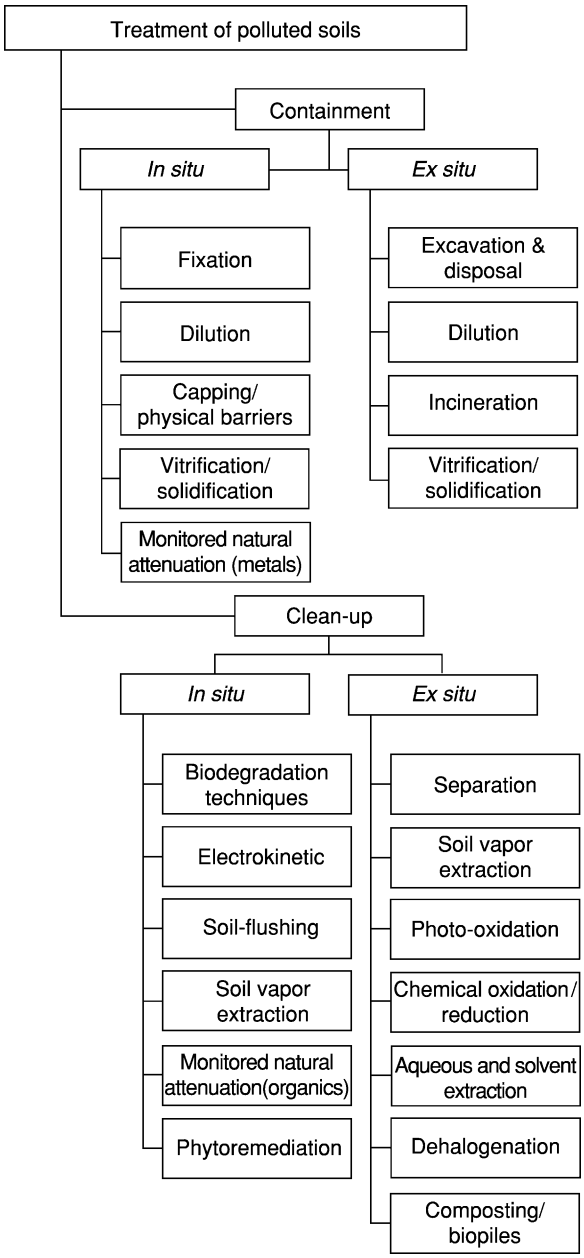


Figure 1 Decision tree of the alternatives available to treat polluted soils.

polluted soils to an external location) may be unavoidable because of impracticalities in establishing essential treatment infrastructure at the polluted site, or due to the site being positioned in a high-risk area.

Physical Treatment Technologies

Excavation and Off-site Disposal (*Ex Situ*, Requires Ongoing Monitoring)

Applicable to both inorganic and organic pollutants, excavation and off-site disposal entails removing the polluted soil from a high-value or vulnerable site to a disposal facility where potential pollutant exposure routes can be more effectively managed. Some pretreatment of the soil may be required to reduce the concentration of readily mobile pollutants in order to meet disposal-facility restrictions. Precautions associated with the excavation and transport of polluted soils need to be taken. This is a rapid method for managing polluted sites which has been widely employed in the past but potentially just relocates the problem and hence is not a remediation treatment *per se*. With landfill space at a premium, leading to rapidly rising disposal costs, this strategy is becoming less acceptable to industry, regulatory bodies, and the public. Ongoing monitoring of the disposal facility to ensure no off-site movement of contaminants is essential.

Dilution (*In Situ* or *Ex Situ*)

Applicable to both inorganic pollutants, including nutrients, and organic pollutants, dilution of soil pollution can be achieved through soil-mixing, deep tillage, or fracturing. It may be effective in reducing the potential contaminant exposure dose at sites in which pollutants are present in low levels or in which contamination is present in small, dispersed 'hot spots.' This method increases the volume of contaminated soil but may be appropriate for management of soils that have been polluted by nutrient oversupply (e.g., soils receiving dairy/piggery effluents). However, dilution is unlikely to meet with regulatory or community acceptance with respect to other types of pollutants.

Capping and Encapsulation (*In Situ*, Requires Ongoing Monitoring)

Applicable to both inorganic and organic pollutants, containment technologies aim to eliminate exposure of humans and the environment to the polluted soil. Lateral or vertical movement of the contaminant (due to water percolation, erosion, or plant root acquisition) is eliminated through the construction of highly reactive and/or impermeable subsurface barriers and surface caps. Both inorganic and organic pollutants may be managed using the monitored natural attenuation strategy. However, as with excavation and off-site disposal, the soil is not cleaned of pollutants, so this is not a permanent remediation solution. Barriers may fail as a result of weathering and/or chemical degradation, settlement, cracking, erosion, poor drainage away from the surface cap, or penetration by deep-rooted vegetation. There is therefore a need for ongoing monitoring of the site to ensure that the barriers and caps have not been breached. Introduction of barriers constructed of materials that either sorb the contaminant or chemically react with it can provide an extra degree of protection against inadvertent contaminant migration. Mobile contaminants, which might otherwise escape through ruptures in the barrier, may instead diffuse toward it and be chemically immobilized.

Separation–Fractionation (*Ex Situ*)

Applicable to both inorganic pollutants, including radionuclides, and organic pollutants, separation and fractionation is based on the principle that contaminants are frequently associated with specific particle-size fractions within soils and uses sieving or sedimentation to separate and remove the contaminated fraction from excavated soils. For example, the most reactive components of soils are generally found within the clay-size fraction ($<2\ \mu$m) due to the high relative surface area of this fraction. Contaminants entering the soil in soluble forms therefore typically partition to, and hence are concentrated in, the clay fraction. Separation of this fraction from the soil produces a highly polluted residue which requires further treatment (e.g., see Incineration, below) or disposal. However, the benefit is that the mass of material requiring further treatment is significantly reduced compared with the original soil. Contaminants may also be present in the form of discrete particles (e.g., lead (Pb) pellets from firing ranges), allowing for removal from the soil using physical sieving techniques. Immiscible organic pollutants may also be isolated from the soil using gravity separation methods.

Electrokinetic Separation (*In Situ*)

Applicable to both inorganic and polar organic pollutants, electrokinetic technology uses the generation of a low-intensity current between electrodes placed in the soil to mobilize and concentrate anionic and cationic pollutants toward the anode and cathode, respectively. The contaminants are then recovered through processes such as electroplating, precipitation, and coprecipitation at the electrode, concentration on ion-exchange resins or physical extraction of the aqueous phase. Soil types amenable to this treatment are those of low permeability (high clay content)

which are poorly drained. This is a new technology, but limited pilot trials show promising results as of 2003.

Soil-Flushing (*In Situ*)

Applicable to both inorganic pollutants, including radionuclides, and organic pollutants, in soil-flushing an aqueous solution (water or water containing contaminant-solubilizing additives) is circulated through the contaminated soil. The pollutant is desorbed from the soil and retrieved from the fluid. The operation usually takes the form of a continuous cycle, with the purified fluid being reinjected back into the contaminated soil, circulated, reextracted, and repurified. This technology performs poorly in soils with low permeability and with pollutants that partition strongly to the soil surface. For soils in which the pollutant has been present for an extended period of time, natural attenuation processes can also markedly reduce pollutant mobility, requiring many flushing cycles to be completed before pollutant concentrations are significantly reduced. These factors can make application of this technology impractical in terms of the timescales needed to complete remediation and associated costs.

Vitrification (*In Situ* or *Ex Situ*, Requires Ongoing Monitoring)

Applicable primarily to inorganic pollutants, especially radionuclides, in vitrification an electric current is used to melt the soil at temperatures up to 2000°C. The melted soil cools to form a chemically stable, leach-resistant, glass-like solid which encases and entraps inorganic pollutants. Organic pollutants are destroyed through pyrolysis at the high temperatures attained. However, this procedure destroys the soil, thereby limiting future site use. Adoption of this method is likely to be limited to inorganic contaminants under extreme cases where pollutant risks are very high, such as may be the case for radionuclide-polluted soils. Given its use for high-risk pollutants, it is important to institute an ongoing monitoring protocol for the solidified mass to ensure that contaminants are not lost over time through leaching or chemical weathering processes.

Solidification (*In Situ* or *Ex Situ*, Requires Ongoing Monitoring)

Applicable primarily to inorganic pollutants, including radionuclides, solidification involves introduction to the polluted soil of a leach- or weathering-resistant material through either mixing or injection which is then allowed to cure to form a solid product. The material acts to prevent pollutant exposure and migration by physically encapsulating the pollutants within a nontoxic, impermeable coating. A wide range of materials have been used in solidification technologies, including: cement, fly ash, bitumen, emulsified asphalt, and polyethylene. The resulting product is not soil but can often be used as a building or construction material. Ongoing monitoring of the final product is necessary to assess contaminant leakage.

Incineration (*Ex Situ*)

Applicable primarily to organic pollutants and pathogens, excavated soil is transferred to a contained facility and heated under controlled conditions in an incinerator. Organic pollutants are vaporized and collected, or destroyed through pyrolysis. Some volatile metals (e.g., mercury (Hg), cadmium (Cd), Pb) may also be released and captured from the exhaust gases. For soils or residues generated from separation–fractionation treatments containing very high concentrations of inorganic pollutants, it may be economically viable to roast or smelt the soil and recover the metal or nutrient for reuse.

Soil Vapor Extraction (*In situ* or *Ex situ*)

Applicable to organic pollutants, in soil vapor extraction, volatile organic contaminants (VOCs) are mobilized and removed from soils through application of a pressure–concentration gradient using a vacuum. Extraction wells are drilled into the soil (*in situ*) or the soil is excavated and placed over a network of perforated pipes (*ex situ*) and the organic pollutants contained within gases extracted following application of the vacuum are collected and destroyed by processes such as pyrolysis or photo-oxidation. Volatilization of the organic contaminants from the soil can be thermally enhanced by steam or hot-air injection, or electrical or electromagnetic heating techniques.

Photo-oxidation (*Ex Situ*)

Applicable to organic pollutants, vacuum-extracted organic pollutants (see Soil vapor extraction, above) are condensed, mixed with catalyst, and fed into a reactor in photo-oxidation. Ultraviolet light (sunlight or artificial) activates the catalyst, causing oxidation of the organic compounds into simple inorganic products, mainly carbon dioxide and water.

Chemical Treatment Technologies

Monitored Natural Attenuation (*In Situ*, Requires Ongoing Monitoring)

Both inorganic and organic pollutants can be managed by a monitoring natural attenuation strategy. The mobility and biological availability of pollutants can decrease with time of contact of pollutants with

soils due to chemical reactions and/or sorption which occur over the long term. The rates of these reactions depend upon the chemical characteristics of both the soil and the pollutant. The rates are difficult to predict, but are possible to measure. For low-risk polluted soils in which there is evidence that natural attenuation is occurring at a satisfactory rate, it may be appropriate to rank these soils as lower priority in terms of requirement for other remedial treatments. Even if contaminant solubility is not of immediate concern, action to reduce movement of polluted soil through erosion needs to be considered to prevent physical transport of contaminants off-site. Factors affecting the potential reversibility of the natural attenuation process are also unclear, so there is an ongoing requirement for monitoring of any site for which natural attenuation is enlisted as the primary 'treatment' strategy.

Fixation–Stabilization (*In Situ* or *Ex Situ*, Requires Ongoing Monitoring)

Applicable primarily to inorganic pollutants, including nutrients. The availability or toxicity of a pollutant is not necessarily related to its total concentration in the soil. This is because contaminants can exist in different chemical forms in soils, and it is primarily the aqueous and easily exchangeable forms which are both mobile and biologically available. Fixation–stabilization technologies aim to decrease the concentration of the aqueous and exchangeable forms of pollutants in soils through the application of nontoxic reactive agents. Agents that have been tested for this purpose include: lime, soluble phosphate compounds, fly ash, clays such as zeolite and also those rich in iron and aluminum oxides (e.g., red mud, steel shots, red gypsum, etc.), as well as mixtures of these materials. Many such agents are by-products from processing industries and so are inexpensive. The agents act either by encouraging the formation of insoluble precipitates of inorganic pollutants or by increasing the amount of pollutant sorbed to the solid phase, thereby decreasing the aqueous concentration of the pollutant. An advantage of this method is that, providing pollutant mobility and availability can be decreased to a satisfactory level, the soil can retain many components of its original functionality. Limitations to use of this technology include soils contaminated with mixed pollutants for which a chemical strategy to manage one of the pollutants is incompatible with the chemistry required to manage another. For example, increasing soil pH can significantly reduce the mobility of metals such as Pb, zinc (Zn), and nickel (Ni) but could result in an increase in the mobility of other co-contaminants such as arsenic (As) and cyanide. Moreover, as with the natural attenuation strategy, care is needed to prevent physical transport of polluted soil offsite, and an ongoing monitoring program may be necessary to assess the continued stability of the fixation treatment.

Chemical Oxidation–Reduction (*Ex Situ*, may Require Monitoring (Inorganics))

Applicable to some inorganic and organic pollutants, chemical oxidation–reduction uses the addition of oxidizing or reducing agents to convert pollutants to less toxic or inert compounds. For example, both Cr and As exist in more than one valence state. Cr(VI) is more toxic and mobile in soils than Cr(III), and As(III) is more toxic and mobile than As(V). Chemical reduction of Cr and oxidation of As can therefore decrease the pollutant toxicity. However, depending on conditions, there is always a risk of reconversion back to the more toxic valence state. Hence it is necessary to institute an ongoing monitoring program to detect this. Biological degradation processes are usually more cost-effective for organic pollutants than chemical oxidation methods and hence tend to be preferred.

Aqueous and Solvent Extraction (*Ex Situ*)

Applicable to inorganic and organic pollutants, in the process of aqueous and solvent extraction, excavated soil is placed in a pile and leached with an extraction fluid, or soil and extraction fluid are mixed in a reactor vessel. The extraction fluid dissolves the pollutants and the fluid is collected and then treated to remove the pollutants. The excavated soil is often pretreated using physical separation techniques to minimize the volume of material required to be extracted. For inorganic pollutants, an acidic aqueous solution which may contain surfactants is typically used as the extraction fluid. Soil pH is neutralized with base following removal of the pollutant. For organic pollutants, the extraction fluid is likely to be an organic solvent. Traces of the solvent may remain in the soil following extraction, so toxicity of the solvent is an important consideration. Low extraction efficiencies for pollutants that are highly attenuated (see Monitored Natural Attenuation, above) can severely limit the performance of this technology.

Dehalogenation (*Ex Situ*)

Applicable to halogenated organic pollutants, dehalogenation, the removal of halogen groups from organic pollutants, generally renders such compounds less toxic or nontoxic. Reagents that catalyze the removal of halogen groups are mixed with polluted soils, and the mixture is heat treated in a reactor

vessel to facilitate the reaction. Due to the cost of reagents, this technology is generally uneconomical for highly polluted soils or large volumes of polluted soil.

Biological Treatment Technologies

Biodegradation Techniques (*In Situ*)

Applicable to organic pollutants, biodegradation techniques are bioremediation clean-up techniques which aim to decrease the total concentration of pollutant in soil through the degradation of organic compounds by microorganisms. The microorganisms involved, also called degraders, are stimulated to use the pollutants as a source of nutrients and energy (enhanced biodegradation). Generally, this is achieved by optimizing the environmental conditions in terms of oxygen, nutrients, and moisture, and controlling the temperature and pH. For instance, the technology known as bioventing aims to stimulate pollutant biodegradation in unsaturated soils by increasing the soil oxygen content. This is achieved by forcing air through the soil by either extraction (vacuum) or injection methods. Specific strains of degraders can also be introduced to the soil to enhance the degradation process in a technology known as bioaugmentation. Typically for this application, microorganisms from the site under remediation are collected and those that preferentially feed on the contaminants at the site are isolated, cultured, and reintroduced to the site as a means of rapidly increasing the population of degraders. Remediation trials have also been conducted using nonindigenous strains of degraders introduced to a polluted site. However, so far there is no conclusive evidence to indicate that the use of non-native microorganisms has been beneficial in the situations tested.

One of the advantages of biodegradation technologies is that they can typically be implemented at a low cost. However the process may require more time than other available treatments. The degradation rate is controlled by both environmental conditions and by the nature of the pollutant. For example, biological treatment of polyaromatic hydrocarbons (PAHs) is selective, and the less-degradable PAHs (of high molecular weight) may remain unchanged. Furthermore, some compounds may be broken down into more toxic by-products during the bioremediation process (e.g., trichloroethylene (TCE) to vinyl chloride). Therefore, continuous monitoring and additional precautions have to be implemented in these cases. Although not all organic compounds are amenable to biodegradation, bioremediation techniques have been used successfully to treat soils polluted by petroleum hydrocarbons, solvents, pesticides, wood preservatives, and other organic chemicals.

Composting and Biopiles, and Slurry-Phase Biological Treatment (*Ex Situ*)

Applicable to organic pollutants, in this suite of technologies, the soil is excavated and mixed with appropriate amendments in order to stimulate microbial degradation of the pollutants. During composting, the microbial activity induces an increase in temperature that allows a faster degradation rate and can also be effective in reducing pathogen numbers. Piles of the material to be remediated can be aerated with blowers or vacuum pumps (aerated static-pile composting), or through periodic mixing with mobile equipment (windrow composting). Alternatively, the compost is placed in a reactor vessel where it is mixed and aerated (mechanically agitated in-vessel composting).

In the slurry-phase biological treatment, the excavated soil is mixed with water and various additives. The slurry is stirred to keep solids suspended and microorganisms in contact with the soil contaminants.

Limitation to this technology can be due to the large area needed for composting, the necessity to control release of VOCs, and the presence of toxic metals, which may impair microbial activity. However, this type of biological treatment has been used successfully to remediate soil contaminated by explosives and PAHs.

Phytoremediation (*In Situ*)

Applicable to both inorganic pollutants, including radionuclides, and organic pollutants, phytoremediation refers to a group of technologies that use green or higher terrestrial plants for treating chemically or radioactively polluted soils. There are five fundamental processes by which plants can be used to remediate soil contaminated with organic or inorganic pollutants. These processes result in the plant-based containment or removal of the soil pollutant. Phytoimmobilization processes prevent the movement and transport of dissolved contaminants by using plants to decrease the mobility of pollutants in soils. Phytostabilization as defined here is the use of pollutant-tolerant plants to mechanically stabilize polluted soils to prevent bulk erosion and airborne transport of pollutants to other environments. In addition, leaching of pollutants may be reduced due to higher evapotranspiration rates relative to bare soils. Phytovolatilization relies on the joint effect of plants and microbes which, acting in concert, are able to convert some metals and organic pollutants to volatile or more available forms. Examples include removal of Se or volatile organic pollutants from soils. In a

different approach, plant biochemistry has been modified by the insertion of bacterial genes that make the transformed plants able to volatilize Hg. Phytoextraction can be used to remediate both metallic and organic constituents from soil by direct uptake into plants and translocation to aboveground biomass. Two phytoextraction approaches are being developed, both aiming to extract toxic metals from soil. The first makes use of hyperaccumulator plants with exceptional metal-accumulating capacity (natural phytoextraction). The second uses high-biomass crops that are induced to take up large amounts of metals by enhancing the mobility of metals in soil with chemical treatments (chemically assisted phytoextraction). This latter method has recently been criticized because enhancing mobilization of toxic metals, to increase plant uptake, is also likely to increase the risk of leaching to groundwater. Phytodegradation refers to enhanced degradation of organic pollutants in soil induced by the action of rhizosphere microorganisms and by plant-derived enzymes. Furthermore, some organic compounds, depending on their chemical and physical characteristics, can be taken up by plants and partially or totally degraded by plant enzymes or endophytic microorganisms. Phytodegradation is also defined as plant-assisted biodegradation, because plant roots supply nutrients to microorganisms, which enhances their biological activities.

This group of technologies have a series of limitations as follows: the depth of the treatment zone is limited by the depth of the root system, high concentrations of the pollutants may be toxic to plants, the process may be seasonal, it is not effective for strongly sorbed contaminants, and may require long times to complete. On the other hand, these technologies are low-cost, environmentally friendly, and thought to be well accepted by the public.

Selection of Remediation Technologies and Future Directions

Once a need for soil remediation is recognized, the best available technology is selected according to the nature, toxicity, and origin of the contaminant, the present and potential hazard related to the degree of contamination, the chemical and physical characteristics of the soil, the land use, the time available for remediation, the community acceptance, and a cost-benefit analysis. These different criteria vary widely in the technologies described above. For instance, the average cost of remediation with different technologies can differ by a factor of greater than 25 (Figure 2). The development of low-cost, low-input technologies, driven in industrialized countries by economic benefits, is also essential to tackle remediation of polluted soils in developing countries.

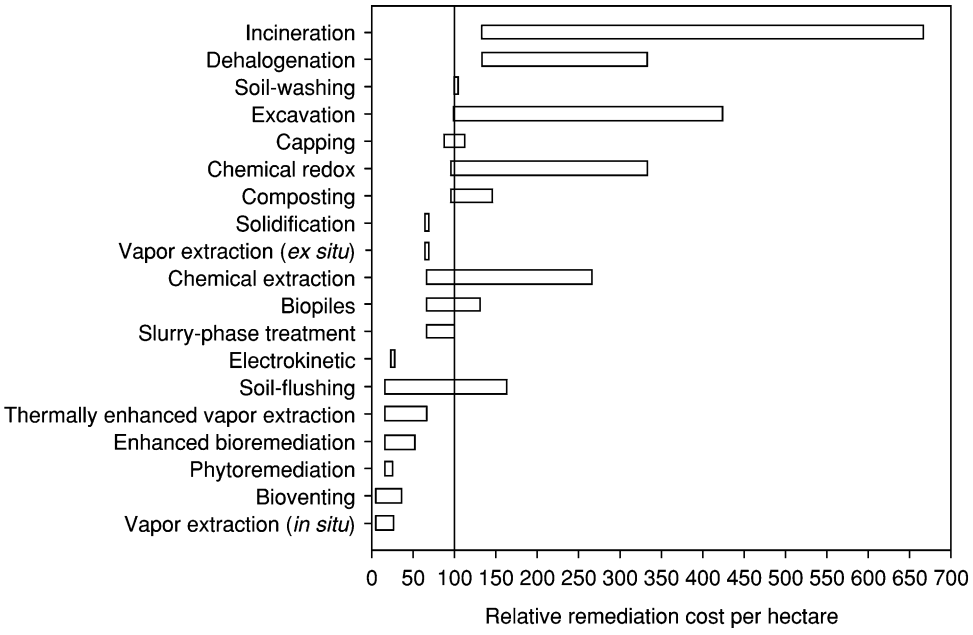


Figure 2 Range of relative remediation costs for different technologies, expressed as a percentage of the cost for capping. Costs are calculated on the basis of the 1 ha area, 0.25 m soil depth, and soil density of 1.2 g cm^{-3}.

Some of the technologies described in this article are well established, whereas others are still being developed. Due to ever-increasing public concern for the risks posed by soil pollution, it is expected that a larger number of technologies will be available in the future. For example, rapid progress in the field of molecular biology will allow a better understanding of the mechanisms involved in bio- and phytoremediation. This knowledge is of fundamental importance in order to optimize these processes. A new field of investigation may also be opened by recent concerns related to the possibility of pathogen-contaminated soils.

See also: **Hydrocarbons**; **Pollutants:** Biodegradation; Effects on Microorganisms; **Salt-Affected Soils, Reclamation**

Further Reading

Adriano DC, Bollag JM, Frankenberger WT, and Sims RC (eds) (1999) *Bioremediation of Contaminated Soils*. SSSA Monographs. Madison, WI: Soil Science Society of America.

Smith LA, Means JL, Means A *et al.* (eds) (1995) *Remedial Options for Metals-Contaminated Sites*. Battelle Memorial Institute. Boca Raton, FL: Lewis Publishers.

Huang PM and Iskandar IK (eds) (1999) *Soils and Groundwater Pollution and Remediation: Asia, Africa and Oceania*. Boca Raton, FL: Lewis Publishers.

Riser-Roberts E (1998) *Remediation of Petroleum Contaminated Soils: Biological, Physical, and Chemical Processes*. Boca Raton, FL: Lewis Publishers.

Wise DL, Tarantolo DJ, Inyang HI, and Cichon EJ (eds) (2000) *Remedial of Hazardous Waste Contaminated Soils*. New York: Marcel Dekker, Inc.

REMOTE SENSING

Contents
Organic Matter
Soil Moisture

Organic Matter

D K Morris, Stennis Space Center, MS, USA
C J Johannsen, S M Brouder, and G C Steinhardt, Purdue University, West Lafayette, IN, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Remote sensing is the collection and interpretation of data obtained from an object at a remote vantage point without making physical contact with the object or subject. Electromagnetic energy, as shown in Figure 1, may be recorded by a variety of remote-sensing methods. The most common include color and color infrared photography, aerial and satellite imaging, and radar sensing.

Aerial photography records energy in the ultraviolet part of the spectrum (which is not used for agricultural analysis because it is absorbed by the atmosphere), the part visible to humans, and a portion of the near-infrared (NIR) spectrum. Optical sensors record energy in a digital format from the visible, reflective infrared, and thermal infrared spectra from both aerial and satellite platforms within the Earth's atmosphere. Microwave wavelengths are another form of energy important to remote sensing. Microwave energy is received on Earth and the responses recorded by a digital instrument. This is known as 'passive energy,' compared with energy from portions of the spectrum that are measured directly and thus called 'active energy.'

For more than three decades, remote sensing has been envisioned as a useful source of data for inventorying and monitoring the Earth's surface for many purposes. One can apply the use of remote sensing to disciplines in agriculture, hydrology, forestry, oceanography, weather, land use, environmental concerns, security purposes, and military operations.

In the 1950s it was demonstrated that infrared aerial photography could be used to detect vegetation stress. This led to a host of applications and, by the 1960s, the US Weather Bureau was launching satellites to observe cloud patterns around the globe. Weather satellites today are used to monitor not only clouds, but also changes in precipitation patterns by radar, lightning strikes through brightness sensors, and vegetation density changes. They also contribute to predictions of where and when hurricanes are expected to strike land areas.

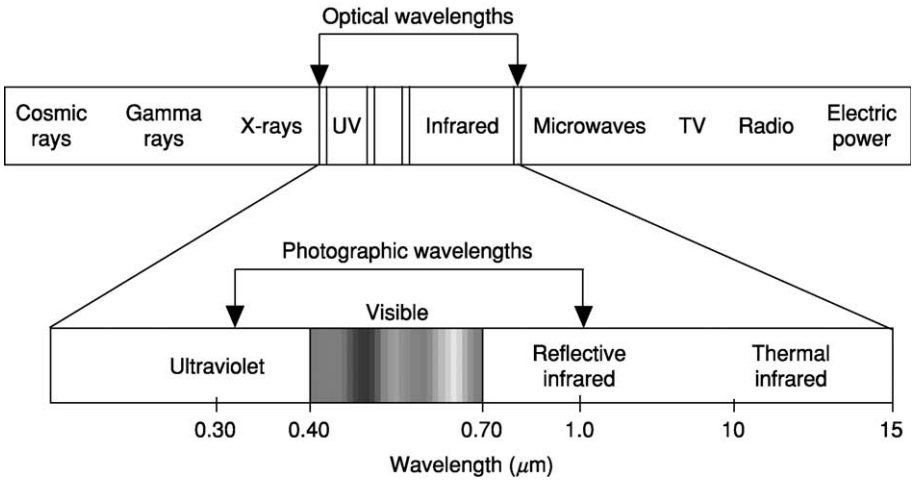


Figure 1 The electromagnetic spectrum showing the wavelengths of energy that are important to remote sensing applications. These include energy measurements from the visible, reflective infrared, thermal infrared, and microwave regions. (Reproduced with permission from Landgrebe DA (2003) *Signal Theory Methods in Multispectral Remote Sensing*. New Jersey: John Wiley.)

However, until recently, satellite data have not been available at a satisfactory temporal frequency for agricultural purposes; that is, they have not had the ability to provide repetitive coverage on a daily basis or 5- to 10-day intervals. The necessary spatial resolution, which is the ability to identify the detail of the measurement at a meter or less, has been provided by commercial satellites only since 1999. Critical needs for remote sensing in agriculture are to provide spatial resolutions of 5–20 m, frequent coverage of 2–5 days, and rapid data delivery to the user.

Aircraft platforms help in rectifying some of the problems associated with satellite acquisition in that spatial resolution can be acquired in the 1- to 5-m range because of smaller sensor viewing areas. Further, aircraft or aerial data collection is not limited to a fixed schedule and/or orbit as is satellite data acquisition. Satellite data are obtained from fixed orbits, provide no data during cloud cover, and have higher cost considerations, all of which users need to consider when acquiring data within a narrow, specific timeframe.

Current trends in the use of remote sensing data in agriculture are driven by goals similar to those of site-specific crop management. The delineation of suitable management units is an important aspect of remote sensing relating to site-specific crop management practices, and remote sensing has great potential to provide soil and crop information for this purpose. One of the most useful aspects of remote sensing is its ability to generate images that show the spatial variation in fields caused by natural and cultural factors without having to depend on costly sampling intervals or geostatistical interpolations. A major challenge is the development of robust image-analysis techniques for delineating soil conditions; another is to provide the proper training to data providers and farmers/producers in the use of this new information.

Early remote sensing research used primarily two approaches: the first was the traditional photointerpretation or visual approach, in which the actual image or photograph was analyzed subjectively by human interpretation. Although this method gave users an overall impression of within-field variation, it also had two crucial shortcomings: first, visual interpretation of black-white or color paper prints depends largely on the interpreter's personal experience and his or her knowledge of field conditions. This leads to the possibility that the results will vary among analysts and over time. Second, the speed of visual interpretation may pose a problem because of the time required to do an exhaustive analysis of the image.

The second approach in the early days of remote sensing, which is more standardized and less labor-intensive than the visual approach, used computer-implemented pattern recognition techniques applied to an image to classify surface soil patterns into different groups or classes. Advances in computer and information technology have overcome concerns related to the large data volume that needs to be analyzed for a successful application.

Mapping of Soils

There are many different soil classification systems used around the world and each is a product of the needs of people, the amount of data available, training, and the technology that was used. Remote sensing techniques are only beginning to show a

positive influence on the way soils are mapped on the Earth's surface. A unified global effort in soil survey and mapping was the FAO/UNESCO Soil Map of the World produced in the late 1960s, with a revised legend in 1988. This effort produced a map at a scale of 1:5 000 000, but did not make use of spatial technologies.

The US Department of Agriculture (USDA) National Resources Conservation Service (NRCS), a partnership of the National Cooperative Soil Survey (NCSS) with local and state agencies and land grant universities within the USA, are generating soil maps and interpretations using Soil Taxonomy. For modern farming practices, these soil surveys typically do not have sufficient resolution to capture natural soil variability in enough detail to support 'precision farming' recommendations, which use technologies such as remote sensing, geographic information systems, and global positioning systems. With the increasing availability of both very high spatial and spectral resolution remote sensing data, the opportunity to redefine soil survey maps is an important benefit.

Another pressing problem related to the use of soil maps, yield-monitoring, or soil samples is that the data are collected at different spatial scales and therefore need to be adjusted in order to make relevant management decisions. Producers have long known that soil variability influences the productive potential of their fields. Because of labor and analysis costs, some highly commercialized producers use a system of soil sampling in small blocks (i.e., grids). These blocks are commonly 0.4 ha (1 acre) to 1 ha (2.5 acres) in size and can be either square or rectangular in shape. This type of sampling scheme requires that soil samples acquired at coarse and/or variant resolutions be interpolated to obtain average soil characterization values at a given finer resolution.

Since natural soil boundaries do not follow straight lines found within the grid system, a producer makes decisions as to which soil is most represented within the grid (Figure 2). Generally, geostatistical analysis is used for this interpolation, based on kriging or inverse-distance weighting. The drawbacks of geostatistical analysis include the need for a large number of samples at close sampling intervals and the assumption of stationary data variation, which is often not the case for soil properties. Remote sensing offers the potential to overcome these drawbacks.

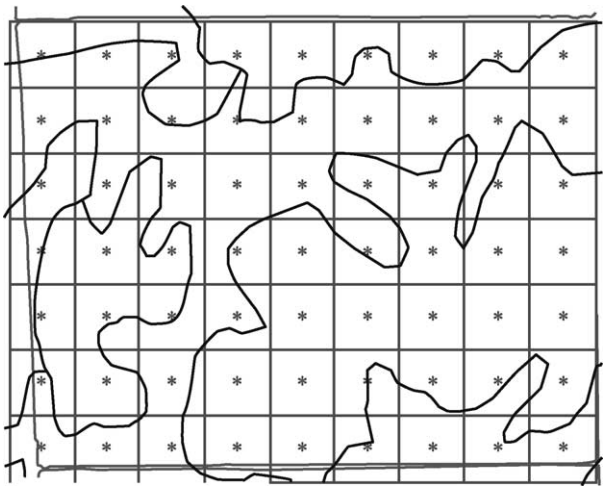


Figure 2 A gridded soil-sampling pattern overlaid on a soil map to illustrate the difficulty in representing the natural soil patterns with this approach.

Remote Sensing for Soil Mapping

Soil scientists have used photography in producing soil maps since the early 1930s. During the first 50 years, much of the interest in photography was related to having a reliable local-to-regional base map. During the 1960s, multispectral remote sensing data were considered of limited use because the evaluated results were compared with existing soil maps generated by conventional methods. The low accuracy attributed to remotely sensed data when compared with existing soil maps was the result of the assumptions that (1) the conventionally prepared soil map is 100% correct, and (2) that the spectral response of the soil surface is determined by the same morphological features used to derive the soil survey maps.

Soil survey mapping units are defined by the soil properties throughout the soil profile that affect management practices, such as drainage, and erosion control; whereas maps of spectral responses are based directly on the spectral properties of only the portion of the soil profile that is exposed at the surface. In laboratory studies, soil reflectance has been successfully correlated with soil properties such as organic carbon, moisture, and total nitrogen under controlled conditions.

With the more recent emphasis on intensive, precision farming, producers have been requesting more detailed field-level soil maps. Current attention is focused on the resolution to be mapped and the soil categories or soil management zones that a producer can use effectively in making management decisions.

We know that (1) soil properties such as organic matter and iron oxides are measured by reflectance or image response, (2) soil surface conditions such as surface roughness influence the response, and (3) remotely sensed imagery of different wavelengths enhances soil patterns so they can be more easily mapped. In summary, the responses of images can be related to soil properties such as organic matter

content, since dark soils (higher organic matter) are less reflective in contrast with light-colored soils (lower organic matter). The vegetation spectral response can also be used to infer various soil conditions such as water-holding capacity, eroded locations, and nutrient differences.

With advances in remote sensing technologies (e.g., digital imaging), such information can now be directly integrated into a geographic information system. The approach commonly taken in studying the feasibility of using imaging devices for soil mapping has been the acquisition and analysis of imagery or other remotely sensed data over a specific area that is then supplemented by limited ground reference data.

There must be an awareness of changes in remote sensing imagery due to surface reflectance, sensor sensitivity, and spectral composition of the sunlight on the area being sensed. For example, if images are obtained over an area every hour throughout the day, the soil surface will have a different soil reflectance during each acquisition caused by the sun's angle. If the surface soil is moist and it is a windy day, the surface soil reflectance will change owing to the drying action (Figure 3).

Automatic pattern-recognition techniques used on aerial sensing data show that different classes of soil surface conditions can be mapped with reasonable accuracy. Since the launch of the Landsat 1 satellite in 1972, much attention has been directed to the relationship of soils and land use. 'Agrophysical units' are defined as geographic areas having definable or comparable agronomic and physical parameters such as soils, which reflect a range in agricultural use and management.

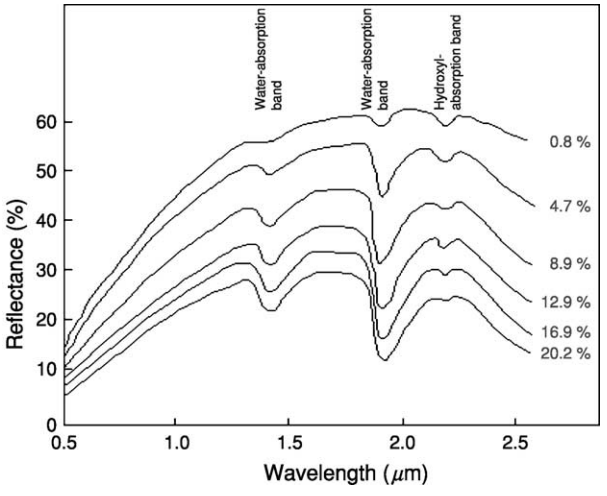


Figure 3 Laboratory spectra of a silt loam soil under different moisture conditions. Soil reflectance decreases in the visible and near infrared region with an increase in soil moisture. This silt loam soil was dried in the laboratory to demonstrate the reflectance change from 0.5 to 2.5 μm. (Adapted from Bowers SA and Hanks RJ (1965) Reflection of radiant energy from soil. *Soil Science* 100: 130–138.)

Early work using visual interpretation of Landsat images provided overlays with acetate transparencies of soil map sheets to investigate relationships of Landsat multispectral data to soil patterns. Although the results vary by geographic regions, there is agreement that the use of satellite or airborne imagery decreases the time required in the soil survey and mapping period and therefore reduces costs. Spectral classification alone cannot distinguish between widely differing soils exhibiting similar spectral responses. By adding ancillary data, a more detailed delineation of soils can be provided as compared to information derived solely from Landsat data. However, these comparisons also reveal some discrepancies between spectral soil maps produced from Landsat data and conventional soil survey maps. One important soil property that has received much focus is soil organic matter.

Organic Matter Detection

Quantitative relationships between soil color and organic matter content are not well quantified because of differences in soil moisture, iron oxides, and other soil properties. However, these relationships are of considerable practical importance in mapping and classifying soils, interpreting soil properties, and designing sensors that can detect soil properties.

Soils of higher organic matter content are generally less reflective or darker in color than those soils with lower organic matter content; therefore, soil color has long been associated with native soil fertility. Soils with thick, dark surface horizons are often separated from other soils in many soil classification systems, which emphasize the importance of these soils both as a medium for plant growth and also as an indicator of land value for agricultural and urban uses.

Accurate and timely information about soil organic matter is essential for agricultural production and environmental research. During the past three decades, high-altitude remote sensing coupled with laboratory-based reflectance spectroscopy has emerged as an important technology for monitoring the Earth's agricultural resources.

Although data from currently operational reflectance sensors such as Landsat, SPOT, and a host of other public and private sensors (Table 1) have proven useful for crop identification and wetland inventory, attempts to characterize soils have attained only limited success. The major difficulty in deriving soil information from remotely sensed data is the

Table 1 Typical spatial resolution values of selected satellite sensors

Satellite & sensor	*Spatial resolution (m)*
DigitalGlobe Quick Bird panchromatic	0.6
Space imaging IKONOS panchromatic	1
DigitalGlobe Quick Bird multispectral	2.4
Space imaging IKONOS multispectral	4
IRS-1C panchromatic	5
RADARSAT	8–100
SPOT panchromatic	10
Landsat 7 panchromatic	15
JERS-1 radar	18
SPOT multispectral	20
Seasat radar	25
Landsat thematic mapper	30, 60, 120
IRS-1B LISS-II	36
IRS-1A LISS-I	72.5
Landsat multispectral scanner	80
Advanced very high resolution radiometer (AVHRR)	1100

complex nature of relationships between soil parameters and reflectance. Organic matter, moisture, texture, cation exchange capacity (CEC), mineral oxides, and surface conditions such as residue cover and soil crusting have demonstrated unique influences on soil reflectance. These relationships are wavelength-dependent; therefore, the soil parameters may affect reflectance differently in different regions of the spectrum.

In glaciated soil regions, high organic soils can be readily differentiated in high-resolution satellite images from soils formed under prairie grass vegetation and soils formed under forest vegetation (Figure 4).

Several problems have been encountered when viewing the soil with an electro-optic sensor. These problems can be reduced if one has collected reference information on ground conditions. Bare soil areas with high-residue content from the previous crop may mask soil spectral responses (Figure 5). The use of high spatial digital terrain models

Figure 4 Soil patterns as seen on an Ikonos color infrared image (24 May 2000) of the Purdue University Agronomy Center for Research and Education (ACRE, W. Lafayette, Indiana, USA) and surrounding area. The dark soils (Mollisols) are higher in organic matter and derived under grassland vegetation, while the lighter soils (Alfisols) have low surface soil organic matter and are derived under forest vegetation. Red shows healthy vegetation; white, streets, roadways, and buildings; dark green to black, areas of high organic matter and high surface moisture of soils; light-green areas, low surface soil organic matter, some of which are muted by high surface residues from the previous crop. © Space Imaging 2000.

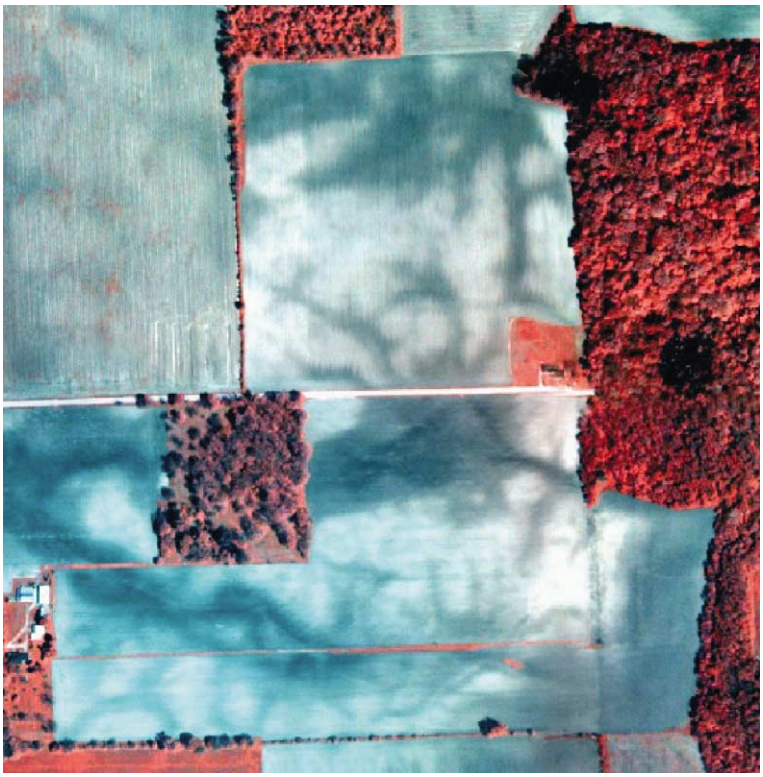


Figure 5 Aerial color infrared image showing soil patterns over an eastern Indiana farming area. The darker areas are higher in organic matter content and were previously wetland areas. The field in the upper left does not have the same soil patterns, since residues from the previous year cover it. Red shows healthy vegetation such as grasses and bushes; maroon, trees (the color resulting from the leaves, branch spacing, and shadows); white, roads and buildings; dark green to black, areas of high organic matter; light green, low surface soil organic matter content. Muted lined patterns masking the normal colors are corn residues from the previous crop. Courtesy of Emerge.

(DEM) produced by remote microwave or laser techniques can provide a better understanding of soil formation and surface moisture movement and aid in interpreting surface soil organic matter content.

Sensing soil organic matter from airborne or satellite platforms requires reference data. Soil types whose classifications are driven by soil color and all of the previously mentioned parameters may have as much variation in soil properties from within a given soil type as they do between different soil types. Many different methods have been developed for the calibration of soil organic matter sensors. Most involve a given sequence of soils of about the same age, related parent material, and similar climatic conditions, but characterized by variations in topography and drainage. Laboratory studies have measured soil reflectance, attempting to quantify the relationship between soil color and organic matter content by comparing the soil spectra with soil Munsell values.

Most studies have concluded that NIR predictions of organic matter and soil moisture become less accurate as one considers soil samples from a wider geographic range. Low correlations between reflectance and organic carbon content have been attributed to the confounding effects of texture, mineralogy, and soil moisture. It is expected that high-resolution (temporal, spatial, and spectral) reflectance data will lead to a better understanding of the relationships between these and other soil properties and reflectance.

Investigations have attempted to link reflectance to specific organic matter components. Typically these studies fractionate organic matter into pools such as humic acid, fulvic acid, and the remaining mineral components; then each fraction is analyzed using a spectrophotometer. It has been shown that even small amounts of organic material may be particularly effective in masking mineral features in the shorter wavelengths. It appears that unfractionated soils and their mineral components are important influences on reflectance with regard to examining organic carbon content in relation to other soil properties.

Table 2 Importance of spectral bands based on response of soil reflectance to organic carbon (OC) content, parent material, and other soil properties

Soils within similar parent materials (northern soils)			*Soils across different parent materials (all soils)*
Responds to OC content differences only (nm)	*Responds to other properties, OC content effect is dominant (nm)*	*Responds to other properties that mask OC content effect (nm)*	*Responds to OC content and minimizes parent material effect (nm)*
1065	405–1055	1775–1885	1955–1965
1085–1105	1075	1975–2205	2215
1125–1165	1115	2225–2255	2265
1955–1965	2325	2275	2285–2295
2215	2375		2315–2495
2265	2425		
2285–2295	2445–2485		
2315			
2335–2365			
2385–2415			
2435			
2495			

Source: Henderson TL, Baumgardner MF, Franzmeier DP, Stott DE, and Coster DC (1992) High dimensional reflectance analysis of soil organic matter. *Soil Science Society of America Journal* 56: 865–872, with permission.

To examine specific effects of organic carbon on soil reflectance, researchers have tested differing spectral response in separable bands in terms of observable soil properties such as sand, silt, clay, iron oxides, magnesium oxides, and organic carbon. The separable bands of the inorganic fractions (in which the extractable organic matter and humic acid have been removed) are compared with the separable bands from the soils that are unfractionated. There is a direct relationship between organic carbon and soil reflectance in the set of bands in question: as soil organic carbon increases reflectance decreases. The organic matter contained within the soil is the physical factor responsible for spectral differences in this set of bands (Table 2).

In Table 2, the first column lists the bands that respond to organic carbon content only. In these bands, higher organic carbon content causes a decreased measure of reflectance. The first three of the 12 optimal bands are in the NIR portion of the spectrum (700–1300 nm), with the others in the middle-infrared (MIR) region (1300–2500 nm). The second column lists bands in which soil reflectance responds to both organic carbon and other soil properties. These bands include the entire visible range of the spectrum (400–700 nm) and some in the NIR and MIR region; they are useful because the organic carbon content masks the effect of the mineral soil properties. The wavelengths in column 3 should be avoided when classifying organic carbon content. Although the soil curves are separable in the bands listed, the spectral response in these bands is unrelated to organic carbon content, since no direct relationship between organic carbon content and the magnitude of the spectral response in those wavelengths exists.

The Future

The hyperspectral format contains up to several hundred wavelength bands and provides the potential for differentiating more specific soil properties. Note that current studies have used only three to four bands of data, which are very broad bands. The hyperspectral bands would be very narrow in terms of the wavelengths covered. The spectral mapping of specific soil properties will provide further assistance to soil pedologists in the development of soil surveys.

Recent advances in precision farming technologies require an accurate representation of yield-limiting variables, which include soil properties. Producers have used soil type maps, as they are related to soil organic matter, to help in management practices such as fertility application, herbicide placement, and seed spacing. When coupled with the adoption of global positioning satellites and geographic information system technologies, accurate soil organic matter maps allow for greater precision in performing management practices.

Advances in remote sensing technology are providing a new way to 'see' soil properties and agricultural fields. The provision of geographically accurate soils data assists producers to adjust their application rates of fertilizers and pesticides according to soil properties. This helps producers to increase profits

and to provide environmentally safer approaches to producing food.

See also: **Classification of Soils**; **Organic Matter:** Principles and Processes; **Remote Sensing:** Soil Moisture; **Statistics in Soil Science**

Further Reading

Baumgardner MF, Silva LF, Biehl LL, and Stoner ER (1985) Reflectance properties of soils. *Advances in Agronomy* 38: 1–44.

Ben-Dor EN, Goldlshleger N, Benyamini Y, Agassi M, and Blumberg D (2003) The spectral reflectance properties of soil structural crusts in the 1.2- to 2.5-μm spectral range. *Soil Science Society of America Journal* 67: 289–299.

Buol SW, Hole FD, McCracken RJ, and Southard RJ (1997) *Soil Genesis and Classification*. Ames, IA: Iowa State University Press.

Ding G, Novak JM, Amarasiriwardena D, Hunt P, and Xing B (2002) Soil organic matter characteristics as affected by tillage management. *Soil Science Society of America Journal* 66: 421–429.

Fernandez RN, Schulze DG, Coffin DL, and Van Scoyoc GE (1988) Color, organic matter, and pesticide adsorption relationships in a soil landscape. *Soil Science Society of America Journal* 52: 1023–1026.

Henderson TL, Baumgardner MF, Franzmeier DP, Stott DE, and Coster DC (1992) High dimensional reflectance analysis of soil organic matter. *Soil Science Society of America Journal* 56: 865–872.

Hoffer RM and Johannsen CJ (1969) Ecological potentials in spectral signature analysis. In: Johnson P (ed.) *Remote Sensing in Ecology*, pp. 1–16. Athens: University of Georgia Press.

Jensen JR (2000) *Remote Sensing of the Environment: An Earth Resource Perspective*, 2nd edn. Englewood Cliffs, NJ: Prentice-Hall.

Johannsen CJ, Baumgardner MF, Willis PR, and Carter PG (1998) *Advances in Remote Sensing Technologies and Their Potential on Agriculture*. 1st International Conference on Geospatial Information in Agriculture and Forestry, Paper 2–1, Proceedings vol. 1: 6–11.

Kristof SJ (1971) Preliminary multispectral studies of soils. *Journal of Soil and Water Conservation* 26(1): 15–18.

Lillesand TM, Kiefer RW, and Chipman JW (2004) *Remote Sensing and Image Interpretation*, 5th edn. New York: John Wiley.

Moran SM, Inoue Y, and Barnes EM (1997) Opportunities and limitations for image-based remote sensing in precision crop management. *Remote Sensing of Environment* 61(3): 319–346.

National Research Council (1997) *Precision Agriculture in the 21st Century: Geospatial and Information Technologies in Crop Management*. Washington, DC: National Academy of Sciences.

Shonk JL, Gaultney LD, Schultz DG, and VanScoyc GE (1991) Spectroscopic sensing of organic matter content. *Transactions of the American Society of Agricultural Engineers* 34(5): 1978–1984.

Stoner ER and Baumgardner MF (1981) Characteristic variations in reflectance of surface soils. *Soil Science Society of America Journal* 45: 1161–1165.

Soil Moisture

T J Jackson, USDA Agricultural Research Service, Beltsville, MD, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Soil moisture mapping using remote sensing is becoming a reality. Soil moisture has been difficult to measure and map using conventional ground-based technologies. Microwave remote sensing-based soil-moisture retrieval has been demonstrated using tower and aircraft instruments; however, the implementation of this approach on satellites has been limited by both technology and availability. Developments in both science and associated technologies now make the exploitation of the microwave region for soil moisture mapping feasible. A review of the basic principles, operational retrieval algorithms, aircraft applications, and current and future satellite missions is presented.

Microwave Remote Sensing of Surface Soil Moisture

Microwave sensors, especially those operating at lower microwave frequencies (<6 GHz), are of particular value in soil moisture measurement because they respond to the amount of water in the soil. When this technique is used in a remote sensing instrument, the surface layer of soil contributes to the microwave sensor measurement.

Microwave remote sensing can provide a direct measurement of the surface soil moisture for a range of vegetation-cover conditions within reasonable error bounds. Two basic approaches are used, passive and active. In passive methods, the natural thermal emission of the land surface (or brightness temperature) is measured at microwave wavelengths, using very sensitive detectors. In active methods or radar, a microwave pulse is sent and received. The power of the received signal is compared to that which was sent to determine the backscattering coefficient.

The microwave region of the electromagnetic spectrum consists of frequencies of 0.3–30 GHz. This region is subdivided into bands, which are often

referred to by a lettering system. Some of the relevant bands that are used in Earth remote sensing are: K (18–27 GHz), X (8–12 GHz), C (4–8 GHz), and L (1–2 GHz). Within these bands, only small ranges are protected for scientific applications, such as radioastronomy and passive sensing of the Earth's surface. Communications and other uses can contaminate the unprotected regions of this electromagnetic region. This contamination is called radiofrequency interference and it is becoming a greater problem each year.

A general advantage of microwave sensors (in contrast to visible and infrared) is that observations can be made through cloud cover, because the atmosphere is nearly transparent, particularly at frequencies <10 GHz. In addition, these measurements are not dependent on solar illumination and can be made at any time of the day or night. Instruments can be mounted on trucks, aircraft, and spacecraft for repetitive large-area observations.

Microwave sensors operating at very low microwave frequencies (<6 GHz) provide the best soil moisture information. At low frequencies, attenuation and scattering problems associated with the atmosphere and vegetation are less significant, the instruments respond to a deeper soil layer, and a higher sensitivity to soil water content is present.

As noted above, there are two microwave remote sensing approaches, passive and active. Each has unique capabilities and limitations that must be matched with specific types of applications. Currently, passive microwave remote sensing offers more robust and well-validated retrieval algorithms. However, it is limited by spatial resolution at satellite altitudes. Very large antenna sizes are required to achieve high spatial resolutions. Getting these into space at a reasonable cost has been a limitation for quite some time. Currently, products are limited to 50 km resolution; however, future systems will provide 10 km products. Passive systems can provide daily global coverage. Active microwave techniques can provide very high spatial resolution data from space (10–20 m) but suffer from weak soil moisture algorithms. The temporal repeat of these systems is also very limited, typically over 30 days.

Reflectivity and Soil Moisture

Active and passive sensors provide different types of measurements, which are determined by various physical phenomena. However, both types of sensor provide information on the surface reflectivity. The essential relationship needed to utilize microwave remote sensing is a link between reflectivity and soil moisture.

Assuming that the Earth is a plane surface with surface geometric variations and volume discontinuities much less than the wavelength, only refraction and absorption of the media need to be considered. This situation permits the use of the Fresnel reflection equations as a model of the system. These equations predict the surface reflectivity (r) as a function of dielectric constant (k) and the viewing angle (θ), based on the polarization of the sensor (horizontal-H or vertical-V).

$$r^{\mathrm{H}} = \left| \frac{\cos\theta - \sqrt{\mathrm{k} - \sin^2\theta}}{\cos\theta + \sqrt{\mathrm{k} - \sin^2\theta}} \right|^2 \qquad [1]$$

$$r^{\mathrm{V}} = \left| \frac{\mathrm{k}\cos\theta - \sqrt{\mathrm{k} - \sin^2\theta}}{\mathrm{k}\cos\theta + \sqrt{\mathrm{k} - \sin^2\theta}} \right|^2 \qquad [2]$$

The Fresnel equations apply when the two media at the interface each have uniform dielectric properties. From the reflectivity, the dielectric constant of the soil can be estimated. The dielectric constant of soil is a composite of the values of its components: air, soil, and water. Although the dielectric constant is a complex number, for soil mixtures the imaginary part is small and can be ignored for computational purposes without introducing significant error. Values of the real part of the dielectric constant for air and dry soil particles are 1 and 5, respectively. Therefore, a very dry soil has a dielectric constant of about 3. Water has a value of about 80 at low frequencies. The basic reason microwave remote sensing is capable of providing soil moisture information is this large difference between the dielectric constants of water and the other components. Because the dielectric constant is a volume property, the volumetric fraction of each component is involved in the computation of the mixture dielectric constant. The computation of the mixture dielectric constant (soil, air, and water) has been the subject of several studies, and different theories exist as to the exact form of the mixing equation.

The dielectric constant of water referred to above is that of free water, in which the molecules are able to rotate and align with an electrical field. However, not all the water in soil satisfies this condition. Some water in the soil has different properties. The initial water added to dry soil below a 'transition' water content is held more tightly by the soil particles and has the dielectric properties of frozen water (~3).

Based upon the discussion above, four components must be considered in computing the dielectric constant of soil: air, soil particles, bound water, and free

water. Thus, in order to interpret the data correctly, some knowledge of the soil texture is needed. All dielectric mixing models incorporate this information. Based on an estimate of the mixture dielectric constant derived from the Fresnel equations and information on the soil texture, volumetric soil moisture can be estimated using an inversion of the model.

Various theories describe the reflection resulting from a soil profile with uniform or varying properties. The computations involve a nonlinear weighting that decays with depth. Some modeling studies suggest that this dominant depth is one-tenth the wavelength. Field experiments suggest that the contributing depth is about one-quarter the wavelength. At a frequency of 6 GHz this depth would be about 1.5 cm and at 1.4 GHz about 5 cm.

Passive Microwave Techniques

Passive microwave remote sensing utilizes radiometers that measure the natural thermal radio emission within a narrow band centered on a particular frequency. The measurement provided is the brightness temperature in degrees Kelvin, T_B, which includes contributions from the atmosphere, reflected sky radiation, and the land surface. Atmospheric contributions are negligible at frequencies <10 GHz, and the following section assumes this to be the case. Galactic and cosmic radiation contribute to sky radiation and have a known value that varies very little in the frequency range used for observations of soil water content (T_{sky} ~4 K). The observed brightness temperature of a surface is equal to its emissivity (e_{obs}) multiplied by its physical temperature (T). The observed emissivity is equal to 1 minus the reflectivity, which provides the link to the Fresnel equations and soil moisture for passive microwave remote sensing.

Based upon the above and neglecting atmospheric contributions, for a specific frequency (f), polarization (p), and viewing angle (θ) the equation for T_B is:

$$T_{B_{fp\theta}} = e_{\mathrm{obs}_{fp\theta}} T + (1 - e_{\mathrm{obs}_{fp\theta}}) T_{sky} \qquad [3]$$

The second term of eqn [3] is about 2 K. A typical range of response for a soil is 60 K; therefore, the reflected sky contribution can be dropped for computational purposes. Eqn [3] can be rearranged as follows:

$$e_{\mathrm{obs}_{fp\theta}} = T_{B_{fp\theta}} / T \qquad [4]$$

If the physical temperature is estimated independently, the emissivity can be determined from T_B. The physical temperature can be estimated using surrogates based on satellite surface temperature, air-temperature observations, or forecast model predictions. As noted above, the derived emissivity is for a specific instrument configuration (frequency, polarization, and viewing angle).

For natural conditions, varying degrees of vegetation type and density are likely to be encountered. The presence of vegetation has a major impact on the microwave measurement. Vegetation reduces the sensitivity of the retrieval algorithm to changes in soil water content by attenuating the soil signal and by adding a microwave emission of its own to the microwave measurement. This attenuation increases with increasing microwave frequency, which is another important reason for using lower frequencies. Attenuation is characterized by the optical depth of the vegetation canopy. A widely used correction for these effects uses information on the vegetation type (parameter b) and the vegetation water content (W_{veg}) as follows:

$$e_{\mathrm{surf}_{fp\theta}} = 1 + [e_{\mathrm{obs}_{fp\theta}} - 1]\exp[-2b W_{veg} \cos\theta] \qquad [5]$$

The corrected emissivity represents that of the soil surface (e_{surf}). Methods have been developed for estimating the vegetation-type parameter (typically derived from land cover) and vegetation water content using visible/near-infrared remote sensing.

After performing the vegetation correction, one can obtain an estimate of the emissivity of the soil surface, which includes the effects of surface roughness. These effects must also be removed in order to determine the equivalent smooth-surface soil emissivity, which is required for the Fresnel equation inversion. Techniques are available for doing this that use statistical characterization of the surface heights and land-use information.

Figure 1 is a simulated set of curves showing the relationship between soil moisture and emissivity for bare soil and vegetated conditions. Vegetation decreases the sensitivity to soil moisture. Errors in the model, parameters, and the precision of measurements become more important and at some point will overwhelm the information in the observation.

Active Microwave Methods

Active microwave sensors or radars measure the sent and received power to yield a variable called the backscattering coefficient (σ^0). The backscattering coefficient is related to the surface reflectivity. As described previously, the reflectivity is then used to determine surface soil moisture. For active techniques, the link between the measurement of the backscattering coefficient and the surface

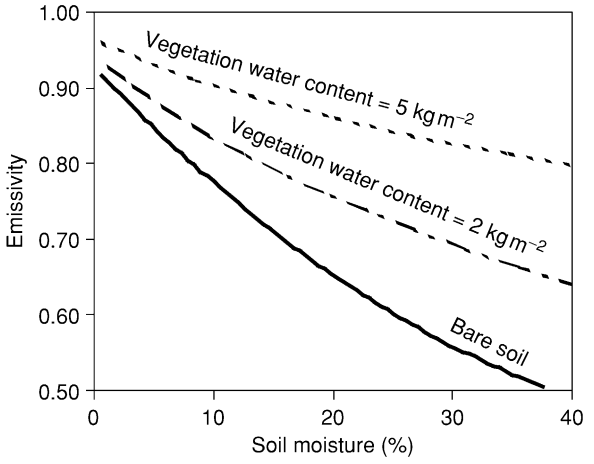


Figure 1 Relationships between soil moisture and emissivity for bare soil and vegetation. The configuration is a passive microwave instrument operating at 1.4 GHZ, horizontal polarization, and a 40 degree viewing angle.

reflectivity is more complex than for passive methods. The geometric properties of the soil surface and vegetation have a greater effect on these measurements and simple correction procedures are difficult to develop.

The signals sent and received by a radar are usually linearly polarized, either horizontal (H) or vertical (V). Possible combinations are HH, VV, HV, and VH. Advanced multipolarization systems can make all of these measurements simultaneously.

For bare soils, all models that relate the backscattering coefficient to soil moisture require at least two soil parameters, the dielectric constant and the surface-height standard deviation. This means that, in order to invert these models, the surface-height standard deviation must be accurately estimated.

For a given sensor configuration of frequency and viewing angle, different results are obtained at different polarizations. Therefore, most approaches to determining soil moisture with active microwave methods utilize dual polarization measurements. With two independent measurements of two dependent variables, both the dielectric constant and the surface height standard deviation can be solved for. Algorithms incorporating this approach for bare soils are available in the literature. A robust approach to account for vegetation has not yet been developed and validated.

Active microwave sensors on aircraft and spacecraft typically employ synthetic aperture radar (SAR), which utilizes the motion of the platform to synthesize larger antennas. Exceptional spatial resolutions with footprints of about 20 m can be achieved from satellite altitudes.

Aircraft Platforms and Application Example

Aircraft-based microwave instruments are especially useful in studies requiring the mapping of large areas. In most cases, they offer better spatial resolution (as low as several meters) than satellite systems, as well as more control over the frequency and timing of coverage. Many of the recent soil moisture studies have utilized the electronically scanned thinned-array radiometer (ESTAR). ESTAR is an L-band H-polarized passive microwave instrument that provides image products. It is also a prototype for a new type of sensor system that uses synthetic aperture antenna technology to solve the high-altitude spatial resolution problem described earlier.

Several active-microwave aircraft systems are in operation. One of the most widely used for studies related to soil moisture is the AIRSAR, operated by NASA Jet Propulsion Laboratory. The AIRSAR flies on a DC-8 aircraft and provides multipolarization data at frequencies of 5.2, 1.2, and 0.4 GHz (C, L, and P-band). It is very efficient for mapping large areas and provides exceptional spatial resolution, on the order of several meters.

Perhaps the most informative and widely studied aircraft data set is that obtained from Washita'92. This was a large-scale study of remote sensing and hydrology conducted using ESTAR over the Little Washita watershed in southwestern Oklahoma, USA. Passive and active microwave observations were made over a 9-day period in June 1992. The watershed was saturated with a great deal of standing water at the outset of the study. During the experiment, no rainfall occurred and observations of surface soil-water content exhibited a dry-down pattern over the period. Observations of surface soil-water content were made at sites distributed over the area. Significant variations in the level and rate of change in surface soil-water content were noted over areas dominated by different soil textures.

Passive microwave observations were made on 8 of the 9 days of the study period. The ESTAR data were processed to produce brightness-temperature maps of a 740-km^2 area at a 200-m resolution on each of the 8 days. Using the retrieval algorithm of soil water content, these brightness-temperature data were converted to images of soil-water content. Gray-scale images for each day are shown in Figure 2. These data exhibit significant spatial and temporal patterns. Spatial patterns are associated with soil textures, and temporal patterns are associated with drainage and evaporative processes.

Follow-on analyses revealed that these types of observations had potential in determining the

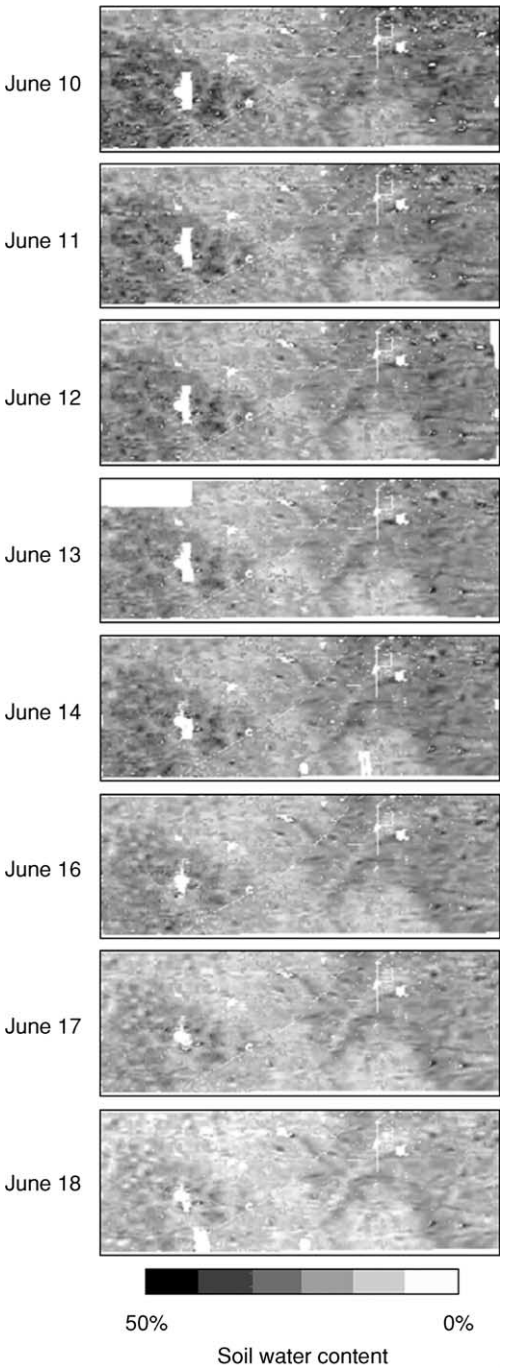


Figure 2 Maps of soil water content for the Little Washita watershed (Oklahoma, USA) derived from passive microwave data for June 1992. Spatial resolution is 200 m.

hydraulic properties of the soils. Like soil moisture, these data can be obtained at the field point scale; the data are not available over large areas, do not take into account spatial variability, and are costly and difficult to obtain. Temporal observations of soil moisture have been used in several investigations to estimate hydraulic properties. These studies constrained the problem by minimizing the role of the surface flux, or by specifying it, in order to estimate the hydraulic properties of the integrated soil profile.

It has been hypothesized that the change in soil-surface moisture following a wetting event and 2 days later could be used to determine the average hydraulic conductivity of the soil profile. Using simulations for soils of different textures, it has been observed that a relationship exists that is strongest for sandy soils and weakens as clay content increases. In addition, it has been observed that a measurement of the 0–5 cm layer provides as much information as a 0–30 cm observation on the deep soil moisture profile. These relationships were affected by evaporation; however, the effect was more significant for soils with lower hydraulic conductivity and with a shallow depth of observation.

This concept has been applied utilizing remotely sensed data from Washita'92 to derive spatially distributed information. As noted above, investigators hypothesized that surface soil-moisture changes could be used to identify soil-texture classes. A conclusion was that even if this method were not highly accurate for each and every sensor resolution element (pixel), the results would significantly improve the definition of the spatial domain of the soil hydraulic properties. In their opinion, this result was more valuable than the accuracy of individual point estimates.

Passive Microwave Satellite Observing Systems and Products

Special Satellite Microwave/Imager

Currently, all passive microwave sensors on satellite platforms operate at high frequencies (>6 GHz) (Table 1). Of particular note, due to the longevity of its data record, is the Special Satellite Microwave/Imager (SSM/I) package on the Defense Meteorological Satellite Platforms. These satellites have been in operation since 1987 and provide high frequencies and two polarizations. Interpreting data from the SSM/I to extract surface information requires accounting for atmospheric effects on the measurement. When one considers the atmospheric correction, the significance of vegetation attenuation, and the shallow contributing depth of soil for these high frequencies, it becomes apparent that the data are of limited value for estimating soil-water content. Spatial resolution of the SSM/I is very coarse (Table 1).

There have been few attempts at generating standard land-surface products using SSM/I data. To date the only standard soil-moisture-related product produced from SSM/I data is the Soil Wetness Index (SWI).

Table 1 Microwave satellite or sensor systems

Satellite	Period of coverage	Frequency (GHz)	Polarization	Spatial resolution of lowest frequency	Repeat frequency (days)
Passive					
SSM/I	1987–present	19.4, 22.2, 37.0, 85.5	H and V	70 km	1–2
TMI	1998–present	10.7, 19.4, 21.3, 37.0, 85.5	H and V	60 km	1
AMSR-E	2002–present	6.9, 10.7, 18.7, 23.8, 36.5, 89.0	H and V	75 km	2–3
Windsat	2003–present	6.8, 10.7, 18.7, 23.8, 37	H and V (U in some channels)	50 km	2–3
SMOS	(2007)	1.4	H and V	50 km	2–3
HYDROS	(2010)	1.4	H and V (passive) HH, VV (active)	40 km (passive) 3 km (active)	2–3
Active					
ERS	1991–present	5.3	VV	30 m	35
Radarsat-1	1995–present	5.3	HH	7 to 100 m	24
Envisat	2002–present	5.3	HH and VV or HH and HV or VV and VH	30 m	35
Radarsat-2	(2005)	5.3	HH, VV, HV and VH	3 to 100 m	24
ALOS	(2005)	1.27	HH or VV and HV or VH	10 to 100 m	46

This index is intended to provide information on significantly wet soil conditions (areal extent of flooding), which can be more reliably detected than variations at lower levels of soil moisture.

A few studies have attempted to extract actual soil moisture from SSM/I data, with some degree of success. For the limited validation data set available the approach performed relatively well. However, part of this may be attributable to the limited conditions and light vegetation conditions evaluated.

Tropical Rainfall Measurement Mission Microwave Imager

Another current option is the Tropical Rainfall Measurement Mission (TRMM) Microwave Imager (TMI). It is a five-frequency, dual-polarized, and passive-microwave radiometer. The lowest TMI frequency is 10 GHz, about half that of the SSM/I. The TMI has a higher spatial resolution (Table 1) as compared to the SSM/I. TRMM only provides coverage of the tropics, which includes latitudes between 38°N and 38°S for the TMI instrument. However, a unique capability of the TMI is its ability to collect data daily, and in many cases more often, within certain latitude ranges. The data from the TMI have been used to generate validated soil-moisture products for 5 years over the southern USA.

Advanced Microwave Scanning Radiometer

Several new multifrequency passive-microwave satellite systems were launched in 2002 and 2003. As opposed to the previously available systems, these offer a lower-frequency channel operating at C band (~6 GHz), which should provide a more robust soil moisture measurement. These satellites were the NASA Aqua, Japanese ADEOS-II, and the US interagency Coriolis satellite. The ADEOS-II satellite stopped operations in October of 2003 and will not be discussed here.

Aqua includes an instrument called the Advanced Microwave Scanning Radiometer (AMSR). AMSR has a 6.9 GHz (C-band) channel with a nominal 60-km spatial resolution. This instrument should be able to provide information about soil-water content in regions of low vegetation cover, less than 1 kg m^{-2} vegetation water content. Aqua collects data at nominal times of 1.30 a.m. and p.m. local time.

As opposed to previous passive-microwave satellite missions, the Aqua project will include soil moisture as a product. The algorithm planned for use by NASA is the multiple-channel approach. The Japanese Aerospace Exploration Agency will also produce a soil-moisture product using one of four algorithms under consideration. Preliminary studies indicate that there is widespread radiofrequency interference (RFI) in the C-band channels. Therefore, it is likely that the most useful channels for soil moisture will be those operating at the slightly higher X band.

The Coriolis satellite includes the Windsat instrument, which is a multifrequency passive microwave radiometer system. It is similar to AMSR with some differences in frequencies and more polarization options. It is a prototype of one component of the next generation of operational polar orbiting satellites that the USA plans to implementing by the year 2010.

Future Passive Systems

Programs are underway to develop and implement space-based systems with a 1.4 GHz (L-band) channel that would provide improved global soil-moisture information. Toward that goal the European Space Agency (ESA) is developing a sensor system called the Soil Moisture Ocean Salinity (SMOS) mission and NASA has initiated a mission called Hydros.

SMOS will utilize two-dimensional synthetic aperture radiometry at L band and provide 40-km H and V polarizations. The ability of SMOS to obtain multiple-incidence angle observations of the surface is the key element of the soil-moisture retrieval algorithm. The launch of the mission is anticipated in 2007.

Hydros will use an active and passive L-band instrument design for mapping at resolutions ranging from 3 to 40 km. Soil moisture will be estimated by Hydros using a microwave radiometer and radar data separately and in combination, taking advantage of the simultaneous, coincident, and complementary nature of the measurements. Hydros is scheduled for 2010.

Active Microwave Satellite Observing Systems

Current Active Systems

At present, several radar satellites are in orbit. ESA has operated a satellite synthetic aperture radar (SAR) series called ERS since 1991, which provides C-band VV data. It includes both a SAR and a scatterometer. Although numerous investigations have been conducted that attempt to utilize ERS SAR data, few results have been reported in the area of soil-moisture estimation. This is due to the limitations of using a single mid-frequency channel single-polarization SAR with an exact repeat cycle of 35 days.

The most recent satellite in the ERS series is called Envisat (launched in 2002) and it has a C-band advanced synthetic aperture radar (ASAR) with multiple polarization capabilities. It also offers the option of varying the incidence angle to allow for different viewing angles and more frequent coverage if angle is not a critical parameter in the application. Data from ASAR are just beginning to be made available to investigators.

The Canadian Space Agency operates a C-band satellite SAR called Radarsat, which offers HH polarization and has a variable viewing angle and a wide swath (large range of incidence angles). These choices offer more frequent temporal coverage of a particular region of the Earth if angle is not important.

There have been a several of attempts at using the single-channel SAR data to retrieve soil moisture. The general consensus from these studies is that there are too many physical variables that have to be known in order to derive soil moisture. These variables include soil moisture, surface roughness, and vegetation. Site-specific studies and creative selection of data sets have resulted in some degree of success.

Future Active Systems

Japan will include an L-band SAR called PALSAR on the Advanced Land Observing System (ALOS) in 2005 or later. PALSAR will have a multipolarization mode as well as varying incidence angles. Canada is developing Radarsat-2, which will be similar to Radarsat-1 but will be fully polarimetric. The satellite will include a mode in which the spatial resolution is 3 m. The expected launch is in 2005.

As noted above, progress in SAR soil moisture mapping has been limited by the available data, single channel. The new generation of multipolarization SAR systems will provide at least one additional measurement. There is hope that more robust soil moisture retrieval methods can be developed as these data become available.

List of Technical Nomenclature

θ	Viewing angle (degrees)
σ^0	Backscattering coefficient (decibels)
b	Vegetation parameter ($m^2\,kg^{-1}$)
e	Emissivity (dimensionless)
f	Frequency (GHz)
k	Dielectric constant (dimensionless)
p	Polarization (dimensionless)
r	Reflectivity (dimensionless)
T	Physical temperature (degrees Kelvin)
T_B	Brightness temperature (degrees Kelvin)
VWC	Soil moisture (volumetric percent)
W_{veg}	Vegetation water content ($kg\,m^{-2}$)

See also: **Evaporation of Water from Bare Soil**; **Evapotranspiration**; **Remote Sensing:** Organic Matter; **Water Cycle**

Further Reading

Altese E, Bolognani O, Mancini M, and Troch PA (1996) Retrieving soil moisture over bare soil from ERS-1 synthetic aperture radar data: sensitivity analysis based on a theoretical scattering model and field data. *Water Resources Research* 32: 653–661.

Bindlish R, Jackson TJ, Wood E *et al.* (2003) Soil moisture estimates from TRMM microwave imager observations over the southern United States. *Remote Sensing of Environment* 85: 507–515.

Dubois PC, van Zyl J, and Engman ET (1995) Measuring soil moisture with imaging radars. *IEEE Transactions on Geoscience and Remote Sensing* 33: 915–926.

Jackson TJ (1993) Measuring surface soil moisture using passive microwave remote sensing. *Hydrological Processes* 7: 139–152.

Jackson TJ and Schmugge TJ (1989) Passive microwave remote sensing system for soil moisture: some supporting research. *IEEE Transactions on Geoscience and Remote Sensing* GE-27: 225–235.

Jackson TJ and Schmugge TJ (1991) Vegetation effects on the microwave emission from soils. *Remote Sensing of Environment* 36: 203–212.

Jackson TJ, Le Vine DM, Swift CT, Schmugge TJ, and Schiebe FR (1995) Large area mapping of soil moisture using the ESTAR passive microwave radiometer in Washita '92. *Remote Sensing of Environment* 53: 27–37.

Mattikalli NM, Engman ET, Jackson TJ, and Ahuja LR (1998) Microwave remote sensing of temporal variations of brightness temperature and near-surface soil water content during a watershed-scale field experiment, and its application to the estimation of soil physical properties. *Water Resources Research* 34: 2289–2299.

Njoku EG and Kong J (1977) Theory for passive microwave remote sensing of near-surface soil moisture. *Journal of Geophysical Research* 82: 3108–3118.

Njoku EG, Jackson TJ, Lakshmi V, Chan TK, and Nghiem SV (2003) Soil moisture retrieval from AMSR-E. *IEEE Transactions on Geoscience and Remote Sensing* 41: 250–261.

Satalino G, Mattia F, Davidson MW *et al.* (2002) On current limits of soil moisture retrieval from ERS-SAR data. *IEEE Transactions on Geoscience and Remote Sensing* 40: 2438–2447.

Shutko AM (1986) *Microwave Radiometry of Water and Terrain Surfaces*. Moscow: Nauka Publishing House.

Ulaby FT, Moore RK, and Fung AK (1986) *Microwave Remote Sensing: Active and Passive*, vol. III. *From Theory to Application*. Dedham, MA: Artech House.

Wagner W, Lemoine G, and Rott H (1999) A method for estimating soil moisture from ERS scatterometer and soil data. *Remote Sensing of Environment* 70: 191–207.

Wang JR and Schmugge TJ (1980) An empirical model for the complex dielectric permittivity of soils as a function of water content. *IEEE Transactions on Geoscience and Remote Sensing* GE-18: 288–295.

Wigneron JP, Waldteufel P, Chanzy A, Calvet JC, and Kerr Y (2000) Two-dimensional microwave interferometer retrieval capabilities over land surfaces (SMOS mission). *Remote Sensing of Environment* 73: 270–282.

RHIZOSPHERE

A C Kennedy and L Z de Luna, USDA Agricultural Research Service, Pullman, WA, USA

Published by Elsevier Ltd.

Introduction

The rhizosphere is a narrow zone of soil surrounding and influenced by the root of vascular plants. This zone is characterized by intense biological activity owing to the release of root exudates, which stimulate or inhibit rhizosphere organisms. The interactions between the soil, the plant, and organisms that comprise the rhizosphere characterize the complexity and the dynamics of this region. Plant/microbe and plant/faunal interactions abound in the rhizosphere. These interactions range from symbiotic relationships such as N_2 fixation and mycorrhizal associations to pathogenic interactions. Management strategies such as bioremediation and biological control may be successful when rhizosphere ecology is considered. A greater understanding of the rhizosphere and its effects on organisms that inhabit this area will allow for manipulations that benefit plant production and the environment.

Hiltner coined the term 'rhizosphere' in 1904 to describe the zone of soil under the influence of plant roots. His observations were initially based on the interactions between symbiotic N_2-fixing bacteria and the legume root, but were then expanded to include all interactions. This zone of soil ranges from only a few hundred micrometers to greater than 5 mm from the root surface (**Figure 1**). Within this region increased root exudation and changes in physical and chemical soil properties may increase the activity and populations of most soil microorganisms. The rhizosphere is constantly in a state of dynamic flux. Several overlapping gradients in this zone vary with time and distance from the root, and influence rhizosphere colonization by soil organisms.

These gradients and the many interactions within them make the understanding of this zone difficult. The most dramatic gradient in the rhizosphere is formed by nutrients released by the root. Plant roots excrete various organic substances into the

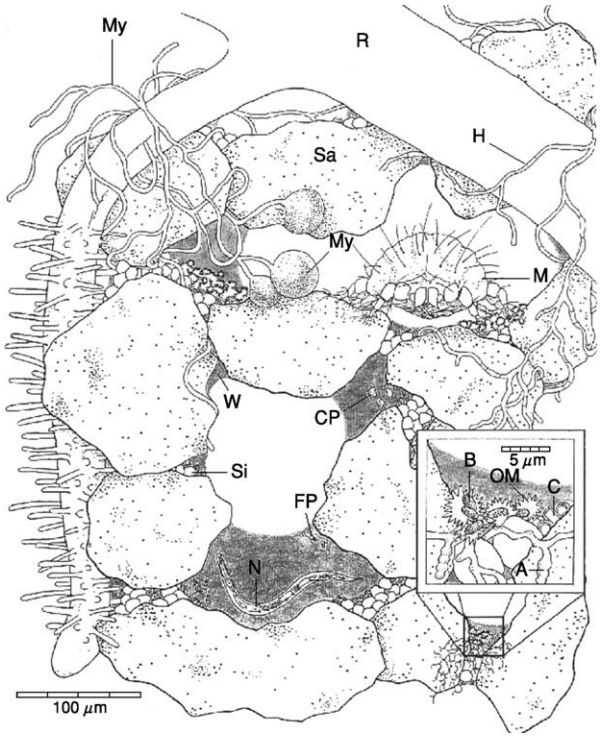


Figure 1 The rhizosphere habitat. R, root; B, bacteria; A, actinomycetes; My, mycorrhizal spores and hyphae; H, hyphae of saprophytic fungi; N, nematode; CP, ciliate protozoa; FP, flagellate protozoa; M, mite; Sa, sand; Si, silt; C, clay; OM, organic matter; W, water. Original drawing by Kim Luoma: Reproduced from Sylvia DM, Fuhrmann JJ, Hartel PG, and Zuberer DA (1998) *Principles and Application of Soil Microbiology*. Upper Saddle River, NJ: Prentice Hall.

rhizosphere, and these substances provide a rich source of nutrients for the microbial community. The greatest exudation arises from those cells involved in cell elongation and lateral root formation. Carbon compounds decline exponentially as the distance from the root increases. Increased exudation can result from extremes in temperature, water stress, P deficiency, increased light intensity, and increased populations of microorganisms. Herbicides, pathogens, foliar treatments, and symbiotic associations may also influence exudation rate. Decreased exudation from the root can result from N deficiency and decreased light intensity. There may also be a relationship between exudate pattern and microbial populations colonizing the root surface (rhizoplane) or the cell layers of the root itself (endorhizosphere).

Plant exudates of high molecular weight, which are primarily released by root tips, form a sheath of slime on the external surface of the root called mucilage. In contrast, mucigel contains plant mucilage and other plant products, as well as bacterial cells and their products. These secretions are important because they help to form and maintain stable soil aggregates. Most soils are extremely low in readily available C, and C-containing substances leaked by the root are of major importance to microbial growth. Exudates can be sugars, amino acids, vitamins, tannins, alkaloids, phosphatides, and other substances such as growth factors, fluorescent substances, nematode cyst or egg-hatching factors, and fungal growth stimulants and inhibitors. Sugars and amino acids, the most-studied root exudates, provide readily available nutrients to the rhizosphere inhabitants and may be a factor in governing susceptibility or resistance of plants to root-infecting fungi.

Plant transpiration can induce severe changes in the water potential at the root surface, which may cause water stress for the organisms near the root as well as reducing the movement of necessary nutrients. Dry soils may inhibit the growth of the root and enhance the release of exudates. Waterlogged soils on the other hand reduce oxygen availability and may increase exudation due to root injury. These conditions may influence the quantity of exudates released and the growth of the microbial and macrofaunal populations in the rhizosphere. Root exudates and the release of H^+, HCO_3^-, or organic compounds, as well as nitrogen nutrition influence rhizosphere pH. Rhizosphere pH can fluctuate as much as one or two pH units over very short time periods. These changes in the soil pH can also affect the growth and survival of the soil biota both positively and negatively. In general, bacteria are negatively affected by low pH, while fungi are not as sensitive to extremes in pH.

Although there is an increase in microbial activity in the rhizosphere, the distribution of those organisms on the root surface is quite variable. Electron micrographs show that microorganisms colonize only 7–15% of the actual root surface. Bacteria are found on the root surface in small colonies at the junctions of epidermal cells or at tears in the root surface (Figure 2). Colonization of the root surface and the rhizosphere is dependent upon factors involving the organisms, the plant, and the environment (Table 1). Microorganisms are thought to move down the root by three different mechanisms: (1) motility of the organism (active transport); (2) water movement on the surface of the root (passive transport); and (3) movement of microorganisms on the root apex as the root cells elongate (passive transport).

Organisms in the Rhizosphere

Bacteria

Rhizosphere bacteria play vital roles in plant nutrition, growth promotion, and disease interactions. Several studies have indicated that bacteria are the most numerous inhabitants of the rhizosphere, although they account for only a small portion of the total

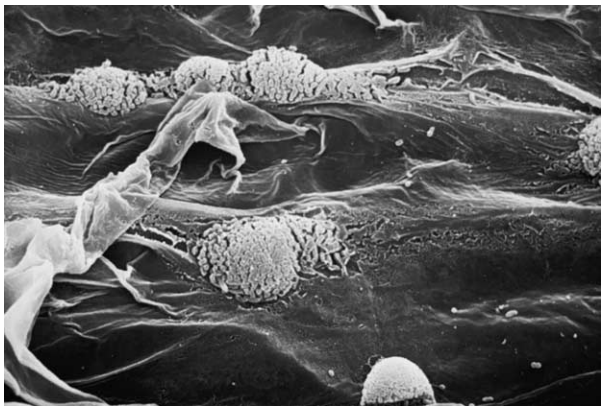


Figure 2 Colonization of the root by bacteria. Note the clumping of colonies at intersection of cells. (Reproduced from Begonia MFT, Kremer RJ, Stanley L, and Jamshedi A (1990) Association of bacteria with velvetleaf roots. *Transactions of the Missouri Academy of Science* 24: 17–26.

Table 1 Factors influencing rhizosphere colonization

Microbial/macrofaunal characteristics
Versatile nutrient requirements
Fast initial growth rate
Enhanced cellulase or other enzyme production
Antibiotic production
Chromophore influence on metal nutrition
Unique physiological attributes
Tolerance to antibiotics, fungicides, or other chemicals
Plant characteristics
Plant species/cultivar
Plant and root age
An unoccupied niche in soil, rhizosphere, or rhizoplane
Altered plant genetics
Foliar treatments
Soil-applied pesticides
Exudation patterns
Cropping sequence and rotation
Environment
pH
Soil type
Soil texture
Soil porosity and compaction
Soil moisture
Soil atmosphere
Temperature
O_2 availability
Fertility, organic matter, and available nutrients
Pesticide and herbicide application

Adapted from Kennedy AC (1998) The rhizosphere and spermosphere. In: Sylvia DM, Fuhrmann JJ, Hartel PG, and Zuberer DA (eds) *Principles and Applications of Soil Microbiology*, pp. 389–407. Upper Saddle River, NJ: Prentice-Hall, with permission.

biomass due to their small size. Typically, a 1-g sample of rhizosphere soil contains approximately 10^8–10^{12} bacterial cells; Gram-negative bacteria predominate in the rhizosphere; of these, the most efficient bacterial root colonizers belong to the genus *Pseudomonas*. Actinomycetes are filamentous chemoorganotrophic bacteria that form asexual spores. Populations of these bacteria are usually higher in rhizosphere than in nonrhizosphere soil. There are indications that actinomycetes may comprise higher percentages of the total bacterial numbers than fluorescent pseudomonads in the root zone of some plants. Although some actinomycetes may cause plant diseases, many species have the potential to produce antibiotics that could impact plant–microbe interactions, such as nodulation in legumes or biocontrol efforts against plant pathogens.

Fungi

Fungi are eukaryotic, nonphotosynthetic, spore-forming organisms that are primarily responsible for residue decomposition and organic matter accumulation through the mineralization of cellulose, lignin, and other organic materials. In the rhizosphere and rhizoplane, fungi may function as symbionts, pathogens, nutrient sinks, and a food source for other soil microorganisms. Fungal mycelia also bind soil particles together to form aggregates.

Fauna

Animals that spend part or all of their lives in close association with plant roots and have significant effects on the structural and functional properties of the rhizosphere make up the rhizosphere fauna. Soil fauna are classified into three types based on body size. Single-celled animals such as protozoa belong to the microfauna, while rotifers, nematodes, and microarthropods are classified as mesofauna. Earthworms, millipedes, centipedes, and insects comprise the macrofauna.

Microfauna (protozoa) Protozoa are eukaryotic, unicellular organisms, most of which are microscopic in size, although some members may attain macroscopic dimensions. While the group exhibits differences in feeding behavior, all require a water film for metabolic activity. It is estimated that there are between 10^4 and 10^5 protozoa per gram of soil. Populations of protozoa are generally higher in rhizosphere soil (R) than bulk soil (S), with large numbers occurring in hotspots such as around dead plant roots. Protozoa feed on bacteria and fungi and thus contribute to the mineralization of N, leading to an increase in N uptake by plants. When rhizosphere protozoa prey on bacteria, up to 60% of the excess N and P is excreted in forms that can be readily utilized by plants. Plants may also benefit from the suppression of pathogenic bacteria and fungi by protozoa in the rhizosphere. Detrimental effects of

protozoa include a suppression of *Rhizobium* populations in the rhizosphere, resulting in reduced root nodulation of *Phaseolus vulgaris*.

Mesofauna (nematodes and microarthropods) Nematodes are microscopic, unsegmented, round worms, usually 0.5–2.0 mm in length. There are five trophic groups of nematodes: bacterivores, which feed on bacteria; fungivores, which feed on fungal mycelia; predators, which feed on other nematodes and small invertebrates; omnivores, which feed on a variety of foods; and herbivores, which are plant parasites. Nematodes play a significant role in mineral cycling. Their consumption of bacteria and other food sources results in excretion of N, mainly in the form of ammonia, and enhanced mineralization of P and S in some systems. Nematodes are also very important in agriculture, because some species are parasitic on crops. Feeding by these plant-parasitic nematodes results in wilting, nutrient-deficiency symptoms, yield losses, and other problems linked to damaged root systems.

Microarthropods such as mites (Acari: Arachnida) and springtails (Collembola: Insecta), are a major fraction of the mesofauna present in soils; however, they generally do not harm plants. Although some soil mites prey on nematodes and other small arthropods, they are generally fungal feeders. A few species pierce roots and suck plant sap, ingest roots, or feed on decomposing plant matter.

Table 2 Activities of organisms in the rhizosphere

Decomposition of plant residue and organic material
Humus synthesis
Mineralization of organic N, S, and P
Increased plant nutrient availability of P, Mn, Fe, Zn, Cu
Symbiotic mycorrhizal associations
Production of organic chelating agents
Oxidation–reduction reactions
P solubilization
Biological dinitrogen fixation
Free-living bacteria and blue-green algae
Associative microorganisms
Symbiotic legume and nonlegume
Plant growth promotion
Production of plant growth hormones
Protection against root pathogens and pseudopathogens
Enhanced nutrient-use efficiency
Plant growth inhibition
Pathogenicity by root pathogens
Growth reduction by deleterious rhizobacteria
Biological control
Reduction of plant diseases
Inhibition of soil nematodes and insects
Reduction in weed growth
Bioremediation
Degradation of industrial and agricultural chemicals
Volatilization of metal contaminants
Hyperaccumulation of metals in plant tissues
Enhanced drought tolerance of plants
Improved soil aggregation

Adapted from Kennedy AC (1998) The rhizosphere and spermosphere. In: Sylvia DM, Fuhrmann JJ, Hartel PG, and Zuberer DA (eds) *Principles and Applications of Soil Microbiology*, pp. 389–407. Upper Saddle River, NJ: Prentice-Hall, with permission.

Interactions in the Rhizosphere

There are a myriad of interactions occurring in the rhizosphere (Table 2). The most studied of these are N_2-fixation, mycorrhizal associations, plant growth promotion, and plant growth inhibition. In recent years, bioremediation and biological control have also been investigated.

Dinitrogen-Fixation Symbiosis

Dinitrogen-fixing bacteria can make otherwise unavailable N_2 available to the plant. Rhizobia and related bacteria form nodules on the roots of plants. This interaction can be mutually beneficial as the plant provides nutrients and protection for the bacteria and the bacteria provide N for the plant. Other N_2-fixing organisms form associations with plants. These bacteria colonize the root surface and interior cells, and intercellular spaces between the cortex and endodermis and in the xylem cells of grasses. Along with plant-available N, they may provide other plant growth-promoting compounds also.

Mycorrhizal Associations

Mycorrhizal fungi form a mutualistic symbiosis with plants and infect roots without causing root disease. These fungi can be found in the rhizosphere of most plants and form associations with all gymnosperms and more than 83% of dicotyledonous and 79% of monocotyledonous plants. Mycorrhizal fungi can form structures either on the outside (ectomychorrhizae) or inside (endomycorrhizae) of plant roots. The fungal hyphae allow the roots to contact a greater volume of soil. Some types of mycorrhizal fungi increase the solubilization of nutrients such as phosphorus. They assist the plant in increasing nutrient absorption, especially in stressed environments (e.g., phosphorus- and water-deficient soils), selective ion uptake, and provide protection from extremes in the environment. Plant exudation patterns may be altered after colonization by these fungi, thus affecting the rhizosphere microbial and macrofaunal communities. The fungi may also protect plant roots from invasion by pathogens. Extraradicle hyphae of endomycorrhizae secrete glomalin, a glycoprotein

that aggregates soil particles, increasing water-stable aggregates and improving soil structure. This association can increase the survival and growth of a plant, especially in adverse or low-nutrient environments, and may have potential in the revegetation of disturbed sites.

Plant Growth-Promoting Organisms

Plant growth-promoting organisms are those organisms that in some way enhance the growth of plants. Plant growth-promoting organisms enhance seed germination and plant growth and are often specific bacteria or fungi. Many different mechanisms are responsible for plant growth promotion. These organisms have been used in biological control programs to protect plants against plant pathogens, as biofertilizers to fix atmospheric N_2, and in phytostimulation, which directly improves plant growth through the production of plant growth regulators. Some may make nitrogen more available to plants and have the potential to reduce application of inorganic fertilizers. These organisms may produce plant growth-stimulating compounds such as gibberellic and indoleacetic acid. The reduction in pathogen load may also be a mechanism of plant growth promotion. Iron-chelating compounds called siderophores can function to make iron more available to an organism and less available to the plant pathogen, thus impairing the pathogen. Since iron is a key nutrient in metabolism and a cofactor in enzyme activity, microbial growth may be influenced by these siderophores. The microbial populations that colonize the interior of the root and form intimate associations with the root are considered endophytes. Endophytic bacteria may increase plant growth, confer disease resistance, and aid the plant in withstanding stresses such as drought. The success of endophytes in colonizing the rhizosphere may be the ability of the microorganism to compete with other bacteria on the external or internal portions of plant roots. Introduction of plant growth-promoting organisms may change the overall composition of the microbial community, especially in the rhizosphere.

Plant Growth-Inhibiting Organisms

Organisms that inhibit plant growth colonize the rhizosphere and can alter plant development. Plant growth inhibition can occur with or without disease symptoms. The rhizosphere harbors pathogens such as fungi and bacteria that attack germinating seeds and plant roots. Pathogens may produce extracellular enzymes that liquefy cell walls, thereby allowing degradation of the cell contents. Some pathogens produce toxins that affect the permeability of the cell membrane or inhibit an enzyme, leading to interruptions in a metabolic pathway. These microbial toxins may affect many species across different plant families or they may be host-specific. Polysaccharides produced by some pathogens may block vascular tissues and cause wilting. Excessive production of auxins and cytokinins, resulting from integration of a tumor-inducing bacterial plasmid into the plant genome, leads to gall formation in the roots. Feeding by plant-parasitic nematodes exposes root cells to destructive enzymes and toxic metabolites, which ultimately result in decreased water and nutrient uptake in plants. Puncture wounds from the feeding process could also become infection courts for the entry of other microbial pathogens into the roots. Deleterious rhizobacteria (DRB) inhibit plant growth without causing obvious disease symptoms. These organisms are present on the root surface and may occur in the intercellular spaces of cortical cells. DRB do not parasitize the plant, but survive on organic compounds released by plant root cells. Their detrimental effects on plant growth are due mainly to the production of metabolites that disturb plant physiological processes, including phytotoxins, plant growth regulators, volatile substances, or antibiotics. Their colonization and inhibition are often found to be species- or cultivar-specific, and root exudates may play a part in some plant–DRB interactions.

Bioremediation

Industrial development and intensification of agriculture in the past several decades have contributed to environmental contamination, releasing hazardous wastes into terrestrial ecosystems. Due to their toxic and carcinogenic nature, and tendency to bioaccumulate in humans and other organisms, the timely and cost-effective clean-up of contaminated areas has become a priority. Plant–microbial systems have shown potential in bioremediation, a process that utilizes the natural role of plants and microorganisms in transformation, mineralization, and complexation of organic and inorganic pollutants. In contrast to conventional soil remediation technologies, bioremediation is less costly, utilizes relatively simple technology, and is minimally disruptive to the contaminated site. Bioremediation may be facilitated by plant root exudates that show high binding affinities to certain pollutants or by rhizosphere microorganisms that mineralize contaminants into their nontoxic derivatives.

Plant roots, together with their associated microbial communities, may enhance mineralization of hazardous compounds in the soil. The 'rhizosphere effect' on bioremediation may occur in several ways. First, it is thought that carbon substrates from roots stimulate increases in xenobiotic degrader populations and activity. Second, the presence of dense populations of diverse microorganisms in the rhizosphere may enhance bacterial conjugation, which could result in the evolution of new pathways of degradation. Third, plants may provide a substrate that increases the production of the enzymes involved in mineralization of the hazardous compound. Rhizosphere organisms can influence the degradation of recalcitrant compounds in the soil (Table 3). Rhizosphere microorganisms may also facilitate the

Table 3 Plant species shown to facilitate microbial degradation of hazardous chemicals in the rhizosphere

Plant		
Common name	*Species*	*Chemical*
African clover	*Trifolium africanum*	2,4-D[a]
Alfalfa	*Medicago sativa*	PAHs[b]
Bahia grass	*Paspalum notatum*	TCE[c]
Beans	*Phaseolus vulgaris*	Diazinon[a], Parathion[f], Temik[d]
Cattail	*Typha* sp.	Surfactants[g]
Corn	*Zea mays*	Atrazine[h], Temik[d], fuel oil hydrocarbons
Cotton	*Gossypium hirsutum*	Temik[d]
Flax	*Linum usitatissimum*	2,4-D[a]
Goldenrod	*Solidago* sp.	TCE[c]
Lespedeza	*Lespedeza cuneata*	TCE[c]
Loblolly pine	*Pinus taeda*	TCE[c]
Meadow brome	*Bromus biebersteinii*	2-Chlorobenzoic acid
Peas	*Pisum sativum*	Diazinon[e]
Prairie grasses	*Andropogon gerardi, Sorghastrum nutans, Panicum virgatum, Elymus canadensis, Schizachyrium scoparius, Bouteloua curtipendula, Agropyron smithii, Bouteloua gracilis*	PAHs[b], TCE[c]
Reed	*Phragmites communis*	VOCs[i], 4-chlorophenol
Rice	*Oryza sativa*	Propanil[j], *p*-nitrophenol, Alachlor[k], carbofuran[l] residues
Ryegrass	*Lolium perenne*	Anthracene, TCE[c], PAHs[b]
Salt marsh plants	*Distichlis spicata, Juncus gerardi, Phragmites australis, Spartina alterniflora, Sporobolus airoides*	PAHs[b]
Sorghum	*Sorghum vulgare*	PAHs[b]
Soybean	*Glycine max*	TCE[c], surfactants[g]
Sugarcane	*Saccharum officinarum*	2,4-D[a]
Summer cypress	*Kochia scoparia*	Atrazine[h]
Tall fescue	*Festuca arundinacea*	PAHs[b]
Tobacco	*Nicotiana tabacum*	Maleic hydrazide[m]
Wheat	*Triticum aestivum*	Diazinon[e], Mecoprop[n], 2,4-D[a], MCPA[o]
Wild rye	*Elymus dauricus*	2-Chlorobenzoic acid
Zinnia	*Zinnia angustifolia*	Mefenoxam[p]

[a]2,4-Dichlorophenoxyacetic acid.
[b]Polycyclic aromatic hydrocarbons.
[c]Trichloroethylene.
[d]2-Methyl-2(methylthio)propionaldehyde *O*-(methylcarbamoyl)oxime.
[e]*O*-*O*-Diethyl-*O*-(2-isopropyl-6-methyl-4-pyrimidyl) phosphorothioate.
[f]*O*-*O*-Diethyl-*O*-*p*-nitrophenyl phosphorothioate.
[g]Dodecyl linear alkylbenzene sulfonate, dodecyl linear alcohol ethoxylate, dodecyltrimethylammonium chloride.
[h]2-Chloro-4-ethylamino-6-isopropylamino-*s*-triazine.
[i]Volatile organic compounds (benzene, biphenyl, chlorobenzene, dimethylphthalate, ethylbenzene, naphthalene, *p*-nitrotoluene, toluene, *p*-xylene, bromoform, choloroform, 1,2-dicholoroethane, tetracholoroethylene, 1,1,1-trichloroethane.
[j]3′,4′-Dichlorpropionanilid-*N*-(3,4-dichlorphenyl) propanamid.
[k]2-Chloro-*N*-(2,6-diethylphenyl)-*N*-(methyoxymethyl)acetamide.
[l]2,3-Dihydro-2,2-dimethyl-7-benzofuranyl methylcarbamate.
[m]1,2-Dihydro-3,6-pyridazinedione.
[n]2-(2-Methyl-4-chlorophenoxy)propionic acid.
[o]2-Methyl-4-cholorophenoxyacetic acid.
[p][(*R*)-*N*-(2,6-dimethyl-phenyl)-(methoxyacetylamino) propionic acid methyl ester].

removal of metal contaminants such as mercury, selenium, and zinc from the soil. Bacterial isolates and consortia have been extensively studied in bioremediation, but fungi, actinomycetes, mycorrhizal associations, and even rhizosphere fauna have shown degrading abilities. Clearly, the greatest challenge in bioremediation technology is the elucidation of mechanisms involved in remediation of contaminated sites by plants and their associated organisms. This knowledge will allow for the selection and engineering of the best plant–microbe combinations and optimization of the remediation process.

Biological Control

The rhizosphere is the location for biological control interactions that make use of pathogens, parasites, or other predators to reduce the population or activity of another organism. Three major strategies for biological control are classical, inundative, and integrated management. The classic approach involves the use of natural enemies for release, dissemination, and self-perpetuation on target pests. The inundative approach is most often used in the rhizosphere and involves the addition of a virulent strain of a biocontrol agent to suppress pests. In this case, the biocontrol agent is not self-sustaining and must be applied to the target host every season. The integrated management approach is a broad strategy that involves management practices to conserve or enhance native enemies of pests.

Biological control can be used in the management of pathogens, nematodes, and weeds. The use of biological control in disease management allows for alternatives to chemical control (Table 4). The interactions of organisms in the rhizosphere and the ecological significance of these interactions that occur in the root and seed environment assist in the success of biocontrol strategies. The reduction in plant pathogens such as *Fusarium oxysporum*, *Gaeumannomyces graminis*, *Pythium*, and *Phytophthora* species due to suppressiveness of some soils is thought to be due to physiochemical and microbial soil factors. The suppression is specific and may have various mechanisms. The mechanisms can include antibiosis, production of siderophores or volatile compounds, parasitism, competition for nutrients, and competition for ecologic niches and induced disease resistance. Crop monoculture can, with time, result in a decrease in disease due to the development of disease-suppressive soils. Other management practices that may alter the rhizosphere environment and be considered as biocontrol approaches are crop rotation, soil amendments, cover cropping, fumigation, or soil solarization.

Fungal pathogens of protozoa and nematodes can be used in biological control of these plant pests. Fungal mycelial growth into the cortical cells is thought to assist in controlling clubroot disease caused by protozoa. Plant-parasitic nematodes alter the exudate patterns of the plant as they feed. This in turn changes the microbial community of the rhizosphere at these sites, which affects the colonization of the fungal and bacterial antagonists of the nematodes. One of the more interesting and most studied fungal interactions for biological control is the use of nematode-trapping fungi to reduce nematode populations.

Table 4 Examples of plant diseases for which potential biological control agents have been identified

Disease	*Host plant(s)*	*Antagonist*
Damping-off (*Pythium* spp.)	Beet	*Actinoplanes* sp.
	Corn	*Burkholderia cepacia* PHQM 100
	Cucumber	*Pseudomonas aureofaciens* 63-28
		Serratia plymuthica; *Pythium oligandrum*; *Trichoderma longibrachiatum* CECT 2606
	Bird's foot trefoil	*Pseudomonas fluorescens* VO61
	Eggplant	*Cladorrhinum foecundissimum*
Sheath blight (*Rhizoctonia solani*)	Rice	*Pseudomonas fluorescens* VO61
Root rot (*Rhizoctonia solani*)	Wheat	*Bacillus* sp.
Take-all (*Gaeumannomyces graminis* var. *tritici*)	Wheat	*Pseudomonas fluorescens* Q8r1-96
Fusarium wilt (*Fusarium* spp.)	Chickpea	*Bacillus subtilis* GB03
	Cucumber	*Paenibacillus* sp. 300
	Tomato	*Pseudomonas* spp.; *Fusarium oxysporum* Fo47[a]; *Fusarium solani*[a]; *Fusarium* spp.[a]
	Radish	*Pseudomonas putida* WCS417
Root rot (*Phytophthora* sp.)	Pepper	*Trichoderma harzianum* 2413
Verticillium wilt (*Verticillium dahliae*)	Eggplant	*Talaromyces flavus*
	Pepper	*Pythium oligandrum*

[a]Nonpathogenic strain.

These predatory fungal species produce many different types of trapping structures such as constricting rings and adhesive knobs or hyphal meshes. However, the factors that induce significant predaceous behavior need to be studied further.

Soil microorganisms that have potential in weed management have been increasingly studied in recent years. Fungal and bacterial pathogens of weeds that attack weed seeds or cause root damage are in various stages of development as biological control agents. These organisms may reduce weed populations through weed inhibition or mortality. Deleterious rhizobacteria that are host-specific have been isolated from many grass and broadleaf weed species. Bacteria have been shown to inhibit winter annual grass weeds and velvetleaf (*Abutilon theophrasti* Medik.). Rhizosphere fungi may also be used to control weeds. Propagules of *Fusarium oxysporum* f. sp. *erythroxyli* applied to soil results in vascular wilt symptoms and death of cocoa plants. The saprophytic, soilborne fungus *Gliocladium virens* produces viridiol, a steroidal phytotoxin that causes severe root necrosis in susceptible weed seedlings. Isolates belonging to *Fusarium* spp. have also shown promise for biological control of leafy spurge. Future development of biological control strategies require ecologic investigations of these systems and will enhance pest-management programs.

Summary

The rhizosphere is a dynamic region governed by complex interactions between plants and the organisms that are in close association with the root. The composition and pattern of root exudates affect microbial activity and population numbers, which in turn have an impact on the nematodes and microarthropods that share this environment. Beneficial or harmful relationships exist between rhizosphere organisms and plants, which ultimately affect root function and plant growth. In addition, the rhizosphere may include organisms that do not directly benefit or harm plants but clearly influence plant growth and productivity. A better understanding of the soil–root and soil–seed interface is needed to manage microorganisms, increase plant growth, and reduce the impact of plant production and agriculture on the environment. The benefits of studying the rhizosphere include the use of plant growth-promoting organisms and the suppression of plant diseases and weeds using biocontrol agents. Rhizosphere organisms can also be used to enhance the formation of stable soil aggregates and as bioremediation agents of contaminated soils. With greater understanding of the ecology and biota in the rhizosphere, this zone of increased nutrients, biotic activity, and interactions can be manipulated to improve plant productivity and environmental quality.

See also: **Bacteria:** Plant Growth-Promoting; Soil; **Fauna**; **Fungi**; **Microbial Processes:** Environmental Factors; **Nematodes**; **Pollutants:** Biodegradation; **Root Exudates and Microorganisms**

Further Reading

Anderson TA and Coats JR (1994) *Bioremediation Through Rhizosphere Technology*. ACS Symposium Series 563. Washington, DC: American Chemical Society.

Andow DA, Ragsdale DW, and Nyvall RF (1997) *Ecological Interactions and Biological Control*. Boulder, CO: Westview Press.

Crossley DA Jr, Mueller BR, and Perdue JC (1992) Biodiversity of microarthropods in agricultural soils: relations to processes. *Agriculture, Ecosystems and Environment* 40: 37–46.

Curl EA and Truelove B (1986) *The Rhizosphere*. Berlin, Germany: Springer-Verlag.

Hornby D (1990) *Biological Control of Soil-borne Plant Pathogens*. Oxford, UK: CAB International.

Keister DL and Cregan PB (1991) *The Rhizosphere and Plant Growth*. Dordrecht, The Netherlands: Kluwer Academic Publishers.

Kennedy AC (1998) The rhizosphere and spermosphere. In: Sylvia DM, Fuhrmann JJ, Hartel PG, and Zuberer DA (eds) *Principles and Applications of Soil Microbiology*, pp. 389–407. Upper Saddle River, NJ: Prentice-Hall.

Lynch JM (1990) *The Rhizosphere*. Chichester, UK: John Wiley.

Paoletti MG (1999) *Invertebrate Biodiversity as Bioindicators of Sustainable Landscapes*. Amsterdam, The Netherlands: Elsevier.

Papendick RI and Campbell GS (1975) Water potential in the rhizosphere and plant and methods of measurement and experimental control. In: Bruehl GW (ed.) *Biology and Control of Soil-Borne Pathogens*, pp. 39–49. St. Paul, MN: American Phytopathological Society.

Pinton R, Varanini Z, and Nannipieri P (2001) *The Rhizosphere*. New York: Marcel Dekker.

Upadhyay RK, Mukerji KG, and Chamola BP (2000) *Biocontrol Potential and its Exploitation in Sustainable Agriculture,* vol. 1, *Crop Diseases, Weeds, and Nematodes*. New York: Kluwer Academic/Plenum Publishers.

Waisel Y, Eshel A, and Kafkafi U (1996) *Plant Roots: The Hidden Half*. New York: Marcel Dekker.

Whipps JM (2001) Microbial interactions and biocontrol in the rhizosphere. *Journal of Experimental Botany* 52: 487–511.

RICHARDS, LORENZO A.

W R Gardner, Texas A & M University, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Most soil physicists, past or present, would agree that L.A. Richards was one of the pre-eminent soil physicists of the twentieth century. Although he was primarily an experimentalist, his name will always be associated with the equation describing the movement of water in unsaturated soil, one of the most useful equations in soil physics.

Richards was born in Fielding, Utah, on 24 April, 1904, and died in Riverside, California on 12 March, 1993. Fielding is a small agricultural community some 25 miles (40 km) west of Logan, which is the site of Utah State University. This proximity was fortunate for Ren (as he was known to his friends), because it made it possible for him to attend Utah Agricultural College (UAC) (as it was then called), as did three of his brothers, Bert L. Richards, Matthias C. Richards, and Sterling J. Richards. He married Zilla Linford, whose family was also replete with scientists. Among them were her brothers Maurice, Leon, Hooper, and Henry. Ren, Sterling, and Leon all majored in physics. Ren and his brother Sterling eventually ended up in Riverside, California, working in soil physics.

All three studied physics at the UAC under Willard Gardner, who received the eighth PhD in physics from the University of California in Berkeley. From this association came Ren's first scientific publications entitled 'The usefulness of capillary potential in soil moisture and plant investigations.' This paper summarized the current understanding of capillary potential as described by Edgar Buckingham and later expanded by Willard Gardner. It went on, however, to describe the principles and construction of the tensiometer and proposed a tension plate apparatus. The latter was intended for use on small samples of soil in the laboratory. He included in the paper water retention data as a function of capillary potential for four soils, some of the first data published. Following the work of Buckingham and Gardner, Richards also extended Darcy's law to unsaturated soils:

$$V = -\mathrm{K}\ \mathrm{grad}\Phi \qquad [1]$$

where V is the flux density of the water, K is the permeability, and Φ is the capillary potential. Possibly influenced by his older brother Bert, who was a plant scientist, Ren went on to consider the availability of water to plants as it moved through the soil to plant roots. He pointed out that the term 'availability' involves two notions: namely the ability of plant roots to absorb the water with which it is in contact and the readiness or velocity with which the water moves to replace that which has been used by the plant. The first was dependent upon the capillary potential or the security with which the water was held by the soil combined with the osmotic potential or salinity. He was to return often to the problem of the measurement of this property in his later career. The second factor or rate of flow of water to the plant roots should be described by the appropriate solution of eqn [1]. Eqn [1], when combined with the equation of continuity:

$$q = -\rho\partial\theta/\partial t \qquad [2]$$

yields:

$$\begin{aligned}\partial\theta/\partial t = {} & \partial(D\partial\theta/\partial \mathrm{x})\partial \mathrm{x} \\ & + \partial(D\partial\theta/\partial \mathrm{y})\partial \mathrm{y} \\ & + \partial(D\partial\theta/\partial \mathrm{z})\partial \mathrm{z} \\ & + \partial \mathrm{K}/\partial \mathrm{z}\end{aligned} \qquad [3]$$

where θ is the soil water content and D is the soil water diffusivity and ρ is the bulk density of the soil.

Eqn [3] is known as Richards' equation. It was not widely attributed to him until towards the end of his scientific career, but will undoubtedly always carry his name. The equation can also be written in terms of the matric or capillary potential ψ. There seems to be little doubt that eqn [3], or at least the capillary form of the equation, was known at Logan during the time that Ren was a graduate student there; however, it was not published until Richards included it as an important element in his PhD thesis from Cornell. It is appropriate that he be given the credit for it. His thesis, published in 1931, was titled 'Capillary conduction of liquids through porous medium.' In this Richards used two parallel porous fired ceramic plates to apply a differential pressure across a sample of soil and measured the flux between the plates. By varying the pressure across the plates he could maintain a variable degree of unsaturation in the soil. He had suggested the use of porous plates to evaluate the capillary potential of soil samples in his master's thesis. His PhD was apparently the first actual use of such plates.

After graduation from Cornell University with his PhD, Richards, like most physicists of that day, had difficulty in finding a job to his liking. Finally in 1935

he was offered a position at Iowa State University with a joint appointment in both the Physics and the Agronomy Departments. Here he began to apply his ideas to real problems and began to make his mark as an experimental physicist.

At that time the tensiometer was largely a novelty. At Logan he was introduced to the principle. The first attempts usually consisted of a glass funnel to which a porous plate was attached across the mouth of the funnel. Securing the plate to the funnel without leakage proved almost impossible. It was most fortuitous for the future of soil physics that Zilla Richards developed into a skilled ceramist. Ren and Zilla's collaboration eventually resulted in ceramic tubes which could be sealed very effectively to copper (and later acrylic) tubing and remain leakproof for days and eventually weeks. The diffusion of air through the ceramic then became the major limitation to their use, and this could be dealt with by periodic refilling with de-aired water.

Richards studied the use of tensiometers in field soils and in lysimeters and demonstrated their practicality and utility. During this time he also explored a number of other physical properties of soils, including the use of thermojunctions for the measurement of the relative humidity of soils. This particular problem was to occupy much of his time for the remainder of his career. Ren devised a method for measuring the flow of water in soils. He and his only student, M.B. Russell, studied the heat of wetting of soils and he and other colleagues pursued the use of tensiometers in the field.

In 1938 Ren was offered a position at the US Salinity Laboratory. Here he established a soil physics program which attracted visitors from all over the world and found an ideal scientific home for his researches for the remainder of his career. At this laboratory he continued to perfect the tensiometer and developed the pressure-membrane and pressure-plate apparatus. The pressure-membrane apparatus was initially conceived as a means of extracting water from soil by applying pressure to a soil sample supported on a porous membrane, with the entire apparatus contained in a brass chamber vented to the atmosphere below the membrane. However it soon became apparent that by increasing the air pressure by increments to 15 atmospheres and measuring the water remaining in the soil sample the capillary potential could be measured. As anticipated by Richards, this proved to be a significant advance.

Richards' work on soil physics was interrupted in 1943 by World War II when Ren joined the staff of the Jet Propulsion Laboratory at the California Institute of Technology. Here he worked on sea-borne rocket launchers. He worked with the diligence and insight that characterized all his research, and he was able to see this project through to completion and traveled to the Pacific Theater of Operations to observe the launchers in action. He eventually obtained several patents on his design. These he added to most of the early patents on tensiometers and pressure plates.

Upon his return to the Salinity Laboratory at the war's end, he completed his work on the pressure-membrane apparatus. After satisfying himself that the sausage casing which he initially used for the membrane was permeable to water but not to salts, he began to employ them widely. The design had one drawback. The membrane was subject to punctures in soil samples that contained small stones and eventually developed leaks. Eventually Zilla succeeded in making increasingly large clay plates, which could withstand pressure up to 15 000 cm (15 atmospheres), the presumed limit for plant growth. These plates were 30 cm in diameter, rendering them suitable for large samples or a large number of small samples. Replacement of the sausage by ceramics made the equipment usable by almost anyone. There rapidly developed such a demand for ceramic cups and plates that Richards set his sons up in a business known as the LARK Instrument Company. This stood for L.A. Richards' kids, who by participating in the manufacture of the ceramics helped pay their way through college.

In 1947 through 1949 Ren published the definitive papers on the measurement of soil moisture tension (capillary pressure). There has been little need for improvement in the equipment since that time. The demand for tensiometers and pressure plates continued to grow until Richards' assistant Percy Skaling set up his own business in order to meet this demand. Skaling was very careful to follow the design of the equipment exactly as described by Richards, except for very minor changes to facilitate manufacture. When colleagues would complain that their own equipment did not seem to work so well as Richards claimed, his response was always the same. He inquired if they had built it exactly as shown and had followed the experimental procedure exactly. They almost always confessed that they had attempted to make improvements in some way. He would tell them to go back into the laboratory and follow his design and instructions to the letter and then if it still did not work he would listen to them. They seldom came back, because he had already tried and discarded almost all variations in favor of the final design.

Ren learned much about soil water properties that is now taken for granted and is often rediscovered at regular intervals. More than any other individual L.A. Richards made soil physics an experimental science, in both field and laboratory. He made it possible to measure with considerable accuracy one of the most fundamental physical properties of soils,

the relation between the soil water content and the capillary (matric) potential. He showed that this relationship was not single-valued but, rather, exhibited hysteresis. The relationship was dependent upon the previous history of wetting and drying and depended in particular upon the number of such cycles a soil sample had been through. To facilitate field measurements of the soil moisture gradient, he developed a multiple tensiometer to measure the matric potential at several depths simultaneously.

Richards always had his eye on the main problem. In a laboratory filled with the best equipment available, most of it of his own design, and with a thorough understanding of the limitations of this equipment, he set himself the task of quantifying most of the known soil physical properties. He often suggested that, at best, we could only measure some of the soil properties on some soils, some of the time. It is still true.

A good estimate of the upper limit of available soil water for plant growth was needed in order to characterize the salinity of a soil. For about 48 h after irrigation the soil continues to drain, albeit ever more slowly. He had already devised the content of 'leaching requirement' for a soil, i.e., the water content at which the drainage rate is equal to the irrigation rate, thus establishing a steady-state soil water content. When the leaching requirement is met, the accumulation rate of salts in the soil no longer continues to increase and the salt content remains constant over the long term. Many attempts were made to determine the water content at which the leaching rate would equal the application rate but, since the drainage rate was more nearly a function of the conductivity of the soil than the water content, these attempts were not very satisfactory. Richards and his colleagues carried out a series of field experiments, and finally suggested that the field capacity concept, i.e., the upper limit of available water, was at best an elusive or approximate concept and had done more harm than good. He concluded that field profiles should be characterized by a curve rather than a single point.

At the other end of the plant growth continuum, the wilting point, the soil water content had been related to the cessation of the elongation of sunflower leaves. Richards developed a freezing-point method to provide a much simpler means of estimating the wilting point of soils.

From his observations of the decreasing rate of outflow from the pressure plate apparatus, Richards concluded that it should be possible to estimate the unsaturated conductivity of the soil as a function of soil water content. A number of individuals attempted this and eventually W.R. Gardner, with Richards' help, succeeded in achieving this by means of a series of successive linear approximations.

Upon completion of the development of the tensiometer and pressure plate apparatus, Richards returned to the thermojunction or thermocouple method of measuring the vapor pressure of soil moisture. This method had the advantage over the tensiometer in that it used the vapor pressure of the soil moisture to measure the total (matric plus osmotic) potential well beyond the tensiometer range and also make measurements on plant or other wet material. This was especially desirable for work in arid regions where the tensiometer was of limited utility. It also had the advantage over the pressure plate apparatus in that it was not limited to soil samples but could be used *in situ* on plants as well as soil. Richards had become familiar with early attempts with this method at Logan, where M.D. Thomas had published work on it. While a student, Ren's brother-in-law Leon Linford had assisted Thomas with the measurements. Later at Iowa State, Ren had begun work on the method but with limited success. At Riverside he devised a constant-temperature bath with a precision approaching one-ten-thousandth of a degree in the hope of measuring the difference between a small droplet of water and temperature of the soil moisture with which it had been equilibrated to within a few hundredths of a degree. Eventually it was with the availability of high-gain electronic amplifiers that Richards was able to eliminate the need for such stable temperatures and also the need for a very precise galvanometer. He finally succeeded in achieving a goal that he had been pursuing for some 30 years. While his colleagues were quick to apply this technology to several problems, Ren was reluctant to use it until it could be applied to plant leaves *in situ* and continued the pursuit of greater sophistication.

While Richards was working on the fundamental problem of measuring the vapor pressure of soil water, he always had two or three lesser problems which he addressed while he thought about the number-one problem. He devised the saturation extract and the portable electrical resistant bridge, which gave a very good approximation of soil salinity. He invented a device for the measurement of the modulus of rupture for soil samples. This was useful in estimating the effect of sodium salts on the crusting of soils, a common problem with alkali soils.

Richards guarded his research time jealously, but at the same time possessed a strong sense of loyalty to the laboratory that afforded him virtually unlimited freedom. Therefore, he accepted the responsibility of editing *Diagnosis and Improvement of Saline and Alkali Soil* better known as Handbook 60. Released in 1954, this handbook became the 'bible' of saline and alkali soils and proved to be one of the most popular and useful US Department of Agriculture

publications. It summarized the information and techniques developed during the first 15 years in the life of the Salinity Laboratory. It is a remarkable document that has stood the test of time, in that many of the concepts and equipment are still valid as a first approximation. A listing of the chapter headings gives a good idea of the scope of the Handbook: (1) Origin and nature of saline and alkali soils; (2) Determination of the properties of saline and alkali soils; (3) Improvement and management of arid and semiarid regions in relation to salinity and alkalinity; (4) Plant response and crop selection for saline and alkali soils; (5) Quality of irrigation water; (6) Methods of soil characterization; (7) Methods of plant culture and analysis; (8) Methods of analysis of irrigation waters.

Chapters 1, 2, 3, and 5, particularly, show Richards' influence. One particular example is the characterization of saline and alkali soils. The Russian school of soil science dealt with saline and alkali soils as though they always were and always would remain so. Handbook 60 assumed that soils could be 'reclaimed' if properly treated and managed. Although Richards was far from the first to espouse this idea, the Handbook gave this concept validity by showing how to achieve and sustain this goal The fact that it has not always been achieved is usually due to improper management, or some impediment to proper management, a *sine qua non*. To put it in Richards' own words, "the question for this laboratory to answer is whether or not irrigated agriculture can be made permanent."

Some 60 years later it would be difficult to improve upon the flow diagram that Richards and his collaborators devised for the diagnosis and reclamation of saline and alkali soils. This Handbook solidified Richards' position not only as the philosopher of the Laboratory but as the philosopher of soil physics. He understood what was important and how it differed from what was merely useful. While he worked from time to time on the latter, he never strayed for long from the former.

As significant as was the equation which bears his name, and the perfection of the tensiometer and invention of the pressure plate, they were eclipsed by Ren's collaboration with C.H. Wadleigh on the chapter in *Soil Water and Plant Growth*. For some time there had been an ongoing controversy between F.J. Veihmeyer at the University of California at Davis on the one side and many physicists and physiologists on the other. The debate centered on the issue of whether soil water was equally available to plants between the upper and lower limits of soil water. Veihmeyer said yes and almost everyone else said no. Wadleigh and Richards researched the American and English literature thoroughly and combined it with their own researches and concluded unequivocally that Veihmeyer was essentially wrong. The problem was in the measurement of soil water. In very sandy soils, especially in the case of fruit trees, Veihmeyer's approximation was not far wrong. In the general case, it was not good enough. Although Veihmeyer remained unconvinced, Richards' and Wadleigh's view has prevailed.

Richards has also had a profound effect upon the terminology of soil physics. In 1940 he published some comments on the permeability units for soils. In 1952 he chaired a committee of the Soil Science Society of America that recommended adoption of units of permeability and infiltration. In 1956, almost in passing, he suggested that the term 'soil suction' be abandoned in favor of the 'matric potential.' He eschewed the term 'capillary potential' because it implied that capillarity was the only physical mechanism influencing the water potential. Such was Richards' influence in the field of soil physics that the term was accepted almost by acclamation and with little public debate.

In 1951–1952 Richards was elected President of the Soil Science Society of America. In this position he succeeded in achieving a long-sought goal. He was able to persuade the Society, which heretofore had published only the papers presented at the Annual Meeting, to change from proceedings to a true journal which published peer-reviewed papers submitted by its members and meeting set standards. It was to be several years before it changed its name to the *Journal of the Soil Science Society of America*, but Richards had made this move inevitable. In 1965, Richards was elected President of the American Society of Agronomy. In this role he provided vigorous leadership to the society and organized and codified the business of the Society with a numerical scheme somewhat analogous to the system used by libraries. This system was sufficiently flexible and easily used so that it was adopted immediately by the Soil Science and the Crop Science Societies and continues in use long after only a few members know where it came from. Well beyond his official retirement Ren continued to work on an automated irrigation system and numerous related patentable devices.

See also: **Tensiometry**; **Water Content and Potential, Measurement**

Further Reading

Blair GY, Richards LA, and Campbell RB (1950) Rate of elongation of sunflower plants and freezing point of soil moisture in relation to permanent wilt. *Soil Science* 70: 431–439.

Buckingham E (1907) *The Mechanics of Soil Moisture*. US Department of Agriculture Division of Soils Bulletin 10.
Gardner W (1920) The capillary potential and its relation to soil moisture constants. *Soil Science* 10: 357–359.
Gardner WR (1956) Calculation of capillary conductivity from pressure plate outflow data. *Soil Science Society of America Proceedings* 29: 317–320.
Gardner WR and Ehlig CF (1962) Impedance to water movement in soil and plant. *Science* 138: 532–533.
Gardner WR and Ehlig CF (1965) Physical aspects of internal water relations of plant leaves. *Plant Physiology* 40: 705–710.
Ivie JO and Richards LA (1957) A meter for recording slow liquid flow. *Review of Scientific Instruments* 8: 86–89.
Klute A and Richards LA (1962) Effect of temperature on relative vapor pressure in soil: apparatus and preliminary measurements. *Soil Science* 93: 391–396.
Neal OR, Richards LA, and Russell MB (1938) Observations on moisture conditions in lysimeters. *Soil Science Society of America Proceedings* 2: 35–44.
Richards LA (1931) Capillary conduction of liquids through porous medium. *Physics* 1: 315–333.
Richards LA (1938) Wet thermojunctions for measuring relative humidity. *Iowa Academy of Science Proceedings* 45: 175–178.
Richards LA (1940) Concerning permeability units for soils. *Soil Science Society of America Proceedings* 5: 49–53.
Richards LA (1941) A pressure membrane apparatus for soil solution. *Soil Science* 51: 377–386.
Richards LA (1942) Soil moisture tensiometer materials and construction. *Soil Science* 53: 241–248.
Richards LA (1948) Porous plate apparatus for measuring moisture rentionion and transmission by soil. *Soil Science* 66: 105–110.
Richards LA (1949) Methods of measuring soil moisture tension. *Soil Science* 68: 95–112.
Richards LA (1953) Modulus of rupture as an index of surface crusting of soils. *Soil Science Society of America Proceedings* 17: 321–323.
Richards LA (1954) Multiple tensiometer for determining the vertical component of the hydraulic gradient in soil. *Soil Science Society of America Proceedings* 18: 7–9.
Richards LA (1955) Water content changes following the wetting of bare soil in the field. *Soil Science Society of Florida Proceedings* 15: 142–148.
Richards LA (1955) Portable conductivity bridge and cells for salinity appraisal. *Soil Science* 80: 55–59.
Richards LA (1956) A soil salinity sensor of improved design. *Soil Science Society of America Proceedings* 30: 333–337.
Richards LA and Neal OR (1937) Some field observations with tensiometers. *Soil Science Society of America Proceedings* 1: 71–91.
Richards LA and Ogata G (1961) A thermocouple for vapor pressure measurements in biological materials and soils systems at high humidity. *Science* 128: 1089–1090.
Richards LA and Ogata G (1961) Psychrometric measurements of soil samples equilibrated on pressure membranes. *Soil Science Society of America Proceedings* 25: 456–459.
Richards LA and Wadleigh CH (1952) Soil water and plant growth. In: Shaw BT (ed.) *Monograph,* vol. 2 *American Society of Agronomy: Soil Physical Conditions and Plant Growth*, pp. 75–251. New York: Academic Press.
Richards LA, Russell MB, and Neal OR (1938) Further developments on apparatus for field moisture studies. *Soil Science Society of America Proceedings* 2: 55–64.
Richards LA, Neal OR, and Russell MB (1939) Observations on moisture conditions in lysimeters. II. *Soil Science Society of America Proceedings* 4: 55–59.
Richards LA, Committee on Terminology (1952) Report of subcommittee on permeability and infiltration. *Soil Science Society of America Proceedings* 16: 85–88.
Russell MB and Richards LA (1938) Heat of wetting of soils. *Iowa Academy of Science Proceedings* 45: 179–185.
United States Salinity Laboratory Staff (1954) In: Richards LA (ed.) *Diagnosis and Improvement of Saline and Alkali Soils*. USDA Handbook 60.
Veihmeyer FJ and Hendrickson AH (1933) *Some Plant and Soil-Moisture Relations*. American Soil Survey Association Bulletin 15.

ROOT ARCHITECTURE AND GROWTH

L E Jackson, University of California–Davis, Davis, CA, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Root architecture and root growth are important for exploration of soil for water and nutrients, as the acquisition of these resources drives plant growth. There are a myriad of root forms, ranging from woody roots to root hairs, and many of their various functions still remain unknown. Several important concepts, however, govern the development of roots and their direct and indirect effects on plant growth: (1) roots are highly plastic, i.e., architecture and growth are highly responsive to environmental conditions and therefore can optimize resource acquisition; (2) tradeoffs exist in root form and function. The

costs of producing roots can yield benefits in terms of the acquisition of soil water and nutrients, and ultimately plant growth and fitness; (3) enormous variation exists among plant species in their root responses to abiotic stress; (4) genetic analysis has revealed that root architectural traits are often multigene or quantitative traits; and (5) roots affect attributes of plant communities and ecosystems, due to their effects on resource acquisition, growth, productivity, water budgets, and nutrient cycling.

Description of Roots and Root Architecture

Root systems are composed of the primary root that originates as part of the developing embryo in the seed, postembryonic, shoot-borne roots, and lateral roots that emerge from all root types. In the monocotyledonous plants such as grasses, seminal root axes (also called primary axes) grow out of the developing embryo, while nodal axes (also called secondary, adventitious, crown, or brace roots) subsequently arise from the shoot and gradually replace the seminal axes. In contrast, for dicotyledonous plants, the entire root system originates as a single root in the embryo, from which laterals arise, and which can later become a taproot, and/or thicken and become woody owing to cambial activity. Root hairs are extensions of epidermal cells and vary in length from 0.2 to 2 mm, depending on the species. They are continually being produced near the elongating root tip and are short-lived. Rhizomes, which occur in both types of plants, are underground, elongate stems that can be distinguished from roots by the presence of leaves, leaf scars, or buds.

Root anatomy changes along the growing root. The root tip consists of four overlapping zones (Figure 1): the root cap that protects the meristem, the meristem itself, a region of elongation, and a region of maturation, where root hairs emerge and vascular tissues form. A cross section of the mature root also has several distinctive zones. The epidermis is the outer layer of the root, with the protective exodermis immediately underneath it. These layers surround the cortex, which stores nutrients. The endodermis is the innermost layer of the cortex. The Casparian strip is a waxy substance that is present in endodermal cell walls. It forces water and nutrients to pass through the protoplasts of the endodermal cells, presumably to increase the selective uptake of nutrients and thus regulate ion absorption and accumulation. Within the vascular cylinder of the root, the xylem consists of

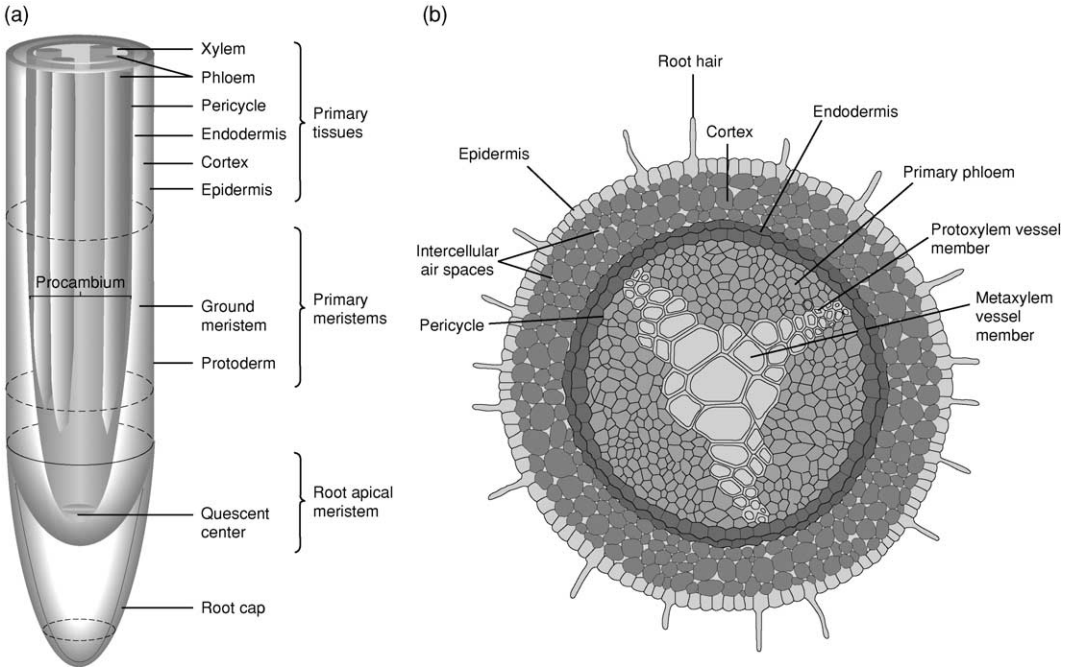


Figure 1 Longitudinal view of a developing root (left) and cross section of a maturing root (right), showing the different tissues, meristems, and differentiation of xylem and phloem elements. The outermost layer is the epidermis and root hairs, and the protective exodermis. In the cortex, there are intercellular air spaces between the cells. The endodermis and pericycle perform specialized functions related to uptake and differentiation, respectively, and surround the vascular cylinder, which contains the xylem and phloem. (Reproduced with permission from Rost TL, Barbour MG, Stocking CR, and Murphy TM (1998) *Plant Biology*. Belmont, CA: Wadsworth Publishing.)

hollow tracheids, which are important for water transport, and phloem elements, which are food-conducting tissue. Within the xylem, protoxylem is the first formed element of the primary xylem. Metaxylem is the part of the primary xylem that differentiates after the protoxylem and before the secondary xylem if any of the latter is formed in a given taxon. Phloem differentiates in the areas between the columns of xylem cells. The pericyle is the outermost boundary of the vascular cylinder. Its long-lived cells are the sites where woody secondary growth originates, as does the bark of woody roots; also lateral roots are initiated here and then grow outward through the cortex and epidermis.

Formation of lateral roots, branching to form clusters of fine roots, differentiation of root hairs, mechanics of soil penetration, and changing root distribution with depth are among the root architectural characteristics that affect nutrient and water acquisition from the soil. In addition, other processes such as nutrient absorption kinetics, root exudates, and microbial activity in the rhizosphere affect nutrient availability and therefore plant uptake of nutrients from the soil. (*See* **Nitrogen in Soils:** Plant Uptake.)

Use of topological models for the description of root architecture has proven to be a useful way to quantify the branching structure of root systems (Figure 2). Topological models consider the number of links and their arrangement: interior links are internodes between two branching points, and external links originate at a branching point and end in a meristem. The number, length, and spatial distribution of internal and external links can be markedly different between even closely related species and can change with developmental age and environmental conditions. In wheat, root axes are produced synchronously with development of specific leaves, and, in other species, vascular connections direct the nutrient and carbon flow between particular shoots and roots. Thus, root topology can be related to shoot development and physiology. Combining topological analysis with measures of root length and allocation, in relation to shoot biomass, provides a robust way of describing the architecture, growth, and response to the environment of a root system (Figure 3).

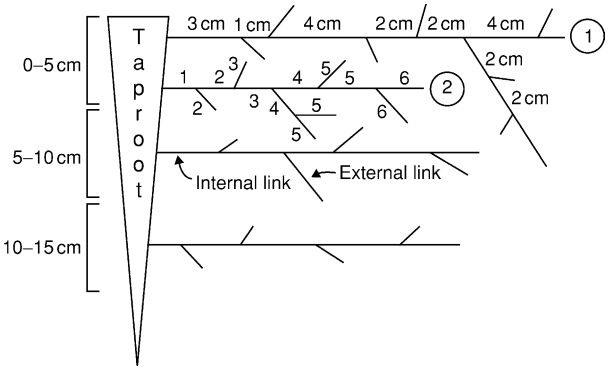


Figure 2 Schematic of the root system of a dicotyledonous species with a taproot, illustrating root topology as described by Fitter. Interior links are between two branching points, and external links originate at a branching point and end in a meristem. The magnitude of the root system or lateral axis is the number of external links (7 for lateral 2), the altitude is the number of links in the longest unique path from the base link to an external link (6 for lateral 2). The numbering on lateral 2 shows all pathlengths. Measurements of length and biomass of the root axes enhance the descriptive information obtained from the topological indices, which theoretically are related to nutrient uptake and growth strategy.

Roots branch in response to endogenous and exogenous factors. The plant hormone auxin has a key role in the initiation of lateral root primordia. It can accumulate at the root apex, which usually exerts some degree of apical dominance, in the developing lateral root primordial, or both. Cytokinins, another type of plant hormone, inhibit lateral root formation. Low nutrient levels, especially of nitrogen, can act as a stress signal that decreases cytokinin content, and this is associated with higher root-biomass allocation. The ratio of auxin and cytokinin concentrations is thought to be important for the regulation of developmental steps for the complete formation of lateral roots. Ethylene, a plant hormone that is involved in root response to stress, can either enhance or repress root growth, depending on its concentration. Plant hormones clearly have important effects on root branching, but the complexity of their interactions is still a major unknown in plant physiology.

Lateral root formation and the relative growth rate of roots are stimulated by nitrate, ammonium, and phosphate, such that highly branched roots and high root-length density are found in enriched soil patches. For immobile ions, root proliferation can increase nutrient uptake from such patches. In contrast, a mobile ion such as nitrate can be readily captured by a root system with low root-length density. It has been hypothesized, however, that localized root proliferation in response to nitrate may be a mechanism by which competing plants can quickly usurp a limiting resource. Water uptake is also enhanced by increased root surface area, even though its transport may be limited by the hydraulic conductivity of the xylem elements in some circumstances.

Tradeoffs in Root Form

Construction and maintenance costs of roots must be considered in the assessment of root architectural patterns. An optimization approach suggests that

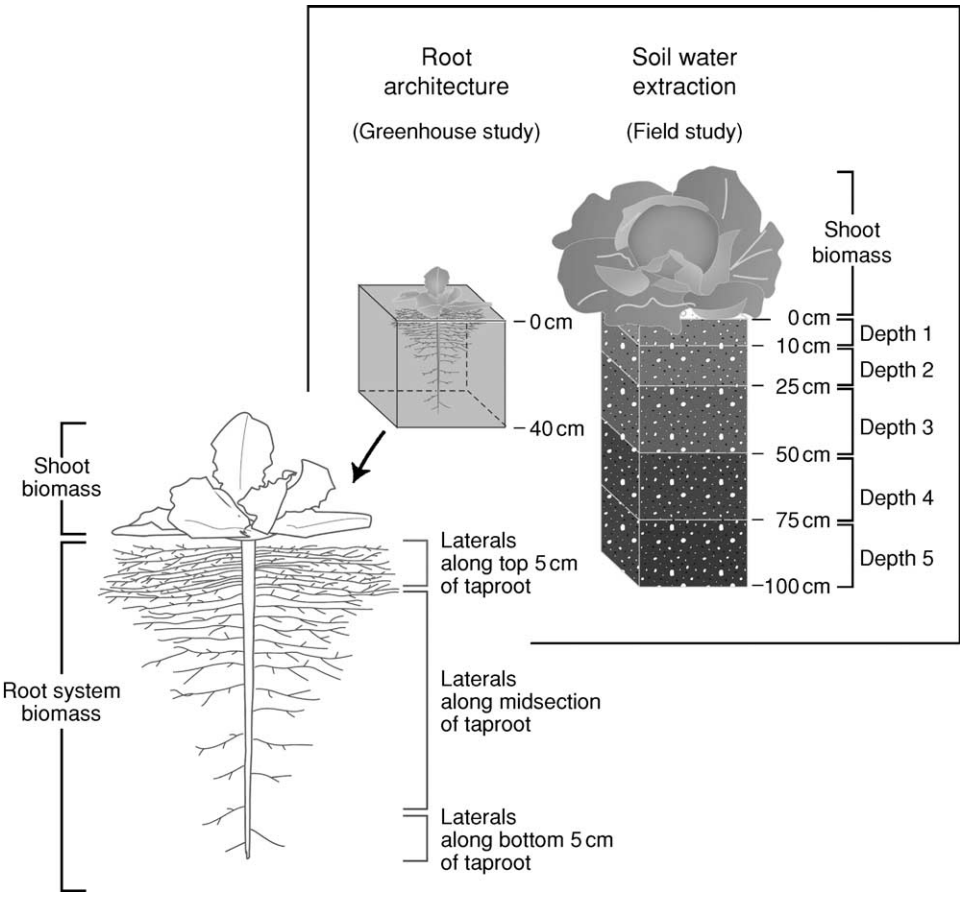


Figure 3 A strategy for examining relationships between root structure and function. In a greenhouse study in pots, root architectural traits were measured, while in a field study, water extraction was measured at different depth increments. When closely related lines of lettuce were tested with both methods, genotypes with higher numbers of laterals near the tip of the taproot or with longer taproots tended to extract more deep water from soil (traits were scaled for biomass), and this resulted in identification of quantitative trait loci (QTL) that could be genetically mapped. (Reproduced with permission from Johnson WC, Jackson LE, Ochoa O, van Wijk R, Peleman J, St. Clair DA, and Michelmore RW (2000) Lettuce, a shallow-rooted crop, and *Lactuca serriola*, its wild progenitor, differ at QTL determining root architecture and deep water extraction. *Theoretical and Applied Genetics* 101: 1066–1073.)

total plant growth will be largest when the root system maximizes nutrient and water acquisition per unit resource supplied by the shoot. Allocation of large amounts of carbohydrates and nutrients to roots may detract from shoot growth, leaf area, and photosynthetic carbon gain, thereby reducing total plant growth, especially if the increased root allocation does not achieve increased rates of nutrient uptake. Under low nutrient availability, root diameter often decreases, and thinner roots are thought to increase the capacity for exploration of soil per unit carbon expenditure than would thicker roots. Roots produced in enriched soil patches can be short-lived, eliminating their maintenance costs when soil depletion causes nutrient acquisition rates to decrease. In contrast, carbon allocation to long-lived storage roots and woody roots, which have high construction costs, limits shoot growth and thus leaf area and the potential for photosynthetic carbon gain; yet long-term beneficial effects for nutrient and water uptake may outweigh these initially negative effects on total growth.

Root Traits for Stress Tolerance

Plants cope with environmental stresses such as drought, soil compaction, salinity, nutrient limitation and imbalances, acid soil, heavy metal toxicity, and waterlogging. Each stress is mediated in part through the root system. Some plant species show highly specific adaptations, involving in many cases root architecture and root growth, to these abiotic stresses, along with certain physiological mechanisms.

Drought tolerance is partially mediated by root characteristics. Some plants have deep root systems that access water that would otherwise be lost to drainage. Another way for roots to exploit deep moisture is hydraulic lift, whereby deep moisture is

transferred at night to upper soil layers with higher root-length density for extraction during the following day. Low root hydraulic conductance is another trait that enables drought avoidance by decreasing the rate at which water is transported to and used by the shoot in dry environments with deeply stored soil water, e.g., Mediterranean-type environments. By restricting water use during the early part of the growing season via low transport rates, more water is available during the latter period of the year when rainfall is absent or minimal. Delayed maturation of the metaxylem vessels in developing roots is one mechanism by which low root hydraulic conductance is achieved. Another example of root-mediated drought tolerance is found in the developmental plasticity of cactus roots, where roots in moist microsites have less lignification and suberization, and higher hydraulic conductivity than roots in dry microsites, thereby apparently maximizing the potential for water uptake from wet patches and minimizing water loss from roots in dry patches.

As soil dries, hydraulic conductivity between the root and soil decreases. Shrinkage of roots also decreases the contact between root and soil. Low transpiration rates ensue, further causing the movement of liquid water by mass flow in the soil to decline. Roots grow through and away from dry sites and toward wetter patches in soil. The root cap is probably the site where this hydrosensing occurs.

Roots in contact with dry soil can be the source of inhibitors of shoot growth and stomatal opening, which in turn limits water loss from the plant. Hydraulic signals that elicit physiological responses to drought move from the roots to the shoots in the transpiration stream, and the hormone abscisic acid (ABA) and root-derived cytokinins appear to play a role in this signaling process. ABA modulates the expression of many genes and physiological processes that protect cells from excessive water loss (*See* **Water-Use Efficiency**), and it also appears to be involved in increasing root growth and consequently root-to-shoot ratio, which would increase water uptake from deeper, moister layers of the soil profile and reduce dehydration when upper layers of the soil profile dry out. Experiments have compared plants with root systems split between well-watered and dry zones of soil, i.e., partial root-zone drying, and a well-watered control. With partial root-zone drying, there can be an increase in water-use efficiency (WUE), along with a decrease in stomatal conductance and total vegetative growth, but little effect on fruit production, presumably because the developing fruit creates a strong sink for carbohydrate partitioning. The mechanisms by which root signals are transported and operate on various plant organs still remains largely unknown, but, even so, partial root-drying may be a potential method to limit water use and still maintain high yields in the commercial production of fruit-bearing crops such as grapes.

In environments where drought is normally a temporary event, maintaining viable roots under drought stress may be advantageous, especially for perennials that benefit from the higher production and maintenance costs of long-lived roots over a period of years. In species from deserts and Mediterranean-type environments, however, where plants intermittently experience severe and long-term soil drought, especially near the soil surface, fine roots are often shed when soil dries, and new root production can occur rapidly after rewetting by rainfall. Many crop species also shed roots in dry surface soil.

Compaction, like soil drying, increases the mechanical impedance to root penetration of soil. In soils with high bulk density, roots become clumped into small pores, since larger pores are scarce, and this limits the soil volume for nutrient and water uptake. While growing through soil, roots must overcome radial, axial, and frictional forces, and there is typically a negative correlation between root elongation rate and penetrometer resistance. The osmotic potential of root cells increases in impeded root tips, possibly as a mechanism to increase cell turgor and soil penetrability. Root impedance by mechanical forces in soil also triggers root signals that reduce leaf expansion and transpiration rates. Reduced leaf-expansive growth and shoot water-loss induced by mechanical stress, even under conditions of high nutrient and water supply, may be responses that prepare the plant for the high probability that these resources may soon become scarce as a result of reduced root growth.

Salinity inhibits root-elongation growth and development in nonhalophytes, and this, in turn, indirectly limits the ability of roots to explore soil for nutrients and water, which decreases the growth rate of the plant. Direct effects of salinity on roots include the osmotic stress from low soil-water potential that can induce both the biosynthesis of costly organic osmolytes to regulate osmotic adjustment in the root and the accumulation of specific ions in the root tissue which can lead to induced toxicities and deficiencies. Salt tolerance is achieved by maintaining a ratio of low levels of Na^+ and high levels of K^+ in the cytoplasm. Elevated Ca^{2+} concentrations in the soil medium can increase the root elongation rate in some plant taxa in saline soils, possibly because there is less displacement of membrane-associated Ca^{2+} by Na^+ than in less salt-tolerant genotypes.

Nutrient limitations are well known to be mediated by continuous root growth into unexplored soil, as well as by high total root surface area, both of which increase the capacity of the root system to take up nutrients. Nutrient bioavailability is mediated by root interception, mass flow, and diffusion (**Figure 4**). As roots grow, they displace soil and intercept available nutrients, for example, adsorbed on to soil particles. This process of root interception, however, is thought to provide only a small proportion of the plant's total nutrient uptake. Mass flow is the transfer of water and dissolved nutrients to the root surface and is driven by transpiration. Rates of diffusion, i.e., transport in soil along the concentration gradient, depend on the mobility and effective diffusion coefficients of specific ions, soil replenishment, and plant uptake. When transpiration rates are high, large quantities of water and dissolved solutes travel to the root via mass flow, but, when soils dry, diffusion of nutrients becomes the main way by which roots access nutrients.

Both old and newer regions of the root are important for nutrient absorption. Even though older parts of the root system are more suberized and may be surrounded by depletion zones, they have been shown to be capable of high potentials for nutrient uptake in laboratory experiments. Since they constitute a large proportion of the root surface area of the plant, they are probably important for a considerable proportion of total nutrient uptake. The highest rates of nutrient uptake per unit root length, however, are found in the region just behind the root apex, where root hairs are most abundant.

Nutrient-poor soils elicit several changes in root morphology, as well as physiology, such as increased capacity for nutrient uptake (*See* **Nitrogen in Soils: Plant Uptake**) but the degree of these responses can be very different among plant taxa. Higher root-to-shoot ratios and lower total root biomass typically occur with nutrient deprivation.

Phosphate (P), an immobile nutrient, does not move readily to the root by either mass flow or diffusion. P-deficient soils tend to cause large changes in a number of root characteristics. Under low-phosphate conditions, roots become thinner and the elongation of laterals increases, especially those at higher orders, i.e., near the meristematic tips of each lateral root axis, and leaf expansion and surface area decline. Also, more root hairs are produced and they have faster elongation rates. Root hairs have their greatest effect on absorption of ions that diffuse slowly in soil. Expansion of root surface area via increased production of root hairs is much less costly than a greater investment in roots.

Acid soils (*See* **Acid Rain and Soil Acidification; Acidity**) decrease root growth, partly due to direct effects of low pH, and also due to the higher availability and toxicity of heavy metals, such as aluminum. Aluminum toxicity decreases both root elongation and cell division, and root cells become shorter and wider, giving roots a stubby appearance. The root tips are the primary sites of aluminum toxicity. In many tropical soils that experience low pH and aluminum toxicity, P is either deficient or is readily fixed to soil particles. Thus, the low root-length density of plants in response to aluminum toxicity is an important limitation to P uptake and total plant growth. Other heavy metals, such as cadmium, copper, and zinc, also decrease root elongation, and root dry mass to a lesser extent, but the mechanisms for these responses are not clearly understood.

In waterlogged or flooded soils, air is displaced from soil pore spaces, and various degrees of oxygen

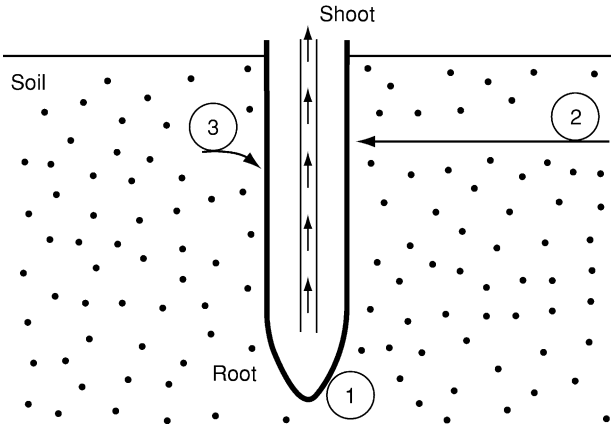


Figure 4 Schematic of the movement of minerals in soil to the root surface: 1, With root interception, the soil volume is displaced by the root volume; 2, With mass flow, the bulk soil solution is transported along the water potential gradient, which is driven by transpiration; 3, With diffusion, nutrient ions are transported along the concentration gradient. (Reproduced with permission from Marschner H (1995) *Mineral Nutrition of Higher Plants*. San Diego, CA: Academic Press.)

depletion typically occur in heterogeneous microsites in the soil. Flooding of a previously well aerated plant can cause immediate cessation of root growth, root respiration, and root death, partly due to the production of toxic substances such as nitrite or acidification of the root cells due to lactic acid produced during anaerobic fermentation. In some plant species after waterlogging, adventitious roots with well-developed aerenchyma, large air spaces in the root cortex, can emerge from the base of the stem, or existing roots can develop aerenchyma (Figure 5). The formation of aerenchyma under hypoxic conditions is associated with the production and accumulation of the plant hormone ethylene. Aerenchyma increases the root porosity and the long-distance transport of oxygen to apical zones of the root system. Pressurized gas transport, driven by temperature gradients between aboveground aerated parts of the plant and submerged root systems, also exists in some flooding-tolerant species.

Mycorrhizal symbioses help overcome some types of stress. Mycorrhizal colonization increases the surface area of the root system by extending hyphae beyond the root surface and consequently increases nutrient uptake, especially of ions with low mobility in soil, e.g., phosphorus, zinc, and copper. Mycorrhizae increase the distance of nutrient exploration away from the root and are important in sites with patchy distribution of nutrients. Water uptake is thought to be increased for some plant species. Arbuscular mycorrhizae (AM) are endophytes with intracellular colonization by Zygomycete fungi, whereas ectomycorrhizae are either Ascomycete or Basidiomycete fungi that form a mantle or sheath enclosing the rootlet. Dependence on AM tends to be highest in plants with coarse, fibrous root systems with few root hairs. Many herbaceous species with more fibrous root systems and high specific root length, i.e., thinner roots, are facultative mycotrophs that become less mycorrhizal with increased phosphorus

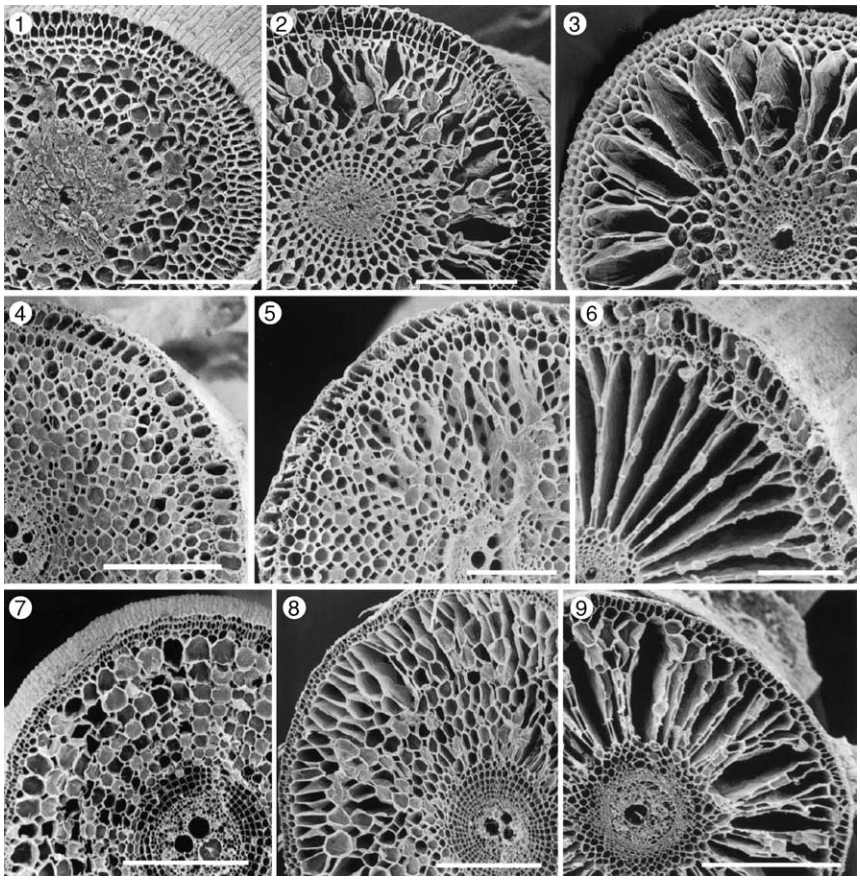


Figure 5 Scanning electron micrographs of transverse sections of roots of wetland plant species at differing distances from the root tip. Micrographs 1–3 show *Sagittaria lancifolia* at 5, 23, and 45 mm from the root tip, respectively. Micrographs 4–6 show *Thalia geniculata* roots at 2, 8, and 32 mm, respectively. Micrographs 7–9 show *Pontederia cordata* 6, 31, and 58 mm, respectively. The development of cortical air spaces (aerenchyma) in roots is an adaptive response to flooding. (Reproduced with permission from Longstreth DJ and Borkhsenious ON (2000) Root cell ultrastructure in developing aerenchyma tissue of three wetland species. *Annals of Botany* 86: 641–646.)

availability. Several studies indicate that root branching and lateral root formation increase with AM colonization. In some plant species, a decrease in specific root length (centimeters per gram of dry weight root) occurs in AM versus noncolonized roots, i.e., mycorrhizal roots are thicker. For ectomycorrhizal roots, the bulky tissue of the mycorrhizal mantle surrounding the roots serves a role in nutrient storage and in root-fungus contact, but does not greatly increase the surface area for nutrient uptake. Instead the extraradical mycelium that extends from the mantle and their rhizomorphs, conglomerations of hyphae with distinct structural shapes, perform this function. Ectomycorrhizal colonization can increase the number of root tips on laterals. Both AM and ectomycorrhizal colonization frequently result in allocation of a lower proportion of total plant weight to root dry weight.

Genetic Analysis of Root Traits

As with many aspects of plant growth and development, root architecture and morphological traits are usually under the control of a complex gene system in which several genes, rather than a single gene, operate together to create a given phenotype. Multigene control of root traits and root responses make the study of root genetics difficult. Also, roots are hard to study due to the large numbers of plants needed for genetic studies, especially in soil in field experiments, where their phenotypic evaluation is most valid.

For rice, a large effort has been invested in understanding root responses to drought. Genetic variation among rice cultivars exists for root thickness, total root length, and root penetration ability, i.e., for proliferation in a soil hardpan. Using quantitative trait locus (QTL) analysis that relates phenotypic variation to specific genetic regions in a genetically segregating population, these traits have been genetically mapped, and several chromosomal regions have been found to affect each of the traits. Like root traits, drought avoidance traits in rice, e.g., relative water content and the extent of leaf rolling that limits transpirational surface area, also give many QTLs that differ between sites and between years at the same site. Yet there is little co-location on the genetic map of drought avoidance and root morphological QTLs, implying that root traits do not seem to be contributing consistently to drought avoidance in rice. Whether this is true or an artifact due to screening techniques remains to be seen. A more straightforward outcome has been achieved for lettuce. In lines derived from a cross between wild and cultivated lettuce, QTLs for deep water extraction and deep rooting traits co-localize (Figure 3), and genetic markers for these QTL regions are now being introgressed into cultivated lettuce, with the hope of creating a deeper-rooted crop that requires less irrigation input.

Single-gene traits for root morphological traits have been identified in some plants, but are best studied in the tiny crucifer *Arabidopsis thaliana*, a current model species used by plant geneticists. Mutants for traits such as failure to initiate lateral roots, auxin-related lateral root growth, and variable lateral root initiation in response to the soil environment exist in this species. Transgenic *Arabidopsis* plants have also been used to demonstrate the function of a specific gene controlling the proliferation of lateral roots in nitrate-rich patches. In normal, nontransgenic plants, the expression of this gene is induced rapidly when nitrate is supplied, but, in transgenic plants in which the gene is repressed, roots no longer proliferate in response to nitrate. The gene may therefore be a key factor in determining the developmental plasticity of roots in response to an environmental cue.

The strong effect of the abiotic and biotic environment in the soil on root development and branching patterns is obviously important for understanding the inheritance of root traits and their adaptation to specific habitats, and for crop improvement for more sustainable use of water and nutrients. For many crops, extensive genotypic variation exists for root-system traits among cultivars, but the exact genetic basis of these traits is difficult to ascertain and therefore hard to implement in crop-breeding programs. Molecular techniques including QTL analysis and marker-assisted selection will help to overcome these problems.

Root Architectural Effects on Plant Communities and Ecosystem Processes

Plant species and ecotypes are adapted to the environments in which they grow. Plant communities of nutrient-poor sites are typically composed of plants with slow growth rates, high nutrient concentrations in tissue, and high root-to-shoot ratios. Even with a high nutrient supply, they have low rates of nutrient uptake and growth. In contrast, natural vegetation on nutrient-rich sites, e.g., many disturbed sites, often consists of species with the capacity for high rates of nutrient uptake and growth, rapid response to increased availability of soil nutrients, and low root-to-shoot ratio. These traits ultimately contribute to higher fitness even though the specific respiratory costs for nutrient uptake may be high. Despite these generalities, different ecophysiological strategies can usually be found within the same plant community. For example, in the vegetation of the arid Great Basin in

the USA, sagebrush and one prevalent perennial grass showed no root response to enriched fertile microsites, while another perennial grass species exploited P via rapid root proliferation in the patches.

Spatial and temporal patterns of changes in root growth, architecture, and distribution affect nutrient and water availability in soil, as well as retention and loss of these resources to the atmosphere, runoff, and leaching. These relationships are best understood for crop systems. Like most annuals, crop plants proliferate roots in the top 0.5 m of soil until subject to drought, at which time roots grow increasingly deeper and rely on deeply stored water. Many annual species reduce or terminate root growth at anthesis, and reproductive growth and seed production often rely heavily on remobilization of nutrients and carbon from roots and other vegetative parts. To examine these effects on water and nutrient budgets, models increasingly rely on information on root growth, activity, and phenology, rather than using solely a 'root-sink' term to explain unaccounted-for fluxes in the soil. These include accounting for matric and osmotic potential effects on water and passive nutrient uptake, root-age effects on root water and nutrient uptake activity, as well as the influence of nutrient deficiency and ion toxicity on root growth.

Root distribution often reflects changes in soil strength and drainage channels in the soil profile. For *Eucalyptus* trees, fine roots concentrate in the surface soil and in zones of preferential infiltration, causing rapid drying and nutrient depletion in these areas. During the wet season, some roots grow into compacted zones, utilizing seasonally available water when moisture reduces soil strength in these areas. Thus, both spatial and temporal aspects of root growth help to reduce nutrient loss caused by deep drainage.

Competition between plants in the same plant community affects root distribution, water uptake, and WUE. The water source used by the roots of different species can be traced with deuterium (^2H) and oxygen-18 ($^{18}O_2$), which are stable isotopes that vary among rainwater, soil water, and groundwater. Carbon isotope composition ($\delta^{13}C$) is related to plant intrinsic WUE, which is the ratio between CO_2 assimilation and leaf water vapor conductance. Studies that use isotopic techniques along with measurements of root biomass and leaf gas exchange combine an assessment of root distribution, and short- and long-term physiological activity. In tree–grass mixtures, for example, grasses exert a negative effect on tree root growth, but usually not vice versa. Drought, however, typically has a more negative effect on grass than on trees, which have more conservative leaf gas exchange characteristics and deeper roots, which utilize deeply stored water. Sorting out these complex, competitive relationships has been possible with a combination of techniques, including stable isotopes. An example is tracking the natural abundance of the stable isotope ^2H in soil water, which varies with depth in dry environments, and in plant xylem water, which gives a weighted average of all soil water acquired by functional roots (**Figure 6**).

Successful plant invasion is due in part to root traits, along with a suite of related ecophysiological characteristics. For salt cedar (*Tamarix* sp.), a non-native phreatophytic species in riparian areas of

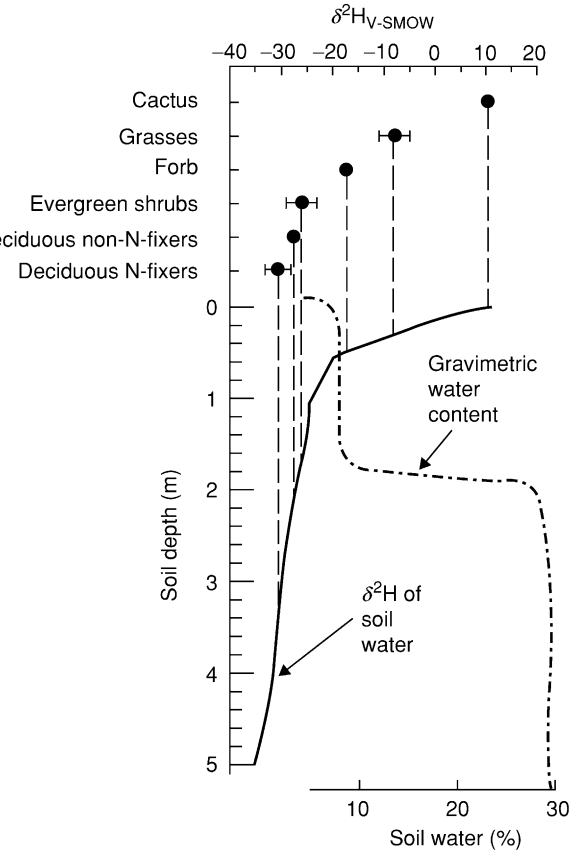


Figure 6 Hydrogen isotope content, δ^2H, in soil water and plant xylem water in a subtropical savanna in Texas, USA. *Prosopis glandulosa*, honey mesquite, is the dominant woody plant. In the soil, δ^2H decreases exponentially with depth, and gravimetric moisture is higher below approximately 2 m depth. Shallow-rooted plant functional groups (grasses, forbs, and cacti) obtain their water from the upper 1 m of the profile, based on their δ^2H. More negative δ^2H values of plant and soil water indicate that tree and shrub roots access water in the zone of high gravimetric content, while roots of grasses, cacti, and forbs do not. Hydrogen isotope composition, δ^2H or $^2H/^1H$, is expressed relative to a conventional standard, V-SMOW (Vienna Standard Mean Ocean Water). (Reproduced with permission from Boutton TW, Archer SR, and Midwood AJ (1999) Stable isotopes in ecosystem science: structure, function and dynamics of a subtropical savanna. *Rapid Communications in Mass Spectroscopy* 12: 1263–1277.)

Table 1 Global land area (10^6 km^2), total root biomass (10^9 Mg), total fine-root biomass (10^9 Mg), live fine-root biomass (10^9 Mg), live fine-root length, and live fine-root surface area for different biomes, as determined by calculations and conversion factors

Biome	*Land area (10^6 km^2)*	*Total root biomass (10^9 Mg)*	*Total fine root biomass (10^9 Mg)*	*Live fine root biomass (10^9 Mg)*	*Global fine root length (10^{15} km, live)*	*Global fine root surface area (10^7 km^2, live)*
Tropical rainforest	17.0	83	9.7	5.7	0.069	12.6
Tropical seasonal forest	7.5	31	4.3	2.1	0.026	4.7
Temperate evergreen forest	5.0	22	4.1	2.5	0.031	5.5
Temperate deciduous forest	7.0	29	5.6	3.1	0.038	6.9
Boreal forest	12.0	35	7.2	2.8	0.031	5.5
Woodland and shrubland	8.5	41	4.4	2.4	0.099	9.9
Savanna	15.0	21	14.9	7.7	0.91	63.8
Temperate grassland	9.0	14	13.6	8.5	1.01	71.2
Tundra/alpine	8.0	10	7.7	2.7	0.059	4.2
Desert	18.0	6.6	4.9	2.3	0.072	9.9
Cultivated	14.0	2.1	2.1	1.1	0.13	5.0
Totals	121.0	294.7	78.5	40.9	2.475	199.2

Reproduced with permission from Jackson RB, Mooney HA, and Schulze E-D (1997) A global budget for fine root biomass, surface area, and nutrient contents. *Proceedings of the National Academy of Sciences of the USA* 94: 7362–7366.

North American deserts, greater root allocation and deeper roots allow higher transpiration and photosynthesis rates under dry, hot conditions than can be achieved by native riparian species. These root traits, along with prolific seed production and vegetative reproduction, help explain the widespread invasive success of salt cedar in this environment.

At a regional scale, root distribution can affect hydrology. An example is the replacement of deep-rooted perennial trees and shrubs, up to 20 m deep, by shallow-rooted pasture and crop plants in the seasonally dry Murray–Darling Basin in Australia. Historically, the native vegetation would have relied on deep soil recharge and groundwater, while the shallow-rooted agroecosystems today occupy only the top meter of soil and utilize far less water than the forest and woodland communities, whose evergreen canopies had higher evapotranspiration rates. The current outcome of the vegetation transition is groundwater recharge and rising water tables, and since the groundwater is now usually saline, secondary salinization.

Improved understanding of the factors controlling root construction, biomass allocation, and distribution will improve ecological, hydrological, and global climate models. Roots are central to many ecosystem processes, including water absorption, which affects groundwater and atmospheric water fluxes, and indirectly influences water availability for soil microbial processes that control carbon dioxide, NO_x, and methane fluxes, as well as nutrient concentrations and fates, which also play a role in these processes. Fine-root biomass and surface area, and distribution through the profile are particularly important for these processes due to their relatively short lifespan, low C:N ratio, which enhances turnover rates, and their impact on nutrient absorption and thus primary production. A recent analysis indicates that substantial differences exist between biomes in total fine root biomass and depth distribution (**Table 1**), and that approximately 33% of global net primary production is used for fine-root production. These findings indicate the need for more physiological and ecological information on the importance of roots in terms of total plant productivity, carbon turnover, and long-term storage of carbon in perennial roots in different ecosystems.

See also: **Nitrogen in Soils:** Plant Uptake; **Water-Use Efficiency**

Further Reading

Araus JL, Slafer GA, Reynolds MP, and Royo C (2002) Plant breeding and drought in C_3 cereals: what should we breed for? *Annals of Botany* 89: 925–940.

Boutton TW, Archer SR, and Midwood AJ (1999) Stable isotopes in ecosystem science: structure, function and dynamics of a subtropical savanna. *Rapid Communications in Mass Spectrometry* 12: 1263–1277.

Drew MC, Jackson MB, and Giffard S (1979) Ethylene-promoted adventitious rooting and development of cortical air spaces (aerenchyma) in roots may be adaptive responses to flooding in *Zea mays* L. *Planta* 147: 83–89.

Eissenstat DM (1997) Tradeoffs in root form and function. In: Jackson LE (ed.) *Ecology in Agriculture*, pp. 173–199. San Diego, CA: Academic Press.

Esau K (1977) *Anatomy of Seed Plants*, 2nd edn. New York: John Wiley.

Fitter AH (1987) Functional significance of root morphology and root system architecture. In: Fitter AH (ed.) *Ecological Interactions in Soil*, pp. 87–106. Oxford, UK: Blackwell Scientific Publications.

Jackson RB, Mooney HA, and Schulze E-D (1997) A global budget for fine root biomass, surface area, and nutrient contents. *Proceedings of the National Academy of Sciences of the USA* 94: 7362–7366.

Johnson WC, Jackson LE, Ochoa O, van Wijk R, Peleman J, St. Clair DA, and Michelmore RW (2000) Lettuce, a shallow-rooted crop, and *Lactuca serriola*, its wild progenitor, differ at QTL determining root architecture and deep water extraction. *Theoretical and Applied Genetics* 101: 1066–1073.

Laclau J-P, Arnaud M, Bouillet J-P, and Ranger J (2001) Spatial distribution of *Eucalyptus* roots in a deep sandy soil in the Congo: relationships with the ability of the stand to take up water and nutrients. *Tree Physiology* 21: 129–136.

Lambers H, Chapin FS III, and Pons TL (1998) *Plant Physiological Ecology*. New York, NY: Springer-Verlag.

Marschner H (1995) *Mineral Nutrition of Higher Plants*. San Diego, CA: Academic Press.

O'Toole JC and Bland WL (1987) Genotypic variation in crop plant root systems. *Advances in Agronomy* 41: 91–145.

Price AH, Cairns JE, Horton P, Jones HG, and Griffiths H (2002) Linking drought-resistance mechanisms to drought avoidance in upland rice using a QTL approach: progress and new opportunities to integrate stomatal and mesophyll responses. *Journal of Experimental Botany* 53: 989–1004.

Rost TL, Barbour MG, Stocking CR, and Murphy TM (1998) *Plant Biology*. Belmont, CA: Wadsworth Publishing.

Smith SE and Read DJ (1997) *Mycorrhizal Symbiosis*. San Diego, CA: Academic Press.

Waisel Y, Eshel A, and Kafkafi U (2002) *Plant Roots: The Hidden Half*. New York: Marcel Dekker.

ROOT EXUDATES AND MICROORGANISMS

B-J Koo and D C Adriano, University of Georgia, Aiken, SC, USA
N S Bolan, Massey University, Palmerston North, New Zealand
C D Barton, University of Kentucky, Lexington, KY, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Plant roots influence the physical, chemical, and biological conditions of the soil in the rhizosphere. The biogeochemical reactions induced by microorganisms at the soil–root interface (i.e., rhizosphere) play an important role in the bioavailability of nutrients and metals to plants. This microenvironment is characterized by distinct physical, chemical, and biological conditions compared with the bulk soil, largely created by the plant roots and its microbial associations. Such associations can include nonsymbiotic and symbiotic organisms such as bacteria and mycorrhizal fungi. The microbial populations are an essential part of the rhizosphere and affect the rhizosphere soil by their various activities such as water and nutrient uptake, exudation, and biological transformations.

Organic acids, sugars, amino acids, lipids, coumarins, flavonoids, proteins, enzymes, aliphatics, and aromatics are examples of the primary substances found within this microzone. Among them, the organic acids have received considerable attention owing to their role in providing substrates for microbial metabolism and for serving as intermediates for biogeochemical reactions in soil.

Nutrients and metals are typically present in the soil solution at low concentrations and tend to form sparingly soluble minerals (except nitrogen, sulfur, and boron), or may be adsorbed to a solid phase through ion exchange, hydrogen bonding, or complexation. The extent to which they are transferred from the soil to the biota (i.e., microbes or plants) is dependent on the biogeochemical interactions and/or processes among the soil, plant roots, and microorganisms in the rhizosphere. At this interface, the presence of root exudates may influence chemical reaction kinetics within the soil environment and subsequently affect biological activities. As such, understanding the role of the rhizosphere on biogeochemical processes within the soil is essential for developing bioremediation technologies of inorganic and organic contaminants.

Rhizosphere Microzone

The rhizosphere, first described in 1904 by Lorentz Hiltner, has been the focus of intensive research for many years because of its importance in plant nutrition and pathogenesis. More recently, research on the rhizosphere has been directed toward its influence on controlling the persistence, mobility, and bioavailability of contaminants in soils.

The rhizosphere generally refers to the portion of soil found adjacent to the roots of living plants. The rhizosphere is subject to the influence of chemicals

excreted by roots of living plants and the microbial community in this microzone. Its domain varies for different plant species and for age and morphology of roots. Depending on plant species, the width of the rhizosphere zone has been shown to extend from 2 to 80 mm away from the root surface. The soil in the rhizosphere supports a typically diverse and densely populated microbial community and is subject to chemical transformations caused by the presence of root exudates and metabolites of microbial degradation.

The rhizosphere microzone is distinguished from the bulk soil zone, more commonly known as the edaphosphere, by enhanced microbial activity and increased concentration of root exudates. Nevertheless, it is rather difficult to separate this zone physically from the root surface or rhizoplane. The rhizosphere effect is expressed quantitatively as the ratio of the number or activity of microorganisms or level of root exudates in rhizosphere soil (R) to that in the edaphosphere soil (E), i.e., the R/E ratio. The R/E ratio for microorganisms and root exudates is often found to range, respectively, from 2 to 20 and from 5 to 100, indicating enhanced microbial activity in the rhizosphere.

Root exudates are known as one of the most important factors affecting microbial growth in the rhizosphere. For example, cell numbers are several orders higher in the root zone than in the background soil lacking plants. The microbial community is more diverse, active, and synergistic than in nonrhizosphere soil. Root exudates can also selectively influence the growth of bacteria and fungi that colonize the rhizosphere by serving as selective growth substrates for soil microorganisms. For example, *Pseudomonad* species are particularly stimulated by enhanced carboxylic acid production from root exudation. Rhizosphere microbial communities can also vary in structure and species composition in different root locations or in response to other environmental factors, including soil type, plant species, or nutritional status.

Microorganisms and Structure of Microbial Communities in the Rhizosphere

Rhizosphere communities are comprised of many cultivable and noncultivable microorganisms. Mycorrhizal fungi that live in symbiosis with higher plants are essential living components of the rhizosphere. Bacteria are also prominent components and may be active participants in the community or simply adventitious members. Therefore, the rhizosphere is the ideal system to study when seeking an understanding of the critical functional interdependence and metabolic capabilities of complex microbial communities containing both prokaryotic and eukaryotic members.

The rhizosphere of land plants hosts a large, diverse community of microorganisms that interact with each other and with the plant's roots. The mycorrhizal fungi are very important, ubiquitous groups within this community. The vegetative mycelium of mycorrhizal fungi and the root tips of more than 90% of land plants, including angiosperms, gymnosperms, and pteridophytes, form a mutually supportive symbiosis. This novel composite organ, the mycorrhiza, is the site of nutrient and carbon (C) transfer between the symbionts. The various mycorrhizal associations allow plants to colonize and grow efficiently in suboptimal environments and, in most terrestrial ecosystems mycorrhizae, not roots, are the chief organs of nutrient uptake by land plants. The arbuscular endomycorrhizae (AM) and the ectomycorrhizae (EM) are the most important mycorrhizal types.

The AM symbiosis is probably the most ancient of all such root-microbe symbioses. Approximately 80% of the higher green plants, including numerous crop species, are symbiotic with AM glomalean fungi, a relatively small group of filamentous fungi with aseptate hyphae and multiple nuclei within distinct cell wall boundaries. In soil, this small monophyletic group of aseptate filamentous fungi forms a very fragile mycelium that can produce large asexual spores filled with numerous nuclei. The extraradical hyphae of most AM fungi improve the uptake of immobile resources, notably inorganic phosphate, since they can acquire these beyond the depletion zone surrounding a root. They can also interfere with pathogens, increasing micronutrient uptake and altering drought-resistance. In some circumstances, AM fungi benefit plants by increasing the uptake of the more mobile inorganic- or even organic-nitrogen (N) compounds.

Only approximately 3% of perennial woody plants develop ectomycorrhizas. Ectomycorrhizal trees, *Betulaceae*, *Fagaceae*, *Salicaceae*, and *Pinaceae*, dominate forest ecosystems in boreal, temperate, and Mediterranean regions. The mycobionts span all of the phyla of true fungi, the *Zygomycota*, *Ascomycota*, and *Basidiomycota*, with at least 15 families within the *Basidiomycota* and *Ascomycota*. The mycelia of EM fungi are usually much more structurally differentiated than those of AM fungi. Many ectomycorrhizal fungi seem to be long-lived and form large genets in stable, established forest ecosystems. The huge diversity of fungal partners optimizes the uptake and mobilization of mineral resources distributed in discrete substrates in soil, litter, and within other soil organisms. EM fungi have better access than plant

roots or AM fungi to organic phosphorus (P) and N compounds and seem well adapted to colonize the heterogeneous organic soil and litter layers of forests.

Both the AM and EM associations form an extraradical- and an intraradical-hyphal network that are functionally different entities. The extraradical hyphae colonize and exploit the soil volume to scavenge efficiently essential micro- and macronutrients. The intraradical hyphal network provides essential nutrient resources (e.g., phosphate and amino acids) to the plant. These nutrient contributions are reciprocated by the host's provision in its roots of a stable carbohydrate-rich niche for the fungal partner, making the relationship one of mutual symbiosis.

In the rhizosphere of plants, mycorrhizal fungi are always accompanied by another important group of microorganisms, the bacteria that also prosper in the organic-rich environment (mostly sugars, amino acids, and organic acids) released from the roots and mycorrhizal fungi. Among them, there is a large category of growth-promoting rhizobacteria that influence plant growth directly or indirectly by releasing a variety of compounds, from mineral nutrients to phytohormones and antimicrobial compounds. The establishment of mycorrhizae on roots is affected by the microbial populations of the rhizosphere, and especially by some bacteria, which can have either a positive or a negative effect on mycorrhiza formation. Plant-associated rhizosphere bacteria also have an important role in establishing and improving plant growth on different soil types, since they affect the availability of essential elements and provide plants with extra N.

Composition of Root Exudates

Root exudates refer to a suite of substances in the rhizosphere that are secreted by the roots of living plants and microbially modified products of these substances. They consist of low-molecular-weight organic compounds that are freely and passively released root-cell material and mucilage associated with roots. Root exudates are a complex mixture of soluble organic substances, which may contain sugars, amino acids, organic acids, enzymes, and other substances (Table 1). The rhizosphere also contains cell lysates, which are released when root cortex cells are ruptured by mechanical damage, microbial grazing, fungal infection, or the emergence of lateral roots that break through the root epidermis. In addition to these substances, rhizodeposition includes the secretion of enzymes, such as acid phosphatase secreted in response to P deficiency.

Although the identified components differ in molecular weight, amount released, and biochemical role, the dominant exudate components have consistently been shown to be organic acids, sugars, and amino acids. Typically, the total organic acid concentration in roots is approximately 10–20 $\mathrm{mmol\,l^{-1}}$ (i.e., 1–4% of total dry weight) for corn (*Zea mays* L.), with the other main organic solutes, namely amino acids (10–20 $\mathrm{mmol\,l^{-1}}$) and sugars (90 $\mathrm{mmol\,l^{-1}}$), present in root cells. Higher concentrations are generally found in rhizosphere soil than in bulk soil. The carboxylic acids that play a significant role in complexing metal ions in soils are not utilized as rapidly as carbohydrates by the soil microorganisms. It appears that microorganisms preferentially utilize the carbohydrates in root exudates. Consequently, organic and amino acids may be the prevalent chemical substances in root exudates.

Typical organic acids found in root exudates are malic, malonic, acetic, citric, fumaric, succinic, lactic, tartaric, oxalic, and others. Organic acids of root exudates seem to have a major role in solubilization of mineral nutrients. Malic and citric acids are major root exudate components for some plant species under aseptic conditions. For example, in corn grown under

Table 1 Composition of root exudates

Compositions	*Substances identified*
Organic acid	Acetic, butyric, citric, glutaric, lactic, maleic, malic, malonic, oxalic, propionic, pyruvic, succinic, tartaric, valeric
Amino acid and amide	α-Alanine, β-alanine, arginine, asparagine, aspartic acid, cystine/cysteine, glutamine, glycine, histidine, lysine, methionine, phenylalanine, proline, serine/homoserine
Enzyme	Amylase, invertase, phosphatase, protease, polygalacturonase
Growth factor	*p*-Amino benzoic acid, auxins, biotin, choline, inositol, *n*-methyl nicotinic acid, niacin, pantothenate, pyridoxine, thiamine
Phenolic acid and coumarin	Caffeic acid, cinnamic acid, coumarin, ferulic acid, salycilic acid, syringic acid, vanillic acid
Sugar	Arabinose, fructose, fucose, galactose, glucose, maltose, oligosaccharide, raffinose, rhamnose, ribose, sucrose, xylose
Others	Nucleotide, flavonone, fatty acids, proteins, sterols, lipids, aliphatics, aromatics, carbohydrates

sterile conditions, exudation of malic and citric acids increases significantly under nutrient deficiency.

Compared with the organic acids, amino acids appear to play a less significant role in complexing or mobilizing plant nutrients. The proteinaceous amino acids are not readily detected in root exudate solutions. The nonproteinaceous amino acids (phytosiderophores) in the rhizosphere have been shown to enhance the mobility of plant micronutrients in soils.

Collection and Quantitation of Root Exudates

Upon reaching a predetermined growth stage, root exudates must be separated from the growing medium. If plants are grown in solution culture, aliquots of the nutrient solution can be purified and concentrated for the assay. The plants can also be transferred from a nutrient solution to another container and, upon a predetermined exposure time, root exudates are collected in water. When plants are grown in solid medium such as soil, root exudates are collected by first removing the plant from the soil then placing it in aerated and sterilized, double-distilled water for 2–6 hours. Subsequently, the exudates are separated and concentrated for analysis. Typically, solutions are filtered using 0.2- or 0.45-μm membrane filters that are then concentrated by rotary evaporation or by lyophilization. The collected material is stored at temperatures ranging from -20 to 4°C until analysis then diluted to a smaller volume for analysis.

Root exudates might be adsorbed on XAD-4, a porous polymer resin, which behaves in a way similar to C-18 high-pressure liquid chromatographic material when used in the reversed-phase mode. Root exudates trapped on the resin are extracted with either distilled water or methyl alcohol for lipids. The system has several advantages over previously used extraction procedures and is easily adaptable to large-scale recovery and extremely low-level concentrations of substances, as is very often the case with germination stimulants and other biologically active exudates.

Only a few attempts have been made to analyze comprehensively the chemical composition of root exudates. This is due to the necessity of employing a wide array of analytical methods for characterizing the different exudate components. Prior knowledge of the presence of substances is essential for selecting the methodologies for proper separation, identification, and quantification of substances from the mixtures. In addition, concentrations of substances in root exudates may be low, further complicating their detection and quantification.

The chemical behaviors of low-molecular-weight organic acids commonly found in root exudates are similar, making their separation in a chromatographic column a challenging analytical problem. A gas chromatographic (GC) approach based on differences in the chemical structure of organic acids has also been utilized. GC with mass spectrometry (GC-MS) analysis has been used to quantify the changes of quantities of root exudate in relation to colonization of roots by mycorrhizal fungi in which the concentrated root exudate solution is injected directly into the instrument. To enhance the resolution, root exudates can be separated into cationic, anionic, and neutral fractions prior to analysis by GC-MS using ion exchange resins. Compounds are then identified by comparing the retention time and mass spectra to respective reference compounds. Organic acids in the root exudates have also been determined by ion chromatography (IC) operating at room temperature. The separation is carried out on an IonPac AS 11 or AS 6 column (250×4 mm) protected by an IonPac AG 11 guard column (50×4 mm).

Factors Affecting Root Exudates

The quantity and composition of root exudates are affected by environmental factors (Table 2). Acidification, chelation, precipitation, and redox reactions in the soil can change the nature and composition of root exudates. For example, nutritional stress

Table 2 Influence of environmental factors on root exudation

Factors	*Effect*
Age of the root	Roots in the initial stages secrete exudates more frequently
Light intensity	High light intensity increases exudation
Microorganism	Presence of microorganisms increases exudation
NH_4^+/NO_3^-	Roots under NH_4^+ have lower organic acid concentrations
Nutrient availability	Nutrient deficiencies increase exudation
Oxygen status	Compositions are different under aerobic and anaerobic conditions
Plant species	Different plants and cultivars have different exudate compositions
Salinity	More organic and amino acids are produced under low salinity
Soil moisture	Relieving drought stress increases exudation
Soil pH	Acidification changes composition of exudates
Soil temperature	High temperature stimulates exudation
Soil texture	Sandy substrate produces greater amounts of exudates
Stress condition	Stress changes the composition of exudates

experienced by plants, such as deficiencies of P, zinc (Zn), or iron (Fe), may increase root exudation. Plant roots alter the redox potential in the rhizosphere soil directly by excreting CO_2, and indirectly through the supply of readily available carbon sources for enhanced microbial respiration. In general, the redox potential is lower in the rhizosphere than in the bulk soil. The decrease in redox potential is likely to have important consequences in the redox reaction of contaminants such as arsenic (As), selenium (Se), Fe, manganese (Mn), and chromium (Cr).

Temperature and photoperiod have profound effects on the quality and quantity of root exudates because they affect photosynthesis, translocation, and respiration in plants. Increase in exudation at high temperature has been reported for many crops. The rate of exudation of certain organic acids by plants, while in vegetative and reproductive stages, may increase with the elevation of temperature and the extension of photoperiod, the mean rate being two or more times higher than the minimum exudation at low temperature with shorter photoperiod.

The imbalance in plant absorption of cation and anion affects the concentration of organic acids in the rhizosphere. In situations where roots take up cations in excess, particularly K^+, negative charges that are provided by organic acids, such as malic, malonic, oxalic, and citric acids, are required to maintain an electrostatic balance. Roots grown in NH_4^+-fertilized fields have lower concentrations of the organic acids than those grown in NO_3^--fertilized fields.

Role of Root Exudates on Plant Nutrient and Metal Dynamics

The enhanced activity of microorganisms, including mycorrhizal fungi, in the rhizosphere is important in relation to the bioavailability and mobility of nutrient ions, metals, and organic contaminants. For example, it has often been shown that the dissolution and release of water-insoluble fertilizer materials such as elemental sulfur (S) and apatite phosphate rocks are enhanced in the presence of plants; this has been related to the increased activity of S-oxidizing microorganisms and the release of organic acids. The enhanced microbial activity is likely to reduce the mobility of nonreactive ions such as nitrate and sulfate through microbial immobilization.

Interactions of Fe, Mn, copper (Cu), and Zn, and the effect of organic acids on P dissolution in soils have been researched. In addition to their role in plant nutrition, root exudates may affect vital soil processes, such as the soil organic matter transformation, cycling of nutrients in soils, and development of stable soil structure. Root exudates (e.g., polysaccharide gum) enhance stabilization of soil structure in the rhizosphere by increasing the strength of bonds between soil particles.

Organic acids such as oxalic, lactic, and succinic exuded by roots are effective in releasing metals from sparingly soluble mineral forms by forming stable chelate complexes with the metal ions. The extent of complexation depends on the organic acids involved (number of carboxylic groups), the concentration and type of metal, and the pH of the soil. Organic acids with only one carboxyl group, such as acetic, formic, and lactic acids, have very little metal-complexing ability. Based on the stability constants summarized in Table 3, malic and oxalic acids have high affinities for metals. These are the most frequently found organic acids in soils. However, this does not always entail efficient complexation at all soil pH levels. For example, using Geochem-PC, a descendant of the multipurpose chemical speciation program and the PC-compatible version of Geochem, the complexation of Fe by malic and oxalic acids is shown to be highly dependent upon soil pH, with little or no complexation at high soil pH.

Biogeochemical reactions occurring in the rhizosphere play an important role in the bioavailability of metals. These reactions can be heavily influenced, notably by the release of protons and organic acids from the roots. The acidifying protons can affect the chemical speciation and mobility of micronutrient metals and heavy metals/metalloids, possibly favoring the formation of ionic species in the rhizosphere. Additionally, organic acids such as those mentioned earlier had the ability to complex metals, with the resultant metalorgano complexes possessing greater mobility. Thus the combined effects of the protons and organic acids may result in chemical speciation and much-enhanced mobility of the metals that could facilitate their uptake by the root or leaching.

Phosphate dissolution rate can be greatly accelerated in soil in the presence of organic acids of root exudates, depending on soil type and concentration of organic acids. There are several possible mechanisms by which P release can occur. The first mechanism involves competitive sorption reactions (direct ligand exchange) in which exuded organic acids compete with P for common sorption sites on soil particles, thus releasing P from the solid phases. For example, citrate may be able to desorb phosphate from sesquioxide surfaces by anion exchange. The second mechanism could involve the modification of the surface characteristics in which organic ligand sorption modifies the surface charge of P sorbents. Anionic organic acids may be sorbed by minerals in the rhizosphere, which may lead to an increase in the net negative charge of clays. Along with this,

Table 3 Stability constants (K) of organic acid–metal complexes

		Log K[b]										
Metal ion	*M:L:H ratio*[a]	*Acetic*	*Butyric*	*Glutaric*	*Lactic*	*Maleic*	*Oxalic*	*Propionic*	*Pyruvic*	*Succinic*	*Tartaric*	*Valeric*
H^+	1:1:0	4.76	4.82	5.43	3.86	6.33	4.27	4.87	2.55	5.64	4.37	4.78
	1:2:0	-	-	-	-	-	-	-	-	-	-	-
Cd^{2+}	1:1:0	1.93	1.25[f]	1.60[c,f]	1.70	2.40[l]	3.89[d,e]	1.19[g]	-	2.72	-	-
	1:2:0	3.15	1.98[f]	-	2.10[g]	3.60[j]	-	1.86[g]	-	-	-	-
Cr^{2+}	1:1:0	1.25[j]	-	-	-	-	3.85[l]	-	-	-	-	-
	1:2:0	2.15[j]	-	-	-	-	6.81[l]	-	-	-	-	-
Cu^{2+}	1:1:0	2.22	2.14	3.16	3.02	3.40[l]	6.23	2.22	2.20	3.28	2.90[g]	2.08
	1:2:0	3.63	2.98[f]	-	4.84	4.90[k]	10.3	3.00[f]	4.90	-	-	-
Ni^{2+}	1:1:0	1.43	2.14	1.60[l]	2.22	2.00[l]	5.16	0.78[g]	1.12[h]	1.60[l]	2.06[g]	-
	1:2:0	-	2.98[f]	-	2.76[g]	-	-	1.26[f]	0.46[h]	-	-	-
Pb^{2+}	1:1:0	2.68	2.13[f]	2.48[c,f]	2.78	2.75[g]	4.91	2.08[f]	2.04[f]	5.64	2.60[g]	-
	1:2:0	4.08	3.74[f]	3.45[c,f]	2.88[g]	4.03[g]	6.76	4.15	3.40[f]	3.73[d]	4.00[g]	-
Zn^{2+}	1:1:0	1.57	0.98[f]	2.45	2.22	2.00[l]	4.87	0.85[g]	1.26[h]	2.52	3.82	-
	1:2:0	1.90[l]	1.65[f]	1.74[g]	3.75	-	7.65	1.23[g]	1.98[h]	2.00[g]	5.00	-

[a]M:H:L ratio indicates the stoichiometry of the complex; M, metal; L, ligand; H, proton.
[b]Temperature (°C) at which K was determined at 25°C unless indicated otherwise; and ionic strength (moles per liter) at which K was determined at zero unless indicated.
[c]30°C.
[d]18°C.
[e]3.0.
[f]2.0.
[g]1.0.
[h]0.5.
[i]0.3.
[j]0.2.
[k]0.16.
[l]0.1.

acidification due to proton exudation accelerates dissolution reaction affecting clay minerals. The third mechanism involves the complexation of metal ions in the solid, which constitute the exchange matrix holding the P. The action of organic chelates in solubilizing phosphates and phosphate minerals is attributed to the formation of complexes with calcium (Ca), Fe, or aluminum (Al), thereby releasing phosphate in water-soluble form. Solubilization of P compounds by organic acids is accomplished by the formation of complexes of organic acid ligands with P-associated metal ions such as Fe, Al, and Ca. The substitution of organic ligands for orthophosphates frees up P for plant uptake. This process depends on the number and position of carboxyl (-COOH) and phenolic (-OH) functional groups in the organic acids.

The pH of the rhizosphere is important in determining contaminant and nutrient mobilization and uptake. Root-induced pH change in the rhizosphere is a known phenomenon and has been shown to have an effect on the bioavailability or solubility of nutrients such as P, Fe, Mn, Zn, and Cu by rhizosphere acidification. Rhizosphere pH can be more than 2 units lower than the bulk soil pH. Several hypotheses have been postulated to explain the differential ability of plant species to affect rhizosphere pH. These include differences in root exudates, especially low-molecular-weight organic acids and the evolution of respiratory CO_2 from differential uptake in cations/anions.

Root Exudates and Biodegradation of Organic Contaminants

Rhizosphere degradation of organic compounds is often a key factor in successfully applying a phytoremediation-based strategy. Secretion of plant exudates not only stimulates overall microbial activity, but some induce specific bacterial- and fungal-degradation pathways. For instance, phenolic root exudates trigger cometabolism of chlorinated organic solvents, including trichloroethylene (TCE), by rhizosphere bacteria. The combined activity of mycorrhizal fungi and bacteria might help to overcome environmental stress imposed on plants by the presence of organic xenobiotics: consortia of mycorrhizal fungi and pseudomonads with toluate-degrading capacity improve the survival of mycorrhizal symbionts and, at the same time, promote toluate degradation in the rhizosphere. Some hydrophobic organic xenobiotics such as polyaromatic hydrocarbons (PAH) are

not taken up efficiently by plants. Phytoremediation of PAHs, therefore, strongly depends on rhizosphere processes such as enhanced degradation by bacteria or fungi.

Root exudates affect the bioavailability of nutrients by directly influencing the chemistry of the rhizosphere soils and indirectly enhancing microbial activity in the rhizosphere soils. Microorganisms utilize sugars, organic acids, amino acids, vitamins, and high-molecular-weight compounds exuded by roots as sources of C and substrates for growth. Microbial metabolites may in turn increase root exudation through stimulation of root activity and growth, thus increasing the surface area over which the exudation occurs. Similarly, Fe deficiency leads to the development of shifted bacterial communities that differ qualitatively and quantitatively in species composition.

Microbial degradation is the major process by which organic contaminant residues are removed from the soil. The rate of degradation of organic contaminants has often been found to be faster in the presence of growing plants. Figure 1 indicates the distribution of the water-soluble C, the microbial activity, and the 2,4-dichlorophenoxyacetic acid (2,4-D) residues in soil samples taken at different distances from the rhizoplane in a microcosm root container. There is an increase in the amount of soluble C in soil samples taken from close to the root surface. The increase in soluble carbon is related to the rhizodeposition of root exudates, which include low-molecular-weight organic acids, carbohydrates, and amino acids. There is a corresponding increase in microbial activity as measured by the amount of oxygen consumed by the microorganisms in this soil section. This enhanced microbial activity is related to an increase in the supply of organic substrate (i.e., root exudates) as a source of C and energy for microbial growth. There is a decrease in the amount of 2,4-D remaining in the soil, indicating that microbial degradation of pesticides occurs during plant growth. The amount of residual 2,4-D remaining in the pesticide cores is less in the soil section close to the root surface than in the other section. Since grass roots are unlikely to absorb considerable amounts of 2,4-D residues, by inference the low amount of 2,4-D residues close to the root core is attributed to the root-induced degradation of the pesticide. The enhanced degradation of pesticide in the soil section close to the root surface is related to the rhizosphere-induced cometabolism of the pesticide that requires the presence of a growth substrate other than the compound to be mineralized. Plant roots excrete a wide range of root exudates, and the availability of these exudates is considered to be the major cause for the existence of an enriched microbial population in the rhizosphere, resulting in accelerated degradation of the pesticide.

It is likely that an increase in the activity of pesticide-degrading microorganisms in the rhizosphere of certain plants is a mechanism by which plants are protected from the toxic effects of pesticides. Selective enrichment of microorganisms is likely to have a significant impact on the rhizoremediation (or

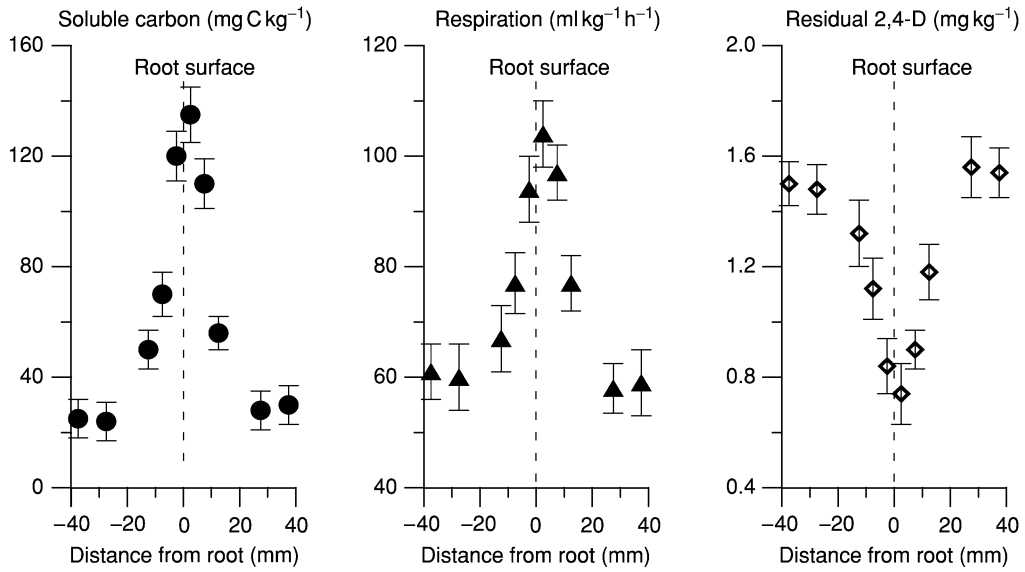


Figure 1 Distribution of soluble carbon, biological activity, and the residual 2,4-dichlorophenoxyacetic acid (2,4-D) in thin sections of soils in the pesticide cores.

more commonly, phytoremediation) of recalcitrant organic contaminants in soils. It has become apparent that the role of the rhizosphere in the natural attenuation of pollutants in soils is being recognized increasingly.

Summary

Plants and their associated microorganisms from complex communities, called rhizosphere communities, have diverse functions and capabilities that are important in many processes related to C management and recycling, biochemical cycling of nitrogen and P, biomass production and conversion to biofuel, and phytoremediation of heavy metals, radionuclides, and organic xenobiotics.

The relationships and interactions between and among microorganisms, plants, and water-borne chemicals are still not well-understood. However, many processes, such as the successful growth of plants, the optimization of biomass production, and carbon sequestration require that we acquire such understanding at the level of diversity, structure, functional interdependence, and capabilities of rhizosphere communities.

Acknowledgments

The authors wish to acknowledge the support provided by Savannah River Ecology Laboratory, University of Georgia, through US Department of Energy contract number DE-FC-09-96SR18546.

See also: **Bacteria:** Plant Growth-Promoting; **Mycorrhizal Fungi**; **Root Architecture and Growth**

Further Reading

Adriano DC (2001) *Trace Elements in Terrestrial Environments: Biogeochemistry, Bioavailability, and Risks of Metals*, 2nd edn. New York: Springer.

Curl EA and Truelove B (1986) *The Rhizosphere*. Berlin, Germany: Springer-Verlag.

Gobran GR, Wenzel WW, and Lombi E (2001) *Trace Elements in the Rhizosphere*. Boca Raton, FL: CRC Press.

Harley JL and Russell RS (1979) *The Soil–Root Interface*. London, UK: Academic Press.

Koo B-J (2001) *Assessing Bioavailability of Metals in Biosolid-Treated Soils: Root Exudates and Their Effects on Solubility of Metals*. PhD Dissertation. Riverside, CA: University of California Press.

Marschner H (1995) *Mineral Nutrition of Higher Plants*, 2nd edn. London, UK: Academic Press.

S

SALINATION PROCESSES

I Shainberg and G J Levy, The Volcani Center, Bet Dagan, Israel

© 2005, Elsevier Ltd. All Rights Reserved.

Origin and Distribution of Saline Soils

Excessive levels of salts occur in large areas of soil around the world and profoundly affect land use. Two major types of soil have been defined within this group, saline and sodic soils. The former refers to a soil that contains excess soluble salts that impair its productivity. The excess salts are comprised largely of the chloride and sulfate anions and sodium, magnesium, and calcium cations in the soil solution. Sodic soils contain excess sodium in the exchange complex that impairs their productivity. Sodic soils may or may not contain excess soluble salts and may contain free sodium carbonate.

Estimates of the areas of the globe covered by salt-affected soils are presented in Table 1. The salt content of soils above which plant growth is affected depends upon several factors, among which are the texture of the soil, the composition of the salt, and the species of the plant. The electrical conductivity (EC) of the saturation extract has been adopted by the US Salinity Laboratory as the preferred scale for general use in estimating soil salinity; and saline soils are defined as soils with EC exceeding $4\,\mathrm{dS\,m^{-1}}$. Based on this classification, the total area of saline soils on the globe is 351.5 millions hectares, with most of the saline soils located in Asia (194.9×10^6 ha) (Table 1). The decision regarding what level of exchangeable sodium in the soil constitutes an excessive degree of saturation is complicated by the fact that there is no sharp change in the properties of the soil as the degree of saturation with exchangeable sodium is increased, and that the effect of sodicity depends on the EC of the soil solution, the soil texture, the soil mineralogy, organic matter content, sesquioxide content, and the soil conditions (e.g., the wetting rate of the soil, the antecedent moisture content, and the aging effectl Thus, in the USA, and subsequently throughout much of the rest of the world, a value of exchangeable sodium percentage (ESP) greater than 15 is the criterion for separating sodic soils, whereas in Australia a value of ESP greater than 6 is used (*See* **Sodic Soils**). This explains the disproportiol large areas of sodic soils in Australia (Table 1). Whatever the definition of sodic soils, it is evident from Table 1 that the area of sodic soils in the world is approximately twice that of saline soils.

The original and direct source of the salts in soils are the primary minerals found in soils and in exposed rocks of the Earth's crust. During the process of chemical weathering, which involves hydrolysis, hydration, solution, oxidation, and carbonation, the salts are gradually released and made soluble. Although weathering of primary minerals is the indirect source of nearly all soluble salts, there are few instances where sufficient salts have accumulated from this source alone to form a saline soil. Saline soils usually occur in areas that receive salts from other locations, and water is the primary carrier. The ocean is the source of salts in low-lying soils along the margin of sea coasts. Sometimes salt is moved inland through the transportation of spray by winds. More commonly, however, the direct source of salts is surface water and groundwater. Water acts as a source of salts when used for irrigation and when groundwater rises close to the soil surface.

Saline soils occur for the most part in regions of arid or semiarid climate. In arid regions, leaching and transportation of soluble salts to the ocean are not as complete as in humid regions. Restricted drainage is a factor contributing to the salinization of soils and may involve the presence of a high groundwater table. Where there is a high water table, upward movement of saline groundwater and evaporation of surface water result in the formation of saline soil.

Properties of Saline and Sodic Soils

Properties of saline and sodic soils are determined by the composition of the soil solution and that of the

solid phase. The composition of the soil solution is determined by the following characteristics:

1. Total concentration of soluble salt, or 'salinity';
2. Concentration of sodium relative to other cations, or 'sodicity';
3. Anionic composition of the water, especially that of bicarbonate and carbonate;
4. Concentration of elements toxic to plants (e.g., boron, B).

Total Salt Concentration

The effect of salt on crop growth is largely of an osmotic nature. Osmotic pressure is a colligative property and is related to the total salt concentration rather than the concentration of individual ionic species. Total salt concentration is determined by total dissolved solids (TDS) in milligrams per liter of salts, by the ionic concentration in millimoles of charge per liter and by the electrical conductivity (EC) in decisiemens per meter. The relationship between salt concentration, either millimoles of charge per liter or milligrams per liter, and EC of various salt solutions is presented in Figure 1a and 1b, respectively. For a mixture of salts, the relationship between TDS and EC is given by:

$$\text{TDS (ppm)} = 640 \times \text{EC (dS/m)} \qquad [1]$$

Table 1 Global distribution of salt-affected soils

	Area (millions ha)		
Continent	*Saline*	*Sodic (alkali)*	*Total*
North America	6.2	9.6	15.8
Central America	2.0	–	2.0
South America	69.4	59.6	129.0
Africa	53.5	27.0	80.5
South Asia	83.3	1.8	85.1
North & Central Asia	91.6	120.1	211.7
Southeast Asia	20.0	–	20.0
Australasia	17.4	340.0	357.4
Europe	7.8	22.9	30.7
Total	351.5	581.0	932.2

Reproduced from Sumner ME, Rengasamy P, and Naidu R (1998) Sodic soils: a reappraisal. In: Sumner ME and Naidu R (eds) *Sodic Soils*, pp. 3–17. New York: Oxford University Press.

Sodicity and Sodium Adsorption Ratio

When soils are leached with salt solutions until equilibrium is attained, a linear relation is obtained between the ratio of the concentrations of exchangeable sodium and divalent cations (ESR, exchangeable sodium ratio) and the ratio of the concentration (in millimoles of charge per liter) of the sodium cation to the square root of the concentration of the divalent cations in the solution (SAR) (Figure 2).

$$\text{SAR} = [\text{Na}]/([\text{Ca} + \text{Mg}]/2)^{1/2} \qquad [2]$$

where [Na] and [Ca + Mg] are the concentration of the soluble ions in millimoles of charge per liter. Analysis of a large number of soil samples from the western USA has led to the empirical equation:

$$\text{ESR} = \text{Na}_{\text{exch}}/[\text{Ca} + \text{Mg}]_{\text{exch}} = -0.0126 + 0.01475 \qquad [3]$$

where Na_{exch} and $[\text{Ca} + \text{Mg}]_{\text{exch}}$ are the concentrations of the exchangeable ions, presented in Figure 2.

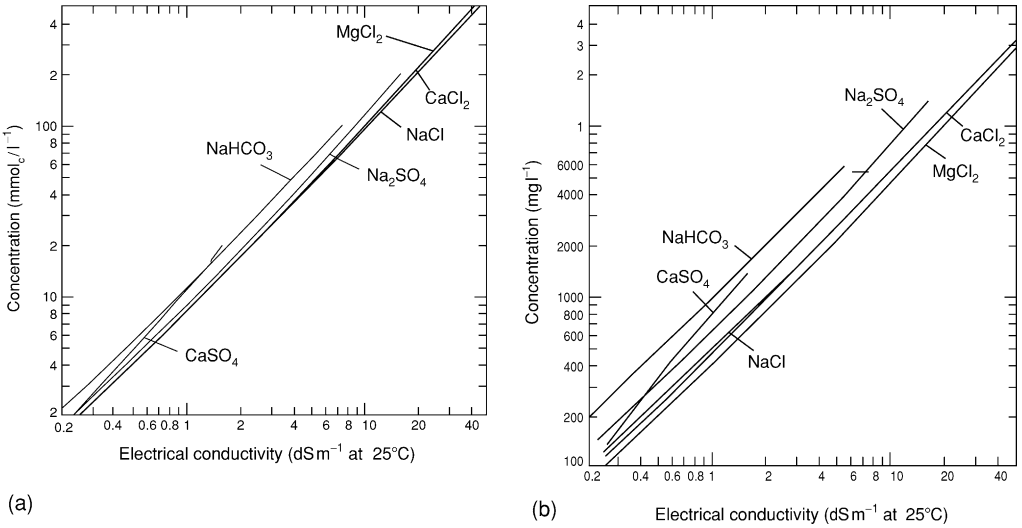


Figure 1 Relationship between salt concentration and electrical conductivity of various salt solutions: salt concentration in (a) millimoles of charge per liter and (b) milligrams per liter. Reproduced with permission from Shainberg I and Oster JD (1978) Quality of Irrigation Water. International Irrigation Information Center, Volcani Center, Israel.

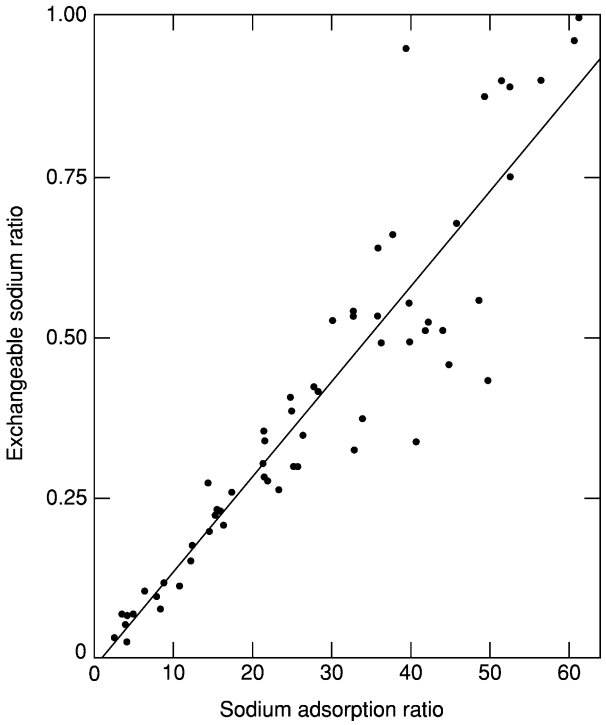


Figure 2 Exchangeable sodium ratio (ESR) of soil samples from the western USA as related to the sodium adsorption ratio (SAR) of soil extracts. Reproduced with permission from Shainberg I and Oster JD (1978) Quality of Irrigation Water. International Irrigation Information Center, Volcani Center, Israel.

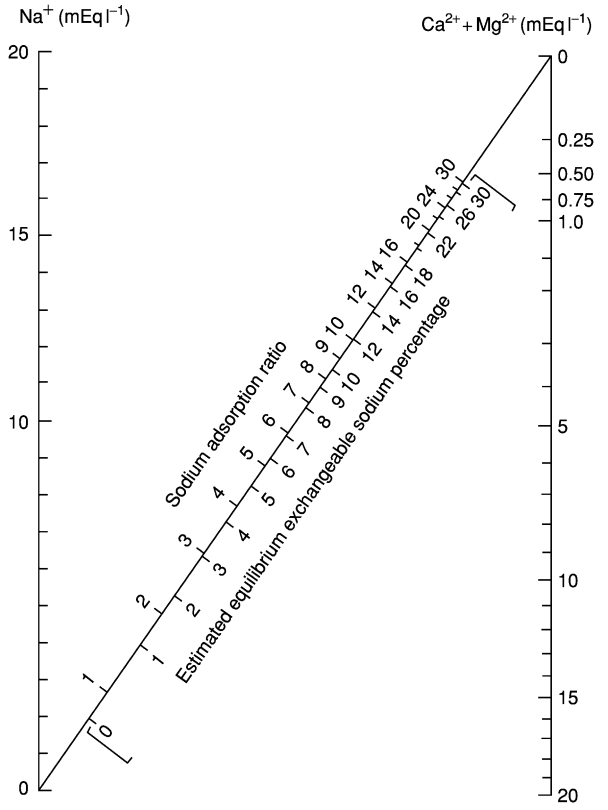


Figure 3 Nomogram for determining the SAR value of irrigation water and for estimating the corresponding ESP value of a soil equilibrated with this water. Courtesy of US Department of Agriculture, Washington.

A nomogram relating the ESP of soils to the SAR of the soil solution is presented in Figure 3.

Anionic Composition

A major factor affecting the final SAR value of soil water is the change in Ca and Mg concentration due to precipitation or dissolution of alkaline earth carbonates. In irrigation waters containing high concentrations of bicarbonate ions, there is a tendency for calcium and, to a lesser extent, magnesium to precipitate in the form of carbonate as the soil solution becomes more concentrated, thus leading to an increase in the SAR of the soil solution and consequently to an increase in the ESP of the soil. Bicarbonate concentration in underground water may be high because of the high partial pressure of CO_2. On exposure to the atmosphere, the excess CO_2 is released, while calcium and magnesium carbonates precipitate. Eaton assumed that all calcium and magnesium would precipitate as carbonates and proposed the concept of residual sodium carbonate (RSC) for evaluating high-carbonate water:

$$\mathrm{RSC} = (\mathrm{CO_3} + \mathrm{HCO_3}) - (\mathrm{Ca} + \mathrm{Mg}) \qquad [4]$$

where the concentration of the ions is expressed in millimoles of charge per liter. From the results of some limited leaching experiments, Eaton concluded that water with RSC greater than 2.5 is not suitable for irrigation. He defined the range of RSC between 1.25 and 2.5 as marginal, and less than 1.25 as safe. Although the RSC is still commonly reported in irrigation water analyses, its use is not justified. A better estimate of the precipitation tendency of carbonates is provided by the so-called saturation index. However, its application to soils leached with bicarbonate water is complicated and limited in its accuracy.

Concentration of Elements Toxic to Plants

Some soluble trace elements may have an inhibitory effect on plant growth. Elements in this category include B, lithium, selenium, and some heavy metals. In terms of saline soils, B is probably the most important. B is essential for plant growth but is exceedingly toxic at concentrations only slightly greater than optimum. B deficiency is found primarily in humid regions or in sandy soils, while toxicity occurs most frequently in arid and semiarid regions. Crops vary in their response to B. The threshold concentration for the most sensitive crop is $0.5\,\mathrm{mg\,l^{-1}}$ (lemon) and for the most tolerant is $6–10\,\mathrm{mg\,l^{-1}}$ (cotton). Since

an appreciable proportion of the B added to soils in irrigation water is sorbed by the soil, and plants respond to the B in soil solution, then conditions affecting equilibrium between sorbed and soluble B are highly relevant to considerations of B toxicity. Owing to adsorption, irrigation water containing marginal levels of B may not be toxic immediately; but after prolonged irrigation with such water, soluble B levels equal or exceed those of irrigation water.

Response of Crops to Salinity

Since the publication of Handbook 60 by the US Salinity Laboratory Staff, the salt tolerance of crops has been reevaluated as a result of new data obtained under conditions of variable soil salinity. It is now thought that most plants are able to withstand higher salinities in the drainage water than previously reported.

Salinity affects plant growth by three mechanisms:

1. Osmotic effects limit the ability of plants to absorb water from the soil solution. The osmotic effect is demonstrated when plant growth is reduced in a similar way by iso-osmotic solutions, and when osmotic and soil-water matric potential have a similar and additive effect on plant growth;
2. Specific ion toxicity results from excessive concentrations of Na or Cl ions, an effect most commonly found in woody species;
3. Changes in soil physical and chemical properties can affect plant production.

Crop yield is not reduced until a threshold salinity level is exceeded. Beyond the threshold level, the yield decreases linearly with rising salinity (Figure 4). The salinity values at zero yields provide an estimate of maximum salinity that plants can tolerate. Plants can be divided into four groups (sensitive to tolerant) according to their response to salinity. Of the field crops, barley, cotton, and sugar beet belong to the tolerant group, wheat and soybean belong to the moderately tolerant group, peanut and rice belong to the moderately sensitive group, and beans and cowpeas belong to the salt-sensitive group.

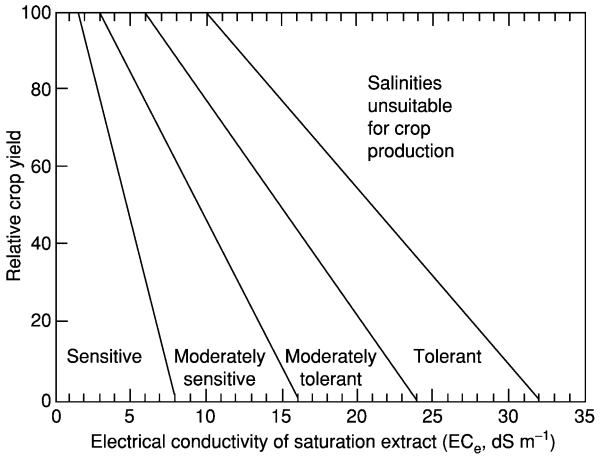


Figure 4 Relative crop yield as a function of the salinity of the soil saturation extract. Reproduced with permission from Shainberg I and Oster JD (1978) Quality of Irrigation Water. International Irrigation Information Center, Volcani Center, Israel.

Plant response to salinity depends also on soil, climate, and plant factors. Soil-water content and frequency of irrigation affect the tolerance of crops to soil salinity. Shortening the irrigation intervals minimizes the deleterious effect of salinity. The salt tolerance of many crops is enhanced when daily drip irrigation is used. Climate can also modify plant response to salinity. Salt tolerance is often reduced under hot, dry conditions; also, crops appear to be more salt-tolerant in areas with air pollution that limits plant growth.

Plant factors such as stage of growth, variety, and rootstock affect plant response to salinity. Rice, barley, wheat, and corn are most sensitive to salinity during the early seedling stage. Beet and sunflower are more sensitive to salinity during germination than in later stages of growth. Salinity effects on the vegetative growth of many plants (e.g., cotton, wheat) are greater than on seed or fiber production. Normally, salinity suppresses top growth more than root growth. The type of rootstock of woody plants may also affect the specific tolerance of fruit crops to salinity.

Response of Soils to Salinity and Sodicity

The physical and mechanical properties of soils (e.g., clay swelling and dispersion, aggregate stability, hydraulic conductivity, runoff, and erosion) are affected by soil salinity and sodicity. The main factor that determines the extent of the deleterious effect of Na on soil properties is the salt concentration in the soil solution. With low soil-solution salinity, as is the case in humid regions or during the rainy season in semiarid regions, even a low sodicity (ESP <5) has a marked and irreversible negative effect on the physical properties of most soils. Conversely, when the salt concentration reaches even a moderate level (10–20 $mmol_c/l$), high sodicity levels ($10<$ ESP <30) cause only small to moderate changes in the physical and hydraulic properties of soils. This is demonstrated in Israel, where irrigation water of 4.6 dS m^{-1} and SAR of 26 are used for irrigation of cotton. During the irrigation season (summer), no deterioration of soil physical properties is observed in spite of the high soil sodicity, because the high salt concentration prevents the adverse effects of sodicity. On the other hand, during the subsequent rainy season, the electrolyte concentration in the soil solution at the soil

surface becomes low, rendering the soil susceptible to crusting, low permeability, and increased runoff and erosion.

The response of soils to sodicity and salinity depends on the soil properties (e.g., texture, clay mineralogy, lime and sesquioxides content). Smectitic soils are more susceptible to sodicity than kaolinitic soils. Calcareous soils are less susceptible to sodicity than noncalcareous soils. Soils rich in sesquioxides are less affected by sodicity than soils with low sesquioxide content. However, the response of soils to sodicity depends also on time-dependent properties such as wetting rate of the dry soil, antecedent moisture content, and aging (time since cultivation or wetting). Fast wetting of dry soils disintegrates the aggregates and increases the susceptibility of the soils to sodicity. The effect of wetting rate depends on the antecedent moisture content of the soils. Less slaking of soil aggregates upon fast wetting occurs when the antecedent moisture content of the soil is high and the moist soil is less susceptible to the deleterious effect of sodicity. Similarly, upon aging of a moist soil, the stability of the soils' aggregates increases (aging effect) and their susceptibility to sodicity decreases.

Reclamation of Salt-Affected Soils

In ideal soils under piston flow conditions, the EC of soil water and applied water would be equal after the passage of one pore volume of rainfall or irrigation water. However, soils are not ideal for two reasons: (1) water flows faster through soil cracks and large pores between soil aggregates than through smaller pores within aggregates. This difference in localized water flow velocities is commonly referred to as 'bypass flow,' or 'preferential flow'; (2) salt diffusion and chemical reactions are not sufficiently rapid to allow salt concentration to be the same among all pores of differing sizes, particularly during periods of rapid water movement.

Salt transport models have provided considerable insight into these processes. However, suitable mathematical models are not available to describe the multiple concurrent processes of water flow, including spatially variable hydraulic properties and all the chemical reactions – exchange, dissolution, precipitation, and CO_2 liquid–gas equilibriums – involved in salt transport on a field scale. Consequently, reclamation guidelines are largely based on experimental relations obtained from field and lysimeter reclamation experiments.

The method of water application and soil type are the primary variables affecting the amount of water required to reclaim saline soils. For saline soils, 60% or more of the salts initially present is leached by a depth of water equal to the depth of soil (Figure 5) under continuously ponded conditions. Sandy loam soils with low saturated water contents have higher leaching efficiency than clay loam or clay soils (Figure 5). Soil pores in sandy soils are more uniform in diameter than in clay loam or clay soils. The finer-textured soils can have large cracks with large pores when dry, and fine pores within aggregates when wet. Such bypass channels reduce leaching efficiency.

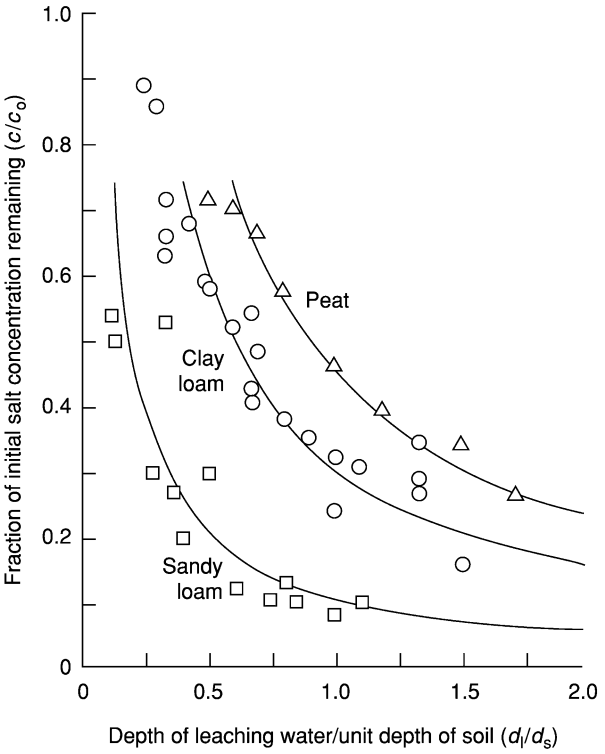


Figure 5 Depth of leaching water per unit depth of soil required to reclaim a saline soil by continuous ponding. (Adapted from Oster JD, Shainberg I, and Abrol IP (1996) Reclamation of salt affected soils. In: Agassi M (ed.) *Soil Erosion Conservation and Rehabilitation*, pp. 315–342. New York: Marcel Dekker.)

The water requirement for leaching can be reduced by intermittent application of ponded water, particularly for fine-textured soils (Figure 6). Under intermittent ponding, unsaturated conditions prevail in the soil profile and the inefficient bypass flow in large cracks and pores is prevented. Similarly, with sprinkler irrigation, reclamation occurs under unsaturated conditions and the efficiency of leaching increases.

Reclamation of Sodic Soils

The reclamation of sodic soils depends on the supply of Ca salts to replace exchangeable sodium and in maintaining adequate soil salinity to counteract the

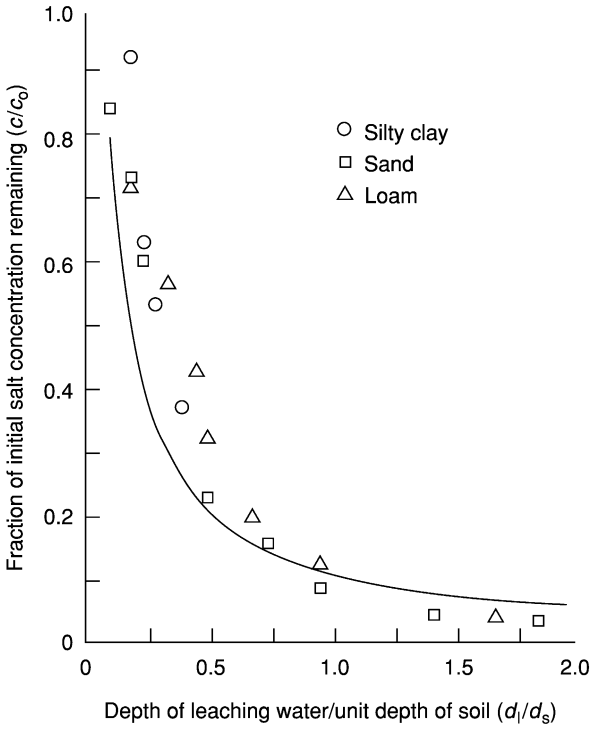


Figure 6 Depth of leaching water per unit depth of soil required to reclaim a saline soil by ponding water intermittently. (Adapted from Oster JD, Shainberg I, and Abrol IP (1996) Reclamation of salt affected soils. In: Agassi M (ed.) *Soil Erosion Conservation and Rehabilitation*, pp. 315–342. New York: Marcel Dekker.)

adverse effect of sodicity on the hydraulic properties of the soils.

Gypsum, either incorporated into the soil or left on the soil surface, is the calcium source most commonly used to reclaim sodic soils and to improve soil water infiltration that has been decreased by low salinities. Sources include mineral deposits and phosphogypsum, a by-product of the phosphate fertilizer industry. When gypsum is incorporated into the soil, the reduction in ESP upon irrigation and leaching is primarily limited to the soil depth interval where gypsum is present. This is a consequence of the greater selectivity of exchange sites for calcium than for sodium. The gypsum required for reclamation can be estimated from the amount of exchangeable sodium in the soil layer to be reclaimed – the gypsum requirement concept.

Since the soil surface is most susceptible to sodicity, reclamation of sodic soils by spreading phosphogypsum at the soil surface is being practiced in Israel. The amounts of gypsum that are applied depend on the depth of rain, assuming that gypsum concentration in the infiltrating rain reaches 50% of gypsum saturation. Thus, in a region with winter rain of 300 mm, the amount of gypsum spread at the soil surface is 3 Mg ha^{-1}, and, in a region with 500 mm of rain, the amount of gypsum applied is 5 Mg ha^{-1}. In the commercial fields in Israel that have been irrigated with saline and sodic water as described above, phosphogypsum is spread annually on the soil surface, following tillage in the autumn, at a rate of 3–5 Mg ha^{-1}. This prevents seal formation and maintains high infiltration rates which, in turn, provide sufficient infiltration of rainfall to leach salts from the root zone. Dissolution of the gypsum provides enough electrolytes to prevent seal formation and maintain high infiltration rates in spite of the fact that the soil surface is not completely reclaimed. This management in the rainy season, coupled with adequate irrigation with the saline-sodic water to meet crop needs during the summer months, results in cotton yields that are similar to those obtained when nonsaline water is used for irrigation.

Summary

Under arid and semiarid conditions, and in regions of poor natural drainage, there is a real hazard of salt accumulation in soils. The processes by which soluble salts accumulate in soils include irrigation with water containing salts, upward movement of moderate-to-saline water from high water tables, and weathering of primary and secondary minerals in the soils. Accumulation of soluble salts in the soil solution imposes stress on growing crops that can lead to decreased yields and, in severe cases, complete crop failure. Use of poor-quality water is increasing due to growing municipal demands for the available supplies of good-quality water and the need to dispose of municipal wastewater and agricultural drainage water. As a result, a better understanding of soil reclamation is becoming an increasingly important component of water and soil management to ensure the long-term sustainability of irrigated agriculture.

See also: **Sodic Soils**

Further Reading

Bower CA, Ogata G, and Tucker JM (1968) Sodium hazard of irrigation water as influenced by leaching fraction and precipitation of carbonate. *Soil Science* 106: 24–34.

Bresler E, Mcneal BL, and Carter DL (1982) *Saline and Sodic Soils*. Berlin, Germany: Springer-Verlag.

Eaton FM (1950) Significance of carbonates in irrigation water. *Soil Science* 69: 123–133.

Keren R (1999) Salinity. In: Sumner ME (ed.) *Handbook of Soil Science*, pp. 1–25. Boca Raton, FL: CRC Press.

Levy GJ, Shainberg I, and Miller WP (1998) Physical properties of sodic soils. In: Sumner ME and Naidu R (eds) *Sodic Soils*, pp. 77–94. Oxford, UK: Oxford University Press.

Mass EV (1990) Crop salt tolerance. In: Tanji KK (ed.) *Agriculture Salinity Assessment and Management*, pp. 262–304. New York: American Society of Civil Engineers.

Oster JD, Shainberg I, and Abrol IP (1996) Reclamation of salt affected soils. In: Agassi M (ed.) *Soil Erosion Conservation and Rehabilitation*, pp. 315–342. New York: Marcel Dekker.

Shainberg I and Letey J (1984) Response of soils to saline and sodic conditions. *Hilgardia* 52: 1–57.

Shainberg I, Levy GJ, Goldstein D, Mamedov A, and Letey J (2001) Prewetting rate and sodicity effects on the hydraulic conductivity of soils. *Australian Journal of Soil Research* 39: 1279–1291.

Suarez DL, Wood JD, and Abrahim I (1992) Reevaluation of calcite supersaturation in soils. *Soil Science Society of America Journal* 56: 1776–1784.

Sumner ME, Rengasamy P, and Naidu R (1998) Sodic soils: a reappraisal. In: Sumner ME and Naidu R (eds) *Sodic Soils*, pp. 3–17. New York: Oxford University Press.

US Salinity Laboratory Staff (1954) *Diagnosis and Improvement of Saline and Alkali Soils*. Agricultural Handbook No. 60, USDA. Washington, DC: US Government Printing Office.

SALINITY

Contents
Management
Physical Effects

Management

D Hillel, Columbia University, New York, NY, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Natural and Induced Soil Salinity

The term 'salinity' refers to the presence in soil and water of electrolytic mineral solutes in concentrations that are harmful to many agricultural crops. Except along seashores, saline soils seldom occur in humid regions, thanks to the net downward percolation of fresh water though the soil profile, brought about by the excess of rainfall compared with evapotranspiration. In arid regions, on the other hand, there may be periods of no net downward percolation and hence no effective leaching, so salts can accumulate in the soil. Hence the combined effect of meager rainfall, high evaporation, the presence of salt-bearing sediments, and – in many places, particularly river valleys and other low-lying areas – the occurrence of shallow, brackish groundwater which gives rise to saline soils.

Less obvious than the appearance of naturally saline soils, but perhaps more insidious, is the inadvertent induced salination of originally productive soils, caused by human intervention. Irrigation waters generally contain appreciable quantities of salts. (For example: Even with relatively good-quality irrigation water containing no more than 0.3 kg salts m^{-3}, applying 10 000 mm water per season would add as much as 3 000 kg salts ha^{-1} per year!) Crop plants normally extract water from the soil while leaving most of the salt behind. Unless leached away (continuously or intermittently), such salts sooner or later begin to hinder crop growth. Where internal drainage is impeded, attempts to leach the soil can do more harm than good, by raising the water table, waterlogging the soil, and causing capillary rise to the surface, where evaporation infuses the topsoil with cumulative quantities of salt.

Overall salinity is generally expressed in terms of the total dissolved solutes (TDS) in milligrams per liter of solution (approximately equivalent to parts per million, ppm). Salinity may also be characterized by measuring the electrical conductivity (EC) of the solution, generally expressible in terms of decisiemens per meter ($dS\,m^{-1}$).

Quantitative criteria for diagnosing soil salinity were originally formulated by the US Salinity Laboratory at Riverside, California (in its *Handbook 60*, first issued in 1954), in terms of the EC of the soil's saturation extract (i.e., the solution extracted from a soil sample that had been presaturated with water). Those criteria are still in wide use today. Accordingly, a saline soil has been defined as having an EC of more than $4\,dS\,m^{-1}$. This value generally corresponds to approximately 40 mmol salt l^{-1}. In the case of a sodium chloride solution, this equals about $2.4\,g\,l^{-1}$.

The clay fraction in saline soils is generally well flocculated. As the salts are leached, however, the

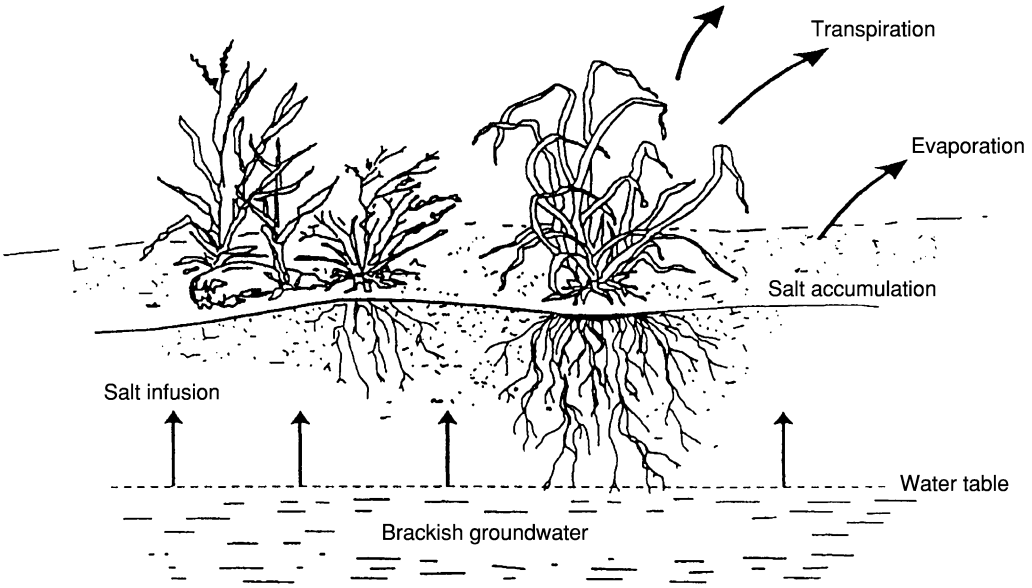


Figure 1 The process of waterlogging and salination. Reproduced with permission from Hillel D (1998) Environmental Soil Physics, Elsevier.

flocs may tend to disperse and the soil aggregates to break down (or slake). This especially occurs where an appreciable concentration of sodium is adsorbed on to the clay particles. The tendency for flocs to disperse and for aggregates to slake and collapse results in the deterioration of soil structure by the clogging of large pores in the soil, and consequently in the reduction of soil permeability. This leads to the associated phenomenon of soil sodicity, also known as alkalinity (Figure 1).

Irrigation Water Quality

Solutes can be added to the soil solution in rainfall (especially along seacoasts, where sea spray mixes with the rain) or in irrigation water infiltrated from above, in groundwater rising by capillarity from below, in the dissolution of salts initially present in solid form within the soil and subsoil, and in the artificial application of agricultural chemicals (e.g., fertilizers, pesticides, and soil amendments). Methods for removal of solutes from the soil include uptake by plants (generally of minor importance), downward transport by percolation and drainage (leaching), erosion of the soil surface by overland flow and by wind, precipitation or adsorption on to the solid phase and conversion to insoluble or unavailable forms, and – for some substances – volatilization of gaseous compounds. To avoid the accumulation of salts to toxic levels, their inputs to the soil must not exceed the rate of their removal from the soil or their conversion to unavailable forms within it. The control of soil salinity must therefore include measures to control both the inputs and the outputs of salts.

Among the various inputs of salts listed above, irrigation water often plays a major role. The quality of irrigation water may affect soil salinity and sodicity, cation exchange, soil acidity or alkalinity, nutrient availability, clay dispersion and flocculation, as well as soil structure, infiltration, and aeration. Clearly, therefore, the composition of irrigation water is an important determinant of soil productivity, crop growth, and agricultural drainage quality. The hazard posed by irrigation water containing salts of varying composition and concentration depends on soil conditions, climatic conditions, crop species and variety, and the amount and frequency of the irrigation applied. In general, irrigation water of EC less than $0.7\,\mathrm{dS\,m^{-1}}$ poses little or no danger to most crops, whereas EC values more than $3.0\,\mathrm{dS\,m^{-1}}$ may restrict the growth of most crops. The major components of salinity are the cations Ca, Mg, and Na, and the anions Cl, SO_4, and HCO_3. The potassium, nitrate, and phosphate ions, though important nutritionally, are usually minor components of soil salinity. In addition, certain constituents (such as boron) may have an important effect on crop growth even though their concentrations are usually too low to have any substantial effect on the soil's total salinity.

Irrigation waters of different sources, locations, and seasons vary greatly in quality (Table 1). Some irrigation waters contain as little as $50\,\mathrm{g\,salts\,m^{-3}}$, and others as much as $3000\,\mathrm{g\,salts\,m^{-3}}$. Since the volume of water applied in irrigation to a crop during its growing season commonly varies between 5000 and $20\,000\,\mathrm{m^{-3}\,ha^{-1}}$, the salt input to a crop may range between 250 and $60\,000\,\mathrm{kg\ ha^{-1}}$.

Table 1 Classification of water quality according to total salt concentration

Designation	*Total dissolved salts (ppm)*	*EC ($dS\,m^{-1}$)*	*Category*
Fresh water	<500	<0.6	Drinking and irrigation
Slightly brackish	500–1 000	0.6–1.5	Irrigation
Brackish	1 000–2 000	1.5–3	Irrigation with caution
Moderately saline	2 000–5 000	3–8	Primary drainage
Saline	5 000–10 000	8–15	Secondary drainage and saline groundwater
Highly saline	10 000–35 000	15–45	Very saline groundwater
Brine	>35 000	>45	Seawater

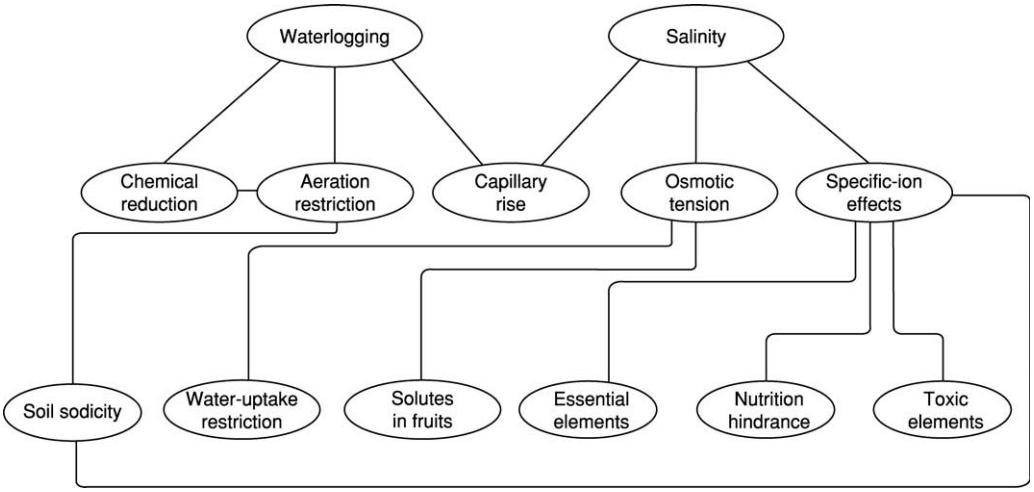


Figure 2 Effects of salinity and sodicity on plants.

Another important criterion of irrigation water quality is the sodium adsorption ratio (SAR). High alkalinity of irrigation water, manifested when the pH value is more than 8.5, generally indicates the predominant presence of sodium ions (often associated with the bicarbonate anion) and poses a danger of soil sodification. Freshly pumped groundwater may have a high SAR even if the pH is less than 8.5, owing to the presence of dissolved CO_2 (samples of such water should be aerated to allow the carbon dioxide to effervesce prior to measurement of the pH).

With high-SAR water, irrigation by sprinkling will increase the soil's tendency to form a surface seal (crust) under the impact of the drops striking the bare soil. Flood irrigation may also cause the breakdown of soil aggregates by air slaking. On the other hand, application of water by drip irrigation at spaced points on the surface or below it, may lessen the physical disruption of soil structure that would otherwise take place under the influence of high-SAR water. High-pH water may cause nutritional as well as structural problems. The addition of gypsum to the water or to the soil surface may help in both respects, by promoting the flocculation of clay and the nutritional balance of the soil solution.

In certain circumstances and with appropriate precautions, brackish water may safely be used for the irrigation of salt-tolerant crops. This includes the secondary use, or reuse, of drainage waters. Especially suitable for this purpose are deep sandy soils that are prevalent in some arid regions, where internal drainage is unrestricted and there is little risk of either groundwater rise or of soil salination and sodification.

The use of brackish water for sprinkling irrigation, however, may cause foliar injury. The degree of injury depends on the following factors: concentrations of ions in the water, sensitivity of the crop at various growth stages, water stress of the plants prior to irrigation, and frequency of irrigation. The potential damage also depends on the prevailing environmental conditions, including the temperature and relative humidity of the atmosphere, which affect the rate of evaporation. Sprinkling at night, when the atmospheric temperature and evaporativity are relatively low, can reduce foliar absorption and injury.

The greater the rate of evapotranspiration and the longer the interval between irrigations, the more concentrated the residual solution in the root zone and the more severe the salt stress imposed on the plants (Figure 2). Hence, when irrigation water is brackish,

a common management strategy is to increase the frequency of irrigation so as to maintain a high content of water and to prevent the rise of salt concentration in the root zone. Thanks to its ready adaptability to high-frequency irrigation, drip irrigation beneath the canopy appears to be the most appropriate method to use with brackish water, especially as it avoids direct foliar exposure to saline water.

Brackish drainage water may also be used in a system of agroforestry. Salt-tolerant trees are capable of thriving when irrigated with brackish water, and also of lowering the water table by the extraction and transpiration of water from deeper layers in the soil, thus reducing the volume and expense of needed drainage in some areas. Among the trees suitable for this type of agroforestry and biodrainage are certain species of eucalyptus, acacia, casuarina, poplar, mesquite, and tamarisk. The harvested wood may be used for fuel, pulp, or construction.

The use of drainage wastewater entails certain distinct hazards. In some places, the concentration of nitrates may become excessive, and contribute to groundwater and surface-water contamination. Heavy metals and various other toxic materials can pose a problem. They tend to accumulate in the soil and thence enter the biological food chain. Especially hazardous are industrial waste products that may be carcinogenic as well as toxic.

Groundwater Drainage

The presence of a high water table can be either a blessing or a curse. The blessing occurs when, in periods of low rainfall or deficiency of water for irrigation, upward capillary flow from the water table augments the water supply to the roots. The curse occurs as the rising water brings up salts from below and thereby salinizes the root zone. In the field, upward capillary rise and downward percolation may take place alternately during the year. Percolation occurs typically during the rainy and early irrigation seasons, when the water requirements of the crop are relatively low and the water supply from above is high. On the other hand, upward flow takes place later in the irrigation season, when the water requirements are high and both rainfall and irrigation are restricted. Over the long term, a net downward flow of salt-bearing water through the root zone is essential to sustainable productivity.

The term 'drainage' can be used in a general sense to denote outflow of water from soil. More specifically, it can serve to describe the artificial removal of excess water, or the set of management practices designed to prevent the occurrence of excess water. The removal of free water tending to accumulate over the soil surface by appropriately shaping the land is termed 'surface drainage' and is outside the scope of this article. Finally, 'groundwater drainage' refers to the outflow or artificial removal of excess water from within the soil, generally lowering the water table or preventing its rise.

In waterlogged soils, gas exchange with the atmosphere is restricted to the surface zone, while, within the profile, oxygen may be almost totally absent and carbon dioxide may accumulate. Under anaerobic conditions, various substances are reduced chemically from their normally oxidized state. Toxic concentrations of ferrous, sulfide, and manganous ions can develop. These, in combination with products of the anaerobic decomposition of organic matter (e.g., methane) can greatly inhibit plant growth. At the same time, nitrification is prevented, and various plant and root diseases (especially fungal) are more prevalent. High-moisture conditions at or near the soil surface make the soil susceptible to compaction and puddling by animal and machinery traffic. Necessary operations (e.g., tillage, planting, spraying, and harvesting) are thwarted by poor trafficability (i.e., the ability of the ground to support vehicular traffic). Furthermore, the surface zone of a wet soil does not warm up readily at springtime, owing to greater thermal inertia and downward conduction, and to loss of latent heat by the higher evaporation rate. Consequently, germination and early seedling growth are retarded. All these phenomena are in addition to the tendency of waterlogged soils, especially in arid areas, to become saline.

The artificial drainage of groundwater is generally carried out by means of drains, which may be ditches, pipes, or 'mole channels,' into which groundwater flows as a result of the hydraulic gradients existing in the soil. The drains themselves are made to direct the water, by gravity or by pumping, to the drainage outlet, which may be a stream, a lake, an evaporation pond, or the sea. In some places, drainage water may be recycled or reused for agricultural, industrial, or residential purposes, as well as for agroforestry or the irrigation of ornamental plants. Because drainage water may contain potentially harmful concentrations of salts, fertilizer nutrients, pesticide residues, and various other toxic chemicals, as well as biological pathogens, it is not enough to provide means to 'get rid' of it; attention must be devoted to the quality of the water to be disposed of and to the long-term consequences of its disposal.

The disposal of salt-bearing effluent may pose a danger to rivers and to groundwater. If the drainage water is returned to a river or an aquifer serving as a water source, or if the drainage is to be reused directly,

Table 2 Prevalent depths and spacing of drainage pipes in different soil types

Soil type	Hydraulic conductivity ($m\,day^{-1}$)	Spacing of drains (m)	Depth of drains (m)
Clay	1.5	10–20	1–1.5
Clay loam	1.5–5	15–25	1–1.5
Loam	5–20	20–35	1–1.5
Fine, sandy loam	20–65	30–40	1–1.5
Sandy loam	65–125	30–70	1–2
Peat	125–250	30–100	1–2

Reproduced with permission from Hillel D (1998) Environmental Soil Physics, Elsevier.

its load of corrosive salts and other pollutants may affect downstream agriculture, households, water utilities, and industry. Where persistent pesticides are applied to irrigated land, their residues can cause further damage to biotic communities in riverine, estuarine, and lacustrine habitats, as well as to public health. Construction of evaporation ponds for disposal of drainage water generally requires the allocation of approximately 10% of the land area. Care must also be taken to avoid leaks and consequent salination of the underlying groundwater. The loss of potentially valuable land, the costs of construction and conveyance, and the environmental impacts may make such ponds impractical in some cases.

Various equations, empirically or theoretically based, have been proposed for determining the desirable depths and spacings of drain pipes or ditches in different soil and groundwater conditions. In the Netherlands, the country with the most experience in drainage, common criteria for drainage are to provide for the removal of approximately 7 mm per day, and to prevent a water table from rising above a depth of 0.5 m from the soil surface. In more arid regions, because of the greater evaporation rate and groundwater salinity, the water table must generally be kept much deeper. In the Imperial Valley of California, for instance, the drain depth ranges from approximately 1.5 to 3 m, and the desired water table depth midway between drains is at least 1.2 m. For medium- and fine-textured soils, the depth should be even greater where the salinity risk is high. Since there is a practical limit to the depth of drain placement, it is the density of drain spacing that must be increased under such circumstances (Table 2).

Leaching Processes

To prevent salts from accumulating in the root zone during repeated cycles of irrigation and evapotranspiration, the obvious remedy is to apply water in an amount greater than evapotranspiration, so as deliberately to cause a fraction of the applied water to flow through the root zone and flush away the excess salts. However, unless the water table is very deep or lateral groundwater drainage is sufficiently rapid, the extra irrigation can cause a progressive rise of the water table. Therefore, the amount of water applied must be optimized to allow leaching without water-table rise. The concept of 'leaching requirement' was first developed by the US Salinity Laboratory in Riverside, California. It has been defined as the fraction of the irrigation water that must be percolated out of the bottom of the root zone in order to prevent average soil salinity from rising above some specifiable level.

According to the standards developed there, the maximum concentration of the soil solution in the root zone, expressed in terms of EC, should be kept below 4 $dS\,m^{-1}$ for sensitive crops. Salt-tolerant crops such as beets, alfalfa, and cotton may give satisfactory yields at values up to 8 $dS\,m^{-1}$. The problem encountered in any attempt to apply such a simplistic criterion is that in the field (unlike the case of plants grown in solution culture or in small containers) the concentration of the soil solution varies greatly in space and time. In addition, the sensitivity of any crop to salinity depends on its stage of growth and on such variables as ambient temperature, atmospheric humidity, soil matric suction, and nutrient availability.

The leaching requirement depends on the salt concentration and composition of the irrigation water, on the amount of water extracted from the soil by the crop, and on the salt tolerance of the crop, which determines the maximum allowable concentration of the soil solution in the root zone. Assuming steady-state conditions of throughflow (thus disregarding short-term fluctuations in soil-moisture content, flux, and salinity), and furthermore assuming no appreciable dissolution or precipitation of salts in the soil and no significant removal of salts by the crop or capillary rise of salt-bearing water from below, the following simple equation is obtained:

$$V_d/V_i = c_i/c_d \qquad [1]$$

in which V_d and V_i are the volumes of drainage and of irrigation, respectively, and c_d and c_i are the corresponding concentrations of salt. Water volumes are normally expressed per unit area of land as equivalent depths of water, and salt concentrations are generally measured and reported in terms of EC. Because the volume of water drained is the difference between the volumes of irrigation and evapotranspiration (i.e., $V_d = V_i - V_{et}$), we can transform the last equation as follows:

$$V_i = [c_d/(c_d - c_i)]V_{et} \quad [2]$$

This is equivalent to the formulation given in the US Salinity Laboratory's Department of Agriculture (USDA) Handbook No. 60:

$$d_i = [E_d/(E_d - E_i)]/d_{et} \quad [3]$$

where d_i is the depth of irrigation, d_{et} the equivalent depth of 'consumptive use' by the crop (evapotranspiration), and E_d and E_i are the electrical conductivities of the drainage and irrigation waters, respectively.

The leaching requirement equation implies that, by varying the fraction of applied water percolated through the root zone, it is possible to control the concentration of salt in the drainage water and hence to maintain the concentration of the soil solution in the main part of the root zone at some intermediate level between c_i and c_d. However, the leaching requirement concept disregards the distribution of salts within the root zone itself, as it is affected by the frequency and spatial variability of irrigation, as well as by its quantity and water quality. In particular, the variation of root-zone salinity is affected by the pattern and degree of soil moisture depletion between irrigations. The less frequent the irrigation regime, the greater the buildup of salt between successive irrigations. In some cases, the commonly recommended leaching fraction may not be sufficient to prevent the reduction of yield below its potential, especially if the climatically imposed evaporation rate is high and the irrigation water is brackish.

With modern methods of high-frequency irrigation, it is possible to maintain the soil solution in the surface zone at a concentration nearly equal to that of the irrigation water. This zone can be deepened by increasing the volume of water applied. Beyond this zone, the salt concentration of the soil solution increases with depth to a salinity level depending on the leaching fraction. High-frequency irrigation not only lowers the concentration of the soil solution in the upper zone (where most roots proliferate), but also tends to minimize the matric suction of soil moisture.

Extensive research has shown that leaching soils at a water content below saturation (e.g., under low-intensity sprinkling or intermittent ponding) can produce more efficient leaching that can be achieved by the once-standard method of continuous flooding. In a soil with macropores – cracks, wormholes, or decayed root channels – much of the water under ponding moves rapidly down those large passageways, bypassing the greater volume of the soil containing the salt, so it is largely ineffective in leaching the micropores of the soil matrix. In contrast, under low-intensity sprinkling, the soil never becomes saturated, so a greater portion of the applied water moves through the soil matrix, thus producing more efficient leaching per unit volume of water infiltrated. However, the processes of infiltration and unsaturated flow under low-intensity sprinkling are inherently slower and require more time than saturated infiltration under ponding.

Nonuniformity of irrigation as well as of soil is a complicating factor. If the leaching requirement is not met throughout the field, soil salinity will prevail in spots, wherever leaching is insufficient. Whether to apply copious amounts of water to the entire field so as to ensure that the leaching requirement is met everywhere, or to accept some reduction in yield in parts of the field, must be determined from an economic analysis of costs and benefits. Such an analysis should take into consideration the danger that saline spots might recur (and perhaps even grow in extent and severity) from year to year. One answer is to apply extra water preferentially to the spots that need it most, but such a strategy requires a flexible irrigation system that would allow controlled variability of water delivery. Although such a specialized irrigation system is likely to be expensive to install and operate, it may well be worthwhile in the long run.

Soil Amendments

The leaching process is enhanced if the applied water contains a sufficient concentration of electrolytes to reduce swelling and dispersion of clay in the soil. Where leaching occurs with water of very low salinity (e.g., rainwater), soil permeability can be increased by the surface application of a slowly soluble electrolytic salt – preferably a substance containing a divalent cation such as calcium, to prevent the sodium ion from dispersing the clay. Such materials, commonly known as soil amendments, can replace exchangeable sodium with flocculation-promoting divalent ions.

The most commonly considered soil amendment for the purpose of improving the structure and the permeability of soils, especially those prone to becoming sodic, are dihydrate calcium sulfate (gypsum) and dihydrate calcium chloride. Gypsum is generally the preferred soil amendment, thanks to its ready availability in many places and to its relatively low cost. It may be derived from mining or be available as a by-product of the phosphate-fertilizer industry. The solubility of pure gypsum is approximately 2.15–2.63 $kg\,m^{-3}$, depending somewhat on temperature. Applied gypsum dissolves in the soil solution until its solubility limit is reached or until its supply is exhausted. The rate of dissolution of applied gypsum depends on its source and degree of granulation. Industrial gypsum generally dissolves more rapidly

than mined gypsum, which generally contains impurities. The amount of gypsum needed to replace the exchangeable sodium depends on the initial exchangeable sodium percentage, the soil's total cation exchange capacity, its bulk density, and the depth of the soil to be treated effectively. If elemental sulfur is added to a sodic soil in lieu of gypsum, it must be oxidized *in situ* to become effective. Upon oxidation, it forms sulfuric acid, which then reacts with lime in the soil to produce gypsum.

Early-Warning Systems

In many cases, irrigation systems are organized and irrigation is begun long before drainage is installed. Indeed, an irrigation project can often function unimpeded for years, even decades, without artificial drainage. In some cases, the land is so well drained naturally that irrigation can thus be continued for a very long time. However, far more typically, the processes of groundwater rise and salt accumulation proceed inexorably, so that sooner or later (especially in ill-drained river valleys, where most irrigation development takes place) the provision of artificial drainage becomes essential.

Granted that a drainage system must be planned in advance at the very outset of an irrigation project, the crucial question is when to begin installing and operating it. If installed too early, the drainage system may lie unused for some time and therefore be both unnecessary and uneconomical, and it may deteriorate in the interim before it comes into use. On the other hand, if installed after waterlogging and salination have advanced, it may be too late to restore productivity economically. These considerations emphasize the importance of having an early-warning system to indicate, before the problem becomes acute, that soil salination is incipient and that the need for drainage is imminent. Soil salinity is normally monitored by a combination of soil sampling, soil-solution sampling by vacuum extraction, and *in situ* salt sensors. Mobile devices with combined electromagnetic induction and four-electrode soil conductivity sensing systems are now under development for monitoring and mapping the distribution of soil salinity over an entire field, but such instruments have not yet entered into general use.

The detection and diagnosis of salinity are difficult in the early stages of its occurrence. Visual inspection of the soil surface may be misleading. For example, the white precipitate formed on the surface of furrow-irrigated or drip-irrigated soil may be due to relatively harmless calcite or gypsum. Visual inspection of crops provides obvious clues to salt stress only after the condition is well advanced. Crop plants suffering salt stress eventually exhibit stunted growth, smaller leaves than normal, and discoloration. Such symptoms first occur in spots, rather than uniformly over the field. Since factors other than salinity (e.g., water stress, nutrient deficiencies, or misapplied pesticides) may produce similar symptoms, visually observed indications of apparent salinity should be checked by chemical analyses of soil, plant, and water samples. One good way to detect the early appearance of salinity is to place plants that are known to be particularly salt-sensitive at regularly spaced intervals throughout the irrigated area. Such interspersed detector plants may reveal early symptoms of physiological stress and thus provide a timely warning of problems that are likely to be exacerbated over time.

An important indicator of the salinity hazard is the depth of the water table. There is cause for concern whenever it approaches within 1.5–2 m of the soil surface. Hence the water table should be monitored regularly by means of observation wells and piezometers (Figures 3 and 4). Both are vertical tubes that

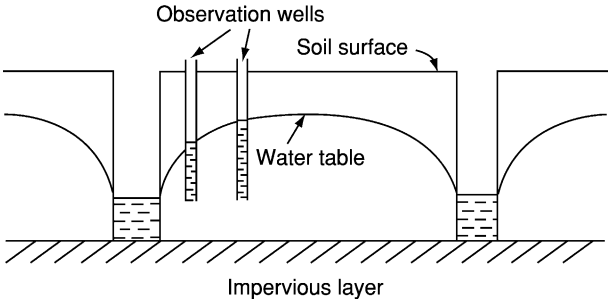


Figure 3 Observation wells to determine elevation of the water table. Reproduced with permission from Hillel D (1998) Environmental Soil Physics, Elsevier.

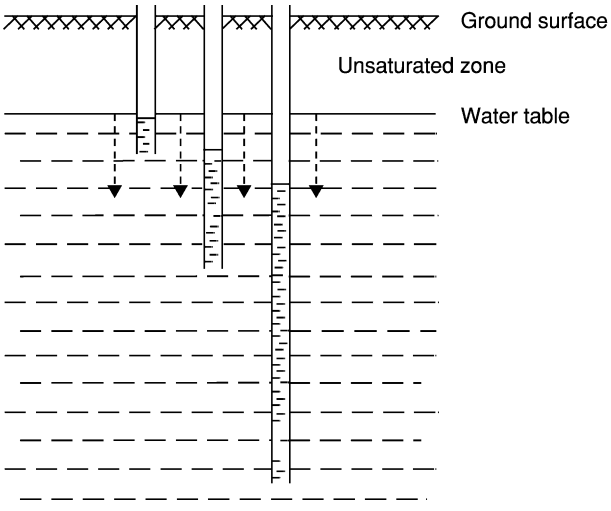


Figure 4 Piezometers to determine vertical pressure gradients below the water table. Reproduced with permission from Hillel D (1998) Environmental Soil Physics, Elsevier.

are inserted into the soil to a depth far below the water table. The difference is that an observation well is perforated to permit free inflow of groundwater along the length of the tube below the water table; in contrast, a piezometer's only opening is at the bottom. Hence a piezometer indicates the hydraulic head (pressure) of the water at the bottom of the tube, rather than the position of the water table. Several piezometers, inserted side by side to different depths, can indicate the vertical gradient of the hydraulic head below the water table. The direction and magnitude of that gradient are indicative of the tendency of the groundwater to rise or fall.

See also: **Solute Transport**

Further Reading

Bresler E, McNeal BL, and Carter DL (1982) *Saline and Sodic Soils: Principles–Dynamics–Modeling*. Berlin, Germany: Springer-Verlag.

Dinar A and Zilberman D (eds) (1991) *The Economics and Management of Water and Drainage in Agriculture*. Boston, MA: Kluwer Academic.

Hanks RJ (1984) Predictions of crop yield and water consumption under saline conditions. In: *Soil Salinity Under Irrigation*. Berlin, Germany: Springer-Verlag.

Hillel D (2000) *Salinity Management for Sustainable Irrigation: Integrating Science, Environment, and Economics*. Washington, DC: The World Bank.

Hoffman GJ (1990) Leaching fraction and root zone salinity control. In: *Agricultural Salinity Assessment and Management*. New York: American Society of Civil Engineers.

Keren R and Miyamoto S (1990) Reclamation of saline, sodic, and boron-affected soils. In: *Soil Salinity Under Irrigation*. Berlin, Germany: Springer-Verlag.

Lauchli, A and Epstein E (1990). Plant responses to saline and sodic conditions. In: *Agricultural Salinity Assessment and Management,* pp. 113–137. New York: American Society of Civil Engineers.

Massoud FI (1981) *Salt-Affected Soils at a Global Scale and Concepts for Control*. Rome, Italy: FAO Land and Water Development Division.

Oster JD and Frenkel H (1980) The chemistry of the reclamation of sodic soils with gypsum and lime. *Soil Science Society of America Journal* 44: 41–45.

Rhoades JD, Chanduvi F, and Lesch SM (1999) *Soil Salinity Assessment Methods and Interpretation of Electrical conductivity Measurements*. FAO Irrigation and Drainage Paper 57. Rome, Italy: FAO Land and Water Development Division.

Richards LA (ed.) (1954) *Diagnosis and Improvement of Saline and Alkali Soils*. USDA Handbook No. 60. Washington, DC: US Government Printing Office.

Shainberg I and Shalhevet J (eds) (1984) *Soil Salinity Under Irrigation*. Berlin, Germany: Springer-Verlag.

Tanji KK (ed.) (1990) *Agricultural Salinity Assessment and Management*. New York: American Society of Civil Engineers.

Yaron D (ed.) (1981) *Salinity in Irrigation and Water Resources*. New York: Marcel Dekker.

Physical Effects

D Russo, The Volcani Center, Bet Dagan, Israel

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Environmental control and agricultural management problems involve the simultaneous transport of water and mixed salts solutions with ions that interact with the soil matrix. The soil solution–soil matrix interactions include phenomena such as cation exchange, anion exclusion, precipitation and dissolution, swelling and reorganization of the soil colloid structures, and, concurrently, rearrangement of the soil pore-size distribution; the latter can affect the soil hydraulic conductivity and retentivity. These interactions, therefore, may considerably affect water flow and solute transport.

The magnitude of the soil solution–soil matrix interactions depends to a large extent upon the types and amounts of inorganic and organic soil colloids. Among soil mineral colloids, the most reactive constituents are the smectite minerals (e.g., montmorillonite, beidellite, monronite), characterized by dioctahedral structure. The presence of these minerals imparts a considerable cation exchange capacity and/or specific surface area to soils from arid and semiarid regions. Furthermore, the smectite minerals are capable of considerable expansion at moderate-to-high exchangeable sodium levels in the presence of relatively low salt soil-solution concentration. Hence, they can impart rather substantial salinity-dependent water retention and conductivity changes to soils, and, concurrently, may considerably affect water flow and solute transport.

Here the emphasis is on the effect of salinity on soil physical properties relevant to water flow and solute transport, in particular, soil water retentivity and conductivity.

Water Retention and Swelling in Salt-Affected Soils

The amount of water retained by the soil depends largely upon the mineral and organic colloid contents of the soil, although the structural arrangement of soil

pores may also be of considerable importance in this respect. In particular, the presence of substantial quantities of soil organic matter and/or smectite minerals results in a considerable water-retention capacity. Soils with high smectite content may also swell considerably more in the presence of high sodium and/or low salt concentrations. Soils high in organic colloids are much more stable and do not tend to swell excessively under adverse chemical conditions. They may disperse, however, at high pH levels or in the presence of low salt concentrations.

The effects of concentration and composition of the soil solution on water retention by a soil containing 20% clay (dominated by montmorillonite) are depicted in Figure 1. The functional relationships (under wetting conditions) are given in terms of the capillary pressure head, h (considered here as a positive quantity), volumetric soil-water content, θ, and equilibrium solution concentration, $C = C_{Na} + C_{Ca}$, where C is concentration in milliequivalents per liter, and composition (expressed in terms of sodium adsorption ratio: $SAR = C_{Na}/\sqrt{(C_{Ca}/2)}$. The points in Figure 1 represent actual measurements for various values of SAR and C; the solid lines represent best visual fits between measured points. Note that the ordinate scale of Figure 1 is shifted upward by an order of magnitude and the data are accordingly translated along this axis, to distinguish between different sets of concentrations. In general, each of the water-retention relationships in Figure 1 can be represented by a single continuous curve. When solution concentration exceeds 50 mEq l^{-1} a single-valued function is obtained for any value of SAR. The same is true for SAR = 0 and 2 mEq $l^{-1} < C < 5$ mEq l^{-1}, as can be seen from the lowest curve of Figure 1. Within the range 2 mEq $l^{-1} < C < 10$ mEq l^{-1} and $0 < SAR < 50$, water content θ depends not only on the pressure head, h, but also on both concentration and composition of the soil solution. For a given h, θ increases as C decreases and as SAR increases.

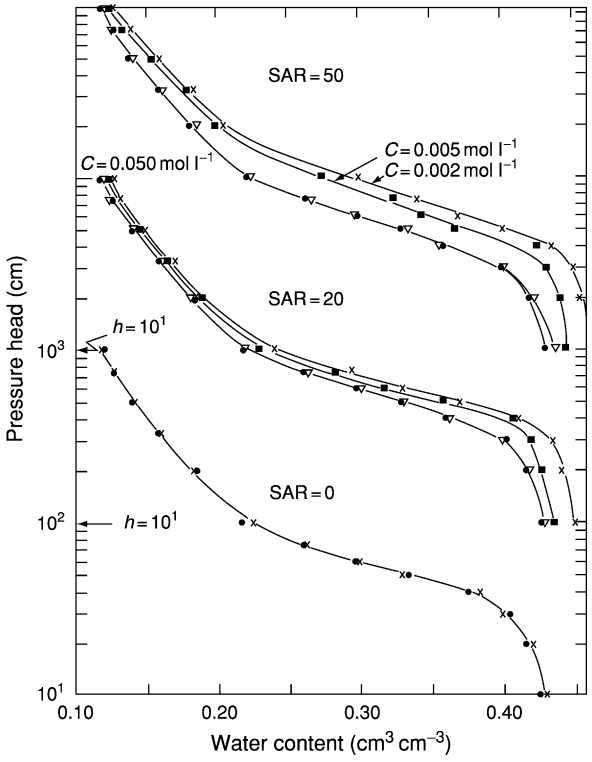


Figure 1 Capillary pressure head, h, as a function of water content, θ, for selected values of the concentration, C, and the sodium adsorption ratio, SAR, of the equilibrium solution. Arrows indicate that the point $h = 10$ cm is shifted and the data are translated along the h-axis. (Reproduced with permission from Russo D and Bresler E (1977) Effect of mixed Na-Ca solutions on the hydraulic properties of unsaturated soils. *Soil Science Society of America Journal* 41: 714–717.)

The experimental findings presented in Figure 1 can be qualitatively explained in terms of electrostatic properties and the porous nature of the soil. It is assumed that the soil pore space can be divided into structural and textural porosity. The latter is regarded as an intrinsic property of the soil, arising from the size and arrangement of basic particles within the soil mass. In these pores the $h(\theta)$ function can be determined quantitatively from the pressure that develops between the clay platelets, within the clay mass between the pores as the soil solution composition and concentration vary.

The classic view of the osmotic pressure of soil colloids is that cations in the diffuse layer between clay particles reduce the free energy of the water relative to that in the external equilibrium solution. This causes water molecules to diffuse into the space between the clay particles, since the cations are electrostatically constrained from leaving the interlayer region. Hence, the particles move apart to accommodate the extra water. If, on the other hand, the clay particles are confined, hydraulic pressure develops between them. For an Na/Ca-Cl soil-water system, this pressure may be determined on the basis of the mixed-ion, diffuse double-layer theory (see eqn [2]).

Equation [2] can be regarded as a reasonably good first-order approximation of the water-retention forces acting within the textural porosity. It predicts that for a given hydrostatic pressure, P, the volume of water retained by the clay mass (which is proportional to the distance between the clay particles) increases as the equilibrium solution becomes more diluted (C decreases) and as the value of SAR increases. In the structural pores, which are sensitive to external forces arising from interactions between clay particles within the textural porosity, the $h(\theta)$ function can be determined from the pressure within the water-filled pores relative to that within the

air-filled pores. In the former pores, at equilibrium, the pressure increases as the pore radius decreases or, for a given pressure, higher values of θ are associated with a larger number of pores having a given corresponding radius. Thus, the functional relationships between h and θ in the structural porosity depend on the total porosity and the pore-size distribution. Changes in the total structural porosity and its distribution result from changes in the volume occupied by the clay mass within the textural porosity. For a given pressure, the space between the clay particles increases as the value of SAR increases and as C decreases. This results in an increase in the amount of water retained by the clay mass, and, concurrently, in the volume that is occupied by the clay mass. For a constant-volume system, changes in the volume of the clay mass are at the expense of the structural porosity and its distribution. Thus, as can be observed in Figure 1, the amount of water retained by the soil for a given capillary pressure head increases with decreasing C and increasing SAR.

Effects of Salts on Soil Hydraulic Conductivity

The retention curves in Figure 1 may be viewed as representatives of the soil's pore-size distribution. The data in Figure 1, therefore, suggest that swelling of the clay fraction within the soil, induced by relatively low C and high SAR, may alter the geometry of the soil pores and thus affect the intrinsic soil permeability, $\kappa(\theta)$. Investigations have confirmed that soil permeability or hydraulic conductivity, $K(\theta) = \kappa(\theta)\gamma/\eta$ (where γ and η are the specific gravity and the dynamic viscosity of water, respectively), is highly affected by the soil-solution concentration, C, and composition, expressed in terms of SAR. It has been suggested that the swelling of clay particles in a confined soil system causes the size of large soil pores to decrease, and that breakdown of domains of clay platelets (tactoids) and the subsequent movement and dispersion of the clay platelets may further block soil pores. Low permeability can result from such geometric restrictions. There is a positive correlation between relative soil permeability and the swelling of extracted soil clays, which, in turn, is affected by the soil-solution concentration and composition. Inasmuch as soil-solution composition may affect the breakdown of tactoids (and the subsequent movement and dispersion of clay platelets) on the one hand, and soil-swelling on the other hand, in much the same way, the relative importance of swelling and dispersion in affecting soil permeability remains essentially unresolved. There are situations (relatively coarse-textured soils associated with relatively high influx rates) in which the demixing of adsorbed ions, which concentrates Na on external particle surfaces, may lead to the breakdown of tactoids and the subsequent movement and dispersion of clay platelets.

The effect of concentration and composition of the soil solution on the hydraulic conductivity function $K(\theta)$ in a soil containing 20% clay (dominated by montmorillonite) is depicted in Figure 2. The functional relationships are given in terms of water

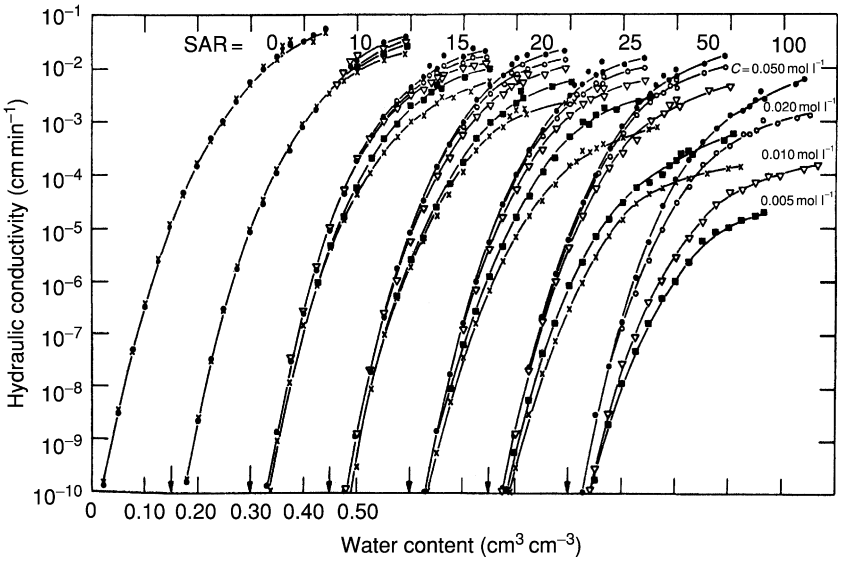


Figure 2 Hydraulic conductivity, K, as a function of water content, θ, for selected values of the concentration, C, and the sodium adsorption ratio, SAR, of the equilibrium solution. Arrows indicate that the point $\theta = 0$ is shifted and the data are translated along the θ-axis. (Reproduced with permission from Russo D and Bresler E (1977) Effect of mixed Na-Ca solutions on the hydraulic properties of unsaturated soils. *Soil Science Society of America Journal* 41: 714–717.)

content, θ, and equilibrium solution concentration, C, and composition, SAR. The points in Figure 2 represent actual measurements for various values of SAR and C; the solid lines represent best visual fits between measured points. Note that the abscissa scale in Figure 2 has been shifted to the right and the data are accordingly translated along the θ-axis to distinguish among different sets of concentrations.

In general, each of the $K(\theta)$ functions in Figure 2 can be represented by a single continuous curve. Furthermore, for the calcium-saturated system (SAR = 0 and $2\,\text{mEq}\,\text{l}^{-1} \leq C \leq 50\,\text{mEq}\,\text{l}^{-1}$), a single-valued $K(\theta)$ is obtained. In the mixed Na-Ca system, however, values of $K(\theta)$ decrease with both θ and C for a given SAR. The differences among the various functions increase as θ becomes larger and as C becomes smaller. These differences are pronounced when SAR and θ are greater than 15 and 0.25, respectively, and when C is equal to or less than $10\,\text{mEq}\,\text{l}^{-1}$. In other words, relatively low values of soil-water content compensate for the negative effects of high SAR and low C on the hydraulic conductivity.

The experimental data depicted in Figure 2 can be explained qualitatively in terms of soil water–clay interactions and the porous nature of the soil. The unsaturated hydraulic conductivity function $K(\theta)$ depends upon the size distribution of water-filled pores and the total water-filled porosity. As mentioned previously, the diffuse double-layer theory for mixed-electrolyte systems predicts that, for a given pressure, the space between clay particles and, concurrently, the amount of water retained by the clay mass increase as SAR increases and C decreases. For a constant-volume system, changes in the volume of the clay mass are at the expense of the quantity and distribution of the soil pores. Thus, for a given pressure head, h, the amount of water retained by the soil increases (Figure 1) and the hydraulic conductivity decreases (Figure 2) as SAR becomes larger and solution concentration, C, decreases. Since the swelling of clay decreases as h increases, the increase in the amount of water retained and the decrease in hydraulic conductivity become smaller as h increases. In addition, both water content and hydraulic conductivity, and, concurrently, water velocity, may considerably decrease with increasing h. Consequently, increasing h may reduce the possibility that movement and dispersion of clay platelets will lead to additional blocking of the pore space available for water flow.

Modeling Hydraulic Conductivity and Water Retentivity in Salt-Affected Soils

In the preceding sections it was shown that in salt-affected soils, both the hydraulic conductivity, $K(h)$, and the water retentivity, $\theta(h)$ may depend on soil-solution concentration and composition. In the present analysis, ion precipitation or dissolution is not taken into account; thus, the simple but credible soil-water system containing only Na, Ca, and Cl ions is considered. The equilibrium solution is quantified in terms of the sum of the molar concentrations of the cations, $C_0 = c_{Na} + c_{Ca}$, and the ratio between the molar concentration of mono- and divalent cations, $R = c_{Na}/c_{Ca}$.

For the modeling of the hydraulic conductivity, $K(h, R, C_0)$, and the water retentivity, $\theta(h, R, C_0)$, functions, it is further assumed: (1) that the soil clay fraction is mostly montmorillonite (having a constant effective surface charge density, Γ_s), which may exist in packets of platelets or domains called tactoids; (2) that the ions on the external surfaces of the tactoids (the 'exchange phase') are distributed according to the planar mixed-ion diffuse double-layer theory; (3) that the number of clay platelets per tactoid depends on the pressure acting on the system and on the exchangeable sodium percentage (ESP); and (4) that, due to the continuity of the pressure in the soil solution, the pressure difference between two adjacent clay particles and the bulk solution (given by the swelling pressure, P (eqn [2]) is equal to the pressure difference across the air–water interface within the soil pores, i.e., the difference between the pore-water pressure, P_w, and the atmospheric pressure, P_a, as given by the capillary pressure, $P_c = P_w - P_a = -\gamma h$.

Calculations of the $b(h, R, C_0)$ and $N(h, R, C_0)$ Functions

For the Na/Ca-Cl soil system considered here, the general equation of the interacting double layer is obtained from the first integration of the one-dimensional Poisson–Boltzmann equation given by:

$$\frac{dY}{dx} = \frac{G}{\sqrt{R+1}}\left[R\left(\exp^{Z_{Na}Y} - \exp^{Z_{Na}Y_d}\right) + \left(\exp^{Z_{Ca}Y} - \exp^{Z_{Ca}Y_d}\right) + (R+2)\left(\exp^{Z_{Cl}Y} - \exp^{Z_{Cl}Y_d}\right)\right]^{1/2} \quad [1]$$

where $Y = \epsilon\psi/kT$ is the scaled electrical potential, Z is the valence of the ion shown subscript, ψ is the electrical potential, ϵ is the charge on the electron, k is the Boltzmann constant, T is the absolute temperature, x is the distance to the charged surface (where $x = s$ and $Y = Y_s$), Y_d is Y at the midplane between two adjacent tactoids (at the plane $x = d$, where $dY/dx = 0$), $G = (8\pi C_0 L\epsilon^2/10^3 DkT)^{1/2}$, L is the Avogadro number, and D is the dielectric constant of the bulk solution.

For dilute solutions, the swelling pressure, P, i.e., the pressure difference between the equilibrium solution (where $Y=0$) and the midplane between two adjacent tactoids (at the plane $x=d$, where $Y=Y_d$), is obtained by substituting the Boltzmann distribution law into the Van't Hoff relation. For the Na/Ca-Cl soil system considered here, with tactoids that are at a distance of $2b$ apart (where $b=d-s$), the swelling pressure, P is given by:

$$P = B\left[R(e^{Z_{Na}Y_d}-1)+(e^{Z_{Ca}Y_d}-1) +(R+2)(e^{Z_{Cl}Y_d}-1)\right] \quad [2]$$

where $B=kLTC_0/10^3(R+1)$. Substituting $P=\gamma h$ in eqn [2], one can obtain, after rearrangement:

$$a_1 e^{Z_{Na}Y_d}+a_2 e^{Z_{Ca}Y_d}+a_3 e^{Z_{Cl}Y_d}+a_4=0 \quad [3]$$

where $a_1=R$, $a_2=1$, $a_3=R+2$, and $a_4=-[a_1+a_2+a_3+10^3\gamma h(R+1)kLTC_0]$.

Thus, the solution of the nonlinear eqn [3] yields Y_d as a function of h, R, and C_0. Estimation of $b(h, R, C_0)$ also requires the knowledge of the scaled electrical potential at the charged surface, Y_s, which can be estimated from the integration of eqn [1] and from the condition of electroneutrality requiring that the surface charge density, Γ_s, be equal to the total space charge density, $\delta(x)$, within the range $s\le x\le d$, thus:

$$\left(\frac{dY}{dx}\right)_{x=s}=\frac{4\pi\pi\Gamma_s}{DkT} \quad [4]$$

Evaluating eqn [1] at the plane $x=s$ (where $Y=Y_s$), using eqn [4], one obtains, after rearrangement:

$$a_1 e^{Z_{Na}Y_s}+a_2 e^{Z_{Ca}Y_s}+a_3 e^{Z_{Cl}Y_s}+a_4=0 \quad [5]$$

where $a_4=-\{[2\Pi 10^3(R+1)\Gamma_s^2/LDkTC_0]+a_1 e^{Z_{Na}Y_d}+a_2 e^{Z_{Ca}Y_d}+a_3 e^{Z_{Cl}Y_d}\}$, and a_i, $i=1, 2, 3$, as before.

The solution of the nonlinear eqn [5] yields Y_s as a function of Γ_s, R, C_0, and $Y_d(h, R, C_0)$. From the values of Y_d and Y_s, the distance from the midplane $x=d$ (where $dY/dx=0$ and $Y=Y_d$) to the charged surface (where $x=s$ and $Y=Y_s$), $b=d-s$, can be obtained by direct integration of eqn [1]. Rearranging eqn [1] yields:

$$\frac{G}{\sqrt{R+1}}\int_d^s dx=\int_{Y_d}^{Y_s}(a_1 e^{Z_{Na}Y}+a_2 e^{Z_{Ca}Y}+a_3 e^{Z_{Cl}Y}+a_4)^{-1/2}dY \quad [6]$$

where $a_4=-[a_i e^{Z_{Na}Y_d}+a_2 e^{Z_{Ca}Y_d}+a_3 e^{Z_{Cl}Y_d}]$ and a_i, $i=1, 2, 3$, as before. Thus a numerical integration of eqn [6] yields $b=d-s$ as a function of h, R, and C_0.

Once $Y_d(h, R, C_0)$ has been estimated from eqn [3], the fractional excess of Na on the exchange phase, Γ_{Na}/Γ_s can be estimated from:

$$\frac{\Gamma_{Na}}{\Gamma_s}=\frac{R\sqrt{C_0}}{\Gamma_s\sqrt{\alpha(R+1)}}\sinh^{-1}\left(\frac{\Gamma_s}{\sqrt{C_0}}\frac{\sqrt{\alpha(R+1)}}{(R+4\cosh Y_d}\right) \quad [7]$$

where Γ_{Na} is the quantity of sodium ionic charges per unit area of the charged surface, and $\alpha=8000\Pi/L\epsilon\kappa T$. Note that, inasmuch as $Y_d=Y_d(h, R, C_0)$, the exchangeable sodium percentage, $ESP=100\Gamma_{Na}/\Gamma_s$ is also a function of h, R, and C_0. For given pressure head, h, and $ESP(h, R, C_0)$, the number of clay platelets per tactoid, $N=N[h, ESP(h, R, C_0)]$, is calculated by using experimental data on the dependence of N on h in a calcium system, and by experimental data on the dependence of N on ESP in clay suspensions.

Calculations of the $\theta(h, R, C_0)$ and $K(h, R, C_0)$ Functions

To estimate the $K(h, R, C_0)$ and the $\theta(h, R, C_0)$ functions, a confined volume element, V, of a homogeneous and isotropic porous medium with uniform porosity, φ^0, is considered. This soil volume, which is large compared with the typical pore size of the porous medium, is subdivided into n equal subvolumes, v_i, $i=1$–n, in such a way that only one value of the average pressure head, h_i, the equivalent radius of the water-filled pores, $r_i^0=2\sigma/\gamma h_i$, (with σ being the surface tension at the air–water interface), and the average number of pores having a unit length, $l_i=\varphi^0 v_i/\Pi(r_i^0)^2$, can be defined in each subvolume, v_i. Using the functions $b(h_i, R, C_0)$ and $N(h_i, R, C_0)$, the water content in each v_i is calculated from:

$$\theta_i(h_i, R, C_0)=\theta_i^0(h_i)+[\nu_i(h_i, R, C_0)-\nu_i^0(h_i)] \quad [8]$$

where θ_i^0 is the water content in a reference state (that is, a soil in equilibrium with a concentrated calcium chloride solution), at pressure head h_i, and ν_i^0 and ν_i are the volumes of water retained by the clay particles in a unit volume of the soil, in the reference state, and in equilibrium with a soil solution of a given R and C_0, given by:

$$\nu_i^0(h_i)=\frac{A_s}{2\,N_i^0(h_i)}\{2b_i^0(h_i)+2b_0[N_i^0(h_i)-1]\}\rho_b \quad [9]$$

and

$$\nu_i(h_i, R, C_0)=\frac{A_s}{2\,N_i^0(h_i)}\{2b_i(h_i, R, C_0)[1+N_i^0(h_i) -N_i(h_i, R, C_0)]+2b_0[N_i(h_i, R, C_0)-1]\}\rho_b \quad [10]$$

respectively, where A_s and ρ_b are the specific surface area and the bulk density, respectively, of the soil at the reference state, the superscript zero refers to the reference state, and $2b_0$ is the thickness of the water film between two adjacent clay platelets within a tactoid ($2b_0 = 9 \times 10^{-8}$ cm, regardless of h, R, or C_0). Similarly, since in the confined soil element the increment in the water retained by the clay fraction is at the expense of the pore space in which water flow takes place, the effective porosity, φ_i, in each subvolume, v_i, is given by:

$$\varphi_i(h_i, R, C_0) = \theta_i^0(h_i) - [\nu_i(h_i, R, C_0) - \nu_i^0(h_i)] \quad [11]$$

and the mean equivalent radius of the pores, r_i, in each subvolume, v_i is given by:

$$r_i(h_i, R, C_0) = \left[\frac{\varphi_i(h_i, R, C_0)v_i}{\pi 1_i}\right]^{1/2} \quad [12]$$

Given the values of $\varphi_i(h_i, R, C_0)$ and $R_i(h_i, R, C_0)$, the relative hydraulic conductivity can be calculated on the basis of the capillary bundle theory. Using a modified series-parallel model of the porous medium, the hydraulic conductivity relative to the reference state is given by:

$$K_i^r(h_i, R, C_0) = \frac{[\varphi_i(h_i, R, C_0)]^2 \sum_{j=i}^{n}[2j-1][r_j(h_j, R, C_0)]^2}{[\varphi_i^0(h_i)]^2 \sum_{j=i}^{n}[2j-1][r_j^0(h_j)]^2} \quad [13]$$

where φ_i^0 is the soil effective porosity in subvolume v_i, in the 'stable' reference state, $h_i < h_{i+1}$, and $i = 1$–n. Thus the soil hydraulic conductivity is given by:

$$K_i(h_i, R, C_0) = K_i^0(h_i)K_i^r(h_i, R, C_0) \quad [14]$$

where $K^0(h)$ is the hydraulic conductivity function of the soil at the 'stable' reference state.

A comparison between the calculated and measured water retention curves in a soil containing 20% clay (dominated by montmorillonite), for different combinations of $C = 10^{-3}\,C_0[1 + (R+1)^{-1}]$ and $\mathrm{SAR} = R[10^3 C_0(R+1)^{-1}]^{1/2}$, is presented in Figure 3. The abscissa scale in Figure 3 is shifted to the right and the data are accordingly translated along the θ-axis to distinguish between the calculated and the measured curves for different sets of concentrations. For calcium-saturated soil (SAR = 0), the experimental retention curves are independent of the solution concentration, whereas the calculated curves are slightly dependent on it. The small discrepancy between the calculated and the measured curves can serve as a measure of the error in the assumptions and

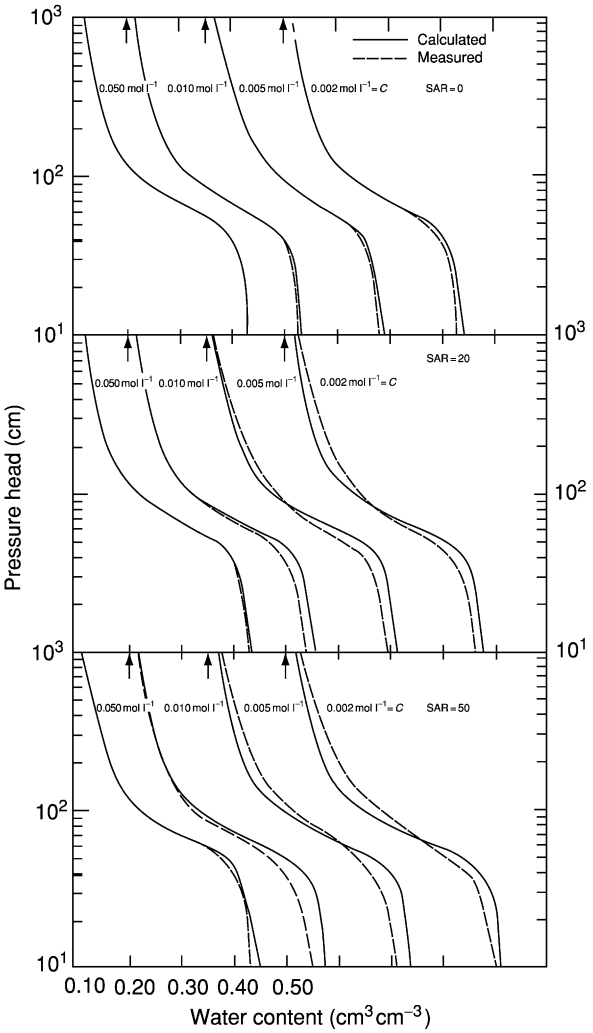


Figure 3 Capillary pressure head, h, as a function of water content, θ, for selected values of the concentration, C, and the sodium adsorption ratio, SAR, of the equilibrium solution. Computed results (solid lines) are compared with measured data (dashed lines). Arrows indicate that the point $\theta = 0.10$ is shifted and the data are translated along the θ-axis. (Reproduced from Russo D (1976) MSc Thesis (in Hebrew), *Analysis of the Saturated–Unsaturated Hydraulic Conductivity in a Mixed Sodium-Calcium Soil System*. Faculty of Agriculture in Rehovot. Jerusalem, Israel: The Hebrew University Press.)

of the ability of the calculations to predict soil-water retention. For the mixed systems, there is reasonably good agreement between measured and calculated retention curves. The agreement improves for lower sodic systems (SAR decreases) and as the system becomes more concentrated.

A comparison between the calculated and measured relative hydraulic conductivity, $K^r = K(\Theta, \mathrm{SAR}, C)/K^0(\Theta)$, where $\Theta = (\theta - \theta_r)/(\theta_s - \theta_r)$ is the effective water saturation, and θ_s and θ_r are the saturated and residual values of θ, respectively, in a soil containing 20% clay (dominated by montmorillonite), for

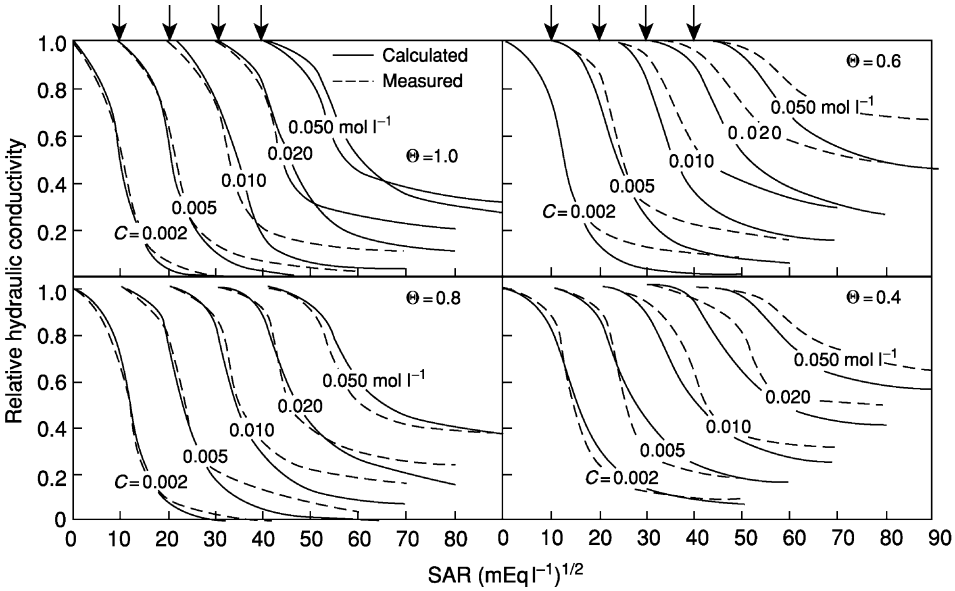


Figure 4 Relative hydraulic conductivity, K^r, as a function of the sodium adsorption ratio, SAR, of the equilibrium solution, for selected values of the solution concentration, C, and degree of water saturation, Θ. Computed results (solid lines) are compared with measured data (dashed lines). Arrows indicate that the point SAR = 0 is shifted and the data are translated along the SAR axis. (Reproduced with permission from Russo D and Bresler E (1977) Analysis of the saturated–unsaturated hydraulic conductivity in a mixed Na-Ca soil system. *Soil Science Society of America Journal* 41: 706–710.)

various combinations of Θ, C, and SAR, is presented in Figure 4. The abscissa scale in Figure 4 is shifted to the right and the data are accordingly translated along the SAR axis to distinguish between the calculated and the measured curves for different sets of concentrations. There is relatively good agreement between calculated and measured data in most cases. For the combinations of Θ, SAR, and C depicted in Figure 4, the calculated K^r deviates by less than twice the value of the measured K^r.

The comparison between calculated and experimental results indicates that a reasonable first-order approximation to the soil-water retentivity and conductivity functions can be obtained for relatively wide ranges of soil-solution concentrations and compositions. This can be done by taking into account the porous nature of the soil, the interactions between the charged solid phase and the ions in the liquid phase within the soil, and the effects of the water pressure and the soil solution composition on the organization and structure of the clay platelets.

Soils of differing textures, containing clay fractions that are mostly montmorillonite, have essentially the same surface charge density ($\Gamma_s \approx 4.5 \times 10^4$ esu cm^{-2}); consequently, they are characterized by essentially the same $\Gamma_{Na}(h, R, C_0)/\Gamma_s$ (eqn [7]) and $\Gamma_{Cl}(h, R, C_0)$ (see eqn [16]) relationships. On the other hand, the $\theta(h, R, C_0)$ and the $K^r(h, R, C_0)$ functions (Figure 5), and, concurrently, the $K(h, R, C_0) = K^0(h)K^r(h, R, C_0)$ function, associated with soils of different texture (see Table 1) may be quite different.

The functions depicted in Figure 5 suggest that the interaction between the soil matrix and the soil solution affects the differences among the various soil types, in their hydraulic properties. In the case of the water retention curves, the differences between the values of θ associated with the various soils generally increase as h and SAR increase and as C decreases. As for the hydraulic conductivity function, the differences between the values of K^r and, concurrently, between values of $K = K^rK^0$, associated with the various soils increase as SAR increases and as h and C decrease. Figure 5 suggests that, because of soil-solution–soil-matrix interactions, the capability of the coarse-textured soil to transmit water and solutes faster than the finer-textured soils is extended to a larger pressure head (smaller degree of saturation) as the soil solution SAR increases and as C decreases.

Both the experimental and the theoretical results presented in the previous sections suggest that, for a given soil solution concentration and composition, the effect of the soil matrix–soil solution interactions on the hydraulic conductivity decreases with decreasing water saturation. This is demonstrated in Figure 6, in which combinations of soil-solution concentration (C) and composition (SAR) required to maintain $K^r \geq 0.75$ (that is, no more than a 25%

reduction in K relative to the 'stable' reference state) are given for various degrees of water saturation, and for soils of several textures (Table 1). Note that the curves in Figure 6 generalize the concept of

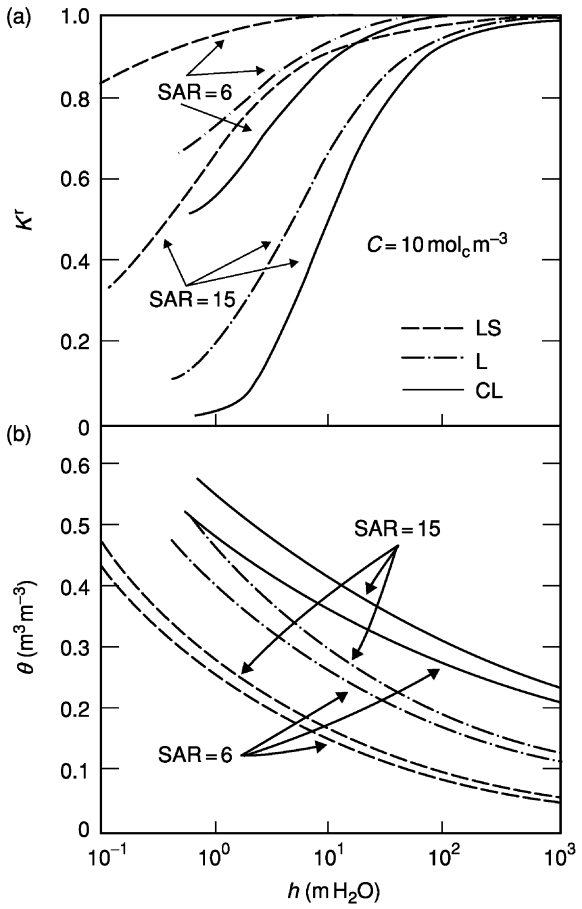


Figure 5 Relative hydraulic conductivity, K^r (a) and water content, θ (b) as a function of the capillary pressure head, h, for selected values of the concentration, C, and the sodium adsorption ratio, SAR, of the equilibrium solution, for three soils of different texture: loamy sand (LS), loam (L), and clay loam (CL). (Reproduced with permission from Russo D (1988) Numerical analysis of the nonsteady transport of interacting solutes through unsaturated soil. 1. Homogeneous systems. *Water Resources Research* 24: 271–284. American Geophysical Union.)

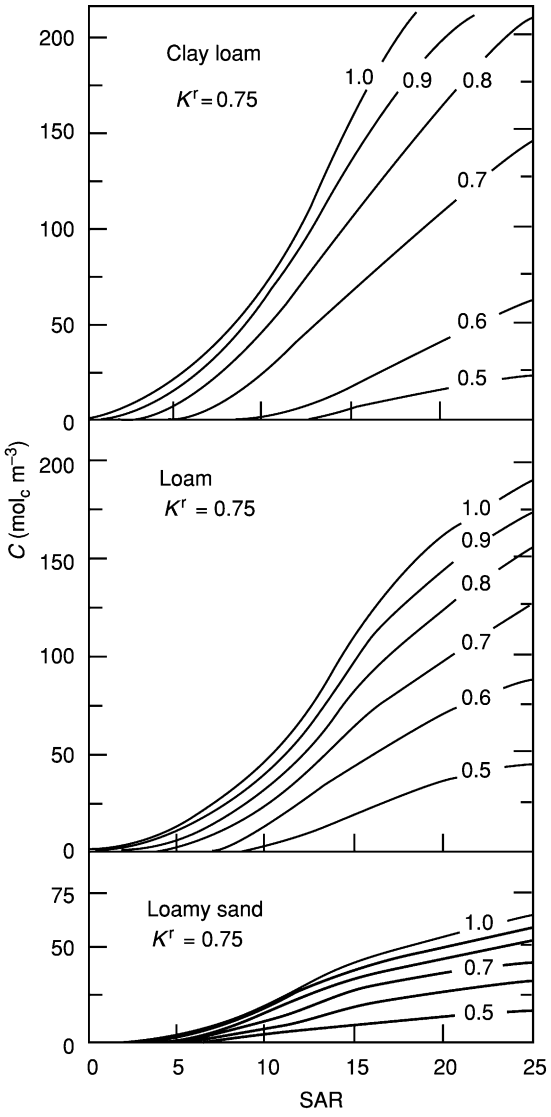


Figure 6 Combinations of the concentration, C, and the sodium adsorption ratio, SAR, of the equilibrium solution, and degree of water saturation, Θ (the numbers labeling the curves) at which 25% reduction in hydraulic conductivity occurs, for three soils of different texture: loamy sand (LS), loam (L), and clay loam (CL). (Reproduced with permission from Russo D (1988) Numerical analysis of the nonsteady transport of interacting solutes through unsaturated soil. 1. Homogeneous systems. *Water Resources Research* 24: 271–284. American Geophysical Union.)

Table 1 Representative values for relevant soil physical properties

Soil texture	*Clay fraction*	A_s ($cm^2\ g^{-1}$)	*CEC* ($mEq\ g^{-1}$)	K_s ($cm\ h^{-1}$)	h_b (*cm*)	β	θ_s ($cm^3\ cm^{-3}$)	θ_r ($cm^3\ cm^{-3}$)	ρ_b ($g\ cm^{-3}$)
Loamy sand	0.06	3.6×10^5	0.056	5.56	9.1	0.30	0.410	0.015	1.53
Loam	0.19	9.5×10^5	0.145	2.50	48.0	0.23	0.451	0.042	1.43
Clay loam	0.34	1.7×10^6	0.248	0.77	63.4	0.15	0.476	0.068	1.33

A_s, soil specific surface area; CEC, cation exchange capacity; ρ_b, bulk density; K_s, soil saturated hydraulic conductivity; θ_s, saturated water content; θ_r, residual water content; h_b, soil bubbling pressure head; β, pore-size distribution index.

Reproduced with permission from Russo D (1988) Numerical analysis of the nonsteady transport of interacting solutes through unsaturated soil. 1. Homogeneous systems. *Water Resources Research* 24: 271–284. American Geophysical Union.

the 'threshold concentration' in terms of soil properties and soil-water status. Consequently, they may be used for water-quality classification in relation to soils of different textures, and for water and soil management. For a given SAR, the solution concentration required to maintain $K^r \geq 0.75$ (that is, the 'threshold concentration') decreases as the soil-water content decreases and as the soil texture becomes coarser. **Figure 6** suggests that a decrease in the water application rate and, concurrently, in water saturation, may improve the efficiency of the reclamation of sodic soils, especially fine-textured ones.

Implications for Flow and Transport

Soil-solution–soil-matrix physicochemical interactions, enhanced by dilute soil solutions with relatively high sodium-to-calcium ratios, can affect the flow parameters considerably (i.e., hydraulic conductivity and water retention functions). Furthermore, these interactions may affect the transport parameters (i.e., retardation and elution factors). In other words, soil-solution–soil-matrix physicochemical interactions may affect both water flow and solute transport.

Generally, for a given soil, and given boundary and initial conditions, the water flow is governed by the soil hydraulic properties. Thus, the effect of the soil-matrix–soil-solution interactions is to reduce the water flow relative to that of a 'stable' reference soil. For a given soil, the retardation of the water movement by the soil matrix–soil solution interaction will increase as either the initial soil solution SAR or the SAR of the applied water increases, as either the initial soil-solution concentration or the solution concentration of the applied water decreases, and as the surface water application rate increases.

On the other hand, the solute transport is governed by the water flow as well as by cation exchange and anion exclusion processes. Thus, in this case, the soil-matrix–soil-solution interactions may act in two opposing directions: the retardation of the solute movement because of reductions in the soil hydraulic conductivity and because of cation exchange, and the acceleration of the solute movement because of anion exclusion. The retardation effect caused by the reductions in the soil hydraulic conductivity and, concurrently, in water velocity, increases with increasing SAR and decreasing C and h, especially in relatively fine-textured soils. The retardation effect because of Na/Ca exchange is given by the retardation factor, R_f:

$$R_f(h, R, C_0) = 1 + \frac{[R+1]\Gamma_{Na}(h, R, C_0)FA_{ad}\rho_b}{Z_{Na}R\,C_0\theta(h)} \quad [15]$$

where F is the Faraday constant, A_{ad} is the specific adsorption surface area of the soil, and Γ_{Na} is the quantity of sodium ionic charges per unit area of the charged surface (eqn [7]). The retardation caused by Na/Ca exchange (expressed in terms of R_f) increases with increasing SAR and decreasing C and h, especially in relatively fine-textured soils.

Anion exclusion can be calculated from the distribution of the scaled electrical potential, $Y(x)$ (eqn [1]) within the domain $s \leq x \leq d$, which, in turn, for given values of $b(h, R, C_0)$ and $Y_s(h, R, C_0)$, is obtained from the solution of eqn [1], by the Runge–Kutta method. Applying the Boltzmann distribution law for the Na/Ca-Cl system considered here, the amount of excluded chloride per unit surface area is calculated from:

$$\Gamma_{Cl}(h, R, C_0) = \frac{|Z_{Cl}|C_0(R+2)}{(R+1)} \times \int_s^d \{1 - \exp[-Y(x; h, R, C_0)]\}dx \quad [16]$$

which, in turn, is used to calculate the elution factor, E_f, given by:

$$E_f(h, R, C_0) = 1 - \frac{[R+1]\Gamma_{Cl}(h, R, C_0)A_{ex}\rho_b}{|Z_{Cl}|C_0[R+2]\theta(h)} \quad [17]$$

where A_{ex} is the specific exclusion surface area.

The acceleration effect due to the chloride exclusion, which is determined by the second term on the right-hand side of eqn [17], increases with increasing SAR and decreasing C and h, especially in relatively fine-textured soils. The net effect of the soil matrix–soil solution interactions on the solute transport is determined, therefore, by the relative magnitudes of all of these processes, which, in turn, depend on the soil type and the boundary and initial conditions imposed on the flow system.

The effect on water flow and solute transport in a salt-affected soil of the conditions at the soil surface pertinent to an irrigation event and a rain event is demonstrated in **Figures 7 and 8**, respectively. The simulated profiles of water content, θ (top), salinity (in terms of electrical conductivity, EC; middle), and hydraulic conductivity, K (bottom) in **Figures 7 and 8** are presented for three different soils ignoring (**Figures 7a and 8a**) and considering (**Figures 7b and 8b**) soil solution–soil matrix interaction. **Figures 7 and 8** clearly demonstrate that in, salt-affected soils, the retardation of water and solute movement by soil matrix–soil solution interaction increases as the soil texture is finer and the clay fraction is larger, particularly during a rain storm (**Figure 8**) associated with diluted solution and relatively high application rate.

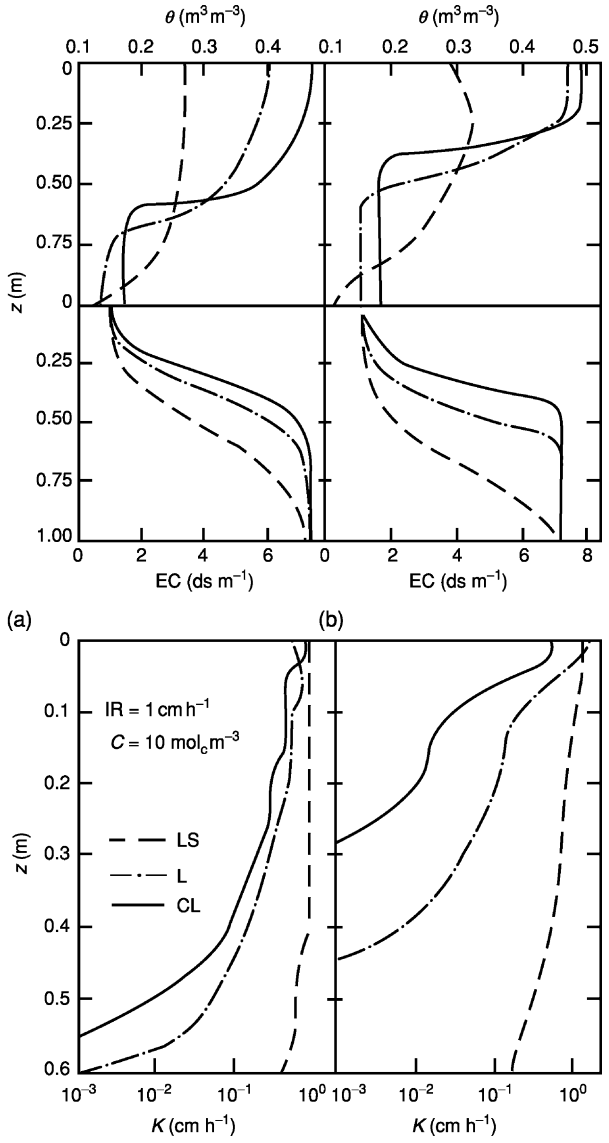


Figure 7 Profiles of simulated soil-water content, θ, soil-solution electrical conductivity, EC, and soil hydraulic conductivity, K, after 12 h of infiltration (irrigation), and for the three different soils (Table 1). Initial profiles are given by $\theta(z, 0) = \theta_s - 0.3$, $C(z, 0) = 70$ mEq l^{-1}, and SAR$(z, 0) = 25$ (mEq l^{-1})$^{1/2}$. (Reproduced with permission from Russo D (1988) Numerical analysis of the nonsteady transport of interacting solutes through unsaturated soil. 1. Homogeneous systems. *Water Resources Research* 24: 271–284.)

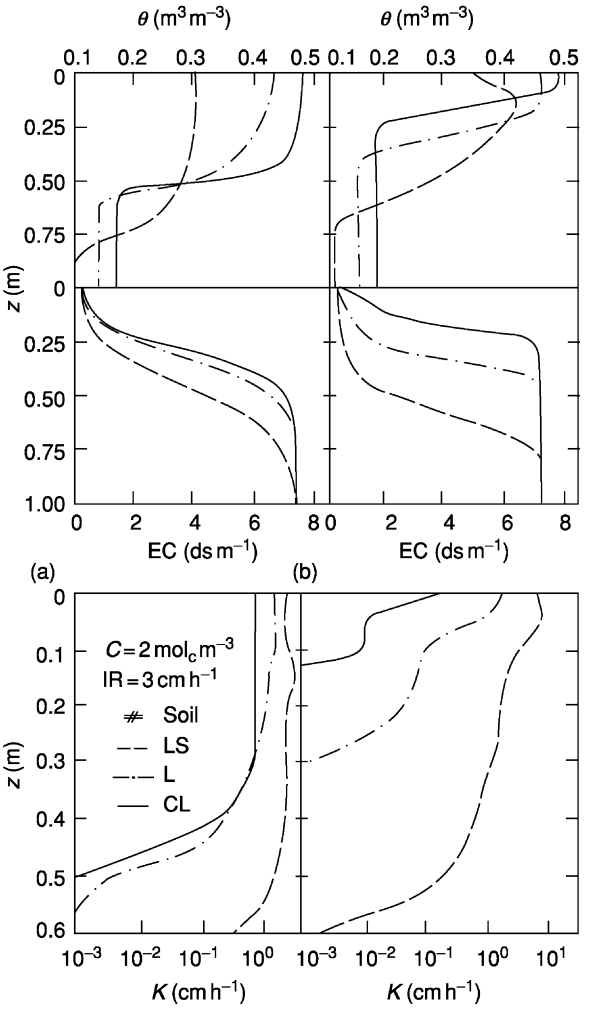


Figure 8 Profiles of simulated soil-water content, θ, soil-solution electrical conductivity, EC, and soil hydraulic conductivity, K, after 4 h of infiltration (rain), and for the three different soils (Table 1). Initial profiles are given by $\theta(z, 0) = \theta_s - 0.3$, $C(z, 0) = 70$ mEq l^{-1}, and SAR$(z, 0) = 25$ (mEq l^{-1})$^{1/2}$. (Reproduced with permission from Russo D (1988) Numerical analysis of the nonsteady transport of interacting solutes through saturated soil. 1. Homogeneous systems. *Water Resources Research* 24: 271–284.)

Summary

Soil solution–soil matrix interactions can alter the geometry of the soil pores and thus may affect the soil hydraulic conductivity, $K(h)$, and water retentivity, $\theta(h)$. The effect of soil solution–soil matrix interactions on $K(h)$ and $\theta(h)$, the retardation factor, R_f, and the elution factor, E_f, and, concurrently, on water flow and solute transport may be significant and generally increases as the soil solution sodium adsorption ratio increases; as both the soil solution concentration and the capillary pressure head decrease; as the clay fraction of the soil increases; and as the soil texture becomes finer.

It is worthwhile emphasizing that the experimental and the theoretical findings discussed here are relevant to soils whose clay fraction is dominated by smectite minerals (e.g., montmorillonite), with solutions containing Na, Ca, and Cl ions only. Furthermore, the model of the hydraulic functions in salt-affected soils presented here is based on a simplified description of the porous medium (the series-parallel model), and of the physicochemical interactions between the soil solution and the soil

matrix (the mixed-ion diffuse double-layer theory). The model concentrates on the change in the soil pore-size distribution caused by the swelling of clay particles in a confined soil system, which, in turn, may be magnified by the breakdown of the domains of the clay platelets (tactoids) when the soil ESP exceeds a critical value. However, the model disregards the subsequent dispersion and movement of the clay platelets which may occur in coarse-textured soils subjected to a relatively large influx rate and, which, in turn, may further block the soil pores. Nevertheless, the data presented here are sufficiently reliable to indicate appropriate trends.

List of Technical Nomenclature

$\alpha = 8000\Pi/LDkT$	(eqn [7]) ($erg\,mol^{-1}$)
β	pore-size distribution index (dimensionless)
Γ_{Cl}	quantity of the excluded chloride per unit area of the charged surface ($mEq\,cm^{-2}$)
Γ_{Na}	quantity of sodium ionic charges per unit area of the charged surface ($esu\,cm^{-2}$)
Γ_s	surface charge density ($esu\,cm^{-2}$)
γ	specific gravity of water ($g\,cm^{-2}\,s^{-2}$)
δ	space charge density ($esu\,cm^{-3}$)
ϵ	electronic charge (4.803×10^{-10} esu electron^{-1})
η	dynamic viscosity of water ($g\,cm^{-1}\,s^{-1}$)
Θ	effective water saturation (dimensionless)
θ	volumetric water content ($cm^3\,cm^{-3}$)
θ_s, θ_r	saturated and residual values of Θ ($cm^3\,cm^{-3}$)
ℓ	average number of pores having a unit length (cm)
κ	permeability (cm^2)
ν	volume of water retained by clay particles in a unit volume of soil ($cm^3\,cm^{-3}$)
ρ_b	bulk density ($g\,cm^{-3}$)
σ	surface tension at the air–water interface ($erg\,cm^{-2}$)
φ	porosity ($cm^3\,cm^{-3}$)
ψ	electrical potential ($erg\,esu^{-1}$)
A_d	specific surface area for adsorption ($cm^2\,g^{-1}$)
A_e	specific surface area for exclusion ($cm^2\,g^{-1}$)
A_s	specific surface area ($cm^2\,g^{-1}$)
a_i(i = 1–4)	coefficients in eqns [3], [5], and [6] (dimensionless)
$B = kLTC_0/10^3(R+1)$	(eqn [2]) ($erg\,cm^{-3}$, where $1\,erg = 1\,g\,cm^2\,s^{-2}$)
b	half-distance between two adjacent tactoids (cm)
C_{Na}	concentration of the ion identified subscript ($mEq\,l^{-1}$)
$C = C_{Na} + C_{Ca}$	($mEq\,l^{-1}$)
$C_0 = c_{Na} + c_{Ca}$	($mol\,l^{-1}$)
c_{Na}	molar concentration of the ion identified subscript ($mol\,l^{-1}$)
D	the dielectric constant of water (dimensionless)
d	position of the midplane between two adjacent tactoids (cm)
E_f	elution factor (dimensionless)
F	Faraday constant (2.893×10^{11} esu mEq^{-1}, where $1\,esu = 1\,g^{1/2}\,cm^{3/2}\,s^{-1}$)
$G = (8\Pi C_0 Le^2/10^3 DkT)^{1/2}$	(molecules electron^{-1} cm^{-1})
h	capillary pressure head (cm)
h_b	air entry value of h (cm)
i, j	indices (dimensionless)
K	hydraulic conductivity ($cm\,s^{-1}$)
K^r	relative hydraulic conductivity (dimensionless)
k	Boltzmann constant (1.38×10^{-16} erg deg^{-1} molecule^{-1})
L	Avogadro number (6.0225×10^{23} molecules mol^{-1})
P	swelling pressure ($erg\,cm^{-3}$)
P_w, P_a, and $P_c = P_w - P_a$,	water, air, and capillary pressures, respectively ($erg\,cm^3$)
$R = c_{Na}/c_{Ca}$	(dimensionless)
R_f	retardation factor (dimensionless)
r	equivalent radius of water-filled pore (cm)
s	position of the plane of the charged surface (cm)

SAR	sodium adsorption ratio $(C_{Na}/\sqrt{C_{Ca}/2})((\text{mEq l}^{-1})^{1/2})$
T	absolute temperature (degrees Kelvin)
V	volume (cm^3)
v_i	subvolume ($\Sigma_{V_i} = V$) (cm^3)
x	distance (cm)
Y	scaled electrical potential (molecule electron^{-1})
Y_d	value of Y at the midplane between two adjacent tactoids (molecule electron^{-1})
Y_s	value of Y at the charged surface (molecule electron^{-1})
Z_{Na}	valence of the ion subscript (electron molecule^{-1})

See also: **Nitrogen in Soils:** Cycle; Nitrates; Plant Uptake

Further Reading

Bear J (1972) *Dynamics of Fluids in Porous Media*. New York: Elsevier.

Blackmore AV and Miller RD (1961) Tactoid size and osmotic swelling in Ca-montmorillonite. *Soil Science Society of America Proceedings* 25: 169–173.

Bresler E (1972) Interacting diffuse layers in mixed monodivalent ionic systems. *Soil Science Society of America Proceedings* 36: 891–896.

Bresler E (1978) Theoretical modeling of mixed-electrolyte solution flows for unsaturated soils. *Soil Science* 125: 196–203.

Bresler E, McNeal BL, and Carter DL (1982) *Saline and Sodic Soils: Principles-Dynamics Modeling*. Berlin, Germany: Springer-Verlag.

Brutsaert W (1967) Some methods of calculated unsaturated permeability. *Transactions of the American Society of Agricultural Engineers* 10: 400–404.

Childs EC and Collis-George N (1950) The permeability of porous materials. *Proceedings of the Royal Society of London, Series A: Mathematical and Physical Sciences* 201: 392–405.

James RV and Rubin J (1986) Transport of chloride ion in a water-unsaturated soil exhibiting anion exclusion. *Soil Science Society of America Journal* 50: 1142–1149.

Marshall TJ (1958) A relation between permeability and size distribution of pores. *Journal of Soil Science* 9: 1–8.

Nielsen DR, van Genuchten MTh, and Biggar JW (1986) Water flow and solute transport processes in the unsaturated zone. *Water Resources Research* 22: 89S–108S.

Russo D (1988) Numerical analysis of the nonsteady transport of interacting solutes through unsaturated soil. 1. Homogeneous systems. *Water Resources Research* 24: 271–284.

Russo D and Bresler E (1977) Effect of mixed Na-Ca solutions on the hydraulic properties of unsaturated soils. *Soil Science Society of America Journal* 41: 714–717.

Russo D and Bresler E (1977) Analysis of the saturated–unsaturated hydraulic conductivity in a mixed Na-Ca soil system. *Soil Science Society of America Journal* 41: 706–710.

Shainberg I (1984) The effects of electrolyte concentration on the hydraulic properties of sodic soils. In: Shainberg I and Shalhevet J (eds) *Soil Salinity Under Irrigation: Processes and Management*, pp. 49–64. New York: Springer-Verlag.

Shainberg I and Kaiserman A (1969) Kinetics of the formation and breakdown of Ca-montmorillonite tactoids. *Soil Science Society of America Proceedings* 33: 547–551.

Salt Balance of Soils

See **Salination Processes**

SALT-AFFECTED SOILS, RECLAMATION

R Keren, The Volcani Center, Bet Dagan, Israel

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Saline and sodic soils exist in arid and semiarid regions and in regions of poor natural drainage. These soils have poor physical properties owing to high sodicity and high pH. Under these conditions, slow water infiltration into the soil, soil erosion, slow internal drainage, poor aeration, compaction, and waterlogging on lower lands commonly occur due to soil swelling and dispersion.

Considerable progress has been made in managing and controlling salinity, sodicity, and alkalinity in irrigated lands. However, because of natural hydrological and geochemical factors, as well as irrigation-induced activities, soil salinity, sodicity, alkalinity, and associated drainage and soil erosion continue to plague agriculture. Control of salinity and sodicity is essential to establish a sustainable, successful agriculture.

Effect of Salts on Soil Properties

Hydraulic Conductivity

Many soils in arid to humid regions have an unstable structure, which makes them difficult to manage because of their tendency to swell and disperse. Since soil permeability decreases with the square of the pore radius, a small reduction in size of the large pores due to swelling and clay migration has a large effect on soil permeability for water and gases. The extent of swelling and dispersion of clays depends on the clay mineralogy, the composition of the adsorbed ions, and the electrolyte concentration (EC) of the soil solution. Whether clay will migrate or seal the soil depends on the extent of swelling. Clay swelling is a continuous process that increases gradually with decreases in EC, whereas clay dispersion is possible only at EC below the flocculation value (FV). The degree of clay swelling before replacing the solution with one that has an EC below the FV of the clay determines whether clay will migrate in the conducting pores, or whether the clay particles will be trapped in the narrow pores (sieving effect) and the hydraulic conductivity (HC) of the soil will further decrease.

Even in arid and semiarid regions where irrigation is practiced, rainstorms occur occasionally. Under such conditions, sodicity damage to the soil will not take place during the irrigation season when the EC in the applied water may be high enough to counter the dispersive effect of the exchangeable Na. The sodicity damage will occur during the rainy season when the EC decreases to values below the threshold concentration determined by the properties of the soil and the exchangeable sodium percentage (ESP). Thus, when soils are exposed to rainwater, susceptibility to low levels of ESP is enhanced. It is obvious, therefore, that any attempt to assess the sodicity hazard in soils, the clay mineralogy, and the total EC of percolating solution should be considered. At any particular ESP, the influence of exchangeable Na is greater with exchangeable Mg as the complementary ion than it is with Ca.

Infiltration Rate

In soils having stable surface structures, decreases in infiltration rate (IR) result from the inevitable decrease in the matric suction gradient which occurs as infiltration proceeds. Decreases in soil IR from an initially high rate can also result from gradual deterioration of soil structure and the formation of a surface seal. When a seal of very low HC is formed, its reduced permeability determines the IR of the soil, and the wetting-front depth has only a slight effect on it. The presence of a seal increases detachment forces owing to runoff, but at the same time increases the shear strength of the soil surface. Thus, the net effect depends on the forces acting between soil particles at soil surface and the energy involved in water-drop impact and shearing forces of the overland flow.

The IR is much more sensitive to ESP than is HC. The higher sensitivity of the soil surface to ESP when the soil surface is exposed to rainwater is explained by: (1) the mechanical impact of the raindrops, which enhances clay dispersion; (2) the absence of the soil matrix, which slows clay transport; and (3) the low EC in the soil surface. Low EC and high ESP cause clay swelling and dispersion, which enhance seal formation and therefore runoff. Adsorbed Mg has a specific effect on soil erosion and IR for montmorillonitic, noncalcareous and calcareous soils.

Reclamation of Salt- and Sodium-Affected Soils

Reclamation of Salt-Affected Soils

Salinity control in the root zone is one of the main problems of irrigation in arid and semiarid regions. In these regions a salinity problem caused or complicated

by poor drainage cannot be adequately controlled unless the water table is stabilized and maintained at a safe depth. This requires open or tile drains, or drainage wells to remove a part of the subsurface saline water and transport it to an acceptable sink for safe disposal.

Effective salinity control depends largely on the initial soil salinity, the content of the soil water during leaching, hydrodynamic dispersion, spatial variation, the salt tolerance of the crops, the depth of the water table and, when applicable, layout of the drainage system. The prevailing concept in earlier reclamation trials was to leach all excess salts from the entire root zone. Under well-drained conditions, however, the aim should be to reduce salinity in the upper soil layer (depending on crop type) to below the threshold value of the growing crop.

Leaching Methods

Continuous and intermittent ponding The efficiency of salts leaching can be defined as the quantity of soluble salts leached per unit volume of water applied. The time factor should also be considered when leaching takes place under conditions of high evaporation. The range of soil pore-size distribution also affects the leaching efficiency. Pore sizes may vary enough to cause substantial changes in hydrodynamic dispersion, especially during ponded leaching. The hydrodynamic dispersion decreases under unsaturated flow, resulting in the more efficient transport of salt. Thus, the efficiency increases as the soil water content decreases during leaching.

On the basis of leaching theory, a water volume of one replacement of pore volume (pv) will reduce soil salinity by one-half, and a pv of 1.5–2.0 of water should remove approximately 80% of the soil salt content. This volume corresponds approximately to an equivalent depth of water per unit depth of soil (assuming a pv of 50%).

Based on field data from various parts of the world, the following empirical formula for salt transport efficiency under one-dimensional leaching is:

$$\frac{C}{C_0} = \frac{k}{(D/D_S)} \qquad [1]$$

where C and C_0 are the salt concentration in the soil at any given time during leaching and the initial concentration, respectively, D and D_S are the depth of leaching water applied and the soil depth to be leached, respectively, and k is an empirical coefficient that ranges from 0.1 to 0.3 for sandy loam and clay soil, respectively, under ponding leaching as shown in Figure 1. Under intermittent ponding with 50–150 mm of water per application or under sprinkler irrigation, the empirical coefficient is approximately 0.1, irrespective of the soil type. This equation applies in the range of $D/D_S > k$. The salt transport efficiency decreases sharply when D/D_S exceed 0.5 or 0.75 for sandy loam and for clay loam to clay soil, respectively.

When natural drainage conditions are limited and an artificial drainage system is required to remove leaching water, leaching efficiency may decrease because the hydraulic gradient is steeper near the drains than away from the drains. Therefore, a larger fraction of the leaching water will flow through the soil next to the drains. This amount depends on the drain spacing (S) and the depth (Y). For example, when $S/Y = 2.5$, only 30% of the area between drains will be leached with maximum efficiency. To increase leaching efficiency, drain spacing and depth should be determined by the use of a drainage equation.

The intermittent ponding method is especially suitable in fields with tile drains, as it allows the water table to draw down, which greatly increases leaching efficiency. When soils develop a surface seal, intermittent ponding may also help water to infiltrate by forming cracks. However, the longer period of wet soil exposure increases water evaporation. Combining intermittent ponding with mulching greatly improves the leaching efficiency.

When fields consist of several types of soils, leaching takes place in sandy soils at rates different to those of more clayey soils. If leaching is continued until the clay soil is sufficiently leached, an excessive amount of water percolates into the sandy soil. This problem is pronounced under ponded leaching, during which ponded water percolates predominantly through the sandy soil. To leach much more efficiently, it is possible to use the stochastic analysis of solute transport in spatially variable fields to divide a field into small sections that need different amounts of water.

Sprinkling As mentioned above, the leaching efficiency increases as the soil water content decreases. Thus, leaching by sprinkling at rates lower than the infiltration rate of the soil will result in a more efficient removal of accumulated salts than will be leached by ponding. Ponded leaching requires three times as much water as sprinkling to reduce soil salinity by the same increment. Sprinklers can be used advantageously when the field is unprepared for ponded leaching or for sandy soils. This method has a high leaching efficiency, except under high evaporation, because the leaching takes place under unsaturated conditions. This method can also be used advantageously when water for leaching is scarce and costly.

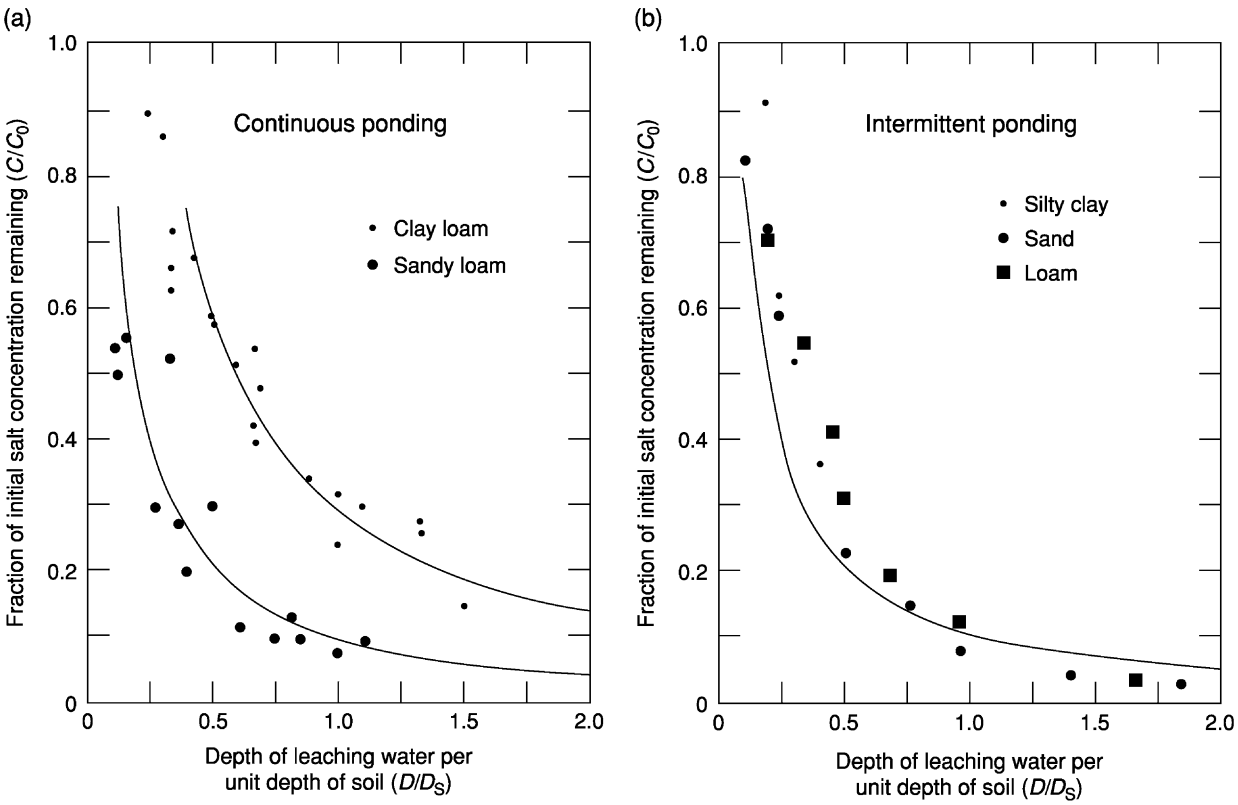


Figure 1 Relationship between fraction of initial salt concentration remaining in soil, C/C_0, and depth of leaching water applied per unit depth of soil, D/D_S: (a) for continuous ponding; (b) for intermittent ponding. (Adapted from Hoffman GJ (1980) Guidelines for the reclamation of salt-affected soils. In: O'Connor GA (ed.) *Second Inter-American Conference on Salinity and Water-Management Technology*, pp. 49–64. Juarez, Mexico.)

In summary, the intermittent-ponding method requires less water than the continuous-ponding method to achieve the same degree of leaching, and the sprinkling method is more efficient than the other methods at removing salt from small pores in the soil profile.

Postleaching operation Prolonged ponding destroys the structure of the soil. Therefore, the last phase of reclamation is drying the soil and, when necessary, taking steps to enhance soil structure, such as applying gypsum or organic matter. Sprinklers or intermittent ponding generally cause less soil-structure destruction and may not require elaborate postleaching measures.

Reclamation of Sodium-Affected Soils

In order to reduce runoff and erosion in sodic soils, reclamation is a common practice. Reclamation of sodic soils involves replacement of exchangeable Na by Ca. The source of Ca for replacing adsorbed Na can be external sources such as gypsum, calcium chloride, and irrigation water containing Ca ions, or the soil itself, involving the dissolution of Ca-containing minerals.

Although chemical amendments may be effective in restoring water penetration after reduction caused by either excessive adsorbed sodium or a low salt concentration in the applied water, they will not be effective if low permeability is due to soil texture, compaction, or water-restricting layers. Thus, reclamation requires a certain flow of water through the soil profile, and an appropriate IR and HC must be achieved.

Reclamation by Adding Soil Amendments

Gypsum Gypsum is the most commonly used amendment for sodic soil reclamation and for reducing the harmful effects of high-sodium irrigation waters because of its solubility, low cost, and availability. Gypsum added to a sodic soil can cause permeability changes by increasing EC and by cation exchange effects. The relative significance of the two effects is of interest for several reasons. If the electrolyte effect is sufficiently high to prevent dispersion and swelling of soil clays, the surface application of

gypsum may be worthwhile. In this case the amount of gypsum required depends on the amount of high-quality water applied and the rate of gypsum dissolution. It is somewhat independent of the amount of exchangeable Na in the soil profile. Conversely, in soils where the EC effect is not enough due to the high ESP level, or in order to obtain permanent improvement, then the amount of gypsum required depends on the amount of exchangeable Na in a selected depth of soil.

The benefit of gypsum for reclaiming sodic soils depends not only on the infiltration characteristics of the soil, but also on the gypsum-dissolution properties. The main mechanism by which gypsum affects the IR of soils exposed to rainwater is by its dissolution and release of electrolytes into the soil solution. The dissolution rate of gypsum is an important factor in IR because of the short contact time between rainwater and gypsum particles at the soil surface. Some of the factors that influence the dissolution rate of gypsum are the surface area of the gypsum particles, the soil water velocity during leaching, and the electrolyte composition of the soil solution.

Gypsum source, application rate, and particle size all have their effect on crust formation and IR (Figure 2). Although the IR for soil without gypsum decreases sharply as the cumulative rainfall increases, the IR remains relatively high in the presence of phosphogypsum (PG), whereas the effectiveness of mined gypsum (MG) on IR is significantly less. The higher efficiency of PG in maintaining high IR is attributed to its high rate of dissolution (Table 1). Application of gypsum on sodic calcareous soil is very effective in lowering surface runoff under field conditions (Table 2). Spreading the gypsum over the soil is more effective than mixing it in the upper 10 cm of soil. Although gypsum affects the ESP of the soil in the upper 0- to 15-cm layer very little (Table 3), it has a large effect on IR (Figure 2). The gypsum has a significant effect on the amount of soil erosion. The effectiveness is greater when the gypsum particles are spread over the soil rather than mixed in the top 10 cm of soil. The presence of gypsum on the soil surface increases soil stability by increasing EC in the percolated solution and by decreasing ESP. This effect of gypsum on soil erosion is due to two reasons: (1) reducing runoff by increasing IR, and (2) changing the erodibility factor. The erodibility of the soil is essentially related to the stability of soil aggregates and to the cohesive forces holding the aggregates together.

In order to obtain permanent improvement, however, the cation exchange effect is most important. The amount of exchangeable sodium to be replaced during reclamation depends on the initial ESP (ESP_i), the cation exchange capacity (CEC; in moles of cation charge per megagram), soil bulk density (ρ_b; in megagrams per cubic meter), the desired final exchangeable sodium fraction (ESP_f) and the depth of soil to be reclaimed (L; in meters). Once the above parameters have been determined, the amount of exchangeable Na to be replaced per unit area of land (Q_{Na}; in moles of cation charge per hectare) can be calculated from:

$$Q_{Na} = 10^4 L \rho_b (CEC)(ESP_i - ESP_f) \qquad [2]$$

The ESP_f value depends on the response of the soil in terms of its physical conditions.

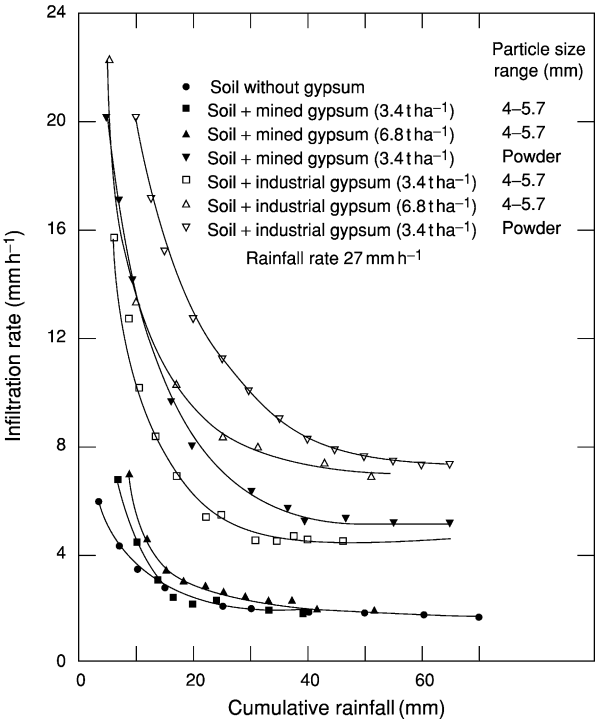


Figure 2 Effect of industrial and mined gypsum (at amounts and fragment sizes as indicated) on infiltration rate of loess soil as a function of cumulative rainfall. (Adapted from Keren R and Shainberg I (1981) The efficiency of industrial and mined gypsum in reclamation of a sodic soil – rate of dissolution. *Soil Science Society of America Journal* 45: 103–107, with permission.)

Table 1 The dissolution coefficients of industrial gypsum, $K_{(PG)}$ and mined gypsum, $K_{(MG)}$, with different fragment sizes

	Particle size (mm)	
Dissolution coefficients (s^{-1})	*1.0–2.0*	*4.0–5.7*
$K_{(PG)} \times 10^4$	198	58
$K_{(MG)} \times 10^4$	19	6
$K_{(PG)}/K_{(MG)}$	10.4	9.7

Adapted from Keren R and Shainberg I (1981) The efficiency of industrial and mined gypsum in reclamation of a sodic soil – rate of dissolution. *Soil Science Society of America Journal* 45: 103–107, with permission.

Table 2 Surface runoff from loess soils at two levels of exchangeable sodium percentage (ESP) during rainstorms as affected by phosphogypsum (PG)

			Surface runoff (% of rainfall)[a]				
			ESP 4.6		*ESP 19.3*		
				PG[b]		*PG*[b]	
Storm no.	*Rainfall (mm)*	*Time between rainstorms (days)*	*Control*	*spread over*	*Control*	*Mixed in*	*Spread over*
1	16	-	0	0	23.1	2.6	0
2	20	40	6.5	0	13.3	3.0	0.7
3	19	20	1.3	0.5	27.3	10.5	3.8
4	59	6	21.5	3.8	41.0	21.2	17.2
5	44	9	13.4	2.1	45.0	22.5	12.4
6	12	1	12.5	2.5	40.0	25	12.5
7	12	15	10.1	2.5	31.7	13.8	6.7

[a]Percentage of annual rainfall (182 mm): ESP 4.6: Control, 12.5%; PG spread over, 2.1%; ESP 19.3: Control, 35.3%; PG spread over, 10.3%; PG mixed in, 16.5%.
[b]The PG rates were 5 Mg ha^{-1} and 10 Mg ha^{-1} for the soils having ESP values of 4.6 and 19.3, respectively.
Adapted from Keren R, Shainberg I, Frenkel H, and Kalo Y (1983) The effect of exchangeable sodium and gypsum on surface runoff from loess soil. *Soil Science Society of America Journal* 47: 1001–1004, with permission.

Table 3 Exchangeable sodium percentage (ESP) in the profile of the sodic loess soil as affected by gypsum treatments

	ESP[a]				
	Soil layer depth (cm)				
Treatment	*0–15*	*15–30*	*30–45*	*45–60*	*60–90*
Control[b]					
Autumn	19.3 ± 4.8	19.0 ± 4.9	16.4 ± 5.1	16.4 ± 4.8	14.2 ± 3.9
Spring	16.9 ± 2.9	19.4 ± 4.0	19.0 ± 2.8	19.3 ± 3.4	19.6 ± 4.0
Gypsum mixed in					
Autumn	17.4 ± 5.0	20.1 ± 4.6	20.3 ± 6.0	20.8 ± 6.0	21.3 ± 6.7
Spring	12.8 ± 5.0	17.7 ± 3.6	19.6 ± 3.8	17.4 ± 2.9	18.0 ± 2.3
Gypsum spread over					
Autumn	16.9 ± 5.8	17.0 ± 5.8	19.3 ± 6.0	20.9 ± 4.5	21.1 ± 4.5
Spring	11.5 ± 4.1	16.3 ± 3.6	19.4 ± 3.5	19.3 ± 7.0	20.4 ± 4.3
CEC ($cmol_c\ kg^{-1}$ soil)	18.2 ± 1.5	17.8 ± 1.3	18.5 ± 2.0	18.9 ± 2.2	18.1 ± 1.1

[a]Mean values with standard deviation (12 and 43 repetitions for ESP and CEC, respectively).
[b]Autumn 1981 (after irrigation with 400 mm of saline water, before gypsum application) and spring 1982 (after the rainy season).
CEC, cation exchange capacity.
Adapted from Keren R, Shainberg I, Frenkel H, and Kalo Y (1983) The effect of exchangeable sodium and gypsum on surface runoff from loess soil. *Soil Science Society of America Journal* 47: 1001–1004.

The amount of gypsum needed to reclaim a sodic soil, in metric tonnes per hectare, can be calculated using eqn [3]:

$$GR = 8.61 \times 10^{-5} Q_{Na} \qquad [3]$$

The efficiency and rate of exchange, namely, the percentage of applied Ca that exchanges for adsorbed Na, varies with ESP, being greater at high ESP values. Removal of Na at ESP levels below 10 is slow, and some of the applied Ca displaces exchangeable Mg so that the efficiency declines to about 30%. Efficiency may also be low (20–40%) in fine-textured soils, because of the slowness of exchange of Na inside the structural units.

Generally, soil water penetration is too low to allow reclamation of sodic soils to be attained in a single leaching. For example, a 50-cm depth of water applied for leaching can only dissolve about 12 Mg ha^{-1} of gypsum (reaching a saturation solution in respect to gypsum). Larger gypsum applications, therefore, will not be effective unless soil water penetration is sufficient for larger water applications. Thus, sodic soil can normally only be reclaimed to a shallow depth the first year (depending on the depth of water applied for leaching), but this will often permit a shallow-rooted crop to be grown after leaching (**Figure 3**). Subsequently annual applications of amendments and leaching volumes can be applied to reclaim the entire profile over a period of years.

In estimating the amount of gypsum to be applied, little attention has been directed to the possibility that

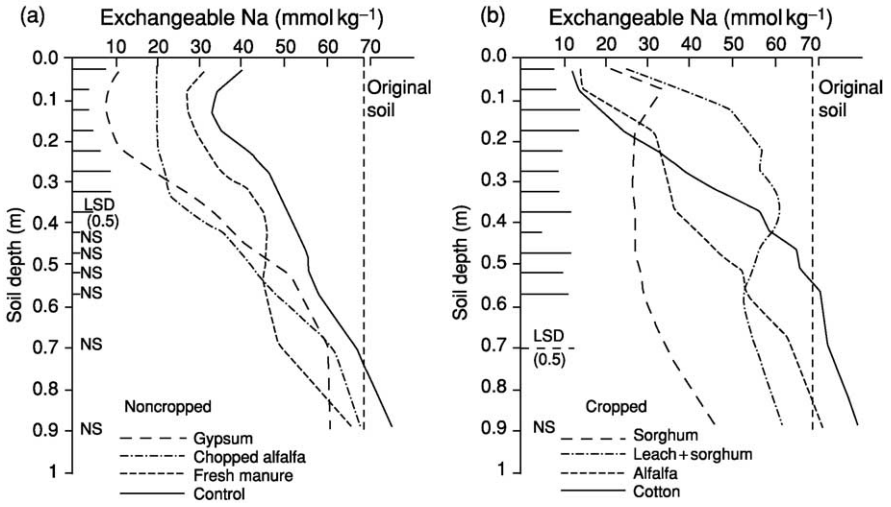


Figure 3 Final distribution of exchangeable Na levels with soil depth resulting from reclamation treatments for (a) noncropped and (b) cropped treatments. LSD, least significant difference; NS, no significance. (Adapted from Robbins CW (1986) Sodic calcareous soil reclamation as affected by different amendments and crops. *Agronomy Journal* 78: 916–920, with permission.)

incorporating large amounts of gypsum powder into the soil may cause temporary reduction in HC because excess small particles of gypsum may block conducting pores in the early stage of the dissolution process.

The soil solution in the gypsum layer reaches equilibrium (in the presence of excess gypsum) only when the contact time between an elemental volume of solution and gypsum particle surface is sufficiently long, or when the surface area of the gypsum fragments per volume of water is sufficiently large. Some of the factors that influence the rate of gypsum dissolution are the surface area of the gypsum particles, the soil water velocity during leaching, and the electrolyte composition of the soil solution.

The gypsum dissolution rate is written as:

$$(\mathrm{d}C/\mathrm{d}t) = k(C_s - C) \qquad [4]$$

where $\mathrm{d}C/\mathrm{d}t$ is the net rate of dissolution, k is the dissolution coefficient, and C_s and C are the solution concentration at saturation and at time t, respectively. Increasing solution flow velocity increases the dissolution rate coefficient, but decreases the contact time between gypsum and the flowing solution; the net effect is a decreasing dissolution rate with increasing soil solution flow velocity.

Integrating eqn [4] yields:

$$-\ln\left(1 - \frac{C}{C_s}\right) = kt \qquad [5]$$

Since the thickness of the stationary film of solution around the gypsum particles changes with the soil water velocity, the dissolution rate coefficient for a given surface of gypsum particle also changes. Thus, the lines obtained from eqn [5] are not linear. The left-hand side of eqn [5] is empirically related linearly to the square root of time:

$$-\ln\left(1 - \frac{C}{C_s}\right) = \alpha t^{1/2} + \beta \qquad [6]$$

where α and β are the slope and intercept, respectively, and t is the time for an increment of solution to leave a given soil depth:

$$t = \frac{L}{V} \qquad [7]$$

where L is the length of the layer of soil-gypsum mixture and V is the soil water velocity. Combining eqns [6] and [5] and introducing eqn [7] yields:

$$k = (V/L)^{1/2}\left[\alpha + (V/L)^{1/2}\beta\right] \qquad [8]$$

The dissolution rate coefficient as a function of water flow velocity can be evaluated for a given surface area of gypsum. Since the permeability of a sodic soil is low, the dissolution rate is relatively high during the initial stage of reclamation. During the reclamation process, however, the permeability increases and the dissolution rate decreases. Thus, the soil regulates gypsum-dissolution rate during reclamation. When it is possible to control water flow velocity in soils (e.g., sprinkler irrigation), reduced velocities are to be preferred for increased gypsum dissolution rate and greater efficiency of cation exchange reaction due to increase in contact time between water and gypsum particles (**Figure 4**).

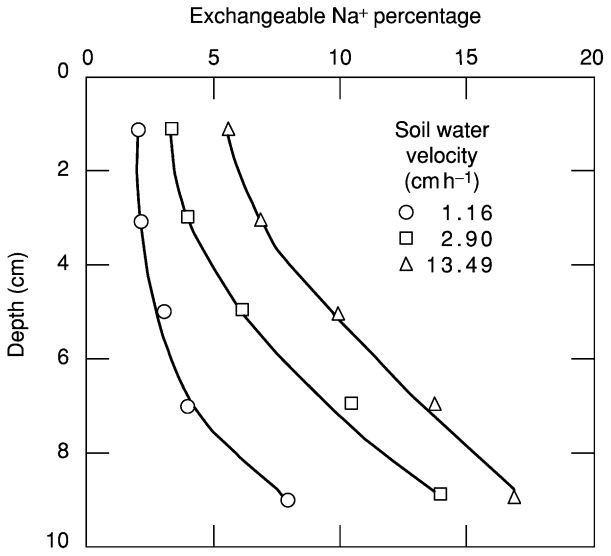


Figure 4 Distribution of exchangeable sodium percentage with soil depth as a function of soil water velocity for a given amount of dissolved gypsum. (Adapted from Keren R and O'Connor GA (1982) Gypsum dissolution and sodic soil reclamation as affected by water flow velocity. *Soil Science Society of America Journal* 46: 726–732, with permission.)

Several chromatographic models have been used for simulating reclamation of sodic soils with gypsum. The models include many of the known chemical reactions occurring in natural soils as a result of $CaCO_3$ and $CaSO_4$ dissolution, salt leaching, and ion exchange reactions. These models have been field-tested with data from sodic-soil reclamation leaching plots. The difference between measured and computed results is no more than the horizontal variations typically found in salt-affected lands. These models provide powerful tools for quantitative predictions of water and gypsum required to reclaim soil profiles to prescribed levels of salinity and ESP.

In another approach, a source term for gypsum dissolution has been introduced into a steady-state solute-transport model. However, considerable difficulties have been encountered in matching calculated concentration–time curves to column experimental data. Miscible displacement models coupled with chemical reaction models provide insight into the processes of reclamation. However, they have not been used extensively in the field, because solving them requires extensive data specific to the site.

It is clear from the literature that the assumption of chemical equilibrium for dissolution of gypsum may not be applicable to soil-reclamation modeling. Another approach is to couple dissolution kinetics of gypsum and exchange-cation reactions to a salt-transport model.

Acids and sulfur These amendments are of value for ameliorating soils containing $CaCO_3$, with which the acid interacts to form gypsum (using H_2SO_4) or calcium chloride (using HCl). In some areas, the availability of acid waste products arising from mining and industrial activities is increasing markedly. Use of these waste products as soil amendments may provide a safe means of disposal and could greatly influence the economics of any amendment proposal. Being highly corrosive, neither H_2SO_4 nor SO_2 should be added to water that flows through metal or concrete irrigation systems. Sulfur (S) requires an initial phase of microbiological oxidation to produce H_2SO_4. Other amendments such as pyrites and polysulfides, which must first be oxidized by soil microorganisms, will be slower-acting than H_2SO_4. Their effectiveness in field experiments has been variable, perhaps being related to the presence or absence of appropriate microbial populations.

Reclamation without Adding Soil Amendments

Calcium carbonate Reclamation of a sodic soil without amendments is possible when a soil source of Ca exists (e.g., calcium carbonate), the drainage through the soil profile is good, and adequate leaching water is present. The rate at which $CaCO_3$ dissolution in water approaches equilibrium is dependent upon a number of factors, which include the surface area-to-solution volume ratio, the ionic composition of the solution, the ion composition of the adsorbed phase, affinity of the clay minerals to cations, the temperature, and the partial pressure of CO_2 in the soil atmosphere. In general, a supersaturation with respect to calcite is observed in solution extracts from calcareous soils.

In soils, the kinetics of $CaCO_3$ dissolution is not a simple diffusion-controlled or first-order reaction. In these calcareous systems, the transfer of atmospheric CO_2 to solution is a limiting step in the kinetics of dissolution. There are several uncertainties in modeling the kinetics of carbonate chemistry: at low pH the rate depends significantly on the thermodynamic transport constant for H^+, which is not well defined, and reaction-site density and controls on pH at the surface interface between the crystal and the bulk solution are also not well understood. Moreover, kinetics of precipitation and dissolution of $CaCO_3$ have not yet been modeled for soil water systems.

Soil $CaCO_3$ may be dissolved slowly to contribute Ca but its solubility is enhanced in saline sodic soils. Moreover, incorporation of fresh organic matter in calcareous soil increases native calcium carbonate mineral dissolution, owing to elevation of CO_2 partial pressure in the soil atmosphere, and therefore

significantly enhances sodic soil reclamation rate (Figure 3).

Successive dilutions of highly saline water When soil physical conditions have deteriorated and HC of the soil is so low that the time required for reclamation or the amount of amendment required is excessive or if a sodic soil is to be leached with a water so low in salinity that water infiltration decreases adversely, the use of the high-salt water-dilution method may be applied. The method involves application of successive dilutions of highly saline water containing divalent cations. This principle has been tested successfully in experiments, and reclamation at various stages of leaching can now be predicted. In the early phase, the high salinity of the water prevents clay dispersion and induces flocculation of the soil colloids, the Ca content provides a source of Ca for exchange with Na, and the sodicity level decreases. On dilution with high-quality water, the sodium adsorption ratio of the irrigation water is reduced by the square root of the dilution factor. When sufficient water is applied at each step of dilution, then, to ensure exchange equilibrium to the desired depth, the total equivalent surface depth of water for reclamation should be about nine times the depth of soil to be reclaimed. The method is particularly effective for soils of the expanding-clay type, which have extremely low soil HC.

Further Reading

Ayers RS and Westcot DW (1985) *Water Quality for Agriculture*. FAO Irrigation and Drainage paper No. 29. Rome, Italy: UN FAO.

Bresler E, McNeal BL, and Carter DL (1982) *Saline and Sodic Soils: Principles, Dynamics Modeling*. Berlin, Germany: Springer-Verlag.

Hillel D (2000) *Salinity Management for Sustainable Irrigation*. Washington, DC: The World Bank.

Hoffman GJ (1980) Guidelines for the reclamation of salt-affected soils. In: O'Connor GA (ed.) *Second Inter-Amercan Conference on Salinity and Water Management Technology*, pp. 49–64. New Mexico, NM, USA: New Mexico State University.

Jury WA, Jarrell WM, and Devitt D (1979) Reclamation of saline-sodic soils by leaching. *Soil Science Society of America Journal* 43: 1100–1106.

Keren R and Miyamoto S (1990) Reclamation of saline, sodic and boron-affected soils. In: Tanji KK (ed.) *Agricultural Salinity Assessment and Management*, pp. 410–431. New York: American Society of Civil Engineering.

Keren R and O'Connor GA (1982) Gypsum dissolution and sodic soil reclamation as affected by water flow velocity. *Soil Science Society of America Journal* 46: 726–732.

Keren R and Shainberg I (1981) The efficiency of industrial and mined gypsum in reclamation of a sodic soil – rate of dissolution. *Soil Science Society of America Journal* 45: 103–107.

Keren R, Shainberg I, Frenkel H, and Kalo Y (1983) The effect of exchangeable sodium and gypsum on surface runoff from loess soil. *Soil Science Society of America Journal* 47: 1001–1004.

Misopolinos ND (1985) A new concept for reclaiming sodic soils with high-salt water. *Soil Science* 140: 69–74.

Nielsen DR, Bigger JW, and Luthin JN (1966) Desalinization of soils under controlled unsaturated flow conditions. In: *Sixth International Congress Communication on Irrigation and Drainage*, pp. 19.15–19.24. New Delhi, India.

Oster JD (1982) Gypsum usage in irrigated agriculture: a review. *Fertilizer Research* 3: 73–89.

Quirk JP (1971) Chemistry of saline soil sand their physical properties. In: Talsma T and Philip JR (eds) *Salinity and Water Use*. New York: Macmillan.

Reeve RC and Doering EJ (1966) The high-salt-water dilution method for reclaiming sodic soils. *Soil Science Society of America Journal* 30: 498–504.

Robbins CW (1986) Sodic calcareous soil reclamation as affected by different amendments and crops. *Agronomy Journal* 78: 916–920.

Shainberg I and Shalhevet J (1984) *Soil Salinity Under Irrigation – Processes and Management*. Berlin, Germany: Springer-Verlag.

Sumner M (1999) *Handbook of Soil Science*. Boca Raton, FL: CRC Press.

Tanji KK and Deverel SJ (1984) Simulation modeling for reclamation of sodic soils. In: Shainberg I and Shalhevet J (eds) *Soil Salinity Under Irrigation – Processes and Management*, pp. 238–251. Berlin, Germany: Springer-Verlag.

SAND DUNES

H Tsoar, Ben Gurion University of the Negev, Beer Sheva, Israel

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Sand dunes are found in all continents and in all world climates. They cover approximately 20% of the world's arid and extremely arid zones but are also found in humid regions of North America and Europe and along most of the coasts of the world. In total they cover an area of more than 5 000 000 km^2, of which 99% is desert. Sand dunes appear in many different shapes as a result of the factors that form them, such as variability in wind direction, grain size, and vegetation cover.

As a soil, dune sand is usually considered to be devoid of any beneficial characteristics for flora and fauna because sand is known to have low fertility, low water-retaining capacity, and sand-sized particles are cohesionless. Wind magnitudes of 4 Beaufort force (5–7 $m\,s^{-1}$) and above can easily erode sand, thus causing another limitation for flora and fauna. As a result, aeolian sand encroachment has become synonymous with desiccation and desertification in arid and semiarid lands.

Dune Sand Texture

Soil particles are grouped according to their sizes into four classes: gravel, sand, silt, and clay. Sedimentologists and geomorphologists use a grade scale consisting of a series of classes in which the boundaries between adjacent size classes differ by a factor of 2. This scale is modified by transforming the grade boundaries into a phi (φ) scale (Table 1), which is defined as:

$$\varphi = -\log_2 d \qquad [1]$$

where d is the grain diameter in millimeters.

Wind should be above the threshold velocity to initiate sand transport. The threshold velocity depends directly on grain size and the cohesiveness of the grains (Figure 1). Grain particles less than 0.10 mm tend to cohere, as the rate of cohesiveness increases with decreasing size. Accordingly, particles of fine-sand size have threshold velocities of erosion that are lower than those of silt and clay by factors of 2–3. The mode size of most dune sand all over the world is in the class of fine sand (0.125–0.250 mm), which has the lowest wind-threshold velocities. Fine sand can be transported most easily on the windward slope of sand dunes under storm winds of average intensity. Coarser grains of medium, coarse, and very coarse sand are unable to climb slopes. This is the main reason for the formation in deserts of sand sheets composed of bimodal sand in which the coarser mode is always greater in grain size than fine sand.

The cohesionless nature of sand causes it to be swept away easily by the wind in a leaping movement, known as 'saltation,' which describes a ballistic path very close to the bed, descending at high speeds and low angles. Less-exposed grains or coarser grains are either pushed in a short-distance movement, ensuing from the scatter of the impact of the saltating grains, known as 'reptation,' or driven forward along the surface in a creep mode through the impact of the saltating grains.

The low cohesiveness and the ensuing form of transportation by saltation have a major implication for plants. The easy erosion of sand by the wind exposes plant roots, causing damage to the plants through grain bombardment, thus making the plants vulnerable to disease. Some plants overcome this by growing minute roots that produce a sticky liquid. Sand particles are glued to these roots, forming a shield which protects them from damage.

The most common adaptational response of animals to low cohesiveness of sand is to increase the surface area of the soles and fingers, either by enlarging or adding scales to the fingers (as do many lizards and skinks), or by growing stiff hairs on the sole (as do many mammals, particularly rodents and carnivores). The other technique used to overcome low cohesiveness of soil is sand-swimming, engaged in by animals such as skinks and some insects, which have relatively long and narrow bodies and short legs that help them to move while winding their bodies between the sand particles.

Table 1 Size classes of grain particles

Grain size (mm)	*Grain size (μm)*	*Grain size (φ)*	*Classification*
>2	>2000	<−1	Gravels
1–2	1000–2000	0 to −1	Very coarse sand
1/2–1	500–1000	1 to 0	Coarse sand
1/4–1/2	250–500	2 to 1	Medium sand
1/8–1/4	125–250	3 to 2	Fine sand
1/16–1/8	62–125	4 to 3	Very fine sand
1/256–1/16	4–62	8 to 4	Silt
1/4096–1/256	0.243–4	12 to 8	Clay

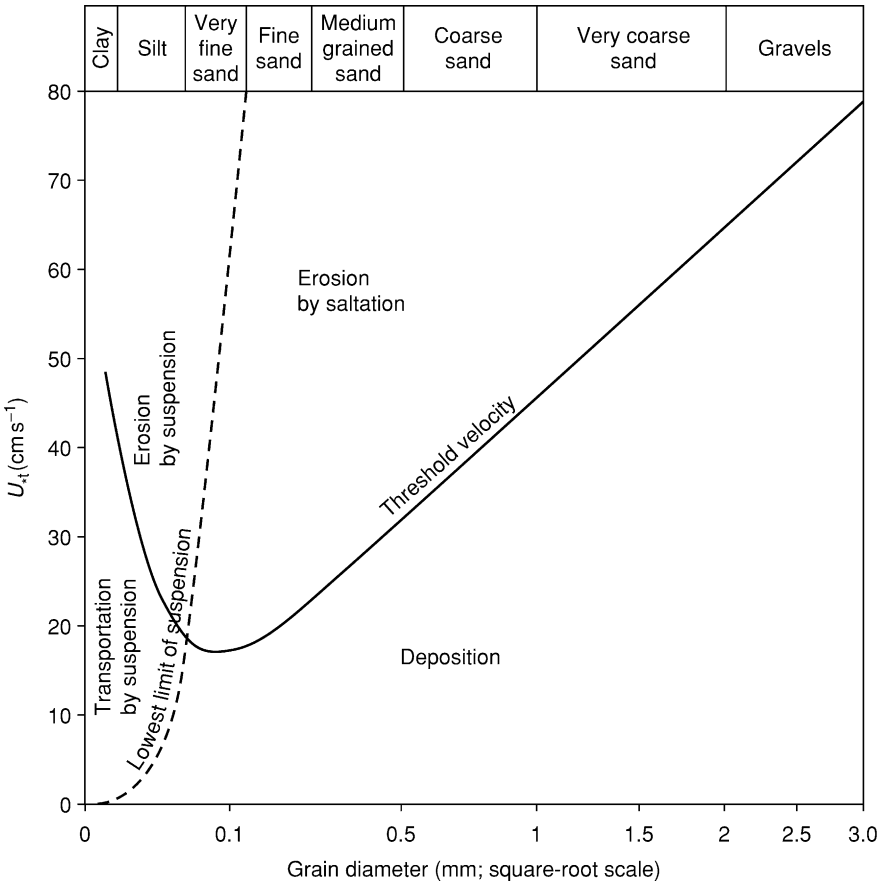


Figure 1 Threshold friction velocity (U_{*t}) curve for quartz grains of different diameters (continuous line). The dashed line separates saltation from suspension. Adapted from Pye K and Tsoar H (1990) Aeolian Sand and Sand Dunes, London: Unwin Hyman.

Particle sizes of clay, silt, and very fine sand are so small that they have low settling velocities relative to the velocities of the vertical component of the turbulent wind flow. As a result, these particles remain in suspension or quasisuspension in their transportation mode and are considered as dust (Figure 1). Dust does not easily accumulate together with sand unless the latter is covered by vegetation, thus serving as a trap for the suspended particle load in the lower atmosphere. Biogenic crust is formed on sand dunes in some desert areas when vegetation spreads and reduces the sand movement. Filamentous cyanobacteria are known to form crusts in the northwestern Negev sand dunes (Figure 2). This crust serves as trap for atmospheric dust, resulting in a feedback mechanism between the amount of dust that is deposited on the surface and the mass of the cyanobacteria. There is a significant correlation in the northwestern Negev between the percentage of fines (silt + clay) and the percentage of organic matter in the crust. For approximately 1.5% of organic matter, there are more than 30% of fines. Rainwash leaches part of the fines to a depth of 30–50 cm, which is the average penetration depth of rain for small, frequent storms in the Negev desert.

The Effect of Dune Sand Texture on Water Infiltration

One of the most important attributes of sand dunes, resulting from the sand texture and the relatively big pore spaces, is low field capacity and a low value for soil moisture availability for plants (Figure 3), as well as a high rate of infiltration and leaching, resulting in the washing away of nutrient elements necessary for plant growth.

The hydraulic conductivity at saturation for sand dunes is very high (between 10^{-3} and 10^{-4} m s^{-1}). Unlike most soils composed of silt and clay, the hydraulic conductivity of sand responds very little to changes in moisture content. The high rate of hydraulic conductivity determines the depth of water infiltration. The maximal depth of infiltration in linear sand dunes in the northwestern Negev, during a year with rainfall that is 10.5% lower than the climatological mean (95 mm), reached 60–90 cm,

Figure 2 Crust on sand dunes in the northwestern Negev desert, where the annual rainfall is 85 mm. The crust is formed by filamentous cyanobacteria. Note an area with exposed sand in the middle lower part of the picture.

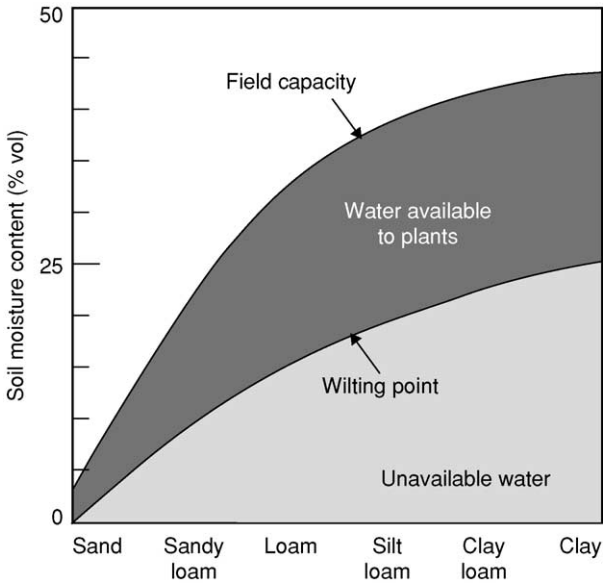


Figure 3 The effect of soil texture on the wilting point, field capacity, and the emanative moisture that is available to plants. Note that sand has the lowest wilting point but also a very low field capacity, which leaves very little amount of moisture available to plants as most of the water infiltrates easily towards the groundwater.

while, in a nearby interdune area covered mostly by silt and clay, the infiltration was limited to a depth of 30–40 cm. During a wetter year, with rainfall 38% higher than the average, the range of infiltration in the same linear dune was in the range of 90–120 cm and in some cases down to 180–210 cm. In the fine-grained interdune layer, it was limited to a depth of 30–60 cm. Deep infiltration to a depth of 180–400 cm occurs in sand only in response to rain that accumulates to 100 mm in less than 2 months.

Sand dunes are known to have an internal laminar structure of varying geometry. The laminae result from two main modes of deposition: grainflow and grainfall. These two types of deposition produce different arrangements of grain packing and distinct concentrations of fine and coarse sand. The internal structure of aeolian sand dunes affects the direction of infiltration into the dune. If the aeolian sand is composed of grainflow deposits that dip at angles of 32–34°, infiltration will deviate by 50–60° from vertical downward movement ('piston infiltration'). The capillary tension caused by closely packed or fine-grained laminae that are near-horizontal or parallel to the surface inclination brings about subsurface lateral water flow. Such a flow can be found in sand dune areas where the mean annual rainfall is more than 200 mm. Areas with approximately 100 mm annual rainfall experience lateral movement only during periods with rainfall much above the average.

The biogenic crust formed on top of ungrazed sand dunes in the western Negev (Figure 2) has a lower hydraulic conductivity than dune sand because of the high proportion of fines. However, the biogenic crust on the Negev sand dunes is not hydrophobic and therefore absorbs rainwater until saturation is reached and the infiltration rate drops to 9–12 mm h^{-1}. Since the prevailing sealing process of bare crust by the impact of intensive raindrops is not effective in the Negev, there is no correlation of runoff yield with rain amount or intensity. Runoff is generated in the Negev on saturated crust by rain intensity

of 9–12 mm h^{-1} lasting for 1–2 min only, and its yields are low to very low.

The Effect of Sand Texture on the Thermal Properties of Sand Dunes

When sand is devoid of or has only meager vegetative cover, its texture and mineralogical composition play an important role in determining the thermal properties of its environment. Heat is absorbed from solar shortwave radiation and is concentrated in the sand's surface to a thickness of approximately 1 mm.

The rate at which modification of temperature takes place within sand (k) is directly proportionate to its thermal conductivity (K) or to its ability to transmit heat, but is inversely proportionate to its specific heat (c) or the amount of heat required to achieve the modification of temperature. This relationship is described in the following equation:

$$k = K/c\gamma \qquad [2]$$

where k is known as the thermal diffusivity, and γ is the bulk density of the sand (approx. 1600 kg m^{-3}). The units of k are meters squared per second. The smaller the value of k, the more slowly heat will diffuse through the substance. The thermal diffusivity of sand is approx. 3.75×10^{-7} m^2 s^{-1}, which is relatively low and induces a very slow penetration of surface temperature changes. This results in a very acute and thin near-surface accumulation of heat. As a consequence, the maximal sand surface temperatures are higher by 10–20°C than the air temperature, at 2 m high, during the afternoon (Figure 4). This steep lapse rate gives rise to local up-rushes of hot air and descent of cooler air, which increases the air turbulence, and the erosivity of the wind. Under certain conditions the ascending hot air develops into whirlwinds known as 'dust devils.' The minimal sand surface temperature is only lower by 1–3°C than the minimal air temperature in the early morning (Figure 4).

These thermal qualities of sand result in three significant, diurnal changes in temperature at deeper levels of dune sand: (1) the maximal temperature is lessened with depth; (2) the diurnal amplitude is degraded with depth; (3) there is a lag between the daily maximum and minimum temperature from the surface to various depths (Figure 4).

The unique thermal qualities of sand dunes have important implications for animals living in this habitat. The most common animal adaptation to high surface temperatures is avoidance: most animals living in sand dune habitats in deserts are nocturnal or are active during the early morning or late evening, when soil temperature decreases to tolerable temperatures. During the rest of the day, most of these animals take shelter below the sand surface, where the temperatures are moderate, either in their own burrows or in holes dug by other animals. Many of the diurnal animals have relatively long legs (lizards, beetles) and can elevate themselves high above the hot surface, thus keeping their bodies at relatively low temperatures while only some parts of their legs are in contact with the hot sand.

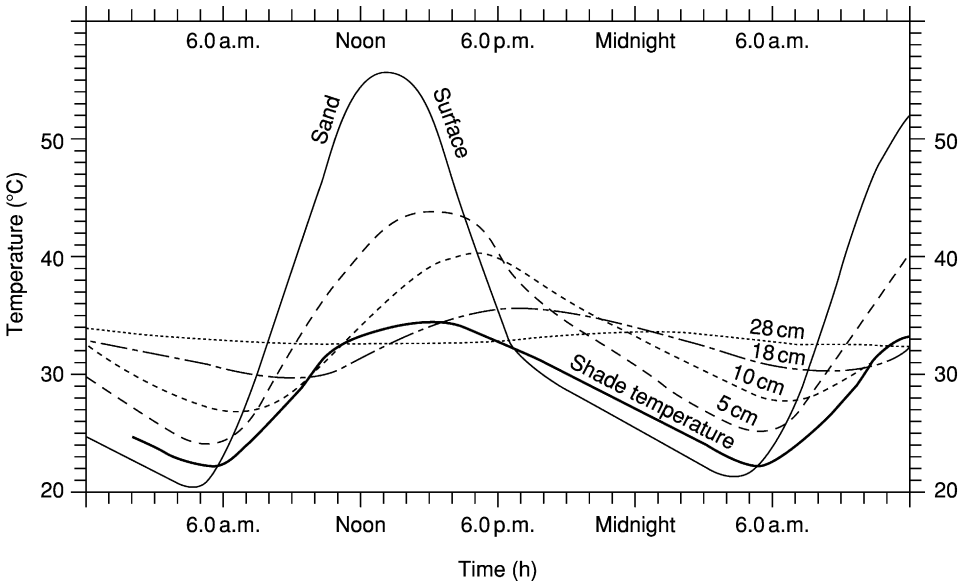


Figure 4 Variation in sand temperature with depth as measured in August. Shade temperature was measured at an elevation of 2 m above the surface. Adapted from Pye K and Tsoar H (1990) Aeolian Sand and Sand Dunes, London: Unwin Hyman.

The Advantage and Disadvantage of Sand Texture on Vegetation Growth in Deserts

Vegetation can grow on sand dunes in arid areas with less than 100 mm of annual mean rainfall (Figure 2). The primary limitation of vegetation growth on dune sand is human impact; the most dominant natural limitation is wind power. Rainfall is a limiting factor only where the annual mean is very low (<50 mm).

Two unique and distinct attributes of dune sand affect vegetation growth on dunes: (1) dune sand has low fertility and low water-retaining capacity (Figure 3); (2) sand grains are not cohesive and therefore are easily carried by the wind (Figure 1). For this reason, plants on sand are sparsely distributed and they are absent where wind erosion is severe, such as the windward slopes of dunes. Accordingly, the amount of vegetation on dunes depends on the wind power. There are arid and humid areas where the wind power is so high that it does not enable vegetation to thrive on these dynamic dunes, such as the Oregon coastal dunes.

The attribute of high hydraulic conductivity of sand relative to silt and clay is advantageous in deserts. In humid areas with abundant precipitation, the main cause of soil-moisture loss in sand is the high rate of infiltration beyond the root layer. In arid and semiarid lands, the situation is quite different. These areas have little precipitation, resulting in infiltration to depths of only 60–120 cm. Perennials can use the moisture at these depths, which are deep enough to protect the sand moisture from the high rate of evaporation typical of deserts during the long dry and hot rainless periods. As a result, moisture retention is assured in desert sand below the upper 30–60 cm. In other desert soils such as loess, which are mostly composed of silt and clay, the infiltration reaches a much shallower depth of 30–40 cm, where evaporation is effective. The delineation between the advantage of coarse-textured sand (low evaporation) over fine-textured soil (silt and clay) versus its inherent disadvantage (high hydraulic conductivity) lies somewhere between a mean of 300 and 500 mm of precipitation.

The moisture content of sand at the wilting point is very low. Therefore, a small amount of water input brings sand moisture to values much above the wilting point, whereas the same amount of water added to loamy soils does not, making the water unavailable for plants (Figure 3).

The effect of soil texture on biomass has been tested on the three soil types (loam, sandy loam, and sand) located in the same area in the Negev desert where the annual mean rainfall is 100 mm. Biomass measurements (annuals only) were taken during two winters with approximately average rainfall (104.6 and 111 mm), but with dissimilar distribution. The highest biomass for the 2 years was formed on the sand soil (647.18 g m^{-2}), with the second highest on the sandy loam (569.95 g m^{-2}), while the loam had the least amount of biomass (115.53 g m^{-2}).

Mobility and Stability of Sand Dunes

There is no direct relationship between the amount of rainfall and the vegetation cover on the dunes. Active dunes with no vegetation cover are found in humid areas (e.g., the Oregon coastal dunes, or the Alexandria coastal dunes in South Africa) while dunes fully stabilized by microphytes and macrophytes are found in the arid part of the Negev desert. Several wind erosion or mobility indices (M) have been developed for various parts of the world for the purpose of climatic changes and their implications on sand dune mobility. All of these indices are based on two factors that increase or decrease dune mobility. The first one deals with the degree of windiness. Most dunes will be mobilized if windiness were to increase. The second factor is the ratio between the mean precipitation (P) and the evaporation (E), known as precipitation-effectiveness index ($P - E$). Based on the above factors, a simple and widely accepted M-index is:

$$M = \frac{W}{(P/\mathrm{PET})} \qquad [3]$$

where W is the percentage of days during the year with sand-moving winds and P/PET is the ratio of mean annual precipitation to mean annual potential evapotranpsiration. The M-index of eqn [3] is calibrated in South Africa. Dunes are fully active when values for M exceed 200 and most dunes are completely stabilized by vegetation when M is less than 50. Because of its simplicity, the M-index of eqn [3] is commonly used by scientists studying the effect of climate change on sand mobility and stability in North America. The sand dunes along the coastal plain of Israel and in the northwestern Negev desert should be active with no vegetation cover according to eqn [3] ($M > 200$). In fact, the northwestern Negev desert dunes in areas of annual mean rain of 100 mm are fully stabilized by vegetation and biogenic crust (Figure 2), and the more humid coastal dunes are partially vegetated.

Dune sand has the lowest moisture content (Figure 3). Thus, chiefly in humid climates, most of the rain is lost to the groundwater and is not available to the plants. Hence the amount of rainfall is not a decisive factor in sand dune stabilization except by increasing the cohesiveness of the sand. The deep infiltration of rainwater in dune sand reduces the effect of

evaporation (PET in eqn [3]) mostly in arid areas relative to other soils composed of fine particles of silt and clay.

Wind power is the most important factor in sand dune mobility because of the noncohesiveness of the sand. Wind above a certain wind power can erode sand to such an extent that it prevents seeds from germinating in the sand and stabilizing it. According to the sand transport equation, the sand flux (q) is directly proportional to the cube of the wind:

$$q = 5.2 \times 10^{-4}(U - U_t)^3 \qquad [4]$$

where q is in metric tons per meter width per hour, U is the wind velocity, measured at a height of 1 m, and U is the threshold wind velocity, which is equal to $4\,m\,s^{-1}$. According to eqn [4], 1 h of wind velocity at $16\,m\,s^{-1}$ can carry the same amount of sand as 72 days with wind velocity of $5\,m\,s^{-1}$, or 27 h of wind velocity at $8\,m\,s^{-1}$. Hence, the factor W of eqn [3] that refers to the wind by the percentage of days during the year with sand-moving winds is a poor representation of the sand-moving power of the wind.

A much better index of wind power is the drift potential (DP) of the wind, which is based on eqn [4]:

$$DP = \Sigma q = \frac{U^2(U - U_t)}{100}t \qquad [5]$$

where $\sum q$ is related to the total sand flux from all wind directions, and t is the amount of time the wind blows above the threshold (in percent). The drift potential (DP) is given in vector units and it is customary to base the calculations on wind velocities measured in knots. An index of wind direction variability is illustrated by the ratio between the resultant drift potential and the drift potential (RDP/DP), where values close to one indicate narrow unidirectional DP, and values close to zero indicate multidirectional DP.

The total yearly average DP and the ratio RDP/DP can explain the mobility and stability of sand dunes mostly because the limiting factor for vegetation on dune sand is wind erosion. When RDP/DP is low, wind energy is distributed on more than one slope of the dune and the energy exerted on each slope is lower. Sand dunes in areas where the annual mean rainfall is $\geq$50 mm are unvegetated and mobile under wind conditions in which M (according to eqn [6]) exceeds 1:

$$M = \frac{DP}{1000 - (750(RDP/DP))} \qquad [6]$$

Table 2 displays the DP, RDP, and M (according to eqn [6]) in various dune sites of the world. It shows that rainfall is not as decisive a factor for dune mobilization or stabilization as is the DP. In the Negev desert, dunes are fully stabilized where the mean yearly rainfall is 90 mm. The dunes in Sinai (east of Al-Ismailiya) and the coastal dunes in Gaza are fully active due to human impact. Dunes that were artificially stabilized, such as along the Netherlands coast, will become active once vegetation is artificially destroyed.

Dune Shape and Dynamics

The tremendous variety of sand dunes in the world deserts makes their classification a difficult task. Sand accumulates into piles of various shapes because of three main causes:

1. Obstruction to the wind flow such as a cliff, boulder, etc.
2. Vegetation cover;
3. Self accumulation of sand into piles.

Classification according to the above three factors is depicted in Figure 5. The first two causes result from the immobilization of sand by vegetation or topographic obstructions. The last one is related to aerodynamics. In addition, other factors, such as

Table 2 Drift potential (DP), directional variability of the wind (resultant drift potential (RDP/DP), *M*-index (eqn [6]), and rainfall of several dune field sites

Location	*Country*	*DP*	*RDP/DP*	*Average yearly rainfall (mm)*	M	*Dunes status*
Al Ismailiya	Egypt (western Sinai)	62	0.47	50	0.10	Fully active
Nizzana	Israel (Negev)	108	0.70	90	0.23	Fixed (fully vegetated)
Gaza	Palestine (coast)	158	0.73	400	0.35	Active (partly vegetated)
Ashdod	Israel (coast)	147	0.71	500	0.31	Semiactive (partly vegetated)
Upington	South Africa (Kalahari)	520	0.66	183	1.03	Stabilized linear dunes
Port Elizabeth	South Africa (coast)	951	0.49	660	1.5	Fully active with no vegetation
Lüderitz	Namibia (coast)	2300	0.85	<100	6.34	Fully active with no vegetation
Ijmuiden	Netherlands (beach)	3999	0.51	768	6.48	Fully stabilized by vegetation; active when vegetation is destroyed

Reproduced with permission from Tsoar H (2001) Types of aeolian sand dunes and their formation. In: Balmforth NJ and Provenzale A (eds.) Geomorphological Fluid Mechanics. Lecture Notes in Physics. Berlin: Springer.

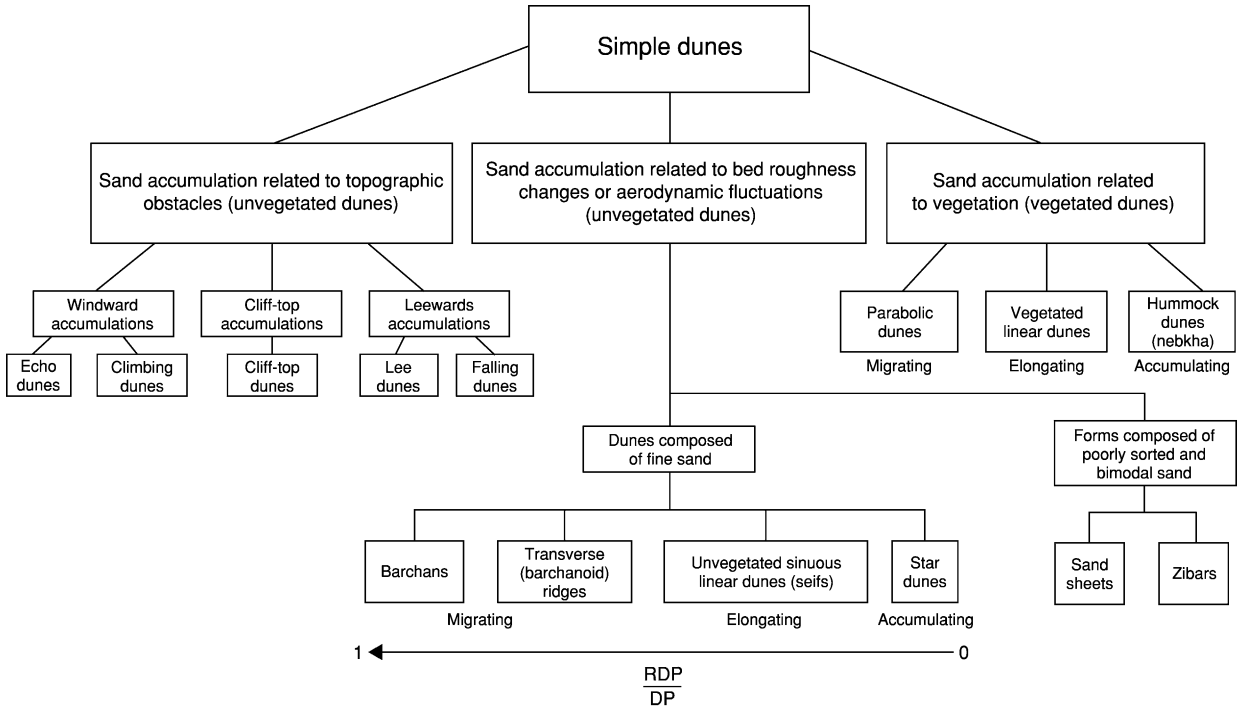


Figure 5 Classification of major dune types based on the way sand is accumulated into dunes and the effect of wind direction variability. RDP, resultant drift potential; DP, drift potential. Adapted from Pye K and Tsoar H (1990) Aeolian Sand and Sand Dunes, London: Unwin Hyman.

variation in wind direction, climatic changes manifested by dramatic shift in wind direction and magnitude, velocity, and frequency of storms, sand availability and thickness of sand cover, and sudden removal of vegetation cover, can have an effect on dune morphology.

From a dynamic point of view, sand dunes can be classified into three distinct groups:

1. Migrating dunes, where the whole dune body advances with little or no change in shape and dimension; best represented by transverse and barchan dunes.
2. Elongating dunes, where the dunes elongate and become extended in length with time; best represented by linear dunes.
3. Accumulating dunes, where the dunes have little or no net advance or elongation; best represented by star dunes.

These three groups are distinguished according to the wind directional variability for dunes without vegetation (Figure 5). The migrating dunes are formed by a wind regime that is unidirectional or close to unidirectional. The wind directions of the elongating dunes are bidirectional with a 70–90° separation between the two directions. The accumulating dunes are formed under multidirectional winds, where the two main directions form an obtuse angle that is approximately 180°.

The migrating and elongating dunes are under the influence of two different types of sand movement and deposition. Migrating dunes such as barchan and transverse dunes have two distinct slopes, a moderate windward and a steeper leeward. The windward slope undergoes erosion and the sand is carried to the lee slope where it is deposited on the upper part of the slope. When the upper lee slope reaches the critical angle of internal friction, a failure begins. The slide of grains from the upper slope downward reduces the slope inclination and the avalanche ceases at the angle of repose (33°), forming a slip-face (Figure 6).

Elongating dunes have a more complicated mechanism of sand transport and deposition because, when winds encounter the dune crest line at an acute angle, the flow at the lee side is deflected to parallel the crest line. Accordingly, the leeward slope is not merely a zone of deposition, but also a zone of erosion by the diverted lee flow. The occurrence of erosion and transport or deposition of sand depends upon the angle of incidence between the wind and the crest line. When this angle is $<40°$, the deflected wind has the power to erode sand and carry it along the lee slope, with little deposition on the upper lee slope. When the angle of incidence is close to 90°, the velocity of the wind in the lee slope drops and deposition largely occurs as it does with migrating dunes. For angles between 40° and 90°, there is a gradual transition in the lee slope, from dominant erosion

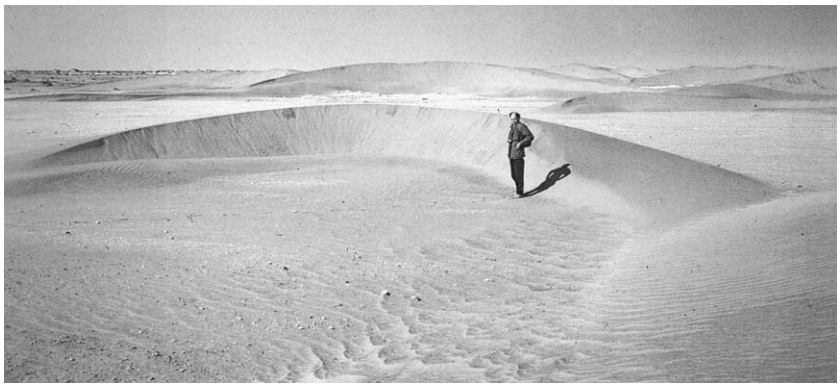


Figure 6 A field of barchans in the Namibian desert looking upwind. Note the avalanche tongues on the slip-face. Reproduced with permission from Tsoar H (2001) Types of aeolian sand dunes and their formation. In: Balmforth NJ and Provenzale A (eds.) Geomorphological Fluid Mechanics. Lecture Notes in Physics. Berlin: Springer.

Figure 7 Oblique aerial picture of seif dunes. Note the dunes' tortuosity and the sharp crest-line. Reproduced with permission from Tsoar H (2001) Types of aeolian sand dunes and their formation. In: Balmforth NJ and Provenzale A (eds.) Geomorphological Fluid Mechanics. Lecture Notes in Physics. Berlin: Springer.

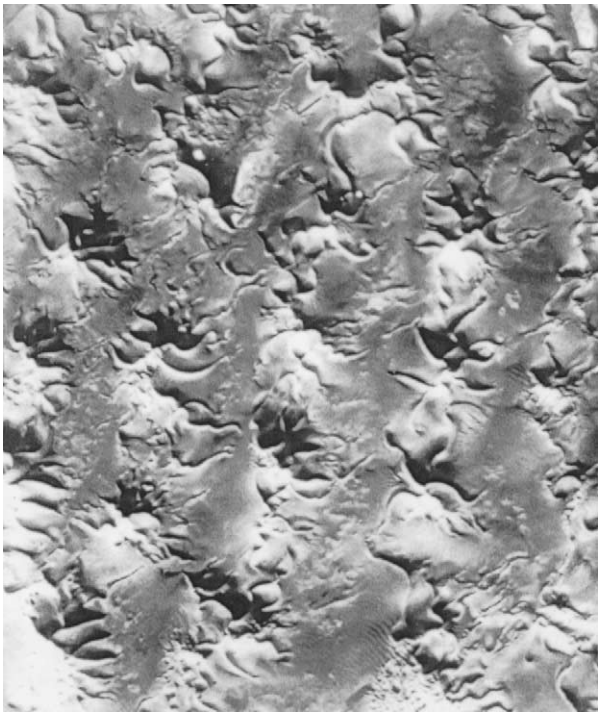


Figure 8 Aerial photograph of star dunes in Erg Oriental in Algeria. Note the radiating arms from the dune central peak. Reproduced with permission from Tsoar H (2001) Types of aeolian sand dunes and their formation. In: Balmforth NJ and Provenzale A (eds.) Geomorphological Fluid Mechanics. Lecture Notes in Physics. Berlin: Springer.

and transport to deposition. For that reason the linear seif dunes tend to meander (Figure 7) and consequently change the angle of incidence and the function of the lee side from an erosional to a depositional zone.

Star dunes as representative types of the accumulating dunes are recognized by their distinctive shape of sinuous arms radiating from a central pyramidal peak (Figure 8). Both mechanisms of migration and elongation act upon the star dunes from several directions. Star dunes are formed in a wind regime with high directional variability (RDP/DP < 0.4), and for that reason they are found mainly in the high latitudes of deserts with marked seasonal changes in wind direction.

The transverse, barchan, seif, and star dunes are formed without the presence of vegetation. There are also several typical vegetated dune types, which are located in arid and humid areas (mostly along the coasts).

Isolated clumps of vegetation act as sand traps and thus lead to the formation of nebkhas (coppice dunes) – hummocks that can reach up to 30 m high and 100 m across (Figure 9). The most effective pioneer plant forming nebkhas is *Ammophila arenaria*, which grows in clumps and can grow up through an accumulation of sand.

The most commonly found vegetated sand ridges on sandy backshores are foredunes that are formed where pioneer vegetation can grow and trap aeolian sand. Foredunes develop into continuous vegetated ridges, which lie parallel to coastlines exposed to onshore, sand-transporting winds (Figure 10). The foredune is a static dune that can only grow upward, and the only dune type that involves the exchange of sand with the beach. Other coastal dunes are mostly transgressive types (barchan, transverse, and parabolic dunes), formed when sand is transferred inland where foredunes are absent, or through blowouts (wind-excavated gaps through which sand is transported landward) in the foredune ridges.

Figure 9 Nebkha (coppice dune) formed by vegetation that traps sand.

Parabolic dunes, as their name implies, are U-shaped, with vegetated arms pointing upwind. Parabolic dunes can be active and transgressive or fully stabilized and inactive (Figure 11). The mechanism of parabolic dune formation in coastal humid areas is due to the fact that vegetation is more easily established at the base of the dune near the water table. Vegetation or dampness along the lower sides of the dune retards sand motion and both are considered to be anchors. Vegetation in parabolic dune formation is said to protect the less-mobile arms against the force of the wind action, thereby allowing the central part to advance downwind. In this way the advancing apex leaves behind trailing ridges that elongate and turn the dune into a hairpin form.

It has recently been found that barchan and transverse dunes turn into parabolic dunes in areas of low wind power where human impact is reduced. According to the profile of barchan or transverse dunes, erosion on the windward slope of the dune diminishes gradually toward the crest, which is an area

Figure 10 Foredune of approximately 5 m height formed on the backshore, where vegetation can thrive.

Figure 11 Parabolic (U-dune) formed by the effect of vegetation on sand transport.

Figure 12 Aerial photograph of vegetated linear dunes. This linear dune elongates in the direction of the dominant, strong wind. Note the straight alignment of the dune (different from the meandering seif of Figure 7) and the Y-junctions, which indicate wind direction from lower left to upper right.

of neither erosion nor deposition. Hence, once human impact stops, vegetation will recover on the barchan crest, thereby starting the process of transformation of these dunes into parabolic ones.

Vegetated-linear dunes belong to the group of elongating sand dunes found in many deserts of the world (Australian deserts, the Kalahari, Indian deserts, and the Negev). Vegetated-linear dunes are low with rounded profiles (Figure 12). They range in height from a few meters up to dozens of meters. Vegetation covers them, sometimes entirely, and sometimes abundantly on the plinth and lower slopes, but is very sparse or absent on the crest. Those that are fully covered by vegetation have become partly or wholly stabilized. Vegetated-linear dunes may run in parallel for scores of kilometers.

An exclusive attribute of vegetated-linear dunes is the tendency for two adjacent dunes to converge and continue as a single ridge. Convergence is in the form of a Y-junction (a tuning-fork shape) commonly open to formative winds. Vegetated-linear dunes are distinguished from seif dunes by: (1) coextension along the strongest dominant-wind direction; (2) the cover of vegetation; (3) straight alignment with no tortuosity; and (4) Y-junctions.

Further Reading

Bagnold RA (1941) *The Physics of Blown Sand and Desert Dunes*. London, UK: Methuen.

Danin A (1996) *Plants of Desert Dunes*. Berlin, Germany: Springer-Verlag.

McKee ED (1979) *A Study of Global Sand Seas*. Washington, DC: US Government Printing Office.

Pye K and Tsoar H (1990) *Aeolian Sand and Sand Dunes*. London, UK: Unwin Hyman.

Saturated and Unsaturated Flow *See* **Hydrodynamics in Soils**; **Vadose Zone:** Hydrologic Processes

SCALING

Contents
Physical Properties and Processes
Transport Processes

Physical Properties and Processes

G Sposito, University of California–Berkeley, Berkeley, CA, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

The physical properties of field soils exhibit a high degree of spatial variability that complicates any attempt to predict their temporal evolution using conventional models based on nonlinear partial-differential equations. This complexity can be reduced by assuming that the varying local values of soil physical properties in a field are connected to one another by arbitrary scale factors, while the form of the governing partial-differential equation remains the same everywhere in the field ('scale invariance'). This assumption turns out to require soil hydraulic properties to be power-law functions of volumetric wetness (water content) and, in particular, it requires a power-law relation between hydraulic conductivity and matric potential. These power-law relationships have often been observed in controlled field experiments, indicating the applicability of scale invariance and opening the way for fractal models of soil physical properties. However, types of relationship other than power-law have also frequently been observed, pointing to the limitation of scaling as a means of reducing the complexity of spatial variability.

Scale Invariance

The physical properties of natural soils are inexorably variable when observed from place to place. Volumetric wetness – to choose perhaps the simplest example – can vary substantially over the span of an agricultural field, even if only the surface horizon of the field soil is examined and care has been taken to provide a uniform input of water prior to examination. The complexity that this kind of variability would bring to any attempt to understand and predict the temporal evolution of volumetric wetness across an entire field is evident, and the same conclusion follows for other important soil physical properties. (*See* **Spatial Variation, Soil Properties**.)

Now suppose that a simplifying hypothesis is made: that the underlying cause of spatial variability in soil physical properties has to do with the vicissitudes of soil texture or structure, but not with the fundamental mechanisms by which the physical properties may change in response to inputs of matter and energy. If this supposition is true, then the underlying physical laws that govern the behavior of these properties must be uniform across a field, even if the properties themselves are not. Moreover, functional relationships between the properties (e.g., the relationship of volumetric wetness to matric potential) should also be uniform, in the sense that each such relationship is quantified by the same parametric equation everywhere in a field, albeit with different (and highly varying) values of the parameters at different places.

That spatial uniformity in the physical laws describing the behavior of soil properties implies a similar uniformity in the functional relationships between these properties may not seem obvious, but this conclusion follows once the physical laws are expressed mathematically in the form of nonlinear partial-differential equations. Fundamentally, imposing the condition of spatial uniformity on a governing partial-differential equation places a strong constraint on the relationships between its dependent variable and the nonconstant coefficients that appear in the equation, with the result that these relationships necessarily must also be uniform.

To make these ideas concrete, let us consider the well-known, nonlinear partial-differential equation that governs isothermal, isohaline, isochoric infiltration processes in a uniform pedon (i.e., the Richards equation in the 'mass picture'):

$$\partial\theta/\partial t = \partial[D(\theta)\partial\theta/\partial z]/\partial z - (\partial K/\partial\theta)(\partial\theta/\partial z) \qquad [1]$$

where θ ($L^3\,L^{-3}$) is volumetric wetness, $t(T)$ is time, $z(L)$ is vertical position, $D(\theta)$ (L^2T^{-1}) is water diffusivity under a gradient in θ, and $K(\theta)$ (LT^{-1}) is hydraulic conductivity. Eqn [1] describes the temporal evolution of volumetric wetness in terms of two empirical transport coefficients, $D(\theta)$ and $K(\theta)$,

whose relationship to θ must be specified in order to apply the equation in a predictive manner. Eqn [1] could be used to model infiltration in an instrumented field plot, with the expectation that $D(\theta)$ and $K(\theta)$ would have widely varying values across an array of such plots, even at the same value of θ. The hypothesis of uniformity requires that eqn [1] apply to observations of $\theta(t, z)$ anywhere in the array, irrespective of the high spatial variability. This latter supposition, in turn, requires that $D(\theta)$ and $K(\theta)$ each preserve the same functional relationship to θ anywhere in the field.

The uniformity hypothesis being imposed on eqn [1] is expressed formally by defining three arbitrary parameters that connect two sets of its dependent and independent variables evaluated at any two places where the equation applies:

$$\theta' = \mu\theta, \quad t' = \delta t, \quad z' = \alpha z \qquad [2]$$

where μ, δ, and α are (positive-valued) scale factors. These scale factors relate $\theta'(t', z')$ in one place to $\theta(t, z)$ in another, the premise being that, since the governing physical law is the same at both places, all of the spatial variability in the volumetric wetness across a field can be captured quantitatively solely by changes in the scales ('magnification' or 'reduction') over which the dependent and independent variables are observed. To be consistent with this premise, eqn [1] must still hold at the 'prime' location, i.e., after it is transformed by multiplying θ on both sides of the equation by μ, then changing t to t' on the left side and z to z' on the right side, using eqn [2]:

$$\begin{aligned}\partial\theta'/\partial t' = {} & (\alpha^2/\delta)\partial[D(\theta)\partial\theta'/\partial z']/\partial z' \\ & - (\mu\alpha/\delta)(\partial K/\partial\theta')(\partial\theta'/\partial z')\end{aligned} \qquad [3]$$

Eqn [3] will have the same mathematical form as eqn [2] only if the scale transformations:

$$\begin{aligned}D'(\theta') &\equiv D(\mu\theta) = (\alpha^2/\delta)D(\theta), \\ K'(\theta') &\equiv K(\mu\theta) = (\mu\alpha/\delta)K(\theta)\end{aligned} \qquad [4]$$

are also imposed on the two transport coefficients. Eqn [4] connects values of the water diffusivity and hydraulic conductivity at any two places. When they hold, the governing partial differential equation for $\theta(t, z)$ will have the same mathematical form under arbitrary scaling anywhere in a field. In this sense, the equation (and the physical behavior which underlies it) is said to be 'scale-invariant.' It is thus immune to the vagaries of spatial variability, these being epitomized instead in the three scale factors (eqn [2]) and the required scale transformations of the transport coefficients (eqn [4]).

Scaling Relationships

It is apparent that eqn [4] cannot be true for just any functional dependence of the two transport coefficients on θ, unless the scale factor μ always happens to have the value 1.0. In other words, if the transport coefficients do not change compatibly with the scaling of eqn [1], then scale invariance of the equation is vacuous. This critical issue can be investigated rigorously with methods in the mathematical theory of Lie groups. This investigation shows that $D(\theta)$ and $K(\theta)$ must have the following functional forms in order that eqn [4] be satisfied:

$$D(\theta) = a(\theta + b)^{\mathrm{m}} \qquad [5]$$

$$K(\theta) = \left[ca^{\beta}/(1 + m\beta)\right](\theta + b)^{1+\mathrm{m}\beta} \qquad [6]$$

where a, b, c are constants to be determined by the units of D and K and the datum of θ, while m and β are empirical parameters related to the scale factors in eqn [2]. Their relation to the scale factors emerges clearly after imposing eqn [4] on eqns [5] and [6]:

$$D'(\theta') \equiv a(\mu\theta + \mu b)^{\mathrm{m}} = \mu^{\mathrm{m}} D(\theta) = \left(\alpha^2/\delta\right)D(\theta) \qquad [7]$$

$$\begin{aligned}K'(\theta') &\equiv \left[ca^{\beta}/(1 + m\beta)\right](\mu\theta + \mu b)^{1+\mathrm{m}\beta} \\ &= \mu^{1+\mathrm{m}\beta}K(\theta) = (\mu\alpha/\delta)K(\theta)\end{aligned} \qquad [8]$$

(In eqns [7] and [8], the arbitrary constant b is assumed to be determined by the datum for θ and, therefore, to scale as θ does.)

It follows that:

$$\left(\alpha^2/\delta\right)^{1/\mathrm{m}} = (\alpha/\delta)^{1/\mathrm{m}\beta} \qquad [9]$$

and, therefore:

$$\delta = \alpha^{(2\beta-1)/(\beta-1)} \qquad [10]$$

Either eqn [9] or [10] then yields the relation:

$$\mu = \alpha^{1/\mathrm{m}(1-\beta)} \qquad [11]$$

for the scale factor of θ. Thus, only one of the three scale factors can be chosen independently once the values of m and β have been measured. (Note also that $m \neq 0$, $m\beta \neq -1$, must be finite and $\beta \neq 1$ or 1/2. These conditions are required in order that eqns [6], [9], and [10] be determinate.)

The physical significance of the scale factor α can be appreciated more incisively after using the definition of the water diffusivity:

$$D(\theta) \equiv K(\theta)\frac{\partial\psi}{\partial\theta} \qquad [12]$$

along with eqns [5] and [6] to derive an equation for the matric potential $\psi(\theta)(L)$ that is consistent with scale invariance:

$$\psi(\theta) = [(1 + m\beta)/cm(1 - \beta)](\theta + b)^{m(1-\beta)} \quad [13]$$

The corresponding equation for $\psi'(\theta') \equiv \psi(\mu\theta)$ then follows from eqn [11]:

$$\psi'(\theta') = [(1 + m\beta)/cm(1-\beta)](\mu\theta + \mu b)^{m(1-\beta)}$$
$$= \mu^{m(1-\beta)}\psi(\theta) = \alpha\psi(\theta) \quad [14]$$

The scale factor α is found to transform the matric potential from one place to another. Similarly, the transformation of the hydraulic conductivity follows from eqns [4], [10], and [11] as:

$$K'(\theta') = \alpha^{-\eta}K(\theta) \quad [15]$$

where

$$\eta \equiv (1 + m\beta)/m(\beta - 1) \quad [16]$$

The right side of eqn [15] is more conventionally expressed in terms of the square of a scale factor ω, defined by:

$$\omega \equiv \alpha^{2/\eta} \quad [17]$$

such that the scale transformation of $K(\theta)$ takes the form:

$$K'(\theta') = \frac{K(\theta)}{\omega^2} \quad [18]$$

Eqns [14] and [18] are known as 'Warrick–Nielsen scaling relationships,' and the exponent η that relates them in eqn [17] is the 'Russo–Jury exponent.'

A further connection between eqns [6] and [13] and conventional practice in soil physics can be made after defining the datum constant b to equal $-\theta_r$, where θ_r is the 'residual volumetric wetness,' a measure of the volume of 'immobile water' that does not participate in flow on the time scale over which eqn [1] is applied. Adding to this the conventional definitions:

$$\psi_s \equiv \psi(\theta_s), \qquad K_s \equiv K(\theta_s) \quad [19]$$

where θ_s is porosity, one can rewrite eqns [6] and [13] in a more convenient power-law form:

$$\psi(S) = \psi_s S^{m(1-\beta)} \quad [20]$$

$$K(S) = K_s S^{1+m\beta} \quad [21]$$

where

$$S \equiv (\theta - \theta_r)/(\theta_s - \theta_r) \quad [22]$$

is the degree of saturation. Eqns [20] and [21] indicate that this scale-invariant variable is a natural one with which to express the relationship between either matric potential or hydraulic conductivity and volumetric wetness. If S is eliminated between eqns [20] and [21], hydraulic conductivity can be connected directly to matric potential:

$$K(\psi) = K_s(\psi_s/\psi)^{\eta} \quad [23]$$

upon noting eqn [16]. The Russo–Jury exponent is now seen as the parameter that determines the power-law dependence of hydraulic conductivity on matric potential, as implied by eqn [17].

Example: Hamra Soil Field Experiment

Eqns [20–23] can be illustrated by their application to a 0.8-ha field comprising an Alfisol (Hamra series) on which 30 instrumented plots were established. Table 1 lists field-wide mean values and coefficients of variation (CVs) for ψ_s, K_s, m, and β as deduced from measurements of $\psi(\theta)$ and $K(\theta)$ made at four depths in the soil profile. Notable in the table are the typical decline in K_s with depth and the relatively large CVs (mean divided by standard deviation) for

Table 1 Mean values and coefficients of variation (CV) of the parameters in eqns [20–23] as measured in the Hamra field soil near Bet Dagan, Israel[a]

	ψ_s (m)		K_s (10^{-5} ms^{-1})		m		β	
Depth (m)	*Mean*	*CV*	*Mean*	*CV*	*Mean*	*CV*	*Mean*	*CV*
0.0	−0.072	0.22	6.12	0.41	1.86	0.68	1.46	0.68
0.3	−0.072	0.23	4.47	0.36	2.50	0.65	1.60	0.65
0.6	−0.074	0.21	2.17	0.65	3.19	0.82	1.69	0.82
0.9	−0.082	0.27	1.98	0.75	3.35	0.81	1.70	0.81
0.0–0.9	−0.074	0.24	3.66	0.69	2.48	0.88	1.60	0.88

m, β, empirical parameters related to scale factors in eqn [2].
ψ_s and K_s, 'air-entry' matric potential and saturated hydraulic conductivity, respectively.
[a]Adapted from Russo D and Bresler E (1981) Soil hydraulic properties as stochastic processes. *Soil Science Society of America Journal* 45: 682–687.

this parameter and the two exponents m and β. The lesser spatial variability of ψ_s parallels that observed for θ_s (all-depth field-wide mean of 0.37, with CV = 0.11). The field-wide mean values of the Russo–Jury exponent also tended to decline with depth: 4.3 ± 1.6, 3.3 ± 0.9, 2.9 ± 0.8, 2.9 ± 0.7; the all-depth mean being 3.3, with CV = 0.36. Other field studies show that η usually lies in the interval (2.0, 4.5) with a typical CV in the range 0.2–0.5.

The complete data set for the Hamra soil has been used to compute values of $\psi(S)$ and $K(S)$ at fixed values of S in the interval (0.750, 0.975) based on eqns [20] and [21]. The resulting all-depth mean values and CVs of the dimensionless ratios, $\psi(S)/\psi_s$ and $K(S)/K_s$, are listed in Table 2 for 10 values of the degree of saturation S. The very large CVs at smaller values of S are immediately apparent. It is this extreme spatial variability that the scale factors α (eqn [14]) and ω (eqn [18]) are designed to mediate. In particular, it should be possible to substitute the all-depth mean values of $\psi(S)/\psi_s$ and $K(S)/K_s$ into the left side of eqns [14] and [18] in order to calculate α and ω anywhere in the field (subject to imposing an all-depth mean value of 1.0 for each scale factor, since only the ratios $\psi(S)/\psi_s$ and $K(S)/K_s$ are used). The result of such a calculation is summarized in Table 3, which indicates that α and ω have virtually identical CVs over the field and that the field-wide mean value of η linking them (eqn [17]) is the same as found from examining values of the matric potential and hydraulic conductivity directly.

Table 2 All-depth, field-wide mean values and coefficients of variation (CV) for the relative matric potential ($\psi(S)/\psi_s$) and hydraulic conductivity ($K(S)/K_s$) in the Hamra soil[a]

S	$\psi(S)/\psi_s$	*CV*	$K(S)/K_s$	*CV*
0.750	25	3.9	0.14	0.92
0.775	13	3.3	0.17	0.86
0.800	7.7	2.8	0.20	0.80
0.825	4.8	2.2	0.24	0.74
0.850	3.3	1.6	0.28	0.68
0.875	2.4	1.2	0.34	0.61
0.900	1.9	0.8	0.41	0.53
0.925	1.5	0.5	0.50	0.45
0.950	1.3	0.3	0.61	0.34
0.975	1.1	0.13	0.77	0.20

[a]Adapted from Jury WA, Russo D, and Sposito G (1987) The spatial variability of water and solute transport properties in unsaturated soil. II. Scaling models of water transport. *Hilgardia* 55: 33–36.

Table 3 All-depth, field-wide mean values and coefficients of variation (CV) for the scale factors α and ω along with the Russo–Jury exponent (eqn [17])[a]

Parameter	*Mean*	*CV*
α	1.00[b]	0.435
ω	1.00[b]	0.446
η	3.36	0.354

[a]Adapted from Jury WA, Russo D, and Sposito G (1987) The spatial variability of water and solute transport properties in unsaturated soil. II. Scaling models of water transport. *Hilgardia* 55: 33–56.
[b]Unit normalization imposed.

Physical Significance of Scaling

Now let us step back to take stock of what the scaling of soil physical properties and processes actually means. The basic premise is that fundamental physical laws governing the behavior of the properties are scale-invariant. The example taken was the Richards equation for infiltration (eqn [1]), which describes volumetric wetness changes when water moves downward in a soil profile whose water diffusivity and hydraulic conductivity are known. Scale invariance of this equation implies that it is equally valid from place to place in a field, with uniform functional relationships between its two transport coefficients and the volumetric wetness. These latter three variables may fluctuate wildly across a field, but they are none the less connected by scale factors that relate their values between any two places. Their spatial variability is thus epitomized solely by differences in the local scales of volumetric wetness, position, and time (eqn [2]) that prevail at different locations, or if the degree of saturation (eqn [22]) is chosen as the dependent variable in the Richards equation, solely by differences in the scales of position and time that attend water movement at different locations in a field. Evidently these differences of scale arise from localized soil textural or structural characteristics, although the precise cause of spatial variability is left unspecified.

Implicit in the scale invariance of the Richards equation is the fact that this equation must also apply on average to an entire field of soil:

$$\partial S/\partial t = \partial[D_m(S)\partial S/\partial z]/\partial z - (\partial K_m/\partial S)(\partial S/\partial z) \quad [24]$$

where $D_m(S)$ and $K_m(S)$ are field-wide mean values of the transport coefficients. Eqn [24] is ultimate testimony to the hypothesis that the mechanism of infiltration is the same irrespective of the scale of observation. Indeed, the dependent variable in eqn [24] is scale-invariant, and the mean transport coefficients are related to local values at any place 'P' through the Warrick–Nielsen scaling relationships:

$$\alpha_P^2 D_P(S)/\omega_P^2 = D_m(S) \quad [25a]$$

$$K_P(S)/\omega_P^2 = K_m(S) \quad [25b]$$

with a corresponding equation for matric potential:

$$\alpha_P \psi_P(S) = \psi_m(S) \qquad [25c]$$

(eqn [25a] follows from eqns [12], [14], and [18].)

In field experiments like the one performed on the Hamra soil, ψ_m and K_m are first calculated, then they are used along with local values of the matric potential and hydraulic conductivity to estimate a set of α_P and ω_P, thus giving pragmatic substance to the reality of eqn [24]. Indeed, using this approach, one need only solve the mean Richards equation for a chosen boundary-value problem and have in hand a set of scale factors $\{\alpha_P, \omega_P\}$ in order to apply the solution at every point in a field – clearly an advantage in predictive applications of the flow equation. However, this computational efficiency can only be realized if there is truly no scale dependence of the mechanism of water flow, a matter for continuing field experimentation to evaluate.

The compatibility conditions in eqns [5], [6], and [13] ensure that the scale invariance of the Richards equation is not a vacuous hypothesis. The matric potential and the two transport coefficients are required to be power-law functions of the volumetric wetness or, equivalently, the degree of saturation. This power-law form has been widely used to describe the functional relationship between matric potential and volumetric wetness in field soils, under which application it is termed the 'Brooks–Corey model':

$$S = (\psi_s/\psi)^{\lambda} \qquad [26]$$

where the Brooks–Corey exponent $\lambda = 1/m(\beta - 1)$ according to eqn [20]. Representative values of the parameters ψ_s and λ have been calculated using hydraulic data on hundreds of soils according to texture. These data show that the 'air-entry value' ψ_s tends to increase, while the Brooks–Corey exponent λ tends to decrease, as the clay content of a soil becomes larger. Eqn [26] can also be derived from a fractal model of soil pore space in which the pore-size distribution is scale-invariant and $\lambda \equiv 3 - D$, where $D < 3$ is the fractal dimension that characterizes the degree of 'lacunarity' of a soil, i.e., the extent to which its pores are space-filling at the expense of the solid matrix. Given the typical range of values observed for the Brooks–Corey exponent ($0.1 < \lambda < 0.7$), it follows that $2.3 < D < 2.9$, with larger fractal dimensions corresponding to clayey soils. Thus, in clayey soils, 'lacunarity,' or the degree to which pores are space-filling, is expected to be larger than in sandy soils, in agreement with the respective porosities of these soils.

A power-law relation is also expected between hydraulic conductivity and volumetric wetness or, more generally, between hydraulic conductivity and matric potential (eqn [23]), if the Richards equation is scale-invariant. This relationship is often observed, but other models of $K(\psi)$ are abundant in the literature of soil physics. For example, the Gardner model postulates an exponential relationship instead of a power-law relationship for $K(\psi)$:

$$K(\psi) = K_s \exp\left[-\frac{\psi_s - \psi}{D_s/K_s}\right] \qquad [27]$$

If this equation accurately describes the behavior of hydraulic conductivity in a soil, the Richards equation cannot be scale-invariant over a field containing that soil. Indeed, application of the scaling relations in eqns [14] and [15] to both sides of eqn [27] is impossible. The reason for the failure is signaled by the presence of the ratio, D_s/K_s, in eqn [27]. This ratio has the dimensions of length and serves as a measure of how quickly $K(\psi)$ declines as ψ decreases below the air-entry value in a 'Gardner soil.' It is, therefore, an intrinsic length scale in the Gardner model, i.e., a length scale that is determined uniquely by the values of D_s and K_s at each separate location in a field soil. In terms of conventional water-desorption methods of measuring the cumulative pore-size distribution in a soil (by combining data on $\psi(S)$ with the Young–Laplace equation to calculate the effective pore radius as a function of the degree of saturation), the length scale D_s/K_s is interpreted as that of the very first pores to empty when ψ drops below the air-entry value, ψ_s. (*See* **Porosity and Pore-Size Distribution**.) The existence of such an intrinsic length scale stands in direct contradiction to the use of an arbitrary scale factor for relating matric potential at two places in a field (eqn [14]). Therefore, in a Gardner soil, the arbitrary scaling of matric potential (and, correspondingly, of all the variables in eqn [1]) is not possible because the intrinsic length scale D_s/K_s preempts that possibility. This example of broken scaling symmetry is valuable, in that it illuminates by contrast the underlying significance of eqns [14–18]: that the flow equation operates everywhere in a field soil without prejudice as to the local scales of length, time, or water content.

See also: **Infiltration**; **Porosity and Pore-Size Distribution**; **Spatial Variation, Soil Properties**

Further Reading

Anderson AN, McBratney AB, and Crawford JW (1998) Applications of fractals to soil studies. *Advances in Agronomy* 63: 1–76.

Baveye P, Parlange J-Y, and Stewart B (1998) *Fractals in Soil Science*. Boca Raton: CRC Press.

Crawford JW and Matsui N (1996) Heterogeneity of the pore and solid volume of soil. *Geoderma* 73: 183–195.
Crawford JW and Sleeman BD (1998) From particles to architecture: fractals, aggregation and scaling in soil science. In: Huang PM, Senesi N, and Buffle J (eds) *Structure and Surface Reactions of Soil Particles*, pp. 41–75. Chichester, UK: John Wiley.
Edwards MP (1994) Classical symmetry reductions of nonlinear diffusion-convection equations. *Physics Letters A* 190: 149–154.
Hillel D (1998) *Environmental Soil Physics*. San Diego, CA: Academic Press.
Hillel D and Elrick DE (eds) (1990) *Scaling in Soil Physics*. Madison, WI: Soil Science Society of America.
Jury WA, Russo D, and Sposito G (1987) The spatial variability of water and solute transport properties in unsaturated soil. II. Scaling models of water transport. *Hilgardia* 55: 33–56.
Millán H and Orellana R (2001) Mass fractal dimensions of soil aggregates from different depths of a compacted vertisol. *Geoderma* 101: 65–76.
Pachepsky YaA, Crawford JW, and Rawls WJ (eds) (2000) *Fractals in Soil Science*. Amsterdam, the Netherlands: Elsevier.
Perrier E, Rieu M, Sposito G, and de Marsily G (1996) Models of the water retention curve for soils with a fractal pore size distribution. *Water Resources Research* 32: 3025–3031.
Rawls WJ and Brakensiek DL (1995) Utilizing fractal principles for predicting soil hydraulic properties. *Journal of Soil and Water Conservation* 50: 463–465.
Rieu M and Sposito G (1991) Fractal fragmentation, soil porosity, and soil water properties. *Soil Science Society of America Journal* 55: 1231–1244.
Russo D and Bresler E (1981) Soil hydraulic properties as stochastic processes. *Soil Science Society of America Journal* 45: 682–687.
Sposito G (ed.) (1998) *Scale Dependence and Scale Invariance in Hydrology*. Cambridge, UK: Cambridge University Press.
Wiltshire R and El-Kafri M (2004) Non-classical and potential symmetry analysis of Richards equation for moisture flow in soil. *Journal of Physics A: Mathematical and General* 37: 823–839.

Transport Processes

R P Ewing, Iowa State University, Ames, IA, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

'Scaling of transport properties' refers to the common and often troublesome observation that the model parameters describing dispersion at one scale are frequently incorrect at another scale. Worse, sometimes one model is applicable at one scale, and a different model at another scale. The word 'scale' denotes extent, most often of space or time, so to say that dispersion scales is to say that different results are obtained from measuring it at different spatial or temporal scales. The most common situation is that one measures dispersion at a small scale and predicts it at a larger scale, then finds that the actual dispersion as measured later at the larger scale is greater than predicted. Further evidence of scaling is the observation that dispersion coefficients compiled from many separate experiments show a strong trend of dispersion increasing with distance. Is this scaling inherent to the dispersion process?

Because prediction of solute transport is generally based upon the advection–dispersion equation (ADE), which uses a diffusion-type description of transport, to say that transport scales often implies that it does not conform to a simple diffusion-like description. Scaling of dispersion is often thought to be an exceptional case, sometimes called 'anomalous dispersion,' implying that dispersion which obeys the ADE is the norm. But as more so-called anomalous transport has been found, it is increasingly accepted as both more general and widespread: 'normal' transport actually represents a special and unusual case. In this article, dispersion that conforms to the ADE will be called 'Gaussian.'

This article discusses the simple case of one-dimensional dispersion of a nonreactive, neutral, conservative solute during steady-state isothermal saturated flow, ignoring any density and viscosity effects of solute concentration. This simple case is already sufficiently difficult, but the reader must remember that real-world situations also include reactive, degrading, sorbing, and/or charged solutes moving nonisothermally and transiently under multiple driving and dispersing forces.

Diffusive–Dispersive Transport

Hydrodynamic dispersion (called 'dispersion' here for convenience) is the result of two mechanisms: mechanical dispersion and molecular diffusion. In simplistic terms, mechanical dispersion is caused by the medium, while diffusion is inherent to the fluid. Mechanical dispersion results from differences in travel times along different streamlines. At the scale of a single pore, these differences result from flow being slower close to a pore wall and faster at pore centers. At a larger scale, streamlines differ in velocity due to heterogeneity of the medium, such as variations in pore size and connectivity, and the presence of structure across multiple scales from pore-scale correlations to large-scale layering and fractures. When flow is rapid, mechanical dispersion

contributes most of the total dispersion, but at slow flow rates dispersion is predominantly diffusive.

Molecular diffusion is the random motion of fluid molecules, so diffusion of solutes is a function of the fluid (solvent and solute). The diffusion equation describes net solute flux as being proportional to, and opposite in direction from, the concentration gradient. In one dimension and in bulk fluid (no solid phase):

$$\frac{\partial C}{\partial t} = D_{\text{aq}} \frac{\partial^2 C}{\partial x^2} \qquad [1]$$

where C is solute concentration (M L^{-3}), t is time (T), D_{aq} is the diffusion coefficient ($\text{L}^2\,\text{T}^{-1}$) of the solute in water, and x is distance (L). In the presence of a solid phase, the molecular diffusion coefficient can be adjusted by the porosity ϕ of the medium and the tortuosity τ of the diffusive path length, giving an effective diffusion coefficient: $D_{\text{eff}} = D_{\text{aq}}\,\phi/\tau$. If solute is introduced at time t_0 in a plane at concentration M per unit area, within an infinite medium having initial concentration $C_0 = 0$, then the resulting concentration at a distance x from that plane at time t is:

$$C(x,t) = \frac{M}{2\sqrt{\pi D t}} \cdot \mathrm{e}^{-\mathrm{x}^2/4\,\mathrm{Dt}} \qquad [2]$$

which is a Gaussian (normally distributed) concentration profile whose width increases in proportion to the square root of time.

Dispersion is often estimated using data from an experiment in which water in the sample is replaced by water with a tracer. If the output is tested for the presence of the tracer, a steady increase in flux concentration is seen as the original water is replaced by water with the tracer. If instead the tracer's resident concentration is measured inside the sample at a single instant, a decrease in tracer concentration is seen from the inlet to the outlet end of the sample. The output of a flux concentration dispersion experiment can be summarized graphically using a breakthrough curve (BTC) (Figure 1). In these curves, the x-axis is presented in units of pore volumes, the amount of water needed to replace the original water in the soil sample precisely once. The y-axis is expressed as the ratio of the measured to the final concentration, C/C_0. This normalization of the two axes allows curves from different experiments (flow rates, column lengths, etc.) to be compared.

Sir Geoffrey Taylor and the Spreading of Solute in a Tube

While diffusion is only one component of dispersion, the physicist Sir Geoffrey Taylor showed that

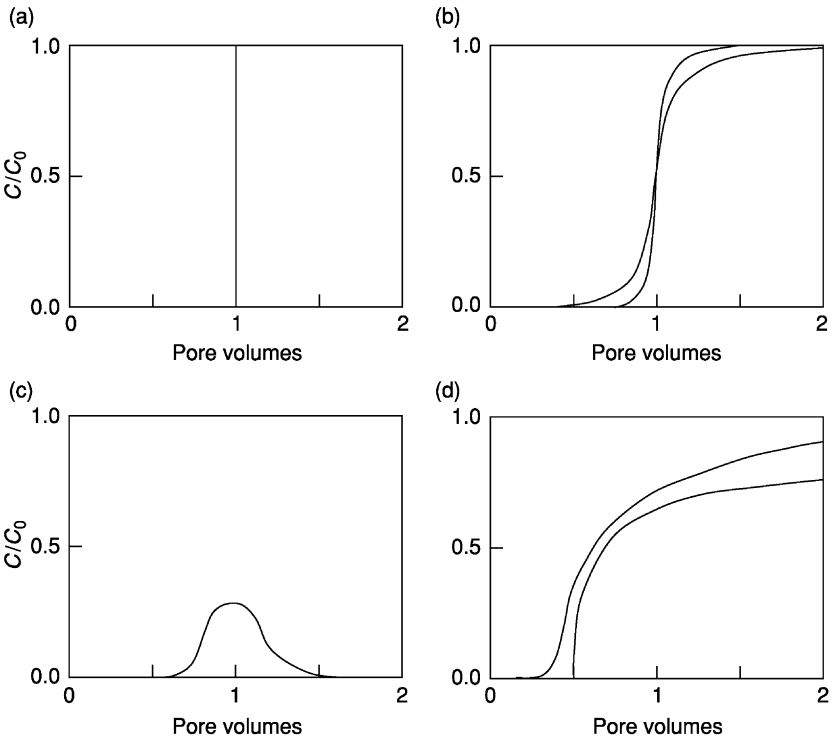


Figure 1 Sample flux concentration breakthrough curves for: (a) a step input with piston flow; (b) a step input with Gaussian dispersion; (c) a Dirac (pulse) input with Gaussian dispersion; and (d) non-Gaussian dispersion. See text for explanation.

dispersion of a solute in fluid flowing steadily through a tube can – under certain conditions – be precisely described by superimposing the mean fluid velocity on to a diffusion-like dispersion term. The main condition is simply that flow be sufficiently slow that solute molecules have time to diffuse freely between fast (tube center) and slow (tube wall) regions, a condition that is often well approximated, at least at the pore scale, under natural conditions in soils and aquifer materials. Taylor's experimental setup – a uniform tube with steady flow – clearly does not match any naturally occurring porous media. Nonetheless, the analogy between dispersion and diffusion has historically been applied to solute movement in porous media generally. The analogy is particularly attractive because the diffusion equation has been widely studied, and operating within this analogy allows use of a wealth of mathematical solutions.

Sir Geoffrey Taylor's three seminal papers on dispersion of solutes in a tube, combined with Rutherford Aris' subsequent analysis, established much of how both engineers and soil scientists now think about dispersion. An earlier paper by Griffith had stated that a drop of dye, placed into water flowing through a glass tube, moves at the mean water velocity, with its dispersion along the tube being symmetrical. Reading Griffith's paper, Taylor was struck by its two 'remarkable' implications: first, movement with the mean water velocity implied that water in the center of the tube, flowing at twice the mean velocity, must pick up dye as it traveled toward the center of the dye slug, then release it after passing the center of the slug; and second, the longitudinally symmetrical dispersion seemed at odds with the radially asymmetrical (parabolic) velocity distribution of water in the tube. Taylor set to work to examine these phenomena both experimentally and mathematically.

Starting by assuming that there was no diffusion at all, Taylor realized that the parabolic velocity profile:

$$v(r) = v_0(1 - r^2/a^2) \qquad [3]$$

(where $v(r)$ is velocity ($\mathrm{L\,T^{-1}}$) at a distance r from the tube center, v_0 is the maximum velocity, a is the tube radius (L), and r is distance from the tube center (L) such that $0 \leq r \leq a$), coupled with a step input of dye, would yield the linear breakthrough curve:

$$C/C_0 = \left(1 - \frac{x}{v_0 t}\right) \qquad [4]$$

where t is time (T), for $0 < x < v_0 t$. He also noticed that the diffusion coefficient D_{aq} ($\mathrm{L^2\,T^{-1}}$) was small in comparison with the observed dispersion coefficient $D(\mathrm{L^2\,T^{-1}})$: apparently longitudinal diffusion (diffusion in the direction of flow) was an insignificant contributor to dispersion. Radial diffusion (normal to the direction of flow), however, could explain the movement of solute molecules between the fast-flowing tube center and the negligible velocities near the wall.

Taylor's analysis of the interaction between radial diffusion and the parabolic velocity profile shows that the dispersion behaves exactly like diffusion, that is, it yields a Gaussian concentration profile the width of which grows in proportion to the square root of time (Figure 2). The theoretical dispersion coefficient D, which agrees well with his observations, is given by:

$$D = \frac{a^2 v_0^2}{192\, D_{aq}} \qquad [5]$$

that is, dispersion is proportional to velocity squared. But this relationship only holds if two conditions are met. The first condition states that $av/D_{aq} \gg \sqrt{48}$, where v is the mean velocity. The quantity av/D_{aq} actually defines the Peclet number (the ratio of advective to diffusive transport), so this condition can be simply restated as $\mathrm{Pe} \gg 7$. Aris, noting that Taylor's D explicitly excluded longitudinal diffusion, added it back into Eqn [5] to obtain:

$$D = D_{aq} + \frac{a^2 v_0^2}{192\, D_{aq}} \qquad [6]$$

thereby eliminating the need for this first condition. Notice that the resulting dispersion coefficient is explicitly the sum of diffusive and mechanical dispersive forces.

The second condition states that $a^2/4D_{aq}t$ must be small; that is, there must be sufficient time for radial

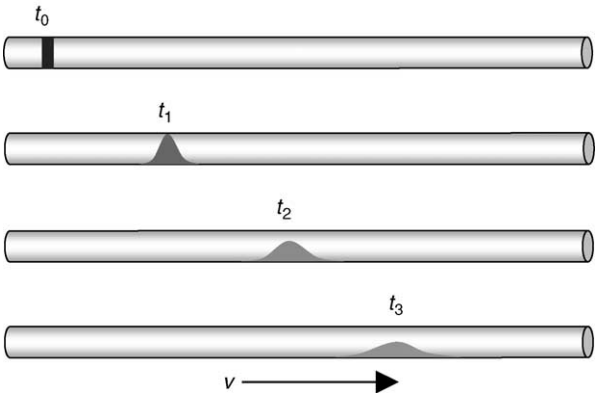


Figure 2 How the solute flows and disperses with time in a tube into which solute is introduced at a single instant in time.

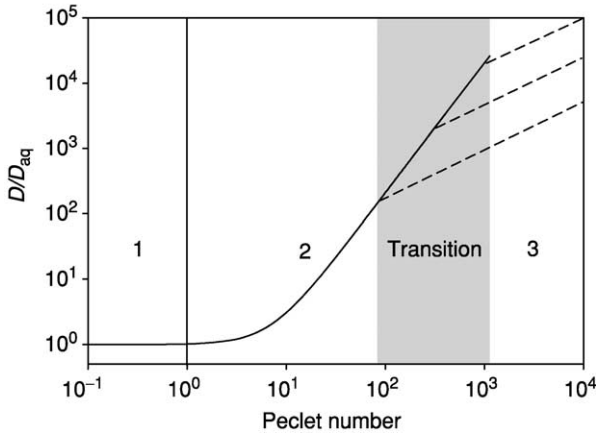


Figure 3 Mechanical dispersion over total dispersion as a function of the Peclet number. Above some Peclet number whose exact value varies with length, dispersion transitions to being linear with time; below that, it scales with the square of time. See text for explanation.

diffusion to move molecules a distance equal to the radius of the tube. When the elapsed time is too short, radial diffusion is insignificant and the dispersion can be approximated by Eqn [4], Taylor's first analysis, based upon the assumption of no diffusion. In terms of the Peclet number, this second condition may be expressed as $\mathrm{Pe} \ll 4L/a$, where L is the distance from the dye injection point at which the measurement is made.

The conditions for Taylor's analysis can be shown by plotting the ratio of total dispersion to diffusion as a function of the Peclet number (**Figure 3**). The solid line represents the Aris equation (Eqn [6]), which shows different behavior in each of three regions; the dashed lines show the behavior that would be seen if the second condition were violated. In the first region ($\mathrm{Pe} \ll 7$), diffusion is an important contributor to total dispersion and so Taylor's original analysis (Eqn [5]) does not hold, although Aris' (Eqn [6]) does. In the second region, both conditions are met and the slope of the line is approximately 2. The Peclet number above which the second condition fails, $\mathrm{Pe} \ll 4L/a$, is shown as a gray region rather than a single line because it changes for different L. In this third, preasymptotic region, dispersion is linearly related to velocity. Because for any given velocity and tube radius the second condition does not hold for some short time or distance, we can refer to short times and/or distances as being 'preasymptotic.'

Solute Spreading in Soils

Superimposing a mean velocity on to the diffusion equation yields the ADE, also called the convection–dispersion equation (CDE). A one-dimensional form of the ADE for nonreactive solutes describes the change in concentration as resulting from the sum of dispersive and advective terms:

$$\frac{\partial C}{\partial t} = D\frac{\partial^2 C}{\partial x^2} - \nu\frac{\partial C}{\partial x} \qquad [7]$$

where D is the dispersion coefficient ($\mathrm{L}^2\,\mathrm{T}^{-1}$) and ν is the mean water velocity ($\mathrm{L}\,\mathrm{T}^{-1}$). Analytical solutions to this equation have been derived for a wide variety of initial and boundary conditions, and numerous computer programs are also available. In a simple case, a point injection of a solute will move at the mean water velocity, while spreading as if it were diffusing with a diffusion coefficient equal in magnitude to the dispersion coefficient D. Non-Gaussian behavior is often seen at early times and/or short distances, but such situations will yield increasingly Gaussian behavior with increasing distance and/or time. Accordingly, non-Gaussian behavior is sometimes called 'preasymptotic,' implying that all dispersion eventually becomes Gaussian.

In the field, dispersion is often observed to increase linearly with fluid velocity. The CDE can in this case be generalized with respect to velocity by defining the dispersivity $\alpha(\mathrm{L})$ as $\alpha = D/\nu$, and replacing the dispersion coefficient D in the ADE with $\alpha\nu$. This dependence on velocity may seem counterintuitive: if dispersion is the sum of molecular diffusion and mechanical dispersion, then mechanical dispersion (which results from the medium) should be independent of velocity, whereas slower flow (having more time) should result in greater diffusion-based dispersion. But as Taylor's studies of dispersion in a tube have shown, D increases with ν within a specific range of Peclet numbers, so dispersion increasing with velocity is not entirely surprising. On the other hand, dispersion in soils is observed to be proportional to ν rather than ν^2, so Taylor's analysis provides only a partial explanation of the whole process.

Unfortunately, the assumption that dispersion can be treated mathematically as if it were diffusion often leads to a poor match between model and data. This is frequently because one or more of the equation's assumptions have been violated: probably either the medium is not uniform (let alone tubular!) or there is insufficient exchange between solute molecules traveling on different streamlines. Because diffusion and flow are not strictly additive with respect to dispersion except under specific conditions, the basic concept of the ADE is not sufficiently general for many situations found in the real world. Much of the recent research on solute dispersion concerns struggles with the poor predictions produced by the ADE, frequently focusing on developing alternative conceptual models.

Scale

Soil science is concerned with processes taking place from the nanometer to the kilometer scale and beyond, and from the near-instantaneous to geologic timescales. It is common to speak of scales in a colloquial manner – 'pore-scale,' 'lab-scale,' 'overnight' – and the mental images that these terms invoke can be useful, although they are not quantitative. Less useful are terms such as 'microscale,' which naturally carry widely different meanings between, and even within, different disciplines. Because soil science is inherently interdisciplinary, such words will have different meanings even to different soil scientists and should be avoided. If quantitative terms are needed, then using the size itself is best: 'meter-scale,' for example.

Some properties are specific to a defined range of scales. To illustrate, suppose one wants to measure the porosity of a sand using computed tomography (CT). If the CT unit measures with a resolution of 0.01 mm, then most voxels (volume elements from the CT scan) will represent either the inside of a solid particle or the center of a pore. In neither case will it be useful to speak of the porosity of an individual voxel: porosity is a property of a volume and becomes meaningless as that volume shrinks to a point. If we coarsen the scale slightly, each individual voxel will sample just a few grains and pores, and so individual measurements will be highly variable. But when porosity is measured over a sufficiently large number of grains and pores (at least 30 is a good guideline), the measured value is stable. This illustrates the concept of the representative elementary volume (REV), the smallest volume over which a given property can be measured with reasonable confidence. Notice that the REV must be larger than the characteristic scale of variability of the medium (if one exists).

At this point we must also make a distinction between scale and level. Scale is strictly about physical extent, while level involves a conceptual difference or a change in properties. The two are easily confused, because changes in physical size often entail changes in level also. For example, a soil core 10 cm deep may come from a single horizon, but 'scaling up' to a 20-cm core may yield a sample with two distinct horizons in it – a difference in level as well as scale. Likewise, scaling up from a 10-cm core to a pedon may introduce fractures or wormholes into the system, and so, rather than studying a soil matrix, one now faces a macropore-matrix or dual-permeability system. The distinction between scale and level may be quite subtle. For example, buoyancy effects may not be evident during fast flow of a weak salt solution through a 1-cm-diameter soil column, but they may be evident during slower flow through a larger-diameter soil column. In the second case, the increase in size (scale) allows the emergence of a previously negligible mechanism (level), so the second system under study is not equivalent to the first. It can be useful to compare relevant dimensionless groups to help discern whether changing the scale also allows new processes to come into play. Relevant dimensionless groups may include the Peclet number (ratio of advective to diffusive forces), the Bond number (ratio of buoyancy to capillary forces), the Reynolds number (ratio of inertial to viscous forces), and the Schmidt number (ratio of kinematic viscosity to molecular diffusivity).

A further difficulty is that many parameters of interest can each be best measured within a fairly narrow range of scales. For example, dispersion is generally estimated from solute concentration, based upon samples at the scale of a liter or less. Measuring solute concentration over a large area is generally accomplished by collecting many small samples, because collecting the entire outflow of a large area is generally impractical, if not impossible. So studies of dispersion generally operate at many scales: the scale of the individual observations (where different observations may themselves have different scales), the scale of the current composite (for example, the inferred pollutant plume based upon some 100 individual core samples), the scale of the desired prediction, and the characteristic scale, if any, of the soil's heterogeneity. While we are mainly concerned here with the disparity between the middle two, any difference in scale may contribute to scaling.

Conceptual Models

Several conceptual models are in common use for predicting dispersion. The ADE, described above, is widely used in both subsurface hydrology and soil science. Conceptually equivalent to flow through a tube (**Figure 4a**), it does not adequately describe the complexities encountered in flow through soils and so frequently fails to predict dispersion correctly. One common failure is finding that the value of the dispersion coefficient increases with the distance over which the solute has moved. A second failure is a non-Gaussian output which persists across many length scales, despite the prediction of the ADE that the dispersion will asymptotically become Gaussian. A common instance of this is 'slow tailing,' in which the breakthrough curve resulting from a step input shows early arrival and a slower-than-Gaussian increase to the final concentration. This has been addressed in some models by the use of a fractional ADE (fADE), which uses a fractional form of the error

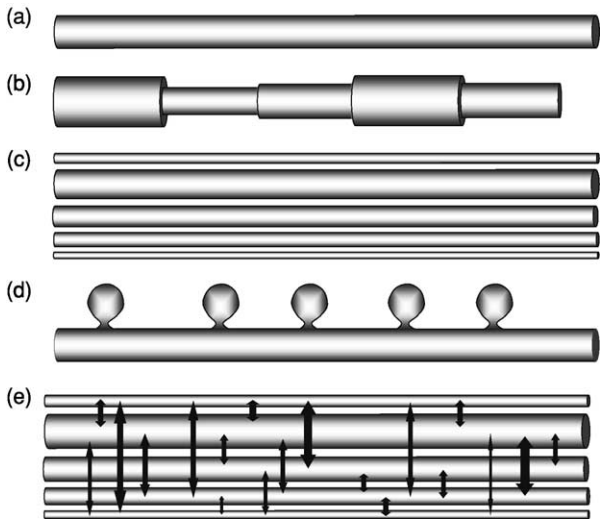


Figure 4 (a) A single tube with constant cross section; (b) a tube with varying cross section; (c) a bundle of tubes of different but constant cross sections; (d) a single tube with dead-end 'pores'; and (e) a bundle of tubes of different cross sections with solute exchange between tubes.

Table 1 Scaling with time (T) of solute mass mean and standard deviation according to four transport models

Model	*Mean*	*Standard deviation*	*Standard deviation/mean*
ADE	T	$T^{1/2}$	$T^{-1/2}$
MIM	T	$T^{1/2}$	$T^{-1/2}$
CLT	T	T	Constant
CTRW ($0 < \beta < 1$)	T^{β}	T^{β}	Constant
CTRW ($1 < \beta < 2$)	T	$T^{(3-\beta)/2}$	$T^{(1-\beta)/2}$

ADE, advection–dispersion equation; MIM, mobile–immobile model; CLT, convective lognormal transfer function model; CTRW, continuous-time random walk model.
Adapted from: Jury WA, Gardner WR, and Gardner WH (1991) *Soil Physics*, 5th edn. New York: John Wiley; Skaggs TH and Leij FJ (2002) Solute transport: theoretical background. In: Dane JH and Topp GC (eds) *Methods of Soil Analysis*, part 4, chap. 6.3, *Physical Methods*. Madison, WI: Soil Science Society of America; Berkowitz B, Kosakowski G, Margolin G, and Scher H (2001) Application of continuous time random walk theory to tracer test measurements in fractured and heterogeneous porous media. *Ground Water* 39(4): 593–604.

function to generate a near-Gaussian distribution with 'fat tails.' This approach, being more general, usually results in a better fit than the original ADE. However, it is somewhat more complex mathematically and only accounts for a subset of cases that are not well fit by the ADE. Interestingly, the ADE is unable to account correctly for dispersion resulting from flow through layered soils, conceptually equivalent to different tubes in series, as shown in Figure 4b: it predicts identical results regardless of which soil comes first, while experimentally the order of the layers affects the final value of the dispersion coefficient.

A second common conceptual model is the so-called streamtube model or transfer function model, in which solute transport is considered to take place through multiple parallel but noninteracting streamlines (Figure 4c). Because mean solute velocity within a tube is proportional to tube radius squared, dispersion results from the differences in velocity among the various streamtubes. Because in principle it is possible to reproduce any possible breakthrough curve by generating the correct streamtube distribution, this method is very flexible. This model is essentially a multiple-tube equivalent of the dispersion that would result from flow in a tube in the absence of diffusion, and so, because their dispersion results from insufficient mixing, streamtube models may be considered to be preasymptotic. Streamtube models are most applicable to media in which flow really does take place through parallel but noninteracting pathways, such as horizontal flow through an aquifer with homogeneous layers of contrasting properties. Some variants of this model allow dispersion within each streamtube, while others assume no local dispersion (i.e., piston flow). One common version, Jury's convective lognormal transfer function model (CLT), specifies the distribution of the solute travel times (or, by analogy, the tube radii) and thus the form of the final BTC.

Knowing the dispersion at one scale only allows correct prediction at a different scale if the chosen model scales the same way as the observations. For example, while the ADE predicts that solute-spreading scales with the square root of time, the CLT (a specific streamtube model) predicts solute-spreading that scales linearly with time (Table 1). Solute movement that is consistent with a streamtube-like mechanism will therefore appear to scale if analyzed using the ADE; likewise, solute transport that is well described by the ADE will scale incorrectly if analyzed using the CLT.

The streamtube models assume zero interaction between streamlines, in contrast to the ADE, which assumes that streamlines do interact. These two models therefore represent extremes along a continuum of streamline interaction. In principle, the streamtube model should better represent early or preasymptotic times, before streamlines have had time to intermix adequately, and the ADE later times. Because in the streamtube model dispersion scales linearly with time (Table 1), invoking a streamtube model at short (preasymptotic) times and the ADE at later times provides a plausible explanation for some scaling behavior: one can argue that the observations took place within the transition between the two models. By this reasoning, the time or distance needed to mix – to make the transition from

streamtube-like dispersion to ADE-like dispersion – should provide an estimate of the soil's REV with respect to dispersion, and so some indication of the characteristic length of the soil's structure.

A third model in common use is the mobile–immobile (MIM) model, also known as the two-region model. Originally developed to explain dispersion in terms of dead-end pores, this model has been widely adopted to describe dispersion in structured soils. The idea is that the soil is composed of mobile regions (such as interaggregate pores) through which most or all of the flow actually moves, closely interconnected with immobile regions (e.g., the interior of aggregates) that exchange mass with mobile regions by diffusion only (Figure 4d). The parameters of the model are therefore the dispersion in the mobile region (identical to the dispersion term in the ADE), the mass exchange coefficient, describing diffusive exchange between the mobile and immobile regions, and the mobile pore fraction. The model can equivalently be thought of as describing chemical sorption to soil particles that comes in two varieties, instantaneous and slow, but, in keeping with the focus here on nonreactive solutes, this view will be ignored. Because diffusive exchange between regions is generally slow relative to advection, a given layer of soil is generally out of equilibrium with the solute flowing through it for some time, so this model is often said to describe nonequilibrium flow. When the immobile region has zero volume, or when the diffusive exchange is instantaneous, the MIM reduces to the ADE; this extreme can be though of as describing a soil of low heterogeneity. As might be expected from its close ties with the ADE, the MIM predicts identical scaling behaviors (Table 1). On a practical level, the MIM often suffers from nonunique solutions; that is, a given BTC may be equally well fit by more than one single set of parameters. This nonuniqueness implies that the equation has too many fitting parameters (three for the MIM versus one for the basic ADE), and casts some doubt on the physical interpretation of those parameters.

Observations and Mechanisms

There is a strong tendency for dispersivity to be greater at larger scales, as documented (for example) in several metastudies, each examining many different field studies of solute transport. Unfortunately, the scatter in the data is great enough that no universal scaling behavior can unambiguously be inferred. While the cause of this trend is not clear, the most common explanation is that, as a solute travels farther, it encounters increasing scales of heterogeneity. This means that at larger scales solute molecules will on average sample a larger range of velocities than they did at a smaller scale. The travel time for each molecule is the sum of many different times at many different velocities; increasing the range of velocities will therefore (assuming random and independent sampling by each molecule) result in a wider distribution of travel times and thereby a larger dispersivity. A fractal medium, with its ever-increasing scale of heterogeneity, results mathematically in dispersivity increasing with distance, i.e., scaling.

But increasing scales of heterogeneity are apparently not necessary for scaling to occur. For example, one laboratory experiment used a two-dimensional flow cell packed with two sands: 16% of the volume consisted of blocks of fine sand randomly placed within a coarse sand matrix, with the hydraulic conductivities differing by a factor of 20. Within approximately 1 m, the fit dispersivity increased from 0.073 to 0.127 m. The skew of the data was fairly constant, suggesting that a change from preasymptotic to asymptotic behavior was not occurring. Similarly, the Borden aquifer tests saw dispersivity increase by a factor of 8 over a distance of some 10 m in what was considered a homogeneous sand. And of course dispersion that occurs primarily through a streamtube-like mechanism, wherein regions that differ in velocity are isolated from each other, will also show scaling of ADE-calculated dispersivity.

Discussion of the 'scale of heterogeneity' also raises the question of the importance of heterogeneity at smaller scales. If much of dispersion is driven by heterogeneity, and that heterogeneity exists at all scales, then presumably the nature of the heterogeneity below the scale of heterogeneity is also important. But because both models and measurements are blind to processes below their respective nonzero-size scales, there are always smaller-scale processes that they cannot take into account, but which still influence the large-scale outcome. An exclusive focus on the largest-scale heterogeneities implicitly assumes that the smaller-scale heterogeneities are either negligible or diffusive in nature, but neither is necessarily the case. The smaller-scale influences can be significant and nondiffusive; for example, the presence of undetected cracks in a soil profile may lead to arrival of solute at the water table well in advance of prediction.

Soils and other geologic porous media are neither purely random nor purely structured; instead they vary in their degree of randomness across most or all scales. But while understanding of soil properties has historically concentrated on the measurements themselves, and only relatively recently on the spatial structure of the properties, generally the topological structure is still ignored. As a consequence,

researchers in the theory of contaminant transport have long considered (for example) what proportion of an aquifer is sand or clay, and even how thick the average sand layer might be, but not the question of whether there is a continuous sand pathway through the system. Continuity and interconnectedness are critical to diffusive and advective exchange across streamlines, and to the range of velocities encountered within a given streamline. They therefore make the difference between the ADE and streamtube models, and so are critical determinants of the scaling behavior of a given soil or aquifer.

In most situations, solute transport in soils or other geologic porous media is subject to all of the mechanisms mentioned so far: diffusive exchange between neighboring streamlines, parallel pathways with little exchange between them, and diffusive exchange between more and less mobile regions of the medium. There are also other important dispersive processes, such as advective (and even turbulent) exchange between streamlines, streamline-specific sorption based upon mineralogy, biodegradation based upon differences in microhabitat, and streamlines with nonconstant velocity, that do not show up in these models. Certainly the extent to which each mechanism contributes to dispersion is difficult to determine in advance, nor can one know in advance whether the length scale at which dispersion is to be predicted is greater than or less than the scale of heterogeneity (if any) of the medium.

Model Selection

An investigator attempting to predict future dispersion from current conditions is faced with the choice of three conceptual models, each based upon quite different mechanisms and assumptions. Prediction therefore involves two tasks: selecting the appropriate model, and fitting it to the data in order to make a prediction. The wrong model choice may result in scaling, such that the predicted behavior may be severely rather than merely slightly wrong. But in fact the choice of models is still quite narrow, in the sense that only two scaling behaviors are available (Table 1), even though intermediate behaviors are commonly observed.

An alternative is to use a new model, originally developed to describe electrical conductivity of amorphous silicon. The continuous-time random walk model (CTRW) has been used to describe solute transport only since the mid-1990s, but it has been shown to be capable of capturing the entire range of behavior seen in the three models discussed above. Each of the three models is a special case of the CTRW, indicating that the CTRW is a generalized transport model that can handle both the extremes and the intermediate cases.

The CTRW conceives of the soil as having a variety of pathways of higher and lower conductivity. What determines the dispersion is not only the range of velocities represented by these pathways, but also the probability with which a solute molecule will transition from one pathway to another (Figure 4e); such transitions can take place both through diffusion and through advective pathways merging and separating. Transition probabilities are given by a joint probability distribution function $\psi(s,t)$, which describes the probability of a particle moving a distance and direction s within time t. The long-time behavior is dominated by transitions at long (asymptotic) times, and asymptotic forms of probability functions tend to decrease monotonically. (In the context of the CTRW, asymptotic behavior is always reached quickly.) The distribution can then be characterized by exponential decay, which leads to ADE-like behavior, and power-law decay in which $\psi(s,t) \sim t^{-1-\beta}$ for $0 < \beta < 2$. Power-law decay yields anomalous transport except for $\beta > 2$. The value of the exponent β controls the rate of decay, and thereby the scaling of the solute mean and standard deviation (Table 1).

A wider range of scaling behaviors is attained with the CTRW than with the original three models combined. Significantly, because the choice of β changes both the shape of the BTC and its temporal scaling, the CTRW implicitly considers temporal scaling to be an inherent aspect of dispersion, rather than a separate phenomenon. This raises the possibility that in the future the concept of scaling as a distinct issue will cease to exist, being replaced by an understanding that dispersion has inherent scale-related behaviors.

See also: **Fractal Analysis**; **Porosity and Pore-Size Distribution**; **Scaling:** Physical Properties and Processes; **Spatial Variation, Soil Properties**

Further Reading

Baveye P and Boast CW (1999) Physical scales and spatial predictability of transport processes in the environment. In: Corwin DL, Loague K, and Ellsworth TR (eds) *Assessment of Non-Point Source Pollution in the Vadose Zone*, pp. 261–280. Geophysical monograph 108. Washington, DC: American Geophysical Union.

Berkowitz B, Kosakowski G, Margolin G, and Scher H (2001) Application of continuous time random walk theory to tracer test measurements in fractured and heterogeneous porous media. *Ground Water* 39(4): 593–604.

Crank J (1975) *The Mathematics of Diffusion*, 2nd edn. Oxford, UK: Clarendon Press.

Gelhar LW, Welty C, and Rehfeldt KR (1992) A critical review of data on field-scale dispersion in aquifers. *Water Resources Research* 28(7): 1955–1974.

Jury WA, Gardner WR, and Gardner WH (1991) *Soil Physics*, 5th edn. New York: John Wiley.

Matheron G and de Marsily G (1980) Is transport in porous media always diffusive? A counter example. *Water Resources Research* 16(5): 901–917.

Molz FJ, Güven O, and Melville JG (1983) An examination of scale-dependent dispersion coefficients. *Ground Water* 21(6): 715–725.

Skaggs TH and Leij FJ (2002) Solute transport: theoretical background. In: Dane JH and Topp GC (eds) *Methods of Soil Analysis*, part 4, *Physical Methods*, chap. 6.3. Madison, WI: Soil Science Society of America.

Sposito G (ed.) (1998) *Scale Dependence and Scale Invariance in Hydrology*. Cambridge, UK: Cambridge University Press.

Taylor GI (1953) The dispersion of matter in a solvent flowing slowly through a tube. *Proceedings of the Royal Society of London. Series A: Mathematical and Physical Sciences* 219: 189–203.

SEPTIC SYSTEMS

R L Lavigne, University of Massachusetts, Leeds, MA, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Since the advent of societal living, humankind has relied on the soil environment to treat a wide variety of its wastewater residues. The use of soil-based treatment systems continues to be a common practice in essentially all countries of the modern world. For centuries people have relied on the capacity of soil to purify waters contaminated from solid-waste disposal facilities, plague-infested communities, municipal sewage collection systems, industry, agriculture, cemeteries, and more commonly small residential, institutional, and commercial wastewater-producing facilities. Given adequate retention or treatment time, and sufficient spacing from surface and groundwater resources, many soils can effectively degrade organics through microbial processes; adsorb, precipitate, or complex most metals; and subsequently release an effluent that is near potable quality. In many parts of the world, soil-based wastewater-treatment systems function in close proximity to public or private drinking-water supplies, with little or no adverse effect on potability. In densely populated areas, however, where appropriate separation and treatment distance are not provided for soil-based treatment systems, potable water contamination from bacteria, viruses, nutrients, metals, and other contaminants has occurred.

Unfortunately, the generally small size of on-site wastewater-treatment systems ('septic systems') has created a global policy where most of them are built without regulations, technical guidelines, or protocols for maintaining and operating them. In highly regulated areas, more stringent controls may exist, but even where septic-system regulations do occur, their requirements are seldom founded on the principles of sound soil science or soil physics. In the USA, municipalities often use septic-system regulations to control development, and all too often the regulations become counterproductive in protecting the environment and public health. A classic example of this paradox is the encouragement of septic systems in highly permeable soils where groundwater and surface-water resources can quickly become polluted.

Technical and Historical Overview

The most common septic systems in use worldwide include outhouses and cesspools. In spite of their simple nature, they are often effective in treating human wastewater and protecting groundwater and surface-water resources. During the 1960s, the Commonwealth of Massachusetts embarked on a region-wide program to install cesspools in populated areas surrounding its Boston water supply known as Quabbin Reservoir. The program was largely successful, and the reservoir continues to provide surface water supplies in the Northeast part of the USA.

Outhouse – Structure and Function

In structure, outhouses are little more than a hole in the ground, usually 1–2 m deep. A wooden structure or 'house' is erected above the hole, with one or more seats. Liquid wastes percolate into surrounding soils, while insect larvae and bacteria degrade the solid component. In time, the hole fills from regular use. The outhouse is then relocated above a new excavation and the process repeats itself. In warm weather, hydrated lime can be added to the waste to reduce odors.

Cesspool – Structure and Function

The advent of indoor plumbing has added relatively large volumes of water to daily domestic wastewater production rates. For much of the world today, showers, flush toilets, laundry, and dishwashing needs result in a per capita production rate of approximately 200 l day^{-1} of wastewater. To accommodate these increased volumes, outhouse holes in the ground have been increased in depth and diameter. Many of these excavations are lined with stone or cribbed with timbers. Properly constructed, the common cesspool can provide decades of service with little or no maintenance needs other than periodic removal of the solids. A secure cover over the top of concrete or wood is required for safety reasons.

The Modern Septic System

For more technically oriented households, the use of strong soaps, bleaches, and cleaning products quickly reduces the ability of fecal bacteria to degrade the solids associated with the wastestream. Cesspools accumulate solids more quickly, minimizing the solid–water interface's ability to effectively conduct water out of the unit. For this reason a separate solids containment unit has evolved that is commonly called a septic tank.

Septic Tank

With mean hydraulic retention times of a day or more, approximately 65% of the solid matter can be removed by simple settling in a septic tank. A baffle or 'T' at the inlet end helps reduce short-circuiting of incoming surges, and a similar structure at the outlet end prevents floating grease and oils from leaving the unit. Although some biological degradation does occur anaerobically, the modern septic tank usually requires cleaning every 2–5 years. Concrete, metal, fiberglass, and other synthetics are commonly used to build septic tanks, but wood structures are also fairly common. The tanks are normally watertight and buried 30–60 cm belowground to facilitate the gravity flow of wastewater from the dwelling. A gas vent from the building sewer is typically hidden in the outside wall of the dwelling, where it normally exits at the roof eve. Anaerobic gases from the septic tank migrate upwards through the building sewer and out the roof vent. During periods of thermal inversion, vent-stack gasses can sink to ground level, giving false signs of septic system failure by way of odor.

Soil Absorption Systems

Once the liquid component of wastewater exits in a septic tank, it must be disposed of in some acceptable manner that is safe for the environment and public health. Soil absorption systems (SASs) satisfy this need in a very effective manner. A SAS can be built in a variety of shapes and configurations, but all SASs are generally designed to do the following:

1. Provide sufficient surface area to permit 100% water exfiltration for many years following the development of a typical soil interface 'scum layer';
2. Provide sufficient separation distance to insure that the wastewater does not enter the groundwater before treatment;
3. Provide sufficient retention time to insure adequate treatment of the liquid component in the SAS, the scum layer (usually 10 cm); and in the underlying unsaturated soils.

Typically a SAS is built in a shallow excavation 50–100 cm deep made in naturally occurring soils. Perforated 10-cm pipes are used to distribute liquid throughout the excavation. The pipes (usually plastic, concrete, or ceramic) are surrounded by clean coarse stone to further enhance the distribution process. A layer of fine stone, synthetic fabric, asphalt, paper, or hay covers the stone bed to prevent infiltration of natural soils that are replaced above. Grass is usually planted over a SAS so that the unit blends in naturally with the surrounding landscape.

From time to time an SAS must be elevated aboveground infill to provide adequate separation distance either to bedrock or the groundwater table. These aboveground units often require that a pump be used to lift wastewater from a 'pump tank' that follows the septic tank. Typical SAS configurations are often referred to as 'beds,' 'trenches,' 'galleries.' Gallery units are often prefabricated using plastic, concrete, or other structurally sound materials. They can also be built using concrete blocks, wood cribbing, or stone. In each case, the SAS is built on a level base to promote the uniform distribution of wastewater over the entire base area. When deeper SAS units can be constructed (i.e., greater than 50 cm), the sidewall area may be utilized to provide part of the required absorption surface area.

SAS Size and Soil Infiltration Rates

Although heuristic methods are often used to size SAS units, it is more common to employ some type of soil-testing to evaluate infiltration rates. These methodologies include: percolation-testing, *in situ* infiltrometer methodologies to determine coefficients of permeability (meters per day), or laboratory hydraulic-conductivity measurements using permeameters. Where testing is required, regulations typically prescribe a loading rate (volume per area) for SAS

units, as well as design-flow rates based on the number of dwelling bedrooms, bathrooms, or inhabitants. Flow-rate estimates are also available for categories of use (e.g., schools, cafeterias, hotels, restaurants).

Percolation testing is an *in situ* procedure requiring that a hole of specific diameter and depth (e.g., $30\,\text{cm} \times 45\,\text{cm}$) be hand dug into the soil strata being considered for a SAS. Once filled with clean water, the rate of water percolating is measured and reported as centimeters per minute or some arbitrary unit of depth versus time. Usually the slower the percolation rate, the greater the area required for loading, thereby increasing the total surface area needed for a SAS unit being designed for a given design-flow rate. In fine soils (silts, fine sands, and clays), a three-bedroom residential structure may require a SAS area several hundred square meters. In coarse-textured soils (sands and loamy sands), the required area for the same dwelling may be as little as $50\,\text{m}^2$. Regardless of the methodologies used, the objectives are to (1) contain 100% of the liquid fraction even during periods of peak loading; (2) provide adequate treatment time to reach acceptable infiltration concentration; and (3) provide appropriate separation distances to protect groundwater and surface-water resources.

Advanced Treatment Systems

Because septic systems have proliferated in some areas in a relatively uncontrolled manner, groundwater and surface water problems have evolved. Usually these problems take the form of nitrate (NO_3^{-2}) contamination in groundwater, or nutrient addition to surface waters that accelerate eutrophication processes. In some instances bacterial and viral contamination has also occurred in wells and surface-water resources. When these situations occur there are usually two paths of action considered. One is to sewer the area to provide municipal treatment. The other is to upgrade SAS units to provide appropriate levels of advanced treatment (e.g., disinfection, nitrate removal, nutrient removal). In the latter half of the twentieth century, many innovative alternative technologies have been developed. Unfortunately, most of these units require some degree of operation and maintenance that individual homeowners may not be able to provide. There has also been increased interest in decentralized sewerage systems where small community systems can be constructed.

Artificial Wetlands

A new variety of SAS incorporates the treatment processes known to be occurring in natural wetlands. These artificial or manmade wetlands can be constructed with or without liners, depending on soil texture and groundwater-protection needs. Top-loading, vertical flow, submerged bed systems are constructed to provide a park-like environment, while simultaneously producing a near drinking-water effluent. Generally these operate with no chemicals, no moving parts, and little or no energy needs other than the sun and gravity. Wetland plants growing in a coarse-textured medium provide oxygen to soil bacteria that in turn function as a biofilm of microbes metabolizing both organic and inorganic wastewater components. A dense stand of grasses and reeds is generally grown in these systems.

Summary

In the USA alone, it has been estimated that approximately 25% of all wastewater produced is treated with on-site septic systems. As rural lands continue to develop, the demand for individual treatment systems will surely rise. Considerable data emerging to show that conventional wastewater-treatment technology is not very environmentally friendly, nor is it always safe for public health.

Large quantities of fossil fuels are consumed each year, providing the energy needs that conventional technology demands. Disinfection processes using chlorine introduce carcinogens into surface waters. Conventional systems seldom are designed to remove nutrients, so that nitrogen, phosphorus, and potassium compounds are being directly discharged into lakes, rivers, and oceans. As these aging wastewater-treatment facilities become more costly to maintain, alternative technologies are likely to play a leading role in the industry. This is especially true for industrially emerging countries in Southeast Asia, Central and South America, and Africa. Properly designed, built, and operated, the conventional septic systems can help to meet the environmental and public health needs in many regions.

See also: **Microbial Processes:** Environmental Factors; Kinetics; **Pollutants:** Persistent Organic (POPs); **Waste Disposal on Land:** Liquid

Further Reading

Salvato JA (1992) *Environmental Engineering and Sanitation*, 4th edn, pp. 489–578. New York: John Wiley & Sons, Inc.

USEPA (2002) *Onsite Wastewater Treatment Systems Manual*, EPA/625/R-00/008. Cincinnati, OH: National Risk Management Research Laboratory.

SHIFTING CULTIVATION

R Lal, Ohio State University, Columbus, OH, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Soil, the three-dimensional upper layer of the Earth's crust that directly or indirectly supports all terrestrial life, is a nonrenewable resource over the human time scale. Further, soil resources of the world are finite, unequally distributed among geographical regions, and prone to severe degradation by misuse and mismanagement. Consequently, soil resources are shrinking with continuous increase in human population, due to degradation and conversion to nonagricultural uses, including urban encroachment. Therefore, judicious management of soil is critical to global food security, the health and welfare of humans and animals dependent upon it, the overall economy, and regional social and political stability.

Soil management implies strategic planning of all inputs into and outputs from the soil ecosystem so that there is a favorable balance of essential components that constitute the basis of soil's life-support system. These components comprise the soil organic matter (SOM) pool, reserves of essential plant nutrients (e.g., N, P, K, Ca, Mg, Zn, Cu, Mo, etc.), water balance and plant-available water capacity (AWC), and effective rooting depth. It is the imbalance caused by drastic differences between outputs and inputs of one or more of these components that sets in motion soil degradative processes. For example, depletion of SOM or biological degradation is caused when removal of crop residue and agronomic output exceeds the input of compost and other biosolids. Soil fertility declines and nutrient depletion occurs when nutrients harvested in plant and animal produce exceed nutrients supplied as chemical fertilizers or organic amendments. Depletion of bases (e.g., Ca, Mg, K) from soil can lead to acidification and elemental imbalance, resulting in excess of Al, Mn, and Fe. Waterlogging and soil salinization are caused by excessive use of irrigation water without provisions for drainage. The problem is aggravated if irrigation water is of poor quality and the subsoil contains a high concentration of water-soluble salts, as is generally the case in soils of arid and semiarid regions. Exposure of a structurally fragile soil to climatic elements (e.g., intense rains, flowing water, and blowing wind) can lead to removal of topsoil, its transport into streams and reservoirs, truncation of the rooting depth, and decline in the soil's productivity and environment moderation capacity.

It is this imbalance in outputs and inputs, due to soil misuse and mismanagement, that causes soil degradation (Figure 1). The latter is defined as a decline in the soil's capacity to perform functions of interest to humans. These functions include biomass productivity; filtering pollutants out of water; biodegradation of contaminants and pathogens; and cycling/storage

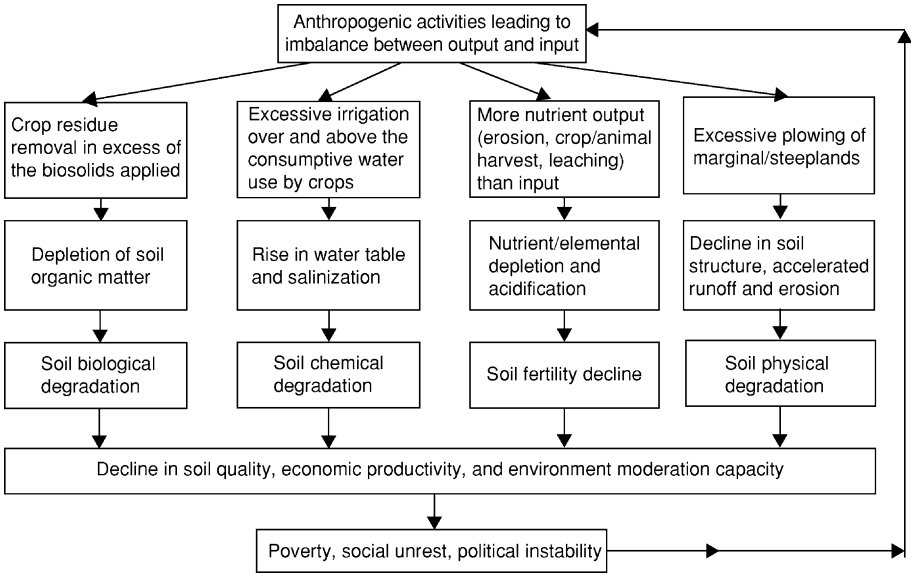


Figure 1 Soil degradation caused by imbalance between outputs and inputs.

of carbon (C), nitrogen (N), and other elements. Degradation severely jeopardizes the soil's capacity to perform these functions.

Soil degradation has been a challenge to humanity since the dawn of agriculture some 10 000 years ago. Severe soil degradation caused some of the once-thriving ancient civilizations to vanish. Indeed, it was the severe erosion caused by deforestation of cedar forests in the Mediterranean Basin that toppled the Phoenicians and destroyed granaries of the Roman Empire. The fertile land of Mesopotamia, present-day Iraq, was converted to desert and shifting sand dunes by deforestation, severe erosion of the hills and watersheds that nurtured the ancient irrigation system, and salinization of agricultural lands. The extinction of the Harappan-Kalibangan civilization in the pre-Aryan Indus Valley has been attributed to erosion-caused desert encroachment and salinization. The collapse of the 1700-year-old Mayan culture in Central America about AD 900 was caused by severe erosion of a thin topsoil by excessive cultivation necessary to support a rapid increase in population. In most cases, soil degradation was caused by the 'tragedy of the commons,' which could have been avoided by rational management and conservation of soil and natural resources.

Therefore, 'conservation' of soil resources is vital to the survival of the human race. The term 'soil conservation' means 'a judicious use of soil so that its productivity and environment moderation capacity are perpetually maintained.' Soil conservation also means 'rehabilitation of degraded soils for restoring their productivity and environment moderation capacity.' In view of the high demographic pressure and conflicting demands on shrinking soil resources, soil management and conservation are more important now than ever before. In fact, the importance of soil management and conservation can never be overemphasized.

Shifting Cultivation

Soil management and conservation have a special significance to regions where shifting, cultivation is practiced. The term 'shifting cultivation' refers to farming or agricultural systems in which a short but variable cultivation phase on 'slash-and-burn' cleared land alternates with a long, equally variable fallow period. Shifting cultivation is the most ancient system of agriculture in which soil fertility is restored by long periods of fallowing rather than by off-farm inputs of fertilizers and amendments, nutrients are recycled between natural vegetation and crops, and ecological balance is maintained by adopting diverse and complex cropping systems rather than monoculture. Shifting cultivation is still being practiced in some parts of the tropics and subtropics, especially in the humid tropics of west and central Africa, Southeast Asia and South/Central America. It is an extensive system of agriculture, with a wide range of variation depending on soil type, vegetation, terrain, climate, and social/ethnic factors. The system is stable where the ratios of the fallow period to cultivation phases are up to 10 or more. The system breaks down, causing severe soil degradation and decline in productivity, when the fallow period is drastically reduced. An important reason for the reduction in fallow period is an increase in the demographic pressure, and an attendant decrease in available soil resources.

Shifting cultivation systems have the following characteristics:

1. They are highly diverse and complex, involving simultaneous growing of 10 or more crops in combination with trees and livestock. Mixed cropping is the general rule.
2. Shifting cultivation systems are labor-intensive with minimal use of purchased off-farm inputs. Restoration/management of soil fertility is based on nutrient cycling through lengthy fallow, biological nitrogen fixation (BNF), and choice of compatible species.
3. Shifting cultivators are generally small-size landholders with a normal size of <2 ha. Therefore, farm operations are performed by the farm family.

There are several attempts to classify shifting cultivation systems. A farming system is designated nomadic herding if the value of the land-use factor, $L \geq 10$; bush fallow or land rotation for L between 5 and 10; rudimentary or sedentary agriculture for L between 2 and 4; compound farming terrace farming or paddy rice cultivation for $L < 2$. Most classification systems are based on 'intensity' or 'frequency' of cultivation. The various categories designate different degrees of intensification as evaluated on the basis:

$$L = (C + F)/F \qquad [1]$$

where C is the number of years of cultivation and F is the number of years fallow. Traditional systems are also classified on the basis of the degree of intensification:

Phase I: Hunting and gathering
Phase II: Subsistence agriculture
Phase III: Early consumer stage
Phase IV: Primary mechanization stage

The true shifting cultivation or the hunting and gathering stage ($L = \geq 10$) is no longer widespread in any region. Yet, subsistence agriculture including

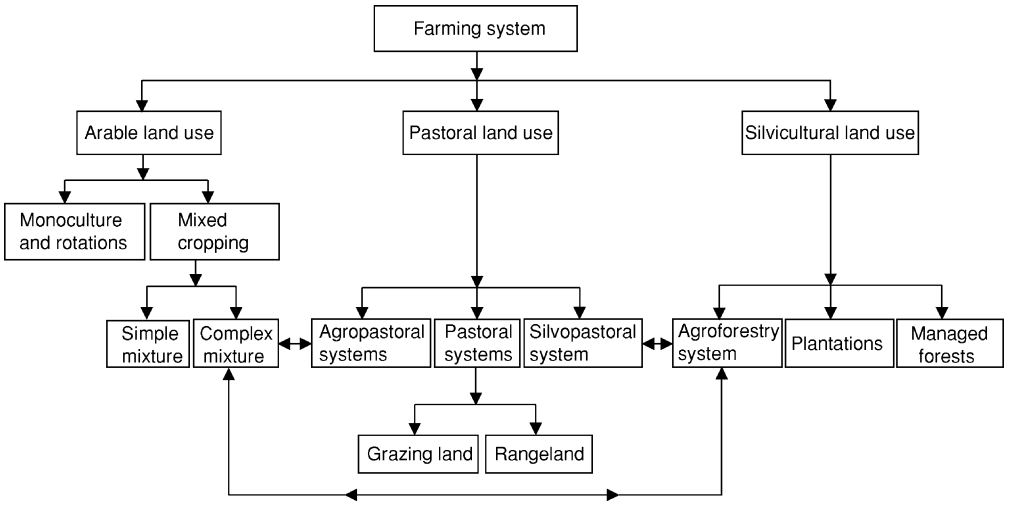


Figure 2 Permanent farming system.

bush fallow rotation ($L = 5$–10) is practiced in some regions of Africa, Asia and South/Central America. Subsistence farmers grow a variety of crops (mixed cropping) in association with selected tree species and animals, and as much as 90% of the produce is consumed on the farm. A majority of the shifting cultivators in the tropics practice rudimentary sedentary agriculture at the early consumer stage ($L = 2$–4). These farmers market less than 50% of their produce, and practice mixed farming. Farmers may also use some machinery and low rates of fertilizers. A predominant type of mechanization use by phase-III farmers involves tractor-based tillage.

In contrast to shifting cultivation, farming systems involving permanent agriculture are outlined in Figure 2. These systems are based on three principal land uses and appropriate two-way or three-way combinations. For example, agropastoral systems are based on a combination of crops and livestock, and agrosilvopastoral systems are based on a combination of crops, livestock, and trees.

Productivity of Complex Farming Systems

Because a large number of crops are grown simultaneously on the same piece of land, and often harvested at different times, it is a challenging task to assess the productivity of such complex systems. Some of the indicators commonly used to assess productivity include the following:

1. Productivity: Productivity can be assessed as biomass (agronomic) yield or income. It can be expressed either as a simple productivity or total factor productivity (TFP):

$$P = p/R \quad [2]$$

$$\text{TFP} = p/\sum_{i=1}^{n}(R_i \times C_i) \quad [3]$$

where P is productivity, p is total production, R is the resource consumed/used, and C is the cost of the resource. Being truly resource-based, it may be difficult to assess the true cost of all resources used for phases I and II of shifting cultivation systems. Therefore, the following indices are used to assess productivity.

2. Land equivalent ratio: Productivity of mixed cropping systems is assessed by computing the land equivalent ratio (LER):

$$\text{LER} = \sum_{i=1}^{n} y_i/(y_m) \quad [4]$$

where y_i and y_m are yields of component crops in the intercrop and monoculture system, respectively, and n is the number of crops involved.

3. Area × time equivalency ratio (ATER): In shifting cultivation systems (phases II and III), there is a wide range and a large number of crop/tree species involved. These species vary widely in their maturity period. The ATER index assesses productivity while considering the crop duration:

$$\text{ATER} = 1/t\left(\sum_{i=1}^{n}(d.y_i)/y_m\right) \quad [5]$$

where d is the growth period of crop in days and t is the total time in days for which the field remained occupied, i.e., the growth period of the longest-duration crop. The numerical value of ATER index

approaches that of LER for a mixture consisting of crops of approximately identical growth periods, i.e., when $t = d$ in eqn [5].

4. Resource-use efficiency: When some off-farm inputs are used (phase IV), the productivity can also be assessed in terms of the resource-use efficiency. This is a relevant option in case of a system with the most limited resource such as P, N, water, soil organic carbon (SOC), or energy use.

5. Carbon dynamics: In view of the rapid increase in atmospheric concentration of CO_2 and other greenhouse gases, assessing C dynamics is an important indicator of the productivity.

$$P = \left(C_o / (\sum_{i=1}^{n} C_i) \right)_t \qquad [6]$$

where P is productivity, C_o is carbon produced in above- and belowground biomass, C_i is carbon input in all components including fertilizers and pesticides etc., and t is time. The higher the ratio, the more sustainable is the system.

Ecosystem-Related Constraints to Agricultural Intensification

Shifting cultivation systems are ecologically viable as long as there is enough land for long (10–20 years) restorative fallow, and expectations of crop yield and the attendant standards of living are not too high. These systems are naturally suited for harsh environments and fragile ecosystems of the tropics. That is why attempts at finding viable alternatives to shifting cultivation have met with only limited success. There are several soil-related and climate-related constraints (Table 1) that must be alleviated. These constraints lead to severe problems of soil degradation upon conversion from shifting cultivation to mechanized intensive agriculture. Consequently, many large-scale mechanized farming schemes introduced in the humid tropics have been a failure.

Table 1 Ecological constraints to finding viable alternatives to shifting cultivation

Soil-related constraints	*Climate-related constraints*
1. Low inherent soil fertility	1. Low level of radiation during the growing season
2. High susceptibility to soil erosion	2. High relative humidity
3. Severe soil physical degradation	3. High storage losses
4. Nutrient/elemental imbalance	4. Problems with grain drying
	5. High incidence of pests and disease
	6. High-intensity rains

The severe problem of soil degradation is caused by indiscriminate and intensive land use based on monoculture regardless of soil capability, introduction of pastures with high stocking rate and uncontrolled grazing, or intensive cropping without input of chemical fertilizers or compost at the required rates. The problem of soil degradation is exacerbated by harsh climate characterized by intense rains of high erosivity and structurally weak soils.

Sustainable Management of Soils Under Shifting Cultivation

Any viable alternative to shifting cultivation system must successfully:

1. maintain soil structure at a favorable level so that risks of soil physical degradation (e.g., crusting, compaction, low infiltration) are minimized;
2. control excessive runoff and accelerated soil erosion so that effective rooting depth is maintained and the nutrient-rich topsoil is not eroded away;
3. strengthen nutrient-recycling mechanisms so that nutrients are not lost out of the ecosystem, and are effectively recycled through incorporation of woody perennials and legumes in the rotation cycle, and judicious use of chemical fertilizers and organic amendments;
4. manage pests through enhancing biodiversity and using a healthy balance between predators and parasites through strategies of integrated pest management; and
5. increase farm income through value-added products and off-farm employment.

While traditional shifting cultivation systems meet these requirements through low cropping intensity and lengthy fallows, new systems have to be identified that produce high yields per unit area of land and per unit time. Sustainable soil management technologies must address specific ecological constraints such as those outlined in Table 1. Any viable alternative must address the synchrony between food security and environment quality without jeopardizing the soil/natural resources, which are already under great stress. Therefore, sustainable soil management systems must:

1. assure food security for a high and increasing population;
2. enhance soil productivity through ecosystem restoration and judicious use of off-farm input;

3. improve environment quality, especially that of water and air, and reduce emissions of greenhouse gases into the atmosphere; and
4. transform subsistence farming into commercially viable agriculture.

Characteristics of an Improved/ Sustainable Soil Management System

Important characteristics of an improved/sustainable soil management system are outlined in Figure 3. The new system must have high energy flux in terms of drastic increases in crop yield brought about by discriminant or judicious use of off-farm input. The objective is to increase nutrient input into the system through integrated nutrient management (INM). The latter involves elemental recycling (e.g., use of crop residue, BNF, use of compost and city sludge, etc.), supplemented by a judicious use of chemical fertilizers. The objective is to increase energy efficiency and flux, because a low-output system can be highly efficient but unproductive, just as a high-output system can be extremely inefficient. Therefore, the goal is to develop a high-output and highly efficient production system. The efficiency can be increased by decreasing the losses. To be sustainable, the system must of necessity be soil-restorative. Most soils under shifting cultivation are being depleted and degraded because of nutrient mining over a long period of time. Some components of an improved system comprise the following:

1. New land development: The emphasis should be toward intensification of existing land so that there is no need for additional deforestation and conversion of a natural to an agricultural ecosystem. If deforestation is absolutely necessary, it must be done in a way to cause as little perturbation in the ecosystem as possible, especially with regards to the following:
   a. water and energy budgets, which are disturbed by deforestation, causing an increase in surface soil temperatures and acceleration of surface runoff with the attendant problem of severe erosion. Thus the objective is to cause a minimal disturbance to the hydrological cycle, especially the surface runoff component, which must be kept under control.
   b. maintenance of soil structure, which is crucial to ensuring a high water infiltration rate, minimizing risks of crusting and compaction, and decreasing risks of accelerated soil erosion. Application of crop-residue mulch and establishment of cover crops are necessary to enhance activity of soil biota and improve soil structure.
   c. increasing plant nutrient reserves, which is important to obtaining high crop yields. This can be achieved by replacing nutrients harvested in crops and those lost by erosion and leaching through use of INM strategies.

Therefore, removal of existing vegetation can be done either manually or by those machines that cause least disturbance to the surface soil (e.g., those with shear blade). Land-use capability assessment and proper planning are essential to any successful development of new land.

2. Intensive use of existing lands: Because of the shortage of potentially suitable new land and the undesirability of converting forest/savanna ecosystems into croplands, an increase in production in several countries must be achieved by more intensive cultivation of land already developed. It is equally important to restore productivity of degraded soils. Therefore intensification of existing arable land is a high priority.

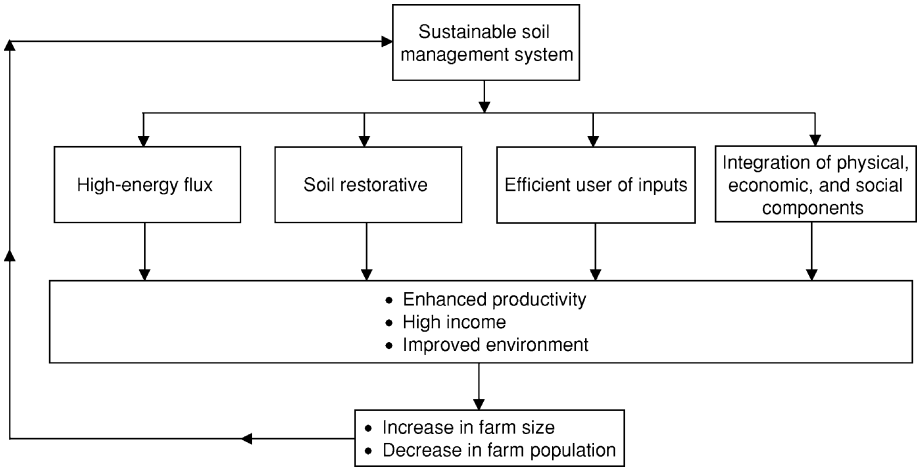


Figure 3 Characteristics of a viable alternative to shifting cultivation systems.

Examples of Sustainable Soil Management Systems

Farming systems, the right combination of several components and subsystems, are soil- and site-specific. Therefore, it is difficult to provide examples of generic farming systems that may be applicable to diverse soils and environments. A logical approach is to identify integral components that can be combined into site-specific farming systems to meet the site-specific needs. Some examples of improved components and subsystems to address specific issues include the following:

1. Erosion management: Conservation-effective techniques for reducing soil-erosion risks involve the following strategies:
   a. using crop residue mulch to minimize the raindrop impact;
   b. improving soil structure through bioturbation for reducing surface crust and soil compaction, and maintaining high infiltration rate;
   c. installing terraces for decreasing the slope length; and
   d. constructing grassed waterways to dispose of excess runoff safely.

Principal techniques of erosion management are outlined in Figure 4. Preventive measures are based on the principles of soil management that reduce the impact of high rainfall erosivity by reducing the raindrop impact and maintaining a high infiltration rate. These measures include mulch farming techniques involving conservation tillage, using cover crops in the rotation cycle, and providing a continuous ground cover through diverse (mixed) cropping systems. Two principal constraints to a widespread adoption of conservation tillage are high cost of herbicides for weed control and lack of appropriate seeding equipment to plant through wet crop-residue mulch. Soil compaction is another impediment to adopting conservation tillage in structurally inert soils of low organic matter content.

Erosion control measures are designed to reduce runoff amount and its velocity by using engineering structures that lead to safe disposal of excess runoff. These structures are grouped under two broad categories: structures to reduce runon (e.g., diversion channels, cut-off drains), and structures to dispose of runoff safely (e.g., terraces, waterways, drop structures). Cultivation of sloping lands (slope gradient > 5%) requires using a combination of erosion preventive and control measures, especially in situations where erosion cannot be effectively controlled by preventive measures alone. In such situations, engineering structures are needed as back-up measures. However, design and installation of these structures require engineering skills and proper planning. Further, such structures are costly to install and maintain. It is also important to realize that failure of engineering structures can lead to deep gullies, mass movement, and often irreversible damage to soil and terrain.

2. Nutrient management: Once soil erosion is controlled and the risks of drought are reduced by water conservation, management of plant nutrients or soil fertility is essential to obtaining high yields. Most soils under shifting cultivation have low inherent soil fertility, and nutrient supply from off-farm sources is essential. The goal is to use an INM strategy (Figure 5), not by reducing external inputs but by minimizing the use of chemical fertilizers through

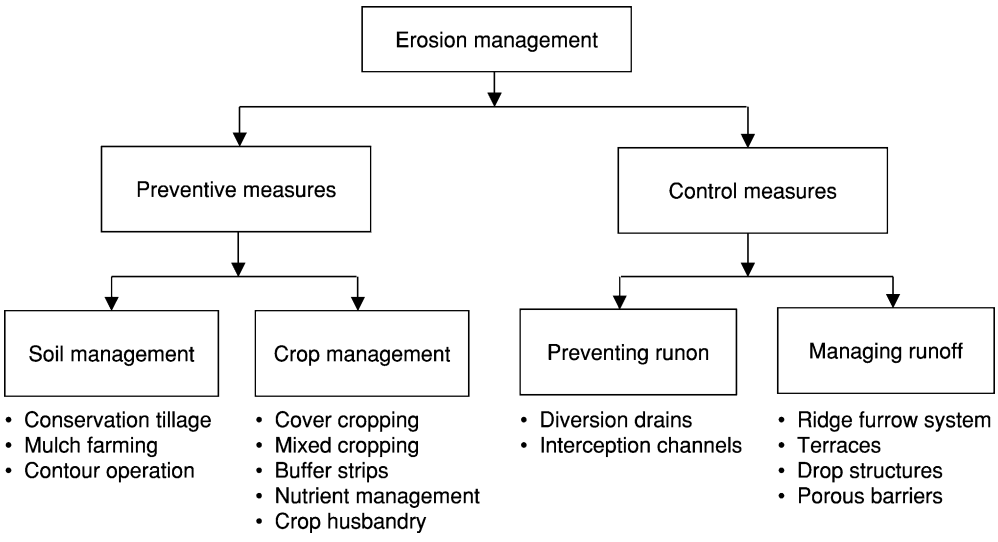


Figure 4 Soil, crop, and runoff management techniques for erosion control.

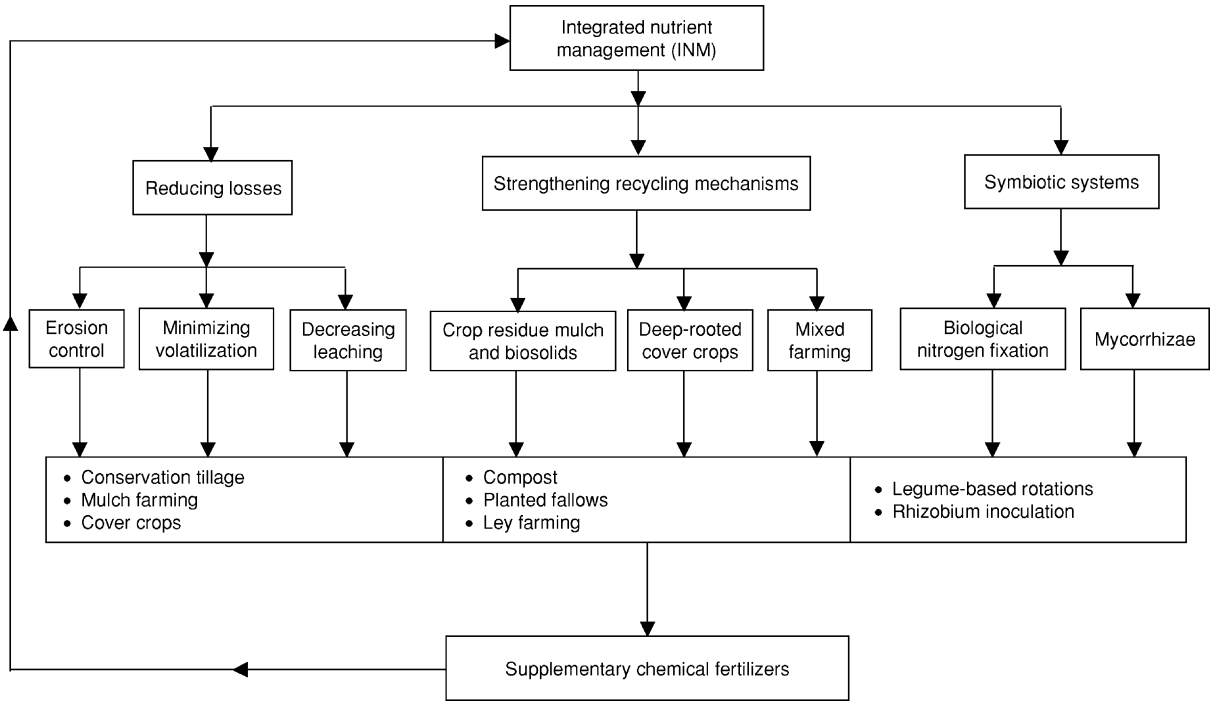


Figure 5 Integrated nutrient-management strategies are designed to enhance nutrient supply without increasing dependence on chemical fertilizers.

decreasing losses, recycling nutrients and increasing BNF. Reducing losses of plant nutrients (in runoff, soil erosion, leaching, and volatilization) is crucial to enhancing nutrient-use efficiency. Equally important is recycling of plant nutrients through the use of biosolids as compost, sludge, and manure. Planting deep-rooted perennials, as in agroforestry systems, is also important to recycling nutrients. Soil N reserves can be enhanced through BNF by incorporating legumes in the rotation. All these measures can decrease the amount of chemical fertilizers required to obtain the desired yield. The objective is not to eliminate chemical fertilizers but to decrease the quantity required and increase the use efficiency of critical elements (e.g., N, P).

3. Water management: There is a strong link between nutrient management and water management. The nutrient-use efficiency cannot be increased without increasing the water-use efficiency. Similar to nutrient management, the overall goal is to decrease losses, strengthen water recycling, and provide supplemental irrigation (Figure 6). The latter may be needed to enhance crop yields in those regions of subhumid and semiarid tropics that are prone to short-term droughts during the growing season. In addition to losses by runoff, those by soil evaporation must be minimized, especially in the tropics, which have high evaporative demand throughout the year. Therefore, mulch farming is an important strategy. Recycling of water, similar to that of plant nutrients, also requires establishing compatible mixtures of diverse crops. Planting shallow-rooted annuals in association with deep-rooted perennials can enhance water-use efficiency. Excess runoff during the rainy season can be stored in small farm ponds and used for supplemental irrigation to prolong the growing season and alleviate drought. If supplemental irrigation involves use of canal (surface water) or tubewell (groundwater), it is absolutely necessary to make provisions for surface/subsurface drainage and reduce risks of a rise in the water table and build-up of excessive salts in the root zone.

Ecological Approach to Management of Soils under Shifting Cultivation

The traditional shifting cultivation systems ($L = >5$) are ecologically compatible with harsh tropical environments because these systems do not cause drastic perturbations in ecosystem functions such as water balance, energy budget, elemental cycling, above- and belowground biodiversity. Therefore, any successful transformation of a resource-based (low input) and extensive (low productivity) system into viable (high energy flux) and commercial (profitable) farming must be based on similar principles of causing the least possible disturbance to ecosystem functions. Yet, the energy and biomass flow must be increased while strengthening mechanisms of recycling and

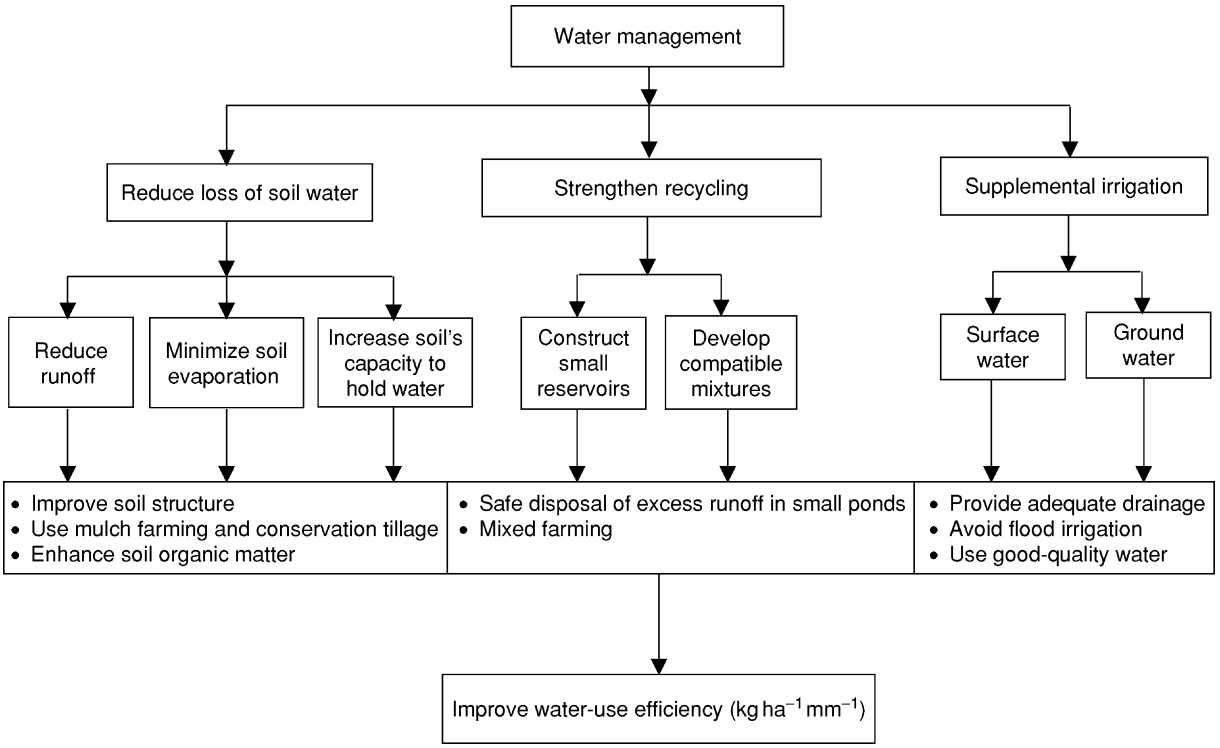


Figure 6 Strategies for water management to enhance water-use efficiency.

minimizing loses through erosion. Essential components of an ecological approach are outlined in Figure 7, and involve science-based interventions. In addition to enhancing production, these technological interventions must minimize risks to the environment. The strategy is to increase productivity while sequestering C within soil and biota, and without releasing chemicals (fertilizers and pesticides) into the environment.

In this regard, the importance of improved germplasm cannot be overemphasized. The strategy is to introduce ecologically compatible crop/tree species and genetically improved varieties. Indeed, some of the soil-related constraints (e.g., toxicity of Al and Mn in the subsoil) may be alleviated by the introduction of improved varieties. Damage by pests can also be reduced through introduction of improved germplasm.

Research and Development Priorities

There is a need to adopt proven technology through on-farm adaptive research with strong participation of the farming community. There is also a need for basic research in addressing new and evolving issues.

1. Terrestrial carbon sequestration: There is a strong link between the cycles of C, N, and H_2O. Harvesting N in the biomass for growing food crops by shifting cultivation systems has been responsible for emissions of CO_2 and other greenhouse gases (CH_4, N_2O) into the atmosphere. In fact, 1.5–2.0 Pg C is annually emitted into the atmosphere by deforestation and land-use change in the tropics. Rather than a source, tropical ecosystems can be a major sink for atmospheric C. Therefore, understanding the cycling of C, N, P, S, and other elements in natural and managed ecosystems, and enhancing their retention within soil and biota are crucial to reducing emissions of greenhouse gases into the atmosphere.

In this regard, restoration of degraded soils and ecosystems is extremely important. Degraded ecosystems are severely depleted of their C and N pools, and have been sources of greenhouse gases. Restoration of these ecosystems would enhance biomass productivity and soil quality, and increase the C pool within the soil and biota. It is important, therefore, to understand elemental cycling in natural and managed ecosystems and how recycling can be enhanced, losses reduced, and use-efficiency increased. Understanding the link between water and elemental cycling is also important to identifying ecologically compatible systems.

2. Soil quality and soil resilience: There is a close link between elemental cycling and soil quality. Rapid and severe soil degradation is linked to depletion of SOC reserves and elemental imbalance, but the exact mechanisms are not widely understood. An answer to the question of how can soil degradative trends be

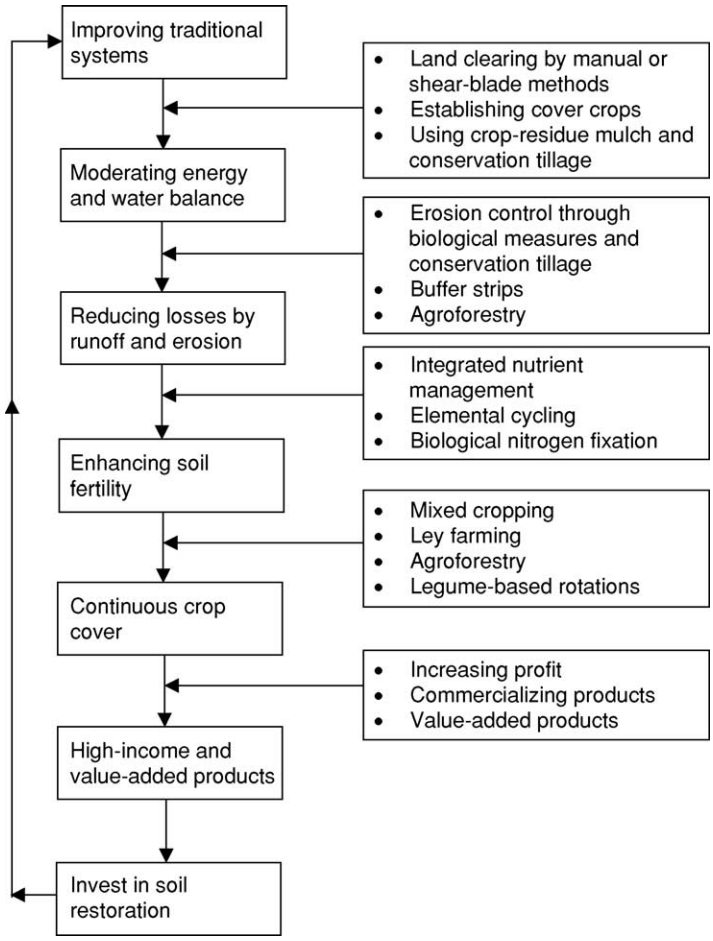


Figure 7 An ecological approach to transforming subsistence agricultural systems.

reversed lies in understanding the relation between soil quality and soil resilience. The strategy is to enhance soil resilience and improve soil quality while improving productivity. Therefore, soil-specific research is needed to understand the mechanisms and processes responsible for enhancing soil resilience and improving soil quality. The maintenance of high levels of the SOC pool may be one such strategy.

3. Agricultural intensification: It is important to develop innovative systems of enhancing the productivity of existing agricultural lands by alleviating soil, biotic, and social/economic constraints. Restoration of productivity of depleted/degraded soils is crucial to achieving food security and improving the environment.

4. Indicators of sustainability: It is important to formulate quantitative criteria of assessing the sustainability of a farming system. Developing quantifiable and scientifically sound indicators of sustainability is a high priority.

5. Modeling: Predictive models are needed to assess the impact of land use and farming systems on soil quality, water balance, micro- and meso-climate, agronomic/biomass productivity, and sustainability. Validation of such models for specific soils and ecosystems necessitates strengthening the database on soil resources, hydrology, vegetation, and climate.

Conclusions

Shifting cultivation and related bush fallow systems have served a useful purpose in preserving the ecological balance in regions of low population pressure. Although subsistence and low in productivity, such systems had minimal adverse impact on the environment. With increasing population pressure and concerns about environmental pollution and contamination by biomass burning and resource degradation, there is a need to identify viable alternatives to these systems. The available research indicates that productivity of these soils can be substantially increased by adopting improved and science-based technologies. Whereas soil and water conservation are achieved by

biological measures based on mulch farming and cover crops, soil fertility can be enhanced by nutrient recycling, BNF, and supplemental fertilizer use. Principal attributes of sustainable systems of soil and water management are erosion control, maintenance of soil structure, increase in SOC pool and biotic activity of soil fauna, and judicious use of chemical fertilizers to create synergistic effects between fertilizers and biosolids.

Major characteristics of sustainable systems are high energy fluxes that increase productivity and use-efficiency of inputs, restorative effects on soil and environments, and increasing C stocks in soil and biota. There is a strong need to conduct basic research in understanding the principles and processes that govern soil degradation, soil quality, and soil resilience. Understanding the link between SOC dynamics on the one hand and elemental cycling (e.g., C, N, P, S) and water balance on the other is critical to developing sustainable systems that are viable alternatives to shifting cultivation.

See also: **Agroforestry**; **Erosion:** Irrigation-Induced; **Tropical Soils:** Arid and Semiarid; Humid Tropical

Further Reading

Dregne HE (1982) Historical perspective of accelerated erosion and effect on world civilization. In: Schmidt BL, Allmaras RR, Mannering JV, and Papendick RI (eds) *Determinants of Soil Loss Tolerance*. ASA Special Publication 45, pp. 1–14. Madison, WI: Soil Science Society of America.

Eckholm EP (1976) *Losing Ground*. New York: Norton.

Hardin G (1968) The tragedy of the commons. *Science* 162: 1243–1248.

Harwood RR (1976) Multiple cropping in tropical Asia. In: Papendick RI, Sanchez PA, and Inplett GB (eds) *Multiple Cropping*. ASA special publication 27. Madison, WI: Soil Science Society of America.

Herdt RW (1993) Measuring sustainability using long-term experiments. Proceedings of the Conference at Rothamsted Experiment Station, 29–30 April 1993, Rothamsted, UK.

Hiebsch CK and McCollum CE (1987) Area × time equivalency ratio (ATER): a method for evaluating the productivity of intercrops. *Agronomy Journal* 79: 15–22.

IPCC (2000) *Land Use, Land Change and Forestry*. Special Report. Cambridge, UK: Cambridge University Press.

Lal R (1994) *Methods and Guidelines for Assessing Sustainable Use of Soil and Water Resources in the Tropics*. SMSS technical monograph 1. Columbus, OH: OSU.

Lal R (1995) *Sustainable Management of Soil Resources in the Humid Tropics*. Tokyo, Japan: United Nations University.

Lowdermilk WC (1953) *Conquest of the Land through Seven Thousand Years*. SCS Agricultural Information Bulletin 99.

Nye PH and Greenland DJ (1960) *The Soil Under Shifting Cultivation*. Farnham Royal, UK: Commonwealth Agricultural Bureau.

Okigbo BN (1984) Improved management production systems as an alternative to shifting intermittent cultivation. In: *Improved Production Systems as an Alternative to Shifting Cultivation*. FAO Soils Bulletin 53. Rome, Italy: FAO.

Oldeman LR (1994) The global extent of soil degradation. In: Greenland DJ and Szabolcs I (eds) *Soil Resilience and Sustainable Land Use*. Wallingford, UK: CAB International.

Olson GW (1981) Archaeology: lessons on future soil use. *Journal of Soil and Water Conservation* 36(5): 261–264.

Ruthenberg H (1971) *Farming Systems in the Tropics*. London: Oxford University Press.

Singh HP (1982) Management of desert soils. In: Kanwar JS (ed.) *Review of Soil Research in India*, pp. 679–699. New Delhi: ICAR.

SITE-SPECIFIC SOIL MANAGEMENT

C J Johannsen and P G Carter, Purdue University, West Lafayette, IN, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

A new approach to intensive agricultural production, called 'site-specific agricultural management,' has gained acceptance by many farmers and producers around the globe since the 1990s. Site-specific management makes use of new tools, known as 'spatial technologies,' which encourage intensive management of soil resources using tools such as global positioning systems (GPS), geographic information systems (GIS), and remote sensing. Other frequently used terms to describe this intensive agricultural system include 'precision agriculture,' 'prescription farming,' and 'variable-rate technology' (VRT). Site-specific management has been defined as:

> an information- and technology-based agricultural management system to identify, analyze, and manage a site relating to the spatial and temporal variability within fields for optimum profitability, sustainability, and protection of the environment

Another definition puts it more simply as:

> using information technologies to tailor soil and crop management to fit the specific conditions found within a field

Services are available to farmers that involve satellites for collecting data and transmitting geographic information, sensors on airplanes to provide measurements as an aid to crop and soil management, sensors on farm equipment to monitor soil and crop parameters, and computer systems for receiving or analyzing digital spatial information. The word 'spatial' is frequently used related to site-specific management and means the area represented around a data point or reference. Spatial data contain information on location, shape, and relationships among geographic features such as soil, vegetation, and hydrology. Therefore data may be collected to represent an entire field, where one sample of 220 t may be the total yield, or data may be collected to represent a small portion of the field, such as an element from a yield monitor of 4 m^2, where the specific yield might be 22 t ha^{-1}.

Successful application of spatial technologies at the farm scale is dependent upon having a defined management system that has the capability to handle large volumes of data in an organized way, and a solid base of experience. It arises at the confluence of three principal technologies: GPS, GIS, and remote sensing.

New technologies make data collection, analysis, and interpretation more beneficial for site-specific farming and are important in improving the soil environment. Some of these technologies include GIS for integrating data layers and data (soil properties, topography ('digital terrain'), slope, and slope aspect) to assist in the analysis of collected data; GPS for relating the location of ground observations and measurements; and remote sensing for defining soil and crop anomalies within a field and assessing their crop production impact. In addition to these technologies, improvements in communications, e.g., the internet, are providing rapid delivery of newly acquired, remotely sensed data and timely transfer of information.

Technologies

GPS was first used during the 1980s by military forces. Composed of 24 satellites and generically identified as a GPS, Navistar blankets the globe with location-finding radio signals that can help determine a person's location to within less then 1-m accuracy. GPS assist farmers to locate areas within a field where application of fertilizer and chemicals is necessary; adjustment of application rates is based on soil conditions and yield data.

The Navistar satellite system produces a low-energy signal containing a 'data message' that includes information about its location, its atomic clock status, and its general condition. These data messages can be 'read' and interpreted by GPS receivers anywhere on the ground, at sea, or in the air. For higher accuracy, simultaneous readings are taken from four or more satellites.

The value of the GPS to a farmer includes: (1) knowing the locations of soil samples, so laboratory results can be compared with a digitized soil map; (2) knowing the locations of soil properties (clay and organic matter content) and soil conditions (relief and drainage) so that fertilizer and pesticides can be applied optimally; (3) the use of lightbar technology which provides guidance in the exact direction for accurately applying and documenting the rates of pesticides; and (4) being able to monitor and record yield data from harvesting equipment within the fields. All of these applications involve more precision in the collection of data and information, which means that the farmer or producer needs a system for keeping track of these data for future decision-making.

GIS technology is also essential to precision farming, because it is the system for keeping track of data and information. GIS is computer software that is used to collect, sort, analyze, map, graph, store, and move data. A GIS records many different types of data in stacked layers that are all precisely aligned geographically. For the farmer this means that information obtained from a field at different times can be easily integrated into a composite profile of each small unit or spatial area of the field. Not only can farmers store yield information by location, but they can also layer the yields on top of their soil map to correlate yield with soil type. A GIS also allows farmers to see costs of production for any part of the field. Accurate data are required and need to be accessed by the GIS software.

Without a GIS, data from soil tests, dates of planting, fertilizer application rates, weather, and yield are often stored as piles of paper on the farmer's desk to sort manually for making decisions. The key attribute of GIS is that all data must be referenced to a common geographic base. This is usually a base map; one of the most commonly used is digital orthophotography, which has geographic references for each pixel (picture data point) such as a longitude and latitude reference. There are more than 40 different reference systems; fortunately, software has been written to convert all of these to a selected base map. Once the base has been selected, all other data

that are collected are referenced to that base. GPS is the most common means of providing the reference points for the collected data.

Remote sensing, the third technology available to precision farming, entails the use of aircraft or satellites as platforms, from which farmers capture images of their fields during the growing season. The images are used to assess whether something unusual has occurred; this information can be used to help improve yield and profit. Remote sensing technology has improved, especially since the late 1990s, to include finer spatial resolution, more wavelengths of recorded light energy, and more frequent coverage over a specific location. Because of improvements in sensors, computer chips, software, and services, intensive agricultural systems may now reap benefits from these technologies.

Farmers can either analyze this remotely sensed information themselves or rely on commercial companies to perform this service for them for a fee. Satellite images from Landsat and SPOT, US and French satellites, respectively, have been used to distinguish locations affected by such problems as weeds, excessive moisture, low soil pH, and imbalance of plant nutrients. Satellites such as the IKONOS, Quickbird, and Orbview satellites have allowed farmers to obtain the equivalent of aerial images of their fields and have already shown benefits for management of intensive crops such as vegetables, orchard fruits, and grapes.

The cost and timeliness of obtaining the images have been the biggest deterrent to date to regular use of such remotely sensed data. Satellites launched during the last few years by Digital Globe, Space Imaging, and others are competing for the agricultural market; however, there are not enough satellites available to provide cloud-free images for critical periods. These multispectral satellites provide images in three to six different frequencies plus black-and-white images at 0.6- to 4-m spatial resolutions.

Remote sensing needs to prove itself in the grain crop areas, since the profit margin of these crops has been low. It has, however, been tested successfully on potato, sugar beet, and cotton fields in some parts of the world. Commercial remote-sensing companies have used weekly images to map the acreage of these crops and monitor their conditions throughout the growing season. The images and interpretations provide growers with an improved bargaining position when contracting with food processors, since they have a prediction of harvest dates and knowledge of how their competitor's crops are performing.

The value of precision farming is that the farmer can perform more timely operations related to better soil management such as tillage, plan more accurate crop-protection programs, adjust seeding rates, and know the yield variations commonly caused by soil differences within a field. The potential value of this system therefore translates into increased net return per acre; improved efficiency of such inputs as seeds, fertilizers, and chemicals; improvement of the soil environment; reduced threat to water quality; and increased managerial skills. This potential, while promised in many farm advertisements, still needs to be proven in a consistent manner and depends on the grower's management style. Farmers need to learn how to use these technologies to improve soil and crop quality based upon conditions found within their fields.

Soil Management

Soil management is a key to the success of site-specific soil management. It starts with a farmer's ability to vary the depth of tillage according to soil conditions and is important in proper seedbed preparation, control of weeds, and fuel consumption, with the potential to lower production costs within an individual field. With a GPS as their guide, farmers using conservation tillage (the practice of leaving residues on the soil surface for erosion control) can adjust tillage depth as they traverse the different soil types. With conservation and reduced tillage, the amount of soil disturbance is minimal, but adjustments according to soil conditions such as texture, moisture content, and pH are important to seed depth and fertilizer placement. The adjustments may contribute to higher yields and safer production at lower cost. This part of site-specific farming is in its infancy. Equipment companies are developing tillage equipment with GPS and selected controls tailored to site-specific farming that will be evolving over the next decade.

The application of chemicals in proper proportions is of environmental and economic concern to farmers. Environmental regulations call for the discontinuance of certain pesticide applications within 30 m of a stream, water body, or well, or within 20 m of an intermittent stream. Using a GPS along with a digital drainage map, a farmer is able to apply these pesticides with VRT in a safer manner (**Figure 1**). Spraying equipment can be preprogrammed to turn off automatically when it reaches the distance limitation or buffer along the drainage feature. Additionally, farmers can preprogram the amount of pesticide to be applied according to the soil characteristics in each area of the field. This saves money and allows for safer use of these materials.

Some companies are marketing sprayers utilizing a sensor positioned ahead of each spray nozzle for applying herbicides in chemical–fallow conditions

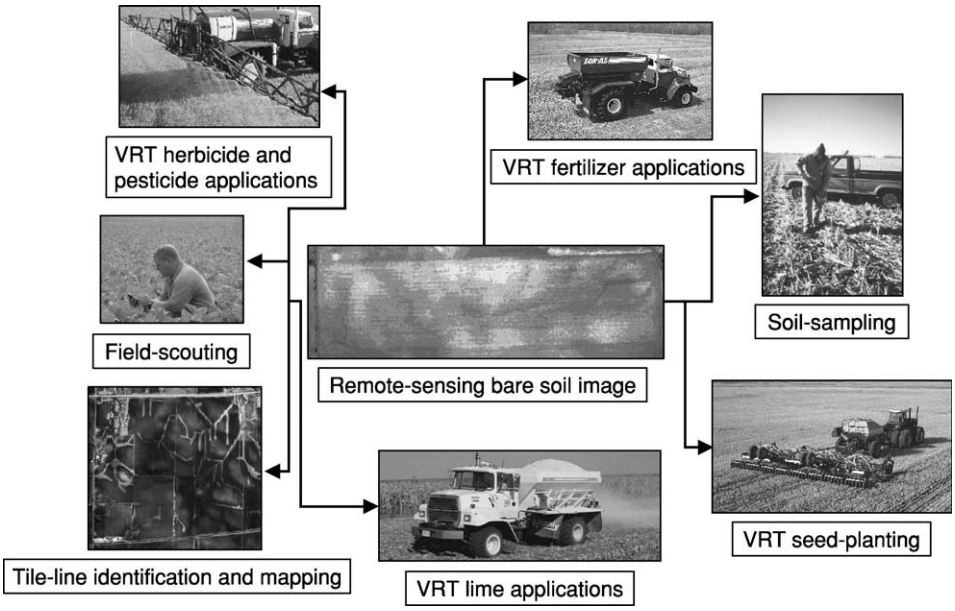


Figure 1 Agricultural global positioning system (GPS) applications as related to a remotely sensed image. The image provides the information to develop variable-rate technology (VRT) application maps.

where a crop is grown every other year. Incident light sensors detect and measure green vegetation, and, when the vegetation, e.g., unwanted plants or weeds, takes up more than 5% of a detector's field of view, a controller automatically turns on the individual spray nozzle. This type of sprayer, which greatly reduces the area sprayed and expenditure on chemicals, is a good conservation tool.

Similarly, nutrient applications of nitrogen and other materials are gaining acceptance owing to effective soil testing, such as taking samples according to soil maps or soil management zones. The application of lime to correct soil acidity by development of application maps based on soil samples has gained acceptance among many farmers.

Proper Seeding

Crops grow best when seeds are spaced such that plants obtain maximum sunlight and moisture and are placed at a proper depth within the soil. This is best accomplished by varying the seeding rate according to soil conditions such as texture, organic matter, and available soil moisture. For example, fewer seeds would be planted in sandy soil than in silt loam soils due to the amount of available moisture. To maximize yield per plant, seeds would be sown at a lower density, which would usually produce larger heads (ears) of harvested seeds because of more soil nutrients and sunlight available per plant.

However, planting fewer seeds runs counter to the practice of increasing seed density to increase yields, a trajectory that when overdone can yield a biomass so dense that some sunlight is blocked and the plants receive insufficient sunlight to produce seeds. The farmer's challenge is to find the right seed density to match all the localized factors affecting plant growth. Since soils can vary significantly across one field, the farmer can change seeding rates while crossing the field and thereby maximize the seeding rate according to the soil conditions.

Equipment modifications needed for VRTs have been designed for corn planters, grain drills, and air seeders (Figure 1). The desired rates are usually determined by reviewing previous yield images and/or soil productivity estimates. A computerized soil map of a specific field on a computer fitted on the tractor along with a GPS can provide farmers with their exact position in the field and guide them in adjusting the seeding rate accordingly. Seed companies are beginning to make recommendations on various seeding rates for individual hybrids, which will help the farmer with the settings on the planter.

Harvesting

A key benefit in the use of site-specific management techniques such as the adjustment of seeds, pesticides, fertilizers, and tillage as a field is crossed is in knowing the precise yields at every location. Harvesting equipment can be equipped with yield-monitoring devices that are coupled to a GPS, which provide a yield measurement for specific areas. A yield amount and location can be obtained every 1–5 s as the field is traversed. With appropriate software, a yield map

Figure 2 Collection of soil and plant measurements in the field for site-specific management: (a) field spectrometer; (b) chlorophyll meter; (c) electrical conductivity; (d) soil-sampling.

can be produced showing the yield variation throughout the field. This allows farmers to inspect the precise location of the highest- and the lowest-yielding areas of the field and determine why the differences exist. It allows the analysis of costs and yield to determine the most profitable practices and rates that apply to each field location. This provides farmers with their own research information that can be used in a practical and profitable way.

For farmers wanting to adopt precision farming, yield monitors are a good place to start. Yield monitors are already in use for measuring corn, soybeans, cotton, and other small grains such as oats and wheat, and monitors for other crops such as potatoes and peanuts are being developed. Farmers are also encouraged to monitor their fields closely by scouting (walking the field) and using remote-sensing techniques so that they have proof of problems or changes. By using a GPS unit they can go to a specific low-yield location and soil sample to determine whether lack of specific nutrients is the cause (Figure 2). With this information, farmers can assess the cost-effectiveness of the additional input. Some low-yield areas might best be ignored for growing crops.

Adoption Considerations

Farmers must know how the technology operates and be computer-literate. If they are working with a crop consultant who can analyze their data and make recommendations for them, that is a start, but being computer-literate is an advantage (Figures 3 and 4).

The number of farmers involved in precision or site-specific management practices around the globe is not known. It is likely that approximately 50% of the farmers in developed countries have had exposure to the use of VRT applications of soil-nutrient amendments and pesticides. Many new combines have yield monitors as standard equipment, which means that the opportunity to learn how to use yield images and/or maps increases with each combine that is purchased. The use of the yield images has caused farmers to request remotely sensed images during the growing season for comparison. Most are hoping to determine the crop yield prior to harvest and in the process are learning that this is difficult to achieve; however, the knowledge of the locations of crop and soil anomalies is beneficial to improving their management skills. Current trends in the use of remote sensing data in agriculture are driven by goals that are part of site-specific crop management. Farmers can adopt precision-farming techniques through cooperatives and through consultants who use GPS technologies (Figures 3 and 4).

The Future

Several companies are starting to market GIS record-keeping systems so farmers can record all of the field operations, including planting, spraying, cultivation,

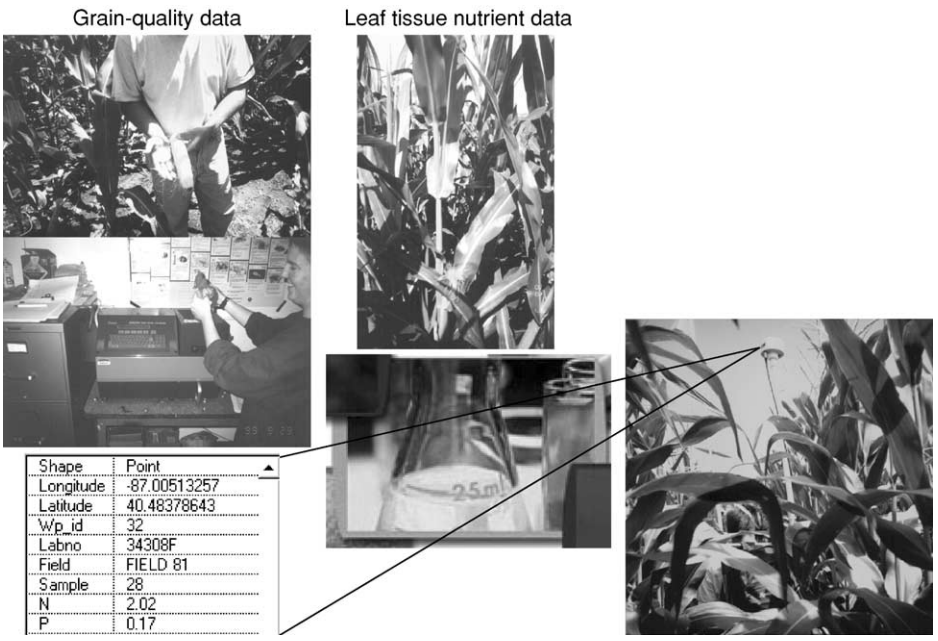


Figure 3 Global positioning system (GPS) technology is used to record the field locations where ground reference plant samples are collected. GPS locations make it possible to relate sample results to remote sensing image data.

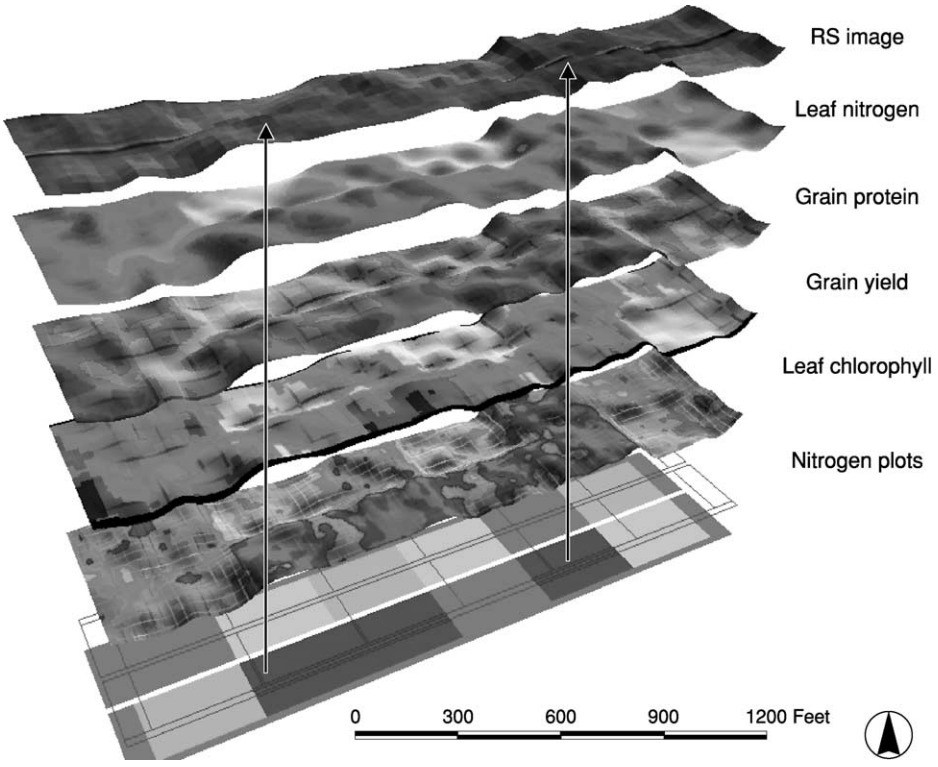


Figure 4 GIS layers in a digital terrain format showing a remotely sensed (RS) image and measured plant properties for a field experiment testing four rates of nitrogen that would provide two inadequate rates, an adequate and an excessive rate within three random replicated corn plots.

and harvest (along with specific information such as type of equipment used, usage rates, weather information, time of day performed, and weather data). Additionally, farmers can record observations such as weed growth, unusual plant stress, coloring, and growth conditions throughout the season. All GPS-collected data can be automatically recorded and used in conjunction with a GIS program.

Remotely sensed data can be analyzed and added to the GIS using soil maps and digital-terrain and field-operations information as ground references. This can be used to guide further field operations such as spraying, fertilizing, and irrigating; plus it is part of a permanent record. A number of new satellites provide data with increased spatial resolution such as 1, 5, and 10 m, which is similar to digital aerial images. This means that remotely sensed data would complement the yield data collected.

Additionally, the spectral resolution will be increased through hyperspectral image data, meaning that, instead of a few broad bands of 100 nm bandwidth, there will be more than 100 narrow bands of 5 and 10 nm bandwidth. This helps to define the colors more sharply, which leads to more detail about leaf pigments, soil organic matter, and other plant and soil properties. Much research is being done to determine precisely what is being measured.

On the horizon are sensors operated from the tractor that will measure soil properties such as organic matter, clay content, and soil moisture in real time. This will have a significant impact on planting operations. It means that, based on measurements from sensors mounted on planters, farmers will be able to adjust seeding, fertilizer application, and chemical application rates as they plant. By recording these inputs and tracking rates by location, they will be able to overlay the yield information and calculate the precise cost of production for any portion of their fields. Precision farming is making a strong impact on the way farmers in highly mechanized systems are able to protect their soil environment and manage their farm operations. The steady growth of remote sensing, GIS, and GPS technologies will continue to encourage site-specific soil management.

See also: **Conservation Tillage**; **Factors of Soil Formation:** Biota; **Geographical Information Systems**; **Remote Sensing:** Organic Matter; Soil Moisture

Further Reading

Johannsen CJ (2001) Glossary of terms for precision farming. In: Lowenberg-DeBoer JM and Erickson KM (eds) *Precision Farming Profitability*, pp. 125–132. Lafayette, IN: Agricultural Research Programs, Purdue University.

Johannsen CJ, Baumgardner MF, Willis PR, and Carter PG (1998) *Advances in Remote Sensing Technologies and their Potential Impact on Agriculture*. 1st International Conference on Geospatial Information in Agriculture and Forestry. Proceeding vol. 1, pp. 6–11. Ann Arbor, MI: Verdian International.

Johannsen CJ, Petersen GW, Carter PG, and Morgan MT (2003) Remote sensing changing natural resource management. *Journal of Soil and Water Conservation* 58: 42–45.

Lowenberg-DeBoer JM and Erickson KM (eds) (2001) *Precision Farming Profitability*. Publication SSM-3, Agricultural Research Programs. West Lafayette, IN: Purdue University Press.

Medlin CR, Hawkins SE, Morris DK, and Parsons S (2001) Variable-rate technologies. In: Lowenberg-DeBoer JM and Erickson KM (eds) *Precision Farming Profitability*. Publication SSM-3, Agricultural Research Programs. West Lafayette, IN: Purdue University Press.

Moran MS, Inoue Y, and Barnes EM (1997) Opportunities and limitations for image-based remote sensing in precision crop management. *Remote Sensing of the Environment* 61: 319–346.

Plant RE (2001) Site-specific management: the application of information technology to crop production. *Computers and Electronics in Agriculture* 30: 9–29.

Robert PC and Stafford J (2003) *Precision Agriculture: An International Journal on Advances in Precision Agriculture* 4: 1–444.

Strickland RM and Parsons S (2001) The basics of geographic information systems. In: Lowenberg-DeBoer JM and Erickson KM (eds) *Precision Farming Profitability*, pp. 97–102. Publication SSM-3, Agricultural Research Programs. West Lafayette, IN: Purdue University Press.

Slash and Burn Agriculture

See **Shifting Cultivation**

Sludge

See **Waste Disposal on Land:** Liquid; Municipal

SODIC SOILS

G J Levy and I Shainberg, The Volcani Center, Bet Dagan, Israel

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Presence of sodium (Na) as an exchangeable cation is found in many soils worldwide. Sodic soils exhibit poor soil–water and soil–air relations, which in turn adversely affect root growth and plant production, increase soil erosion, and make the soils difficult to farm. Problems associated with sodic soils are expected to increase. The need to provide food to an ever-expanding population, coupled with the increasing demand for good-quality water from the urban and industrial sectors, results in poorer-quality water (e.g., treated effluents) and soils being used for food production. Understanding the behavior of sodic soils and its effects on agriculture and the environment is, therefore, essential in order to manage sodic soils properly for securing crop production and maintaining sustainable agriculture worldwide.

Definition of Sodic Soils

Two important parameters are used to define sodic soils. The first is a soil property – the exchangeable sodium percentage (ESP), which describes the level of adsorbed Na in soil:

$$\text{ESP} = (\text{exchangeable Na}/\text{cation exchange capacity}) \cdot 100 \qquad [1]$$

Instead of cation exchange capacity it is also possible to use the sum of exchangeable cations (Ca + Mg + Na + K). The second parameter which indicates the sodicity level of the irrigation water or soil solution is termed sodium adsorption ratio (SAR):

$$\text{SAR} = (\text{Na})/((\text{Ca} + \text{Mg})/2)^{0.5} \qquad [2]$$

where parentheses reflect cation concentration in $\text{mmol}_c\ l^{-1}$. Thus, SAR has units of $(\text{mmol}_c\ l^{-1})^{0.5}$. It must be emphasized that the sodicity hazard of solutions is related to the ratio of Na to the divalent cations present in the water and not to Na concentration solely.

Exchange reactions take place between the soil solution and the exchangeable phase. Thus, soil ESP can be estimated from the SAR of saturated paste extracts using the following empirical relationship:

$$\text{ESP} = 100(-0.0126 + 0.01475 \cdot \text{SAR})/(1 + [-0.0126 + 0.01475 \cdot \text{SAR}]) \qquad [3]$$

or from the nomograph presented in Figure 1.

To date there is no widely accepted critical ESP value which separates sodic from nonsodic soils. This is because the effects of ESP cannot be separated from that of the electrolyte concentration (EC) of the soil solution. The interrelationship between ESP of the soil or SAR of its equilibrium solution and the EC of the soil solution may help estimate whether a sodic or nonsodic behavior will occur.

Origin and Distribution of Sodic Soils

In many sodic soils sodicity originates from the nature of the parent material and pedogenic processes. There are also sodic soils where sodicity arises from anthropogenic processes (human activity), and this is termed secondary sodification. Irrigation with sodic water or without proper drainage, forest clearing, and other land management practices that lead to water logging

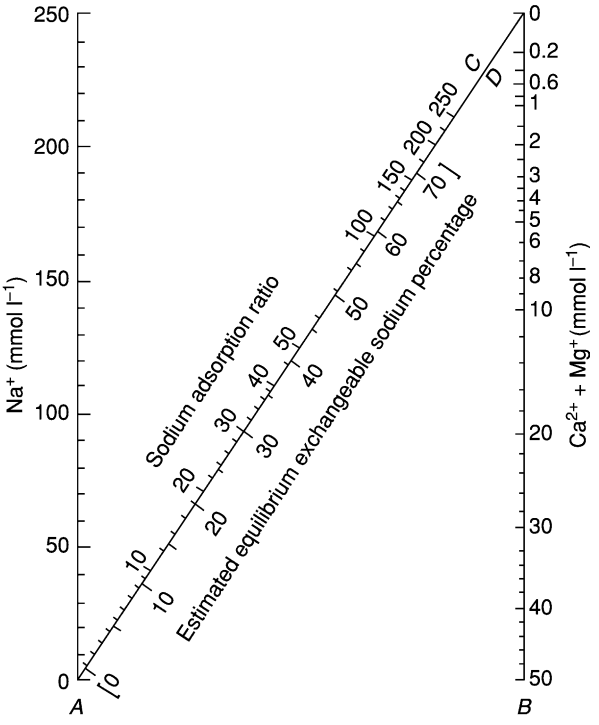


Figure 1 Nomograph for determining the sodium absorption ratio of a saturation extract and estimating the corresponding exchangeable sodium percentage value of soil at equilibrium with the extract. Reproduced from US Salinity Laboratory Staff (1954) *Diagnosis and Improvement of Saline and Alkali Soils*. Agricultural Handbook 60. Washington, DC: USDA.

Table 1 World distribution of sodic soils

Continent	*Country*	*Area of sodic soils (thousand hectares)*
North America	Canada	6974
	USA	2590
South America	Argentina	53 139
	Bolivia	716
	Brazil	362
	Chile	3642
Africa	Botswana	670
	Cameroon	671
	Chad	5950
	Ethiopia	425
	Kenya	448
	Madagascar	1287
	Namibia	1751
	Niger	1389
	Nigeria	5837
	Somalia	4033
	Sudan	2736
	Tanzania	583
	Zambia	863
South Asia	Bangladesh	538
	India	574
	Iran	686
North and Central Asia	China	437
	USSR	119 628
Australia	Australia	339 971

Reproduced from Massoud FI (1977) Basic principles for prognosis and monitoring of salinity and sodicity. *Proceedings of an International Management of Saline Water for Irrigation*, pp. 432–454. Lubbock, TX: Texas Technical University.

are key activities that may yield rapid secondary sodification.

The total area worldwide of sodic soils is estimated at 210 million ha. The area of sodic soils in each continent and in some countries where they form an important proportion of the soils is presented in Table 1. The spatial distribution of sodic soils indicates that they occur in a range of climates. However, the outlined distribution of sodic soils and the conditions promoting their formation are based on a pedological approach. Following the increase in the contribution of human activity to soil sodification, the definition of sodic soils is now based on soil behavior. Many agricultural soils that exhibit sodic behavior may not fit into the classical sodic soil group. Consequently, the sizeable area these soils cover remains unaccounted for.

Mechanisms Responsible for Sodic Behavior

Among the various soil constituents, the fraction that considerably determines the physical behavior of soils is the colloidal clay. Compared with other soil constituents, the clay fraction possesses a large specific surface area (surface area per gram of clay), which, combined with its charge, makes the clay a very reactive soil constituent in physicochemical processes, mainly swelling and dispersion. These two processes largely determine the microstructure of the soil, thus significantly affecting many of its physical properties.

The intensity of swelling and dispersion processes and the magnitude of their adverse effects on soil properties become more severe with the increase in soil sodicity. However, the swelling and dispersion behavior of soil clays differs significantly from that of pure clay systems, possibly because soil clays usually occur as mixtures and because of their association with other minerals, oxides, and organic matter present in the soil. It has been demonstrated that the EC required for preventing dispersion of soil clays is two- to 10-fold higher than that for pure clays. On the other hand, the presence of hydrous oxides or sparingly soluble minerals such as $CaCO_3$ has resulted in lower levels of soil clay dispersivity. Thus, extrapolation from pure clay systems to soil systems may often prove problematic.

Structural Stability of Sodic Soils

Soil structure is commonly referred to as the size, shape, and arrangement of particles (micro- and macroaggregates) and pores in soil. Soil structural stability is therefore the ability of the soil to retain its arrangement of solids and pores when exposed to different stresses. The presence of high levels of sodicity aggravates aggregate susceptibility to breakdown and slaking. However, these effects of sodicity are intricate and should not be considered independently of the EC of the soil solution and soil texture (Figure 2).

Sodicity-related degradation of soil structure leads to deterioration in soil water transmission properties, and increased susceptibility to crusting, runoff, and erosion. Water movement in the soil profile is commonly characterized by its hydraulic conductivity (HC). In cases where a thin seal is formed at the soil surface, water penetration into the soil is characterized by the infiltration rate (IR) rather than HC. Because the properties of the seal differ from those of the underlying soil, and its thickness is difficult to measure, this makes the determination of its HC extremely difficult.

Hydraulic Conductivity

The HC of a soil depends on its properties and those of the percolating fluid. Existing models have failed in adequately predicting soil HC, largely because they do not take into account the effect of fluid properties

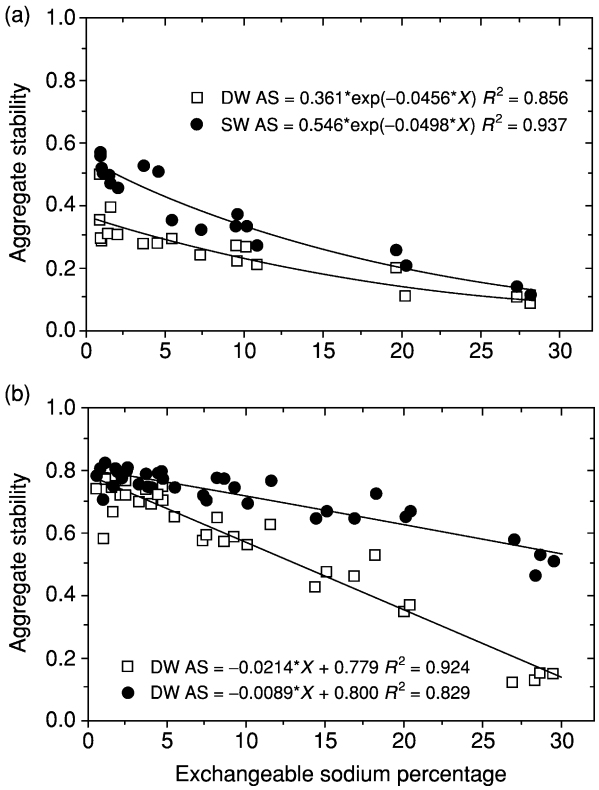


Figure 2 Aggregate stability (AS, expressed as a fraction) as a function of exchangeable sodium percentage and electrolyte concentration for (a) soils with clay content ≤25% and (b) soils with clay content ≥35%. DW, distilled water; SW, saline water. Modified from Levy GJ, Mamedov AI, and Goldstein D (2003) Sodicity and water quality effects on slaking of aggregates from semi-acid soils. *Soil Science* 168: 552–562.

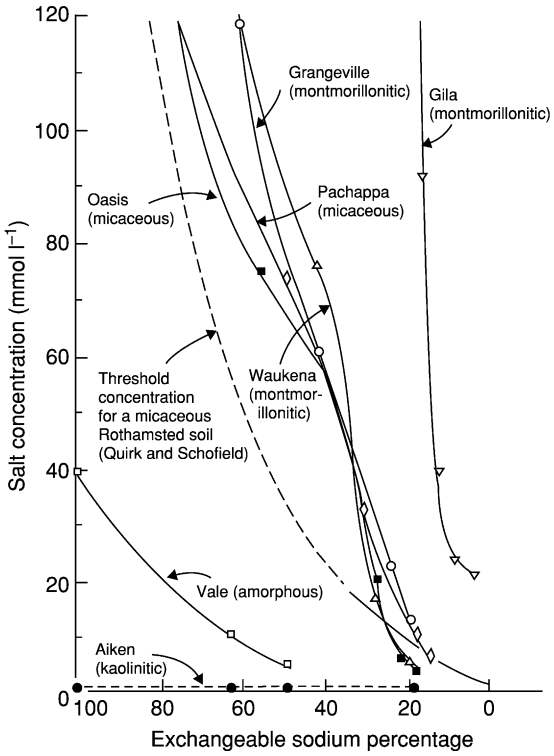


Figure 3 Combinations of salt concentration and exchangeable sodium percentage required to produce a 25% reduction in hydraulic conductivity for selected soils. Reproduced with permission from McNeal BL and Coleman NT (1966) Effect of solution composition on soil hydraulic conductivity. *Soil Science Society of America Proceedings* 30: 308–312.

on the soil matrix geometry. Swelling and dispersion of clay in response to sodic conditions and the level of EC in the percolating solution change the size of the conducting pores and hence affect the HC.

Sodicity and electrolyte concentration The realization that soil HC depends on both the SAR and the EC of the percolating solution led to the development of the 'threshold concentration' concept, which was defined as the EC required to prevent a given decrease in the HC for a given soil ESP or SAR of the percolating solution (Figure 3). This concept has been tested and validated on a large variety of soil types. It explains observations of decreased HC in soils, even Ca-soil and soils of low ESP, exposed to water of low salinity (e.g., rain or snow water). The decrease in HC is because salt concentration in the soil solution is not sufficient to prevent swelling and clay dispersion. For instance, in Australia under rain-fed conditions, soils (especially clayey ones) often waterlog during the rainy season because what little salt is present is leached in the first rain, causing the clay to disperse and the HC to drop sharply. Conversely, it is possible to maintain high HC in an Na-saturated soil provided the EC is sufficiently large. In Israel, experience from commercial fields showed that, for Na-affected soils, irrigation with water having an SAR value as high as 26 with an EC of 4.6 dS m^{-1} exhibited no permeability problems, as the electrolyte concentration in the irrigation water was sufficient to prevent the dispersive effect of Na. However, upon applying electrolyte-free water (to simulate rainwater in winter), soil HC dropped to a small fraction of its initial high value.

A distinction should be made between the effects of clay swelling and those of clay dispersion on the HC of soils. Soils begin to swell as the EC of the soil solution is reduced, but the effect of swelling on HC becomes apparent only at SAR values above 10 in medium- and heavy-textured soils. In light-textured soils the effect of swelling on reducing the conducting pores and hence HC is hardly noticeable. Furthermore, the swelling process is reversible; thus an increase in HC can be noted following an increase in EC of the soil solution. Clay dispersion, on the other hand, can take

place at already low ESP levels provided the EC is below a critical level. When clay dispersion occurs, the dispersed clay particles move down the soil profile and may even cause a complete blockage of the conducting pores, and hence an irreversible change in the HC. Consequently, the effects of reducing EC on the HC are likely to be more severe when clay dispersion, rather than swelling, is the predominant cause.

Soil properties affecting sodic behavior The response of soils to sodic conditions varies among soils, mainly because soil properties have a significant effect on the dependence of the HC on sodicity. A major factor causing differences among various sodic soils in their susceptibility to hydraulic failure when leached with low-EC water is their rate of salt release from mineral dissolution. Mineral dissolution determines the electrolyte concentrations of percolating solutions and hence soil HC. Sodic soils containing minerals ($CaCO_3$ and a few primary minerals) that readily release soluble electrolytes will not readily disperse when leached with distilled water (DW) at moderate ESP values, because a sufficiently high electrolyte concentration ($\sim 3\,mmol_c\,l^{-1}$) can be maintained to prevent clay dispersion. In addition, the ESP level will decrease, because most of the cations released are divalent (Ca and Mg).

Clay content, through its effect on pore size distribution, contributes significantly to the response of soils to sodic conditions. The sensitivity of HC to sodic conditions is generally greater for soils having medium and high clay content. When clay content is high, clay swelling may significantly reduce the size of the water-conducting pores and lead to low HC. Similarly, under dispersive conditions the dispersed clay particles clog the pores and cause a reduction in soil HC. Conversely, in sandy soils, due to the low clay content, clay swelling has only a limited effect on the HC. Conditions supporting clay dispersion may cause the HC first to decline and subsequently to recover. This phenomenon is explained by the change in flow pattern, from a solution flowing through a matrix of sand covered with clay which starts to disperse and reduce HC, to a flow of a clay suspension in a pure sandy matrix having large pores resulting in subsequent increase in HC.

Other soil properties that may affect the response of soils to sodic conditions include:

1. Presence of aluminum and iron oxides: These oxides, commonly found in highly weathered soils, enhance soil stability, via a cementing action, and substantially decrease its susceptibility to sodic conditions.
2. Clay mineralogy: Soils with a high content of 2:1 layer silicates (e.g., montmorillonite) are the most labile ones while those high in kaolinite and sesquioxides are the least labile.
3. Exchangeable Mg: Although in many soils the exchangeable Ca/Mg molar ratio is high, in some soils the reverse is true. The effect of adsorbed Mg on soil hydraulic properties seems to depend on soil type. In calcareous soils, the presence of Mg enhanced the dissolution of $CaCO_3$, producing electrolytes which prevent clay dispersion and HC decay, while in non-calcareous soils Mg causes a decrease in the HC of the soils beyond that of the Na/Ca system. There is the direct ('specific') effect of exchangeable Mg in causing a decrease in HC, and there is the inability of Mg in irrigation water to counter the accumulation of exchangeable Na in the soil.

Sodicity and conditions prevailing in the soil It has recently been recognized that the soil response to sodic conditions depends not only on soil properties but also on the conditions prevailing in the soil (e.g., antecedent moisture content and wetting rate). Soil conditions affect aggregate stability/breakdown and thus the pore size distribution and consequently the HC. Aggregate breakdown results from dispersion and from slaking during wetting, whose rate determines the degree of slaking. It has been observed that the HC of sodic soils leached with DW decreased more steeply and to lower values with the increase in the rate at which the soil was wetted. Fast wetting, which increases aggregate slaking, increases the susceptibility of the soils to sodic conditions. Conversely, slow wetting decreases significantly the effects of sodicity on soil HC, especially in fine-textured soils (Figure 4).

Infiltration Rate

Soil infiltration rate (IR) is defined as the volume flux of water flowing into the profile per unit of soil surface area under any set of circumstances. Often when water is supplied to soil, the IR decreases from its initial high rate due to formation of a thin layer (<2 mm) at the soil surface, termed seal. The discussion to follow will focus on structural seals usually formed by the impact energy of water drops.

Sensitivity of seal formation and IR of a smectitic loamy sand to ESP when exposed to distilled water under laboratory-simulated rain is presented in Figure 5. Even at the lowest sodicity (ESP 1.0), a seal was formed and the IR dropped from an initial value >100 to a final infiltration rate (FIR) of $7.0\,mm\,h^{-1}$. An ESP value of 2.2 was sufficient to cause a further drop in the final IR of the sandy loam to $2.4\,mm^{-1}$. The amount of rain required to approach the FIR (i.e., the rate of seal formation) was

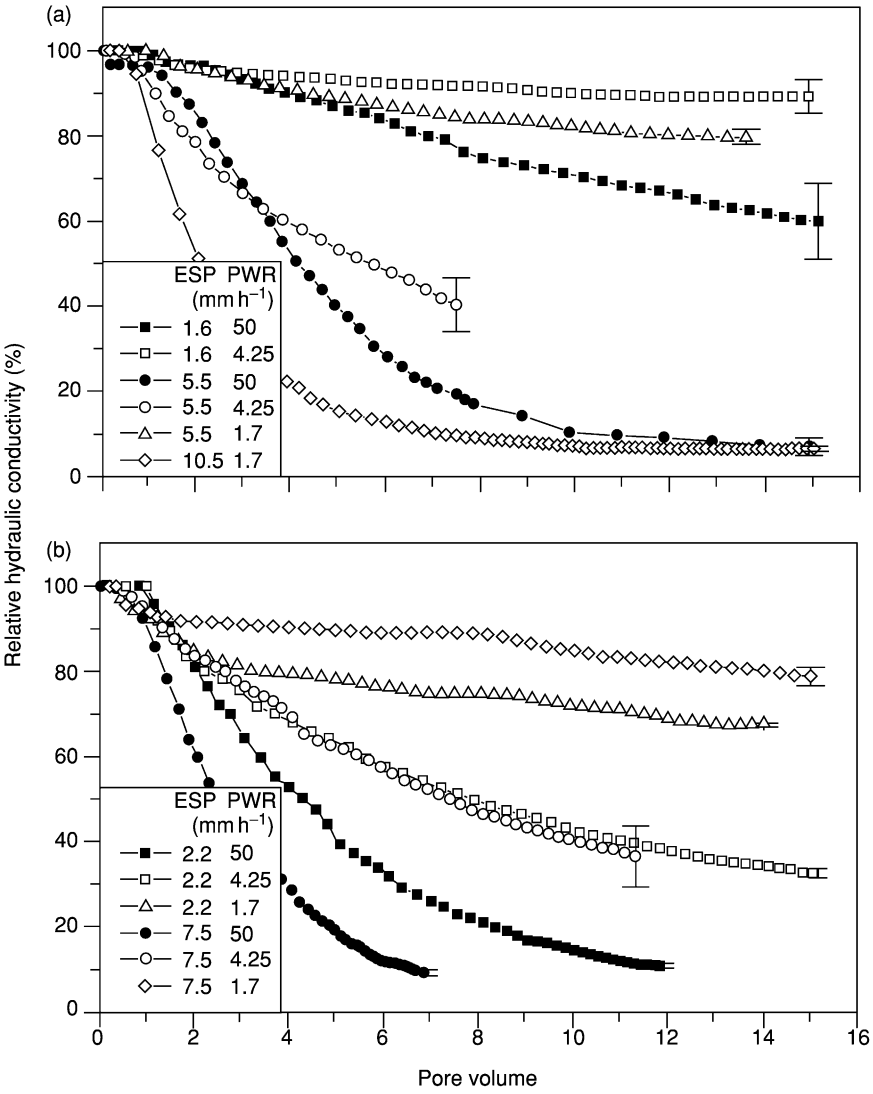


Figure 4 Relative hydraulic conductivity of (a) sandy clay with exchangeable sodium percentage (ESP) of 1.6, 5.5, and 10.5, and (b) clay Y with ESP 2.2 and 7.5 as a function of three prewetting rates (PWR: 50, 4.25, and 1.7 mm h^{-1}) and leaching with deionized water. Reproduced with permission from Shainberg I, Levy GJ, Goldstein D *et al.* (2001) *Australian Journal of Soil Research* 39: 1279–1291. CSIRO Publishing, Melbourne, Australia.

also affected by sodicity (Figure 5). As raindrop impact energy was the same in all the experiments, the differences in IR curves for the various treatments were the result of chemical dispersion of the soil clay caused by sodicity. The high sensitivity of the soil surface to sodicity (even at low ESP values) is explained by the mechanical impact of the raindrops, which enhances chemical dispersion, and the almost total absence of electrolytes in the applied distilled water.

In cases where soil is exposed to overhead irrigation (e.g., sprinklers), IR is sensitive to both the EC and SAR of the applied water. The lower the concentration of the electrolyte solution, the faster the rate at which the IR decreases (Figure 6). For similar electrolyte concentrations, increasing soil ESP leads to a sharper decrease in and a lower final IR (Figure 6). Increasing EC greatly affected IR even at low sodicity. It was reported for Californian soils that final IR increased from 2 to 28 mm h^{-1} as cation concentration in the applied water having SAR <5 increased from 5 to 28 $cmmol_c\ l^{-1}$. These results, and others, indicate that IR is far more sensitive than HC to both SAR and electrolyte concentration of the applied water. Furthermore, when Mg is the complementary cation to Na, then IR in Mg-Na soil may often be lower than that in the Ca-Na soil, especially under conditions where chemical dispersion is dominant in controlling IR (i.e., water drops with low to medium kinetic energy).

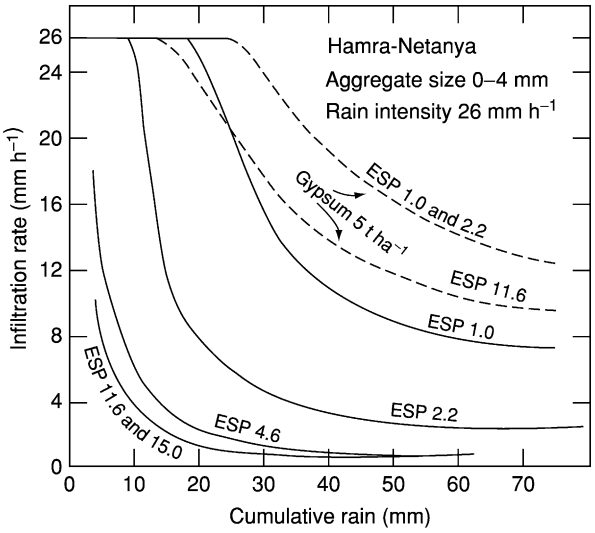


Figure 5 The effects of the soil exchangeable sodium percentage (ESP) and phosphogypsum application on the infiltration rate of the Hamra (Netanya) soil as a function of the cumulative rain. Reproduced with permission from Kazman Z, Shainberg I, and Gal M (1983). Effect of low levels of exchangeable applied phosphogypsum on the infiltration rate of various soils. *Soil Science* 35: 184–192.

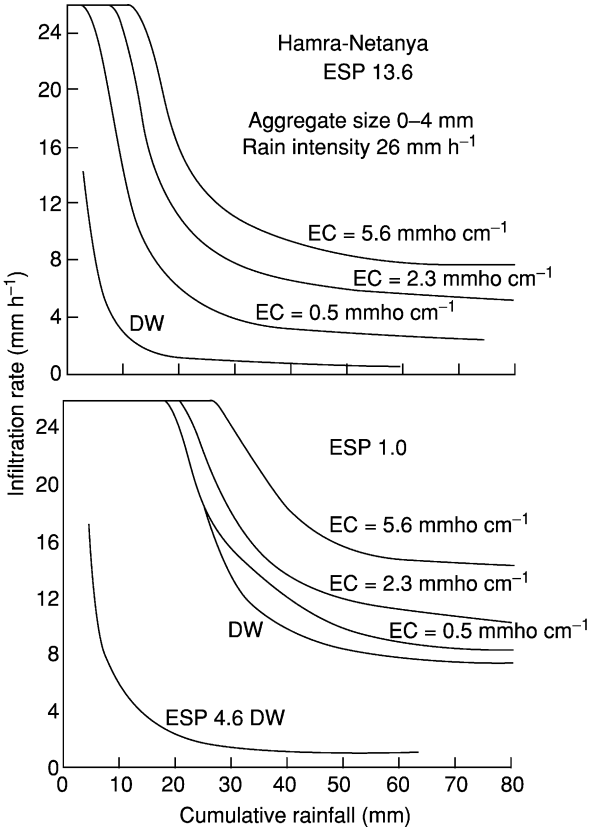


Figure 6 Effect of electrolyte concentration in rain simulation experiments on the infiltration rate of a loess soil. ESP, exchangeable sodium percentage; DW, distilled water; EC, electrical conductivity. Reproduced from Agassi M, Shainberg I, and Morin J (1981) *Soil Science Society of America Journal* 45: 848–851.

Similarly to HC, it has recently been noted that the IR of sodic soils depends not only on soil properties but also on the conditions prevailing in the soil (e.g., antecedent moisture content and wetting rate) because IR is greatly affected by aggregate stability/breakdown. The dependence of IR of sodic soils on the rate of aggregate wetting is closely associated with soil texture. In coarse- and medium-textured soils, sodicity effects on IR are only slightly altered by aggregate wetting rate. In fine-textured soils, decreasing the rate of aggregate wetting prior to exposing them to rain significantly decreases the sensitivity of the IR to sodic conditions (Figure 7).

Sodicity and Soil Productivity

Sodic conditions are expected to affect soil productivity adversely because such conditions can lead to poor aeration, restricted root development, and enhanced root disease. Moreover, sodic conditions can lead to nutritional imbalances and micronutrient deficiencies since they could alter soil redox potential, pH, and dissolved organic carbon concentration.

Nitrogen, phosphorus, and potassium are the most important nutritional elements. Phosphorus and potassium do not pose a nutritional problem in sodic soils. Conversely, sodic soils may suffer from inadequate nitrogen availability. The reasons for this may stem from low rates of nitrogen mineralization, denitrification under restricted aeration conditions, volatilization of ammonia under alkaline conditions, and inhibition of nitrate (NO_3) uptake by chloride (Cl) or sulfate (SO_4) ions which are usually in abundance in sodic soils. Sodic soils may also suffer from impaired uptake of Ca. The ratio of Ca to total cation concentration (TCC) is considered a better indicator of Ca availability to plants than actual Ca concentration. Saturated paste extracts of sodic soils often exhibit Ca/TCC < 0.1, which for many crops indicates Ca deficiency. In addition, high Mg/Ca ratios, often found in sodic soils, can also cause Ca deficiency.

Two approaches exist with respect to studying crop tolerance to sodicity, i.e., in the presence or absence of sodicity-related adverse soil physical conditions. Studies on crop tolerance to sodicity based merely on nutritional factors have indicated that, already at relatively low ESP levels (2–10), some crops show signs of sodium toxicity; however, most field crops are classified as moderately tolerant or tolerant to sodicity (Table 2). On the other hand, studies that have been carried out under conditions where soil structure had been allowed to be affected by sodicity indicated that grasses were tolerant, wheat and barley were moderately tolerant, and pea and beans were

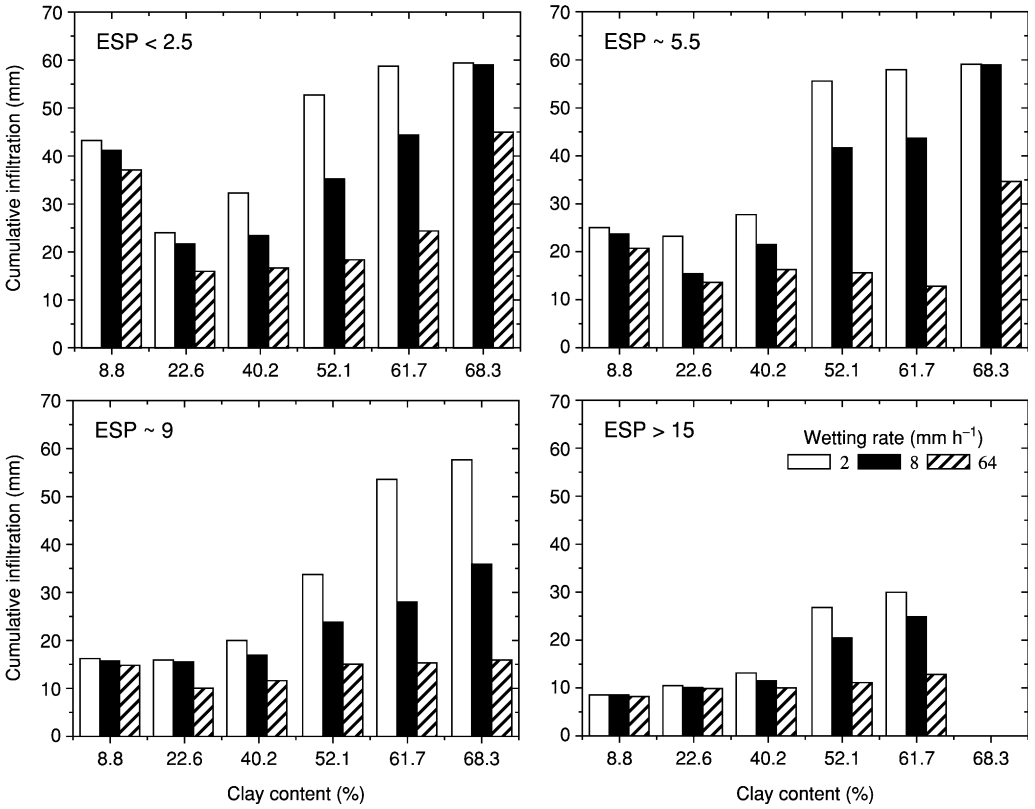


Figure 7 Effects of wetting rate on the cumulative infiltration from a 60-mm rain storm for six soils varying in clay content and for four exchangeable sodium percentage (ESP) levels.

Table 2 Tolerance of various crops to exchangeable sodium

Tolerance to ESP	*Crop*
Extremely sensitive (ESP 2–10)	Deciduous fruits
	Nuts
	Citrus
	Avocado
Sensitive (ESP 10–20)	Beans
Moderately tolerant (ESP 20–40)	Clover
	Oats tall fescue
	Rice
	Dallis grass
Tolerant (ESP 40–60)	Wheat
	Cotton
	Alfalfa
	Barley
	Tomatoes
	Beets
Most tolerant (ESP >60)	Crested and fairway wheat grass
	Tall wheat grass
	Rhodes grass

ESP, exchangeable sodium percentage. Reproduced from Pearson GA (1960) Tolerance of crops to exchangeable sodium. *US Department of Agriculture Information Bulletin* 216.

sensitive to sodicity. It should be borne in mind that the results of such studies could be site-specific, and may thus hinder their applicability to other soil types.

Reclamation of Sodic Soils

Reclamation of sodic soils, which is the mending of soil structure, is essential if productivity is to be maintained. The main concept for reclamation of sodic soils advocates the addition of a source of Ca for replacing exchangeable Na coupled with excess water to leach the Na from the root zone deep into the soil profile. Less common methods include the addition of a variety of organic ameliorants to maintain high levels of organic matter in the soil and thus stabilize its structure.

Chemical Reclamation

Reclaiming sodic soils via the addition of divalent cations can be achieved by using chemical amendments that release Ca ions to the soil, or by applying successive dilutions of high-salt water containing divalent cations.

Chemical amendments Typical amendments are those containing a source of soluble Ca or that dissolve Ca on reaction in the soil. Common amendments include gypsum, lime, $CaCl_2$, sulfuric acid, and sulfur.

Gypsum ($CaSO_4$), from both geological and non-geological sources (industries that produce $CaSO_4$ by-products), is the most commonly used amendment for sodic soil reclamation, primarily because of its low cost, reasonable solubility, and availability. Gypsum added to a sodic soil can increase its permeability by means of both EC and cation-exchange effects. The electrolyte effect is important during application of water with extremely low EC.

The cation exchange effect is important in soils where treating high ESP of the soil profile is important. The amount of gypsum required depends on the amount of exchangeable Na in a selected depth of soil. The amount of exchangeable Na to be replaced per unit land area (Q_{na}, $mol_c\,ha^{-1}$) during reclamation depends on the initial and desired final exchangeable Na fraction, the cation exchange capacity of the soil, its bulk density, and the depth of soil to be reclaimed. Thus, the amount of gypsum (in $Mg\,ha^{-1}$) needed to reclaim the soil, termed 'gypsum requirement' (GR), is calculated as follows:

$$GR = (8.61 \times 10^{-5})Q_{na} \quad [4]$$

The efficiency and rate of exchange, namely the percentage of Ca that exchanges for adsorbed Na, vary with ESP, being much greater at initially high ESP values. A number of GR models exist, and provide a powerful tool for quantitative predictions of water and gypsum required to reclaim a soil to a predetermined level of sodicity.

The use of $CaCO_3$ to reclaim sodic soils is considered to be of limited value, mainly because of its low solubility and low dissolution rate. Naturally occurring $CaCO_3$ may help soil reclamation by maintaining EC in the soil solution at a level high enough to prevent clay dispersion and a subsequent reduction in soil HC. However, because of its low solubility, $CaCO_3$ is not effective in maintaining high permeability at the soil surface during rain.

Successive dilutions of high-salt water In areas where water is not a limiting factor, reclamation of sodic soils can be obtained by leaching the soil with successive dilutions of high-salt water containing divalent cations. In the early stage of the leaching, the high EC in the applied water prevents clay dispersion and induces flocculation of the soil colloids. Simultaneously, the Ca ions in the water decrease sodicity by replacing exchangeable Na. Upon dilution of the saline water with high-quality water, the SAR of the water is reduced. To ensure successful soil reclamation, the depth of water added should be at least 9–10 times the depth of soil to be reclaimed.

Management of Sodic Soils

In order to maintain sustainable agriculture in sodic soils, and keep their productivity, proper management practices must be applied, including the use of soil amendments, tillage, and cropping.

Soil Amendments

Common soil amendments used to maintain high permeability in sodic soils include gypsum that provides electrolytes to prevent clay dispersion and organic polymers that stabilize soil aggregates. Phosphogypsum (PG) and flue gas desulfurization gypsum (FGDG), which are by-products of the phosphate and power generation industries, are readily available and therefore frequently used. Spreading powdered PG at rates of $5\,Mg\,ha^{-1}$ on the soil surface can be effective in slowing down the rate of IR decline and maintaining a higher final IR compared with the control, even in soils of ESP < 2.5 (Figure 5), indicating that some chemical dispersion takes place even at very low ESP values, and possibly even in Ca-saturated soils. The use of PG is also very effective in maintaining high IR in soils with high ESP (Figure 5). The favorable effect of PG on IR should be attributed not only to its effect on the EC of the percolating water but also to: (1) the physical interference with the continuity of the seal; and (2) the partial mulching of the soil surface which protects the soil from the beating action of the raindrops.

Organic polymers used for improving aggregate stability, maintaining high IR, and reducing seal formation are mainly polysaccharides (PSD) and polyacrylamides (PAM). The addition of small amounts of these polymers ($10–20\,kg\,ha^{-1}$), either sprayed directly on to the soil surface or added to the applied water, was noted to be effective in stabilizing and cementing aggregates together at the soil surface, and hence maintaining high IR values in soils with ESP < 5. On the other hand, when added to soils with high ESP (>20), polymers were ineffective in controlling seal formation and maintaining high IR values; in soils with moderate ESP (10–15) the effects of polymers were inconsistent and depended on water quality and the amount of polymer added. The efficacy of anionic polymers in preventing seal formation is enhanced when the soil clay is maintained in a flocculated state and the resulting final IR could be 10 times higher than the control. It is therefore recommended that in soils exposed to rain, polymer application should be accompanied by spreading a source of electrolytes (e.g., gypsum) at the soil surface. Of the polymers currently available and under study, anionic polyacrylamide with high molecular weight ($>10^6$ Da) has been found to be the

most effective in controlling seal formation and maintaining high IR.

Tillage

Tilling sodic soils may result in the improvement of soil physical and hydraulic properties, especially when accompanied by some reclamation measures. Tillage reduces bulk density and increases total porosity (especially macroporosity), and improves internal drainage and saturated and unsaturated HC. Deep tillage (1–2 m) or ripping breaks up hardpans, and cemented layers mix soil layers and generally cause a permanent improvement in soil structure and physical properties. However, deep tillage can sometimes lead to waterlogging and infiltration problems by bringing sodic subsoil to the surface. An alternative to deep tillage in the case of sodic subsoil is the use of ridges instead of bed farming techniques. With ridges, the plants are located about 0.25 m above the neighboring furrows and thus aeration is improved, especially in heavy-textured soils suffering from water logging in the subsoil. Shallow tillage (disking or chiseling) is effective in reducing soil permeability problems associated with surface crusts, hardsetting, and compaction. Shallow tillage requires repetition from time to time.

Sodicity-Related Environmental Hazards

Sodic soils are more susceptible to clay dispersion and surface sealing, which generate runoff flow and subsequently lead to elevated levels of soil erosion, a serious problem from the crop production point of view. In addition, off-site damages of runoff and erosion from cropped fields are severe; in the USA alone the cost of these damages is estimated to be in excess of $1 billion. Sediments can clog drainage ditches and irrigation canals and decrease the capacity of water reservoirs, thus increasing their maintenance costs. In addition, sediments can damage or entirely destroy fish-spawning areas. Sedimentation of rivers and streams reduces their capacity to carry water and increases the danger of flooding and consequent damage to property and to human health.

The presence of suspended sediments in water increases its turbidity (a measure of water clarity). Under certain conditions, common in sodic soils (e.g., high pH and Na levels), organic matter can dissolve in runoff water and enhance clay dispersion and mobilization, and hence increase water turbidity. Turbid water is less attractive for recreational activities and thus this reduces the value of the water. Turbidity also impedes the passage of sunlight through the water, thus negatively affecting biological productivity of the water. High turbidity in water bodies used for domestic consumption requires cleaning the water, which in turn causes an increase in their price.

An additional environmental hazard stemming from sodic soils is the transport of various contaminants by soil colloidal material, whether to groundwater or to surface water bodies via runoff. Metals which are usually fairly immobile in soil have a strong affinity to soil colloids and thus have been shown to move together with the mobile colloids. Similarly, insoluble pesticides and herbicides, and some nutrients (e.g., phosphate), have a high affinity for soil colloids and can thus be transported with the suspended solids to adjacent water bodies and groundwater. The increased concentration of suspended colloids in runoff water due to sodic conditions greatly increases the danger of higher than permitted loadings of these pollutants in groundwater and surface water bodies.

Conclusions

Many soils worldwide contain sodium at levels that may negatively affect their behavior, and should thus be considered sodic soils. The area occupied by sodium-affected soils is expected to grow due to increased use of poor-quality soils and water for agricultural production. The adverse effects of sodic conditions on soil physical and hydraulic properties are aggravated when the soil is exposed to good (i.e., low EC) water and aggregates are slaked. Productivity of sodic soils is likely to decline, not only because of physical soil degradation but also because of nutritional imbalances, especially with respect to nitrogen and Ca. Sodicity enhances runoff and soil erosion, which lead to both on-site (cultivated fields) and off-site (environmental) damages. However, sodium-related soil degradation and loss of productivity are, to a certain degree, reversible as sodic soils can be reclaimed by the addition of amendments or by excessive leaching. Realization of sodicity hazards and consequent adoption of proper management tools allow for a safe and sustainable use of sodium-affected soils, without the danger of them being abandoned.

See also: **Aggregation:** Physical Aspects; **Calcium and Magnesium in Soils**; **Cation Exchange**; **Conditioners**; **Crusts:** Structural; **Flocculation and Dispersion**; **Liming**; **pH**; **Redox Potential**; **Salination Processes**; **Structure**

Further Reading

Australian Journal of Soil Research (1993) Special issue, vol. 31, No. 6.

Australian Journal of Soil Research (2001) Special issue, vol. 39, No. 6.

Bresler E, McNeal BL, and Carter DL (1982) *Saline and Sodic Soils. Principles – Dynamics – Modeling*. Berlin: Springer-Verlag.

Keren R (1995) Reclamation of sodic-affected soils. In: Agassi M (ed.) *Soil Erosion Conservation and Rehabilitation*, pp. 353–374. New York: Marcel Dekker.

Levy GJ (2000) Sodicity. In: Sumner E (ed.) *Handbook of Soil Science*, pp. G27–G63. Boca Raton, FL: CRC Press.

Levy GJ, Mamedov AI, and Goldstein D (2003) Sodicity and water quality effects on slaking of aggregates from semi-arid soils. *Soil Science* 168: 552–562.

Shainberg I and Letey J (1984) Response of soils to saline and sodic conditions. *Hilgardia* 52: 1–57.

Sumner ME and Naidu R (1998) *Sodic Soils*. New York: Oxford University Press.

US Salinity Laboratory Staff (1954) *Diagnosis and Improvement of Saline and Alkali Soils*. Agricultural Handbook 60. Washington, DC: USDA.

SOIL–PLANT–ATMOSPHERE CONTINUUM

J M Norman and M C Anderson, University of Wisconsin, Madison, WI, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

The dominant distinguishing feature of the Earth, in comparison with all other celestial bodies we observe, is the presence of life. Humans, as a part of life on Earth, attempt to understand all of the life system, albeit from an anthropocentric perspective. James Lovelock proposed a perspective that was unsettling to some and appealing to others when he suggested that the entire Earth may be functioning as a living organism. As our integrative understanding of living systems within the soil–plant–atmosphere continuum (SPAC) grows, the idea that a whole ecosystem can function as a living entity becomes more acceptable.

Consider a terrestrial ecosystem on land: The most essential ingredient for life on Earth is water and most of the water is not on the land. However, a rapid, long-distance system for transporting water from oceans to land is available through the atmosphere, much like the xylem system in a tree, which transports water long distances (relative to cell sizes) from the soil to the leaves high above the ground. This long-distance atmospheric transport system is driven unceasingly by the uneven, solar heating of the Earth. On the land, where many organisms reside, water moves much more slowly through the soil than through the atmosphere and this provides an essential storage reservoir that maintains the critical continuous flow of water to plants in the presence of fluctuations in the atmospheric supply. The rapid atmospheric transport system that is so critical to moving water from the ocean to the land also fluctuates wildly over the life span of most organisms. The combination of a low-storage, highly fluctuating system for rapid global transport in the atmosphere and a large-storage, stable, low-flow buffer in the soil redistributes water over the Earth to sustain life on land. This is a highly structured system like an organism. Furthermore, the living systems exert a strong influence on this larger system to enhance their functioning, just as we humans modify our environment to improve our lot.

Water is the most obvious component of the SPAC, and the interaction of this water with energy is the cornerstone of our understanding of how life interacts with its environment. Every schoolchild learns the water cycle: evaporation of liquid water from surfaces supplies water vapor to the atmosphere, where air currents lift this water to heights that cool it and cause condensation of the water vapor back to the liquid form of droplets, which fall back to the Earth to be recycled. The extraction of energy from the surface as a result of evaporation and the release of energy to the atmosphere by condensation represent the primary means by which the energy from the sun ultimately warms the atmosphere. This 'latent heat,' which is associated with the change of phase of water between liquid and vapor states, and not a temperature change, provides the key feedback between water and energy cycles that stabilizes the soil–plant–atmosphere system, permitting it to sustain life.

Materials such as water flow through the ecosystem seamlessly, moving from place to place, serving innumerable functions, and consuming or releasing energy to achieve a phase suitable to the medium that contains it. This is the SPAC; 'continuum' here does not refer to the material storage forms so apparent to our eyes such as plant leaves or roots, lakes, raindrops, the soil itself, or even the air we breathe; rather it refers to the unimpeded flow of many forms of material and energy throughout an elegant system

for sustaining life on land. Animals are as much a part of this system as plants, but in this article plants are our emphasis. This ceaseless flow of material and energy sustains life at every level; from the molecular interactions that sustain and mutate the genetic 'memory,' to cellular processes that make up whole organisms, to the global level of a living, breathing planet.

Living systems cycle matter through the SPAC and themselves to create and maintain their structure and extract energy for powering their life functions. All material necessary for sustaining life is cycled on all spatial levels from subcellular to global. The connectedness and functioning of the SPAC can be illustrated by reference to water, the most critical mass constituent to life on Earth. However, understanding the SPAC through water will necessarily involve some considerations of at least carbon and energy.

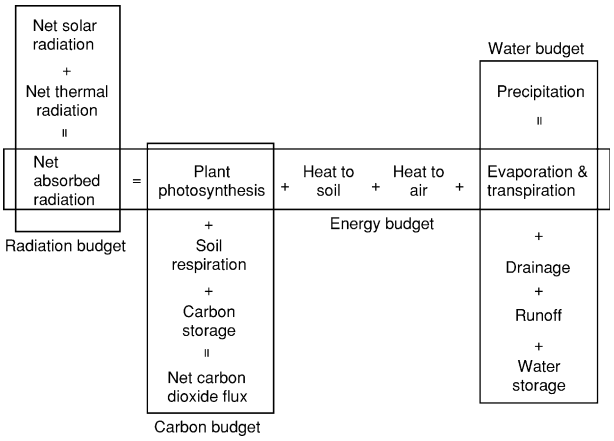


Figure 1 The interlocking energy and mass budgets for water, carbon, and energy in the soil–plant–atmosphere continuum. Altering individual components can cascade through the system and affect many seemingly unrelated factors.

Fundamental Principles

Two general principles are indispensable for studying the SPAC: (1) the conservation principles that take the form of mass and energy 'budgets,' and (2) the transport principles that relate the flow of some quantity to the difference or gradient of other quantities that influence or 'force' the flow and describe the 'state' of the exchange process.

The principles of conservation of mass and energy are the backbone of integrative studies of the SPAC. Because mass and energy can take many forms as they move throughout the SPAC, and even interact with each other, budgets are constructed to quantify the important stores and flows of important life-enabling constituents such as water, carbon, or energy. A budget is simply the application of the conservation principles (mass or energy) to a specific system that must be carefully defined. This system may be defined as a leaf, a community of plants, or even the entire Earth. Like balancing a bank account, a budget is a formal statement of the following: Incoming quantity minus outgoing quantity equals the increase in storage of the quantity in the system. In the case of water (W), carbon (C), and energy (E), the budgets become the following (Figure 1):

$$W_{IN} - W_{OUT} = \Delta W$$

$$C_{IN} - C_{OUT} = \Delta C$$

$$E_{IN} - E_{OUT} = \Delta E$$

The intertwined budgets of water, carbon and energy inflow, outflow and storage within the soil–plant–atmosphere system are shown in Figure 1. Understanding the SPAC through the role of water will involve simultaneous consideration of carbon and energy budgets on a spatial scale appropriate for living systems.

Consider the energy budget: Energy can exist in many forms as it flows through the soil–plant–atmosphere system, and the energy budget represents a way to keep track of the energy, no matter what its form. The rate of energy flow (joules per second = watts) per unit of area (meters squared) is a convenient way to describe the energy budget for a particular land area, because the size of the area does not matter and the emphasis can be on the energy exchanges. The radiant energy from the sun and surroundings can be absorbed by a leaf (this is referred to as the net radiation) and be transformed to the following: (1) sensible heat energy that is convected away from the leaf by moving air that is cooler than the leaf; (2) latent energy in molecules of water that are converted from liquid to vapor inside the leaf as they are transpired (evaporated) from the leaf; and (3) biochemical energy in the form of organic compounds such as sugars and starches created through photosynthesis and other physiological processes to sustain life. These components of the leaf energy budget are shown in Figure 1.

The transport principle simply states that the flow of some quantity such as water is equal to an appropriate 'driving force' divided by a resistance to transport exerted by the various media through which the quantity moves. The transport principle is more difficult to use than the conservation principle because the 'driving forces' and transport resistances depend on the characteristics of the system and mechanism of transport. For example, the diffusion of liquid water

depends on the water potential difference across a hydraulic transport resistance, whereas the transport of water vapor depends on the water vapor pressure difference across a diffusive transport resistance. In addition, movement of liquids or vapor by mass flow is 'driven' by pressure differences across transport resistances. In general, diffusion is an effective mode of transport over small distances, but mass flow can move materials and energy over longer distances. Heat transport is driven by temperature differences. The relevant thermal resistance depends on whether the heat is moving through a solid such as soil (conduction, which is analogous to diffusion), or through a moving fluid such as air (convection, which is analogous to mass flow and is a much more rapid transport mechanism than diffusion). Heat transport by radiation depends on still-different formulations. Thus much research on the SPAC relates to determining appropriate transport resistances and state variables for characterizing the flow of important quantities throughout the system. A complete understanding of this soil–plant–atmosphere system is a formidable task, but such an understanding will improve the ability of humans to make choices that will sustain a diverse and resilient ecosystem, upon which our survival depends. The fundamental concepts of energy and mass conservation and transport relations constitute the theoretical backbone of environmental biophysics (the study of the SPAC).

Water in the Soil–Plant–Atmosphere-Continuum

The smallest spatial scale that is usually associated with the SPAC is a field plot (square meters to thousands of square meters) from the bottom of a root zone (a few meters deep) to a few tens of meters above the top of the vegetation. Although individual studies may use smaller systems such as individual leaves or potted plants to reveal mechanisms, the scale of at least a small field is required to encompass most of the essential components of the natural SPAC and yet be accessible to direct measurement. Researchers generally use formulations nearly identical to those measured in field plots to represent idealized patches of the Earth's surface thousands of square kilometers in size; then they combine these patches to understand better the human influence on global climate processes. Typical time scales for small-scale studies vary from minutes to months or even years. On these spatial and temporal scales, researchers attempt to measure all the important flow and state variables, which are used to validate models that can be applied in systems that cannot be measured directly. After all, society cannot afford to measure everything everywhere, but will support measuring some things in many places and many things in a few places.

The transport and storage of water in the SPAC are shown in Figure 2. The flow of liquid water from the soil into the plant depends on the plant having a lower water potential than the soil; thus the soil water potential determines the availability of soil water for diffusion into the plant. The water potential is the amount of energy (joules) that can be obtained from moving a mass of water (kilograms) to a pool of pure water at atmospheric pressure and specified elevation; and in the soil and plant it is usually negative, indicating that energy must be expended to remove the water. As soils dry, the water potential decreases, making it more difficult for plants to obtain the water they need to grow. The soil water potential is related to the water content of the soil through the 'moisture-release curve,' which varies from soil to soil. Moisture-release curves have been measured for many different soil types. Typically soils can provide an amount of water to plants from storage that is equivalent to 5% (sands) to 20% (silt loams) of the plants' rooting depth. Thus for a plant rooted to a depth of 1 m, soils can provide approximately 50–200 mm of water to the plant. With evapotranspiration rates of 2–10 mm day^{-1}, soils can store enough water to sustain plants for a month or more without rainfall or irrigation.

Once water enters the plant through its roots, it moves through the plant vascular system to the leaves if the leaves have lower water potentials than the roots. Since the transport resistance to mass flow in the stem xylem is usually less than the transport resistance to diffusion through the root, and the flows through both root and xylem must be equal according to conservation principles, the water potential difference from the soil to the root xylem will usually be greater than the water potential difference from the root xylem to the leaf.

Water is lost from plants through pores in the leaf surfaces, called stomata. Evaporation from cells just below the stomatal pores draws water from the leaf cells, which in turn are hydraulically connected to a continuous water column down to the roots. As water is transpired through the stomata, additional water is drawn through the xylem to replace it, very much like suction applied to a straw in a soda. Of course, this evaporation that occurs just below the leaf surface consumes most of the Sun's energy that is absorbed by the leaf, because water vapor molecules carry away latent energy that was not contained in the liquid water that came up from the roots. While the flow of liquid water from the soil to the leaf is proportional to water potential differences, the flow of water

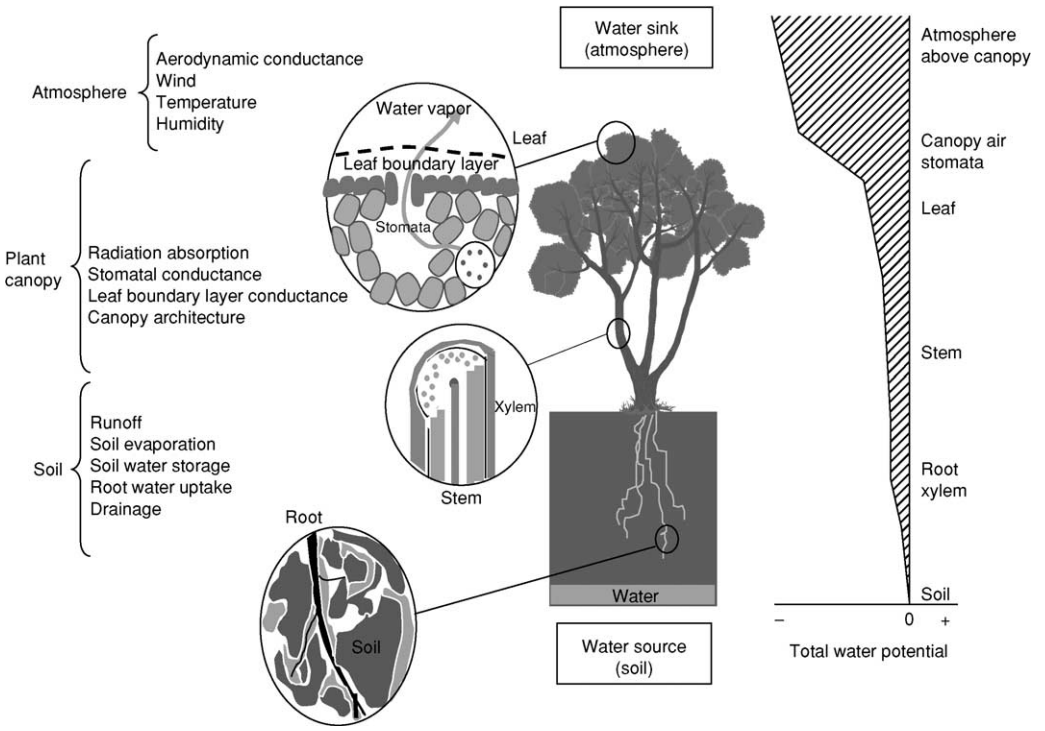


Figure 2 Water movement and storage in the soil–plant–atmosphere continuum. The graph on the right indicates the decrease in water potential that occurs as water moves from the soil to the atmosphere. On the left the three components of the soil–plant–atmosphere system are listed with the major factors for each that should be considered in a model of intermediate complexity.

vapor from inside the leaf to the leaf surface is proportional to water vapor pressure differences. At the interface, a near-equilibrium exists between the liquid water and the water vapor in contact with it. Although different processes are used to describe water movement in different parts of the system, the flow of water is continuous across these discontinuities of processes.

A very small amount of the water that passes from the soil to the atmosphere through the plant remains in the plant as storage as the plant grows (less than 5%). An even smaller amount of the hydrogen and oxygen in the water taken up from the soil remains in the plant in the form of organic molecules synthesized by the plant.

Water that exits the stomata must pass through a thin, still-air layer adjacent to the leaf, called the leaf boundary layer, and be transported through the canopy space by turbulent mixing along the path of decreasing water vapor pressure. Ultimately the water vapor exits the plant canopy, passes through the planetary boundary layer in the lower few thousand meters of the atmosphere, and is lifted high into the atmosphere, where it condenses into clouds and eventually falls back to the Earth as precipitation to repeat the cycle.

Small-Scale Observations and Model Building

In plant-environment studies, research derives and validates the transport principles on small, manageable systems in the laboratory or in field plots where direct measurements are possible. These principles are then applied at the ecosystem and global levels, where direct measurements are not possible, to obtain estimates of the important flow or state variables to quantify human impacts.

Consider the measurements that are necessary to understand the cycling of water in the SPAC. Precipitation falls on a field and either runs off the soil surface to streams or infiltrates into the soil itself (studied in the discipline of hydrology). The infiltrated water may be stored in the soil matrix (soil physics), evaporated from the soil surface (micrometeorology), transpired from plants (plant physiology), or drain below the plant root zone. Water that drains below the plant roots becomes part of the groundwater system (hydrogeology) and may reenter rivers (hydrology). Plants are sustained by the water stored in the soil, and the quantity of this storage depends on the type of soil. Nearly all the water taken up by plants is transpired to the atmosphere at a rate that

depends on solar radiation absorbed, weather variables, and the amount of vegetation. The amount of vegetation is estimated from the fraction of ground area covered by plants (vegetative cover). By keeping track of the daily water budget (Figure 1 and Table 1), the amount of water stored in the soil can be estimated, and in turn the influence of this stored water on direct evaporation from the soil and plant transpiration (together referred to as evapotranspiration) can be quantified. The maximum amount of water available to plants depends not only on soil properties, but also on plant rooting depth. Plant physiologists have measured rooting depths for the major classes of vegetation so that, given vegetation type and soil maps, the amount of water available to the plants for growth and transpiration can be estimated.

Table 1 contains an example of a 40-day-long water budget for a corn crop that begins about a month after emergence, during the period when the crop is growing rapidly. The incoming solar radiation column represents the maximum amount of evapotranspiration that could occur if all the solar radiation were used to evaporate water; of course some

Table 1 Simulation of corn growth and water use from 31 to 70 days after emergence, during a rapid growth phase. The solar radiation is in units of millimeters per day of evaporation that could occur if all solar radiation were to evaporate water. Vegetative cover increases as the crop grows and crop biomass increases. Soil evaporation tends to decrease as the soil surface dries out after rainfall and as vegetative cover increases; evapotranspiration (transpiration plus soil evaporation) exceeds rainfall so soil water storage is gradually depleted

Days after emergence	*Rainfall (mm)*	*Radiation (mm day^{-1})*	*Vegetative-cover fraction*	*Biomass (kg m^{-2})*	*Transpiration (mm day^{-1})*	*Soil evaporation (mm day^{-1})*	*Water stored in soil (mm)*
31	0	11.02	0.17	0.047	1.16	2.91	177
32	0	10.10	0.21	0.061	1.26	2.53	173
33	0	6.76	0.24	0.073	0.97	1.13	171
34	0	11.70	0.26	0.083	1.83	1.53	168
35	0	6.93	0.31	0.101	1.23	0.72	166
36	0	11.17	0.34	0.113	2.13	0.99	163
37	0	7.62	0.38	0.135	1.61	0.57	161
38	0	6.04	0.41	0.151	1.35	0.39	159
39	0	11.63	0.43	0.164	2.72	0.67	155
40	0	10.94	0.47	0.191	2.75	0.53	152
41	20	5.60	0.51	0.219	1.49	0.74	170
42	10	6.53	0.53	0.234	1.78	0.82	177
43	0	11.41	0.55	0.252	3.20	1.45	173
44	11	5.82	0.59	0.284	1.70	0.62	181
45	0	9.43	0.60	0.301	2.81	0.95	178
46	0	11.61	0.63	0.329	3.55	1.07	173
47	0	10.70	0.66	0.364	3.36	0.89	169
48	0	11.03	0.69	0.398	3.54	0.83	164
49	0	11.49	0.71	0.433	3.76	0.78	160
50	0	11.31	0.73	0.471	3.76	0.69	155
51	0	11.13	0.75	0.509	3.75	0.62	151
52	0	10.16	0.77	0.546	3.45	0.52	147
53	0	9.88	0.78	0.581	3.39	0.46	143
54	0	12.13	0.79	0.614	4.19	0.38	139
55	0	10.51	0.81	0.656	3.65	0.24	135
56	0	10.83	0.82	0.693	3.79	0.20	131
57	0	12.19	0.83	0.731	4.28	0.19	126
58	0	12.76	0.84	0.773	4.50	0.17	122
59	0	9.08	0.85	0.818	3.21	0.10	118
60	20	6.70	0.85	0.851	2.38	0.19	136
61	0	11.61	0.86	0.874	4.13	0.32	131
62	0	11.98	0.86	0.916	4.27	0.31	127
63	0	11.04	0.87	0.958	3.94	0.27	123
64	0	11.12	0.87	0.998	3.97	0.26	118
65	0	11.05	0.87	1.037	3.95	0.25	114
66	0	12.14	0.88	1.077	4.35	0.27	109
67	0	10.83	0.88	1.120	3.89	0.23	105
68	0	10.71	0.88	1.159	3.84	0.23	101
69	0	9.87	0.88	1.198	3.54	0.20	98
70	0	8.32	0.88	1.233	2.99	0.17	94

radiation is reflected so the actual evapotranspiration is the sum of the transpiration and direct evaporation from the soil, which is approximately one-third of the maximum possible evapotranspiration. The maximum amount of water that can be stored in the rooting zone of the soil is 200 mm for this soil, and clearly the corn crop is using water faster than the rainfall is replenishing it, so that at the end of 70 days of growth approximately half of the maximum soil water storage has been used up. At this point the water depletion in the soil will begin to cause the growth rate of the corn to decrease unless there is more rainfall. This buffering capacity of soil is crucial to growing crops in continental climates all over the world.

Measurements of water transport and storage are difficult and expensive to make, so they can only be done at a few sites under a few conditions. Models that capture the fundamental conservation and transport principles are therefore essential to applying results of limited SPAC studies to a wider range of conditions than could possibly be measured directly. A wide array (hundreds) of computer models are available for quantifying the role of water in the SPAC and the range of applicability of such models depends on having a minimal, faithful representation of critical factors in each component of the SPAC; namely in the soil, the plant, and the atmosphere. SPAC models can be divided into three classes: simple, intermediate, and complex.

Simple SPAC models are the easiest to create, generally emphasize the knowledge base of one discipline with a narrowly focused purpose, and are the most limited in applicability. For example, the Simple Inverse Yield Model, for precision agriculture, and the CERES Maize model, for crop yield and water-use prediction, neglect the effect of vegetation roughness and wind on evapotranspiration, reflecting their agricultural roots and limited applicability. There is also a simple micrometeorological model of evapotranspiration that omits soil processes, reflecting its atmospheric origins and limited applicability.

Intermediate SPAC models incorporate at least a minimal consideration of key soil, plant, and atmospheric components to provide wide applicability without undue complexity. These intermediate models incorporate atmospheric factors (radiation, wind, temperature, humidity, and precipitation), vegetation factors (plant type, vegetative-cover fraction and height, rooting depth), and soil properties (soil evaporation, root-zone water transport and storage). Examples of Intermediate SPAC models include the Two-Source Model, for remote sensing applications, and the ALEX (atmosphere–land exchange) model, for local site-specific studies.

Complex SPAC models incorporate most of the processes known to be important at small (leaf to field) scales, and can be applied to a wide range of cover types and climatic conditions. Because they are very detailed, they generally involve a large array of coefficients related to specific plant and soil types. For this reason, complex models are often not well-suited to regional scale applications, where these myriad of parameters need to be defined at all locations. The greatest value of complex SPAC models lies in improving our understanding of key processes and using this knowledge to derive intermediate models with the fewer parameters and wider applicability.

An example of this use of a complex SPAC model is the derivation of the Two-Source and ALEX models from the complex Cupid model. The Cupid model predicts plant–atmosphere exchanges of carbon, water, and energy in a very detailed mechanistic way, and it requires dozens of species-specific coefficients related to plant physiology. By studying the behavior of the Cupid model, the ALEX model could be created using only a few carefully chosen physiological coefficients and yet achieve a surprising level of generality compared with Cupid. Not only does ALEX link growth and water use in a way nearly as fundamental as Cupid, but it does so with a widely used 'light-use efficiency' coefficient that is already tabulated for most vegetation types around the world. The result of this process is a robust model suitable for continental-scale applications.

Complex SPAC models have also been used to study the relation between the architecture of vegetation canopies and their remotely sensed signatures from aircraft and satellites. Since remote sensing is the only way to characterize vegetative cover on the regional and continental scales, we must understand how to interpret remotely sensed land observations to accommodate the profound influence of vegetation density on energy and material exchanges in the SPAC. For example, in Table 1, the increased importance of transpiration (and extraction of soil water to great depths by plant roots) over evaporation from the soil (water extraction to only a few centimeters of soil depth) is apparent as the vegetation cover increases from 17% to almost 90% during the 40 days of growth.

Determining the level of complexity appropriate for an application depends on the objective of the endeavor; typically, the simplest model that provides adequate results is the model of choice. Our understanding of energy and water flows in the SPAC is sufficiently well developed to extend this knowledge to larger scales, where direct measurements are only rarely, if ever, possible and evaluate implications of human activities on ecosystem health.

Application to Regional and Global Scales

The impact of humans on the global ecosystem has become so profound that extending our understanding of biophysical processes from small fields to whole regions is becoming critical for our survival. The extension of process-level understanding of energy and material flows in the SPAC to larger scales imposes a severe constraint that is not present in plot-scale studies; namely, that the allowable input quantities are limited to state variables that can be monitored over large areas. Only a few state variables are routinely measured widely in networks across the world and thus can be considered generally available for monitoring the SPAC. For example, near-surface weather variables (air temperature, humidity, wind speed and direction, precipitation, and cloud cover), solar radiation, atmospheric carbon dioxide, soil maps, vegetation-type maps, river and stream monitoring, and satellite images are routinely obtained at many locations in many countries of the world. This means that these are the main quantities that can be assumed to be generally available everywhere (not always true, but a reasonable assumption). All other inputs required for SPAC models must be computed from fundamental principles or from quantitative relations derived from validation experiments at small scales. This pursuit has been exceedingly challenging, but systematic research over decades in numerous disciplines has identified many of the critical processes, and integrative computer modeling has synthesized much of this information so that monitoring water transport and storage on continental scales may now be possible.

Consider an example of a suite of models that has been assembled to explore systematically the energy and material flows in the SPAC over a range of scales from small fields to the continental USA (Figure 3).

At the heart of this suite of models is a pair of Intermediate SPAC models; namely ALEX for computing in the forward direction and the Two-Source Model for inverse computation using satellite observations as inputs. When the sky is cloudless for a few hours in the morning, operational satellites can be used to quantify vegetative cover and monitor surface temperature to estimate the surface flows of

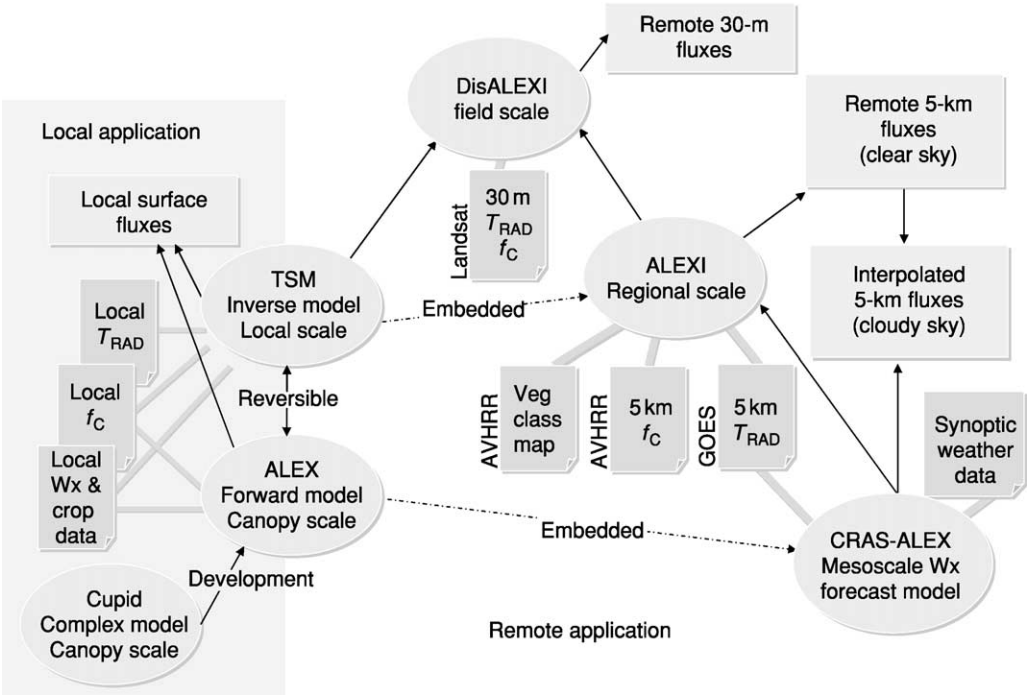


Figure 3 Suite of models for estimating evapotranspiration on large spatial scales (remote application) based on measurements from studies on small fields (local application). Cupid, the Atmosphere–Land Exchange model (ALEX), and the Two-Source Model (TSM) are used on the field scale and require inputs that can be directly measured in field studies. ALEX Inverse (ALEXI) disaggregation ALEXI (DisALEXI), and CRAS-ALEX (Cooperative Institute for Meteorological Satellite Studies (University of Wisconsin) weather forecast model) are used on the 5- to 10-km-grid scale for continental-sized applications, and they require inputs from satellites (Landsat, AVHRR, or GOES) or national weather measurement networks (synoptic weather data). The TSM is used at the local scale, embedded in ALEXI for use at the continental scale, and used in DisALEXI for remote estimates at the large field scale; likewise, ALEX is used at the local scale and also embedded in the weather forecast model for use on the continental scale. GOES, geosynchronous operational environmental satellite; AVHRR, advanced very-high-resolution radiometer satellite; T_{RAD} radiometric temperature; f_C, fraction of vegetative cover on the ground; Wx, weather.

energy and water at the 5- to 10-km scale and larger using the atmosphere–land exchange Inverse ALEXI model (Figure 3). ALEXI is a combination of the Two-Source Model and a meteorological model of the behavior of the lower 2000 m of the atmosphere (the planetary boundary layer). The ALEXI model can be used to update a mesoscale weather forecast model wherever a cloudless morning occurs. Likewise, the weather forecast model can be used to interpolate quantitites used in ALEXI whenever the satellite view is obscured by clouds. Mesoscale weather forecast models capture important feedbacks between land and atmospheric processes so that, as humans change the land characteristics, these models consider how these changes may affect subsequent weather. Meteorologists have refined these mesoscale models for decades and can now capture the transport of quantities from spatial scales of approximately 10 km and time scales of minutes to continental spatial scales over days, months, or even years of time.

The combination of the mesoscale model and ALEXI can yield time-continuous maps of the flow of water and energy in the SPAC at scales of 5–10 km and larger. How can the accuracy of such estimates be validated? Two methods have been used: (1) the use of research aircraft in intensive field experiments, and (2) the application of a disaggregation technique (DisALEXI), which uses data from a high-spatial-resolution satellite (such as Landsat, with a 30-m spatial resolution) to disaggregate 5- to 10-km ALEXI flux estimates to the spatial scale of in situ ground measurements (hundreds of meters) based on micrometeorological methods such as eddy covariance. Not only can the DisALEXI model assist with validation of the ALEXI model, but it can be used to update crop growth and water-use models in agriculture by providing occasional independent estimates of the flow of energy and water on the fine spatial scale of 30 m.

This process of using fundamental principles to infer global implications of plot-scale measurements reminds one of a statement Mark Twain once made: "There is something fascinating about science. One gets such wholesale returns of conjecture out of such a trifling investment of fact." Clearly the power that fundamental principles give scientists to extend the influence of humans is difficult for nonscientists to comprehend.

Institutional Issues

Studies of the SPAC are interdisciplinary (involving research in areas that are not claimed by existing disciplines) and multidisciplinary (requiring collaboration among scientists from numerous disciplines). For example, to understand the influence of vegetation on weather forecasts involves the following activities: meteorologists study the effect of the surface energy and water exchanges on the lower few thousand meters of the atmosphere called the planetary boundary layer; physiologists characterize the dependence of photosynthesis and stomatal processes on atmospheric and soil environmental factors; soil scientists measure the heat- and water-holding characteristics of the soil as well as carbon-exchange processes; micrometeorologists relate vegetation-canopy characteristics and atmospheric variables to the exchange of heat, mass, and momentum with the atmosphere; and hydrologists consider the runoff characteristics of the soil surface and the recharge of groundwater by percolation. The study of vegetative-canopy architecture, within-canopy processes, and the interaction between remote sensing and vegetative cover are key interdisciplinary areas that are not associated with any particular discipline.

Most scientific research is funded by government agencies and conducted within a disciplinary structure of research establishments or universities; therefore, competition among disciplines for funding and the limited perspective of individual researchers mitigate against developing an integrated view of the SPAC. Although much lip service has been paid to the value of interdisciplinary and multidisciplinary research, no sustaining structure in the research establishment is apparent for such integrative activities, and progress continues to be dominated by ad hoc efforts from committed individuals or teams with determination and vision. One strategy that has been most valuable in recent decades is the organization and execution of large, international, multidisciplinary, intensive field experiments, with broad objectives that appeal to scientists in many disciplines. The wide distribution of and open access to data from such experiments have been responsible for tremendous gains in understanding of the SPAC in recent years. Two examples of such experiments are the First ISLSCP (International Satellite Land Surface Climatology Project) Field Experiment (FIFE) in the grasslands of Kansas, USA, and the Boreal Ecosystem Atmosphere Study (BOREAS) in the Canadian boreal forest.

The time-honored reductionist perspective of 'changing a single variable and holding all other factors constant' is not possible in holistic studies of the profoundly interconnected SPAC. Instead the challenge of exploring important processes without the constraints of limiting paradigms and parochial disciplines must be confronted, a formidable challenge for environmental scientists. Efforts to synthesize knowledge from various components of the SPAC into a holistic view continue to be challenging

in a scientific world dominated by technological reductionism. The view of the Earth as a complex, self-organized system gains wider acceptance as our knowledge of the interconnectedness among components of the SPAC improves.

See also: **Energy Balance**; **Evapotranspiration**; **Water Cycle**; **Water-Use Efficiency**

Further Reading

Anderson MC, Norman JM, Diak GR, Kustas WP, and Mecikalski JR (1997) A two-source time-integrated model for estimating surface fluxes using thermal infrared remote sensing. *Remote Sensing Environment* 60: 195–216. [Description of the ALEXI model.]

Anderson MC, Norman JM, Meyers TP, and Diak GR (2000) An analytical model for estimating canopy transpiration and carbon assimilation fluxes based on canopy light-use efficiency. *Agricultural and Forest Meteorology* 101: 265–289. [Description of the ALEX model.]

Campbell GS and Norman JM (1998) *Introduction to Environmental Bio-physics*. New York: Springer-Verlag.

Choudhury BJ, Ahmed NU, Idso SB, Reginato RJ, and Daughtry CST (1994) Relations between evaporation coefficients and vegetation indices studied by model simulations. *Remote Sensing Environment* 50: 1–17.

Foley JA, Prentice IC, Ramankutty N, Levis S, Pollard D, Sitch S, and Haxeltine A (1996) An integrated biosphere model of land surface processes, terrestrial carbon balance, and vegetation dynamics. *Global Biogeochemical Cycles* 10: 603–628.

Gower ST, Kucharik CJ, and Norman JM (1999) Direct and indirect estimation of leaf area index, fAPAR, and net primary production of terrestrial eco-systems. *Remote Sensing Environment* 70: 29–51.

Hillel D (1998) *Environmental Soil Physics*. New York: Academic Press.

Jones CA and Kiniry JR (1986) *CERES-Maize: A Simulation Model of Maize Growth and Development*. College Station, TX: Texas A & M University Press.

Lovelock JE (1987) *Gaia: A New Look at Life on Earth*. New York: Oxford University Press.

Mecikalski JR, Diak GR, Anderson MC, and Norman JM (1999) Estimating fluxes on continental scales using remotely-sensed data in an atmospheric-land exchange model. *Journal of Applied Meteorology* 38: 1352–1369.

Morgan CS, Norman JM, and Lowery B (2003) Estimating plant-available water across a field with an inverse yield model. *Soil Science Society of America Journal* 67: 620–629.

Norman JM and Arkebauer TJ (1991) Predicting canopy light-use efficiency from leaf characteristics. In: Ritchie JT and Hanks RJ (eds) *Modeling Plant and Soil Systems*, pp. 125–143. Madison, WI: American Society of Agronomy.

Norman JM, Kustas WP, and Humes KS (1995) A two-source approach for estimating soil and vegetation energy fluxes from observations of directional radiometric surface temperature. *Agricultural and Forest Meteorology* 77: 263–293.

Sellers PJ, Hall FG, Asrar G, Strebel DE, and Murphy RE (1992) An overview of the first international satellite land surface climatology project (ISLSCP) field experiment (FIFE). *Journal of Geophysical Research* 97(18): 345–371.

Sellers PJ, Hall FG, Kelly RD *et al.* (1997) BOREAS in 1997: experiment overview, scientific results, and future directions. *Journal of Geophysical Research* 102 (D24): 28 731–28 769.

Stewart JB, Kustas WP, Humes KS, Nichols WD, Moran MS, and De Bruin AAR (1994) Sensible hear flux-radiometric surface temperature relationship for eight semiarid areas. *Journal of Applied Meteorology* 33: 1110–1117.

SOLUTE TRANSPORT

M C Sukop, Florida International University, Miami, FL, USA
E Perfect, University of Tennessee, Knoxville, TN, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

The movement of chemicals in soils and groundwater aquifers (solute transport) plays a central role in global biogeochemical cycling. Human activities have placed many new chemicals into the environment and greatly enhanced the local loading of wastes to the subsurface. Obvious applications of an understanding of solute transport in porous media include the efficacy and fate of agricultural chemicals (fertilizers and/or pesticides); the treatment and disposal of wastewaters and wastewater treatment products from human, animal, and industrial sources; and contamination resulting from landfills, mining operations, industrial activities, high-level radioactive waste disposal, etc. Given its great importance in the maintenance of safe water supplies, the nourishment

of plants, regulation of the fate of soil-applied chemicals, and many other critical hydrologic processes, solute transport has long been the focus of intense research in soil science, groundwater hydrology, and chemical engineering. Nevertheless, our *a priori* capability to predict solute transport remains limited.

Pores may contain water, soil gases, and/or introduced materials such as nonaqueous-phase liquids. Chemicals can be transported in any or all of these three fluid phases. The general aspects of transport are similar in each potential phase so the emphasis here and in much of the solute transport literature is on the transport of chemicals in aqueous solution. A solute is a chemical that is dissolved in a solvent to form a solution.

Three mechanisms operate to transport solutes in soils and aquifers. 'Advection' (or synonymously 'convection' in the solute transport literature) refers to the movement of chemicals with soil fluids and groundwater. Advection often dominates the solute transport process. Diffusion is the omnidirectional spreading of a chemical that results from thermal (Brownian) motion of the fluid and solute molecules. Dispersion results primarily from variations in fluid velocity that occur at all scales. Dispersion is anisotropic and enhanced in the overall direction of flow.

Breakthrough curves display the variation in solute concentration as a function of time or distance. These measurements and the advection–dispersion equation are the primary tools used in the evaluation of solute transport.

Solutes can interact strongly with soil surfaces by physical or chemical adsorption and their transport can be appreciably slowed in a process known as retardation. Repulsion or exclusion of solutes from soil surfaces can result in transport more rapid than that of the carrier fluid.

Transport Processes

Advection

In the most simplistic sense, dissolved solutes are carried along with moving soil fluids. Here, 'fluid' refers to either a gas or a liquid phase; simultaneous treatment of interacting gas and liquid phases requires coupling of the treatment described below. A fluid flowing through a porous medium at a flux q (volume of fluid per unit area of soil per unit time) with a solute concentration C (mass of solute per unit volume of fluid) carries with it a solute flux $\mathbf{J}_c = qC$ (mass of solute per unit area of soil per unit time). The mean velocity of the flow v (and the solute, assuming no losses, production, or interaction with the solid phase) can be computed from knowledge of the flux $\mathbf{q}$ and the volumetric fluid content θ via $\mathbf{v} = \mathbf{q}/\theta$. Under steady flow, $x = vt$ (where v is the magnitude of $\mathbf{v}$), and the distance x traveled by an average solute molecule in time t, or the time t needed for the solute to traverse a given distance x, can be calculated. The advective solute flux J_c can also be written in terms of the mean velocity, the concentration, and the fluid content: $J_c = v\theta C$. In some instances, for example when the solute source is continuous and the solute concentration is uniform in the flowing fluid, advection dominates the solute transport process and may be the only process that need be considered.

Diffusion

Diffusion is the spreading of a solute due to thermally induced Brownian motion. It is generally considered to be omnidirectional and driven by concentration gradients according to concentration-independent coefficients, though close study reveals that it is more complicated when solute concentrations are high. Under the typical assumptions, diffusion can be treated with Ficks first law, which gives the diffusive solute flux J_s as a function of the concentration gradient ∇C and the bulk diffusion coefficient D_b, i.e., $J_s = -D_b \nabla C$.

In soils, the diffusion paths are obstructed by solid particles and solutes are constrained to follow the tortuous, connected pore network. The connected network changes as the fluid content changes. The effective diffusion coefficient D_e must therefore be computed from the bulk diffusion coefficient and the soil water content and porosity. A commonly used empirical expression for the estimation of the effective diffusion coefficient is $D_e = D_b\theta^{10/3}/\phi^2$, where ϕ is the porosity.

Dispersion

Consideration of saturated water flow in perhaps the simplest possible pore geometry – a straight capillary tube with circular cross section – provides insight into the processes that take place in the more tortuous and convoluted pore networks that characterize natural porous media. The velocity of fluid in such a tube depends on the pressure drop between the ends of the tube (a distance L apart), the tube's radius r, and the dynamic viscosity η. In general, the velocity is not uniform across the tube's cross section. In the laminar flow regime, generally thought to apply to water movement in most soils, the velocity profile in a capillary tube is parabolic:

$$v(y) = 2v\left(1 - \frac{y^2}{r^2}\right) \quad [1]$$

where $v(y)$ is the velocity as a function of radial distance y from the tube center.

The mean velocity v can readily be determined from Stokes law in the form:

$$v = \frac{r^2 \Delta P}{8\eta L} \quad [2]$$

According to this model, the maximum velocity at the center of the tube ($y = 0$) is $2v$, and there is zero velocity at the tube's inner surface ($y = r$).

A solute introduced at concentration C_0 at the upstream end of the tube at time zero ($t = 0$) will be convected rapidly in the center of the tube and more slowly near the tube's surface (Figure 1). Taking integrated samples from various cross sections along the tube at a later time ($t > 0$) results in the collection of fluid containing no solute and fluid containing solute at the introduced concentration C_0 in various proportions. Dilution of the solute with solute-free fluid in the sample occurs according to the relative cross-sectional areas of the solute paraboloid and the solute-free fluid. Thus, the solute front is spread out or dispersed.

In addition to the dispersion introduced by velocity variations within individual pores, the variable sizes and path lengths of pores in a natural porous medium and mixing at pore intersections also contribute to dispersion. Buoyancy factors may cause fingering and attendant mixing of a dense solution into a less-dense one. If the dispersion arises from such velocity and path-length variations, and hydrodynamic mixing alone, it is referred to as 'hydrodynamic dispersion.' The impact of variable velocities, path lengths, and mixing accumulates through the entire hierarchy of scales pertinent to the particular porous medium under consideration. Unsaturated conditions found in most soils generally add additional complexity to flow fields within individual pores and increase the tortuosity. Hence, the degree of dispersion is usually greater under partially saturated conditions than in the fully saturated case.

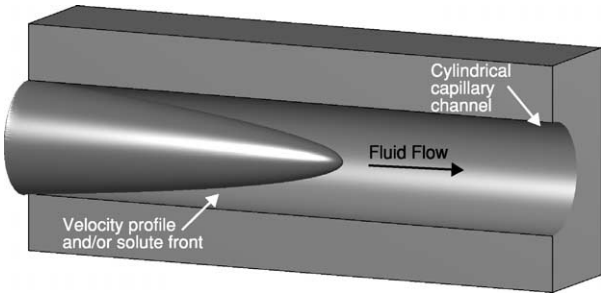


Figure 1 Schematic velocity profile and/or solute front position in an idealized cylindrical capillary tube. Flow from left to right.

Diffusion results in a spreading of solute fronts similar to that induced by dispersion. It is generally not possible to separate out the effects of diffusion and dispersion. They are typically considered to be additive with a lumped diffusion–dispersion coefficient D_d given by $D_d = D_h + D_e$, where D_h is the coefficient of hydrodynamic dispersion. The diffusive–dispersive solute flux $\boldsymbol{J}_d$ is then treated assuming a Ficks law-like formulation: $\boldsymbol{J}_d = -D_d \boldsymbol{\nabla} C$.

The diffusion–dispersion coefficient has been notoriously difficult to estimate *a priori* and is generally determined via inverse modeling. Excellent computer programs for its determination from transport measurements are available. However, this reliance on measurements of the actual transport process for parameter estimation is unfortunate, because the measurements are difficult and time-consuming, and can rarely be carried out at the scale of interest.

Inside a pore, the dimensionless Peclet number ($Pe \equiv vl/D_b$, with l a characteristic length) indicates the relative importance of diffusion and advection; large values of Pe indicate an advection-dominated process, while small values of Pe indicate the dominance of diffusion. Similarly, the dimensionless diffusion–dispersion coefficient $D^* \equiv D_d/D_b$ reflects the relative importance of hydrodynamic dispersion and diffusion. For porous media with well-defined characteristic lengths (i.e., bead diameter in packed beds of uniformly sized glass beads), D^* can be estimated from Pe. Different classes of behavior have been proposed based on the observed relationship between Pe and D^* (Figure 2). Class I is characterized by very slow flow and the dominance of diffusion. Class II is transitional, with approximately equal and additive hydrodynamic dispersion and diffusion. Hydrodynamic dispersion dominates the dispersion process in classes III and above. In class III, the role of diffusion is still nonnegligible, but it ceases to be significant in class IV. Finally, in class V, the velocity is so high that the flow of many fluids is turbulent.

Dimensionality

The dimensionality of the solute transport process is an important factor. In many soil science applications, for example the uniform application of a pesticide to the soil surface, the process can be treated as one-dimensional. However, in other instances two- or

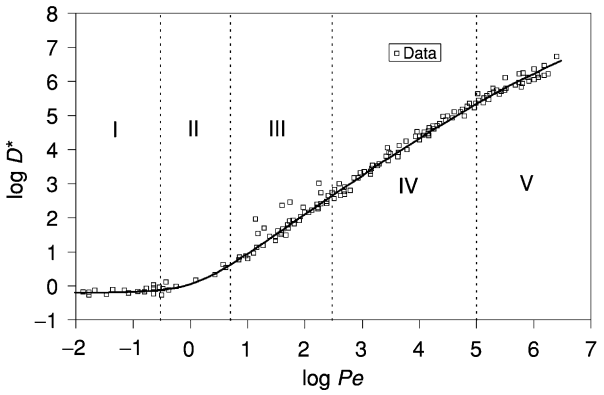


Figure 2 Empirical relationship between dimensionless dispersion coefficient (D^*) and Peclet number (Pe) with data for uniformly sized particle beds. Roman numerals indicate behavioral class (see text). (Adapted from Fried JJ and Combarnous MA (1971) Dispersion in porous media. *Advances in Hydroscience* 7: 169–282, with permission.)

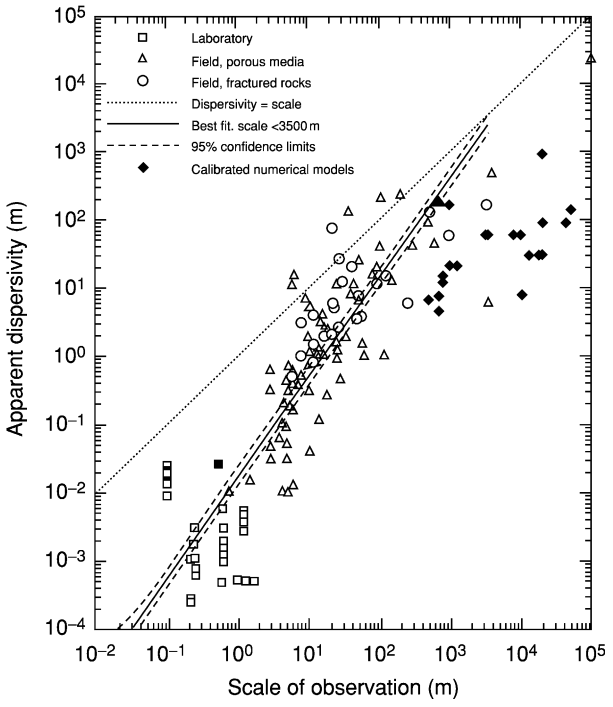


Figure 3 Apparent dispersivity versus scale for laboratory and field measurements. (Adapted from Neuman SP (1995) On advective transport in fractal permeability and velocity fields. *Water Resources Research* 31: 1455–1460, American Geophysical Union, with permission.)

three-dimensional aspects of the transport process cannot be neglected. The problem can be viewed in terms of the geometrical configurations of the source and sink of interest. In the case of uniform soil application, the source is considered planar and the sink may be a planar water table. In contrast, a chemical spill limited to a small area overlying a thick vadose zone may need to be considered a point source in a fully three-dimensional system where solute may be dispersed both horizontally and vertically.

Even in a flow field that is macroscopically one-dimensional, dispersion is an anisotropic process and is enhanced in the direction of flow relative to transverse directions. A fully three-dimensional treatment of dispersion therefore requires a tensorial representation. For simplicity, we will restrict our discussion of solute transport to one-dimensional flow, where the effects of dispersion are embodied in a single scalar, the longitudinal dispersion coefficient, and the related longitudinal dispersivity.

Longitudinal Dispersivity

The relationship between the one-dimensional diffusion–dispersion coefficient and mean pore-water velocity is frequently modeled with the empirical equation $D_d = m + \alpha v^n$, where m and n are fitting parameters, and α is the longitudinal dispersivity. In the hydrodynamic dispersion-dominated regime, the magnitude of m is negligible and n approaches unity. Under these conditions, the longitudinal dispersivity has dimensions of length and can be defined as $\alpha \equiv D_d/v$.

The dispersivity ideally would be a constant for a particular porous medium, and efforts have been made to relate it to other medium properties. In general, it has been observed that under saturated conditions the magnitude of the dispersivity depends upon the scale of the domain under consideration (Figure 3). The reasons for this scale dependency are unclear, but, at a macroscopic scale, the phenomenon may be related in part to spatial variations in the underlying saturated hydraulic conductivity field. New approaches for treating the relationship between α and scale have recently appeared and are described further below.

Another important influence on dispersivity under saturated conditions is soil structure, as influenced by differences in texture. In general, sandy soils have a narrower range of pore sizes and a smaller dispersivity than finer-grained soils. The combined effects of measurement scale and soil texture on α for soil columns (based on empirical relationships between dispersivity and soil hydraulic parameters and power law scaling of dispersivity with length) are shown in Figure 4.

Breakthrough Curves

Plotting the downstream changes in solute concentration against time at a fixed position or against distance from the source at a fixed time results in a

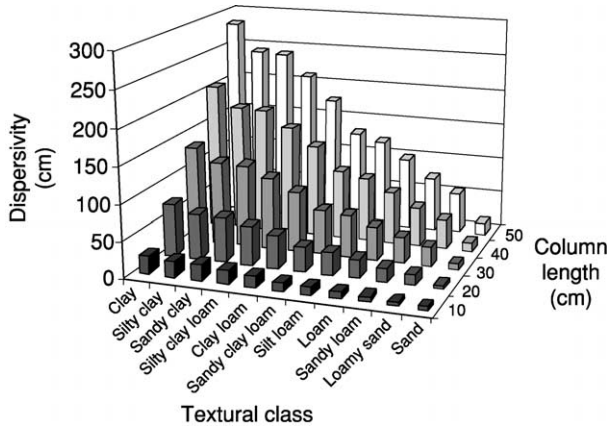


Figure 4 Approximate effects of measurement scale and soil texture on dispersivity for soil columns based on empirical relationships between dispersivity and soil hydraulic parameters and power-law scaling of dispersivity with length.

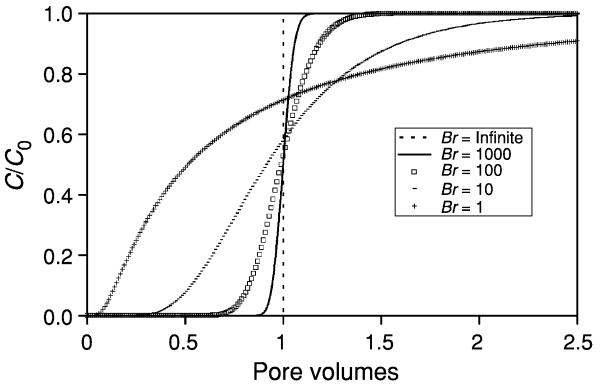


Figure 5 Effect of Brenner number (Br) on relative concentration (C/C_0) in column effluent for a step-change application of solute.

breakthrough curve. Breakthrough curves generated from a continuous source will be approximately sigmoidal (resembling a cumulative Gaussian distribution function) provided the Brenner (or column Peclet) number ($Br \equiv vL/D_d$, where L is the column length) is large enough. Complete breakthrough occurs when the effluent concentration (C) is equal to the influent concentration (C_0), giving a relative concentration $C/C_0 = 1$. If the solute is introduced as a short pulse, complete breakthrough may not be attained and the solute concentrations will eventually fall and return to background levels. In this case, a breakthrough curve will resemble a Gaussian density function for adequately large Brenner numbers. Small Brenner numbers (usually due to short columns or high dispersion coefficients) result in asymmetrical breakthrough curves.

It is common to plot breakthrough curves from soil columns as relative concentration against the number of pore volumes eluted from the column. Because the computation of eluted pore volume incorporates time and column length, it is a means of non-dimensionalizing the x-axis. The pore volume in a column of length L and cross-sectional area A is $LA\theta$. The volume of fluid eluted over time t is $At v\theta$. Thus, the number of pore volumes T eluted at any given time t is $(At v\theta)/(LA\theta)$ or $T = tv/L$.

This representation has instructive and conceptual benefits because, during piston-type flow where the solute front is not dispersed, it is fairly obvious that, after the input of one pore volume of solution containing solute at concentration C_0 into an initially solute-free column, the concentration of the effluent will immediately jump from $C = 0$ to $C = C_0$. This is shown in Figure 5 where it corresponds to the curve with infinite Brenner number. Decreasing the Brenner number (for example by increasing the dispersion coefficient or shortening the column) causes spreading of the solute elution. Some solute travels faster than the mean velocity and elutes before one pore volume of fluid has entered the soil, while some travels slower than the mean and elutes after one pore volume of fluid is injected. When the Brenner number is relatively high (Figure 5; $Br = 1000$ and $Br = 100$), the early- and late-arriving solute masses are approximately equal, and the breakthrough curve is approximately sigmoidal and well described by a cumulative Gaussian distribution. The value of C/C_0 at one pore volume on this curve is approximately 0.5; this is considered a hallmark of an 'ideal' solute transport process by many practitioners, with deviations calling for elaborate modifications to standard theory. Inspection of the gross asymmetry introduced by decreases in the Brenner number (Figure 5; $Br = 10$ and $Br = 1$) indicates that such deviations should be expected under many conditions. This effect is pronounced for solute pulses, and is perhaps somewhat counterintuitive in that smaller Brenner numbers (more dispersion) can actually lead to higher peak concentrations than larger ones (Figure 6; $Br = 1$ vs. $Br = 10$).

Adsorption and Exclusion

Adsorption of solutes on to soil solids generally causes them to move at a fraction of the velocity of the fluid. The most simplistic (and common) treatment of adsorption involves the assumption of a linear adsorption isotherm that relates the adsorbed concentration on soil solids (C_s, mass of solute per unit mass of soil) to the concentration in solution at equilibrium via $C_s = K_d C$, where K_d is the soil/water partitioning or

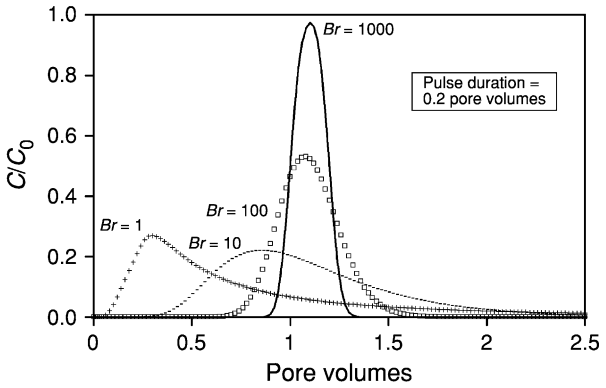


Figure 6 Effect of Brenner number on concentration in column effluent for a pulse application of solute. Pulse duration is 0.2 pore volumes.

distribution coefficient with dimensions $L^3 M^{-1}$. The linear isotherm is primarily applicable to low-solute concentrations and adsorption of a nonspecific physical nature.

K_d can be measured independently in the laboratory by mixing soil and solution and determining the equilibrium partitioning of the solute between the solid and fluid. In practice, however, the K_d of organic solutes is often estimated from the organic carbon partitioning coefficient $K_{oc}(L^3M^{-1})$ using $K_d = K_{oc} f_{oc}$, where $f_{oc}(MM^{-1})$ is the mass fraction of organic carbon in a soil. There are many tables of K_{oc} values available (e.g., Table 1) and f_{oc} is a commonly measured soil property. Solutes with low K_d values have a low affinity for soil surfaces and, for a conservative solute, $K_d = 0$. Other solutes can be strongly bound to soil and have very high K_{oc} and K_d values.

Exclusion, in which a solute is repelled from solid surfaces and hence concentrated in more rapidly flowing fluid away from pore walls, has also been observed. In particular, negatively charged anions are often excluded from negatively charged soil mineral surfaces. Thus, they may elute earlier from a column than the carrier fluid with which they were introduced.

Adsorption and exclusion phenomena are often incorporated into solute transport models through the use of a retardation factor, R, given by:

$$R = 1 + \frac{\rho_b K_d}{\theta} \qquad [3]$$

where ρ_b is the soil bulk density. For conservative solutes, $R = 1$, for sorbing solutes, $R > 1$, and, for excluded solutes, $R < 1$.

Figure 7 shows the effects of solute adsorption and exclusion relative to the breakthrough curve for a

Table 1 Organic carbon partitioning coefficients for nonionizable organic compounds

Compound	*Mean* K_{oc} *(l kg*$^{-1}$*)*
Acenaphthene	5028
Aldrin	48 686
Anthracene	24 362
Benz(*a*)anthracene	459 882
Benzene	66
Benzo(*a*)pyrene	1 166 733
Bis(2-chloroethyl)ether	76
Bis(2-ethylhexyl)phthalate	114 337
Bromoform	126
Butylbenzylphthalate	14 055
Carbon tetrachloride	158
Chlordane	51 798
Chlorobenzene	260
Chloroform	57
DDD	45 800
DDE	86 405
DDT	792 158
Dibenz(*a,h*)anthracene	2 029 435
1,2-Dichlorobenzene (*o*)	390
1,4-Dichlorobenzene (*p*)	687
1,1-Dichloroethane	54
1,2-Dichloroethane	44
1,1-Dichloroethylene	65
trans-1,2-Dichloroethylene	38
1,2-Dichloropropane	47
1,3-Dichloropropene	27
Dieldrin	25 604
Diethylphthalate	84
Di-*n*-butylphthalate	1580
Endosulfan	2040
Endrin	11 422
Ethylbenzene	207
Fluoranthene	49 433
Fluorene	8906
Heptachlor	10 070
Hexachlorobenzene	80 000
α-HCH (α-BHC)	1835
β-HCH (β-BHC)	2241
γ-HCH (Lindane)	1477
Methoxychlor	80 000
Methylbromide	9
Methylchloride	6
Methylene chloride	10
Naphthalene	1231
Nitrobenzene	141
Pentachlorobenzene	36 114
Pyrene	70 808
Styrene	912
1,1,2,2-Tetrachloroethane	79
Tetrachloroethylene	272
Toluene	145
Toxaphene	95 816
1,2,4-Trichlorobenzene	1783
1,1,1-Trichloroethane	139
1,1,2-Trichloroethane	77
Trichloroethylene	97
o-Xylene	241
m-Xylene	204
p-Xylene	313

Adapted from US Environmental Protection Agency, Soil Screening Guidance: Technical Background Document.

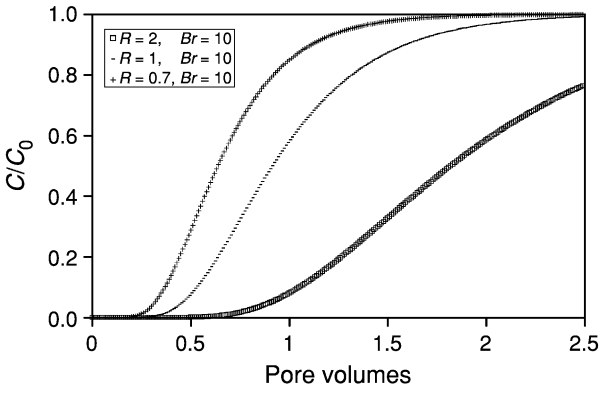


Figure 7 Effects of retardation (R) and exclusion on breakthrough curve at $Br = 10$.

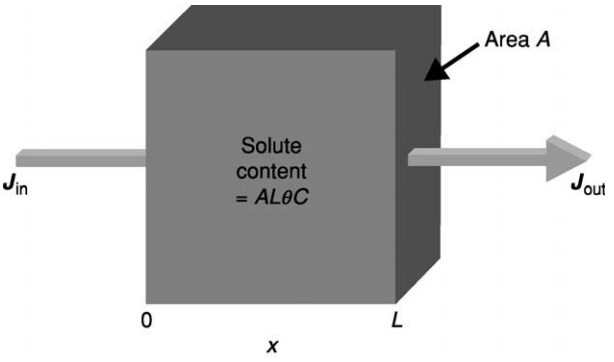


Figure 8 Elemental volume, with one-dimensional solute fluxes. J, flux; C, concentration; L, length; θ, volumetric fluid content.

conservative solute ($R = 1$, $Br = 10$, in Figure 7). The retarded curve ($R = 2$, $Br = 10$, in Figure 7) shows a considerable delay in solute elution. Note also that the degree of solute dispersion is enhanced, even though the dispersion coefficient itself is unchanged. Conversely, the curve representing anion exclusion ($R = 0.7$, $Br = 10$, in Figure 7) shows early elution of the solute and less dispersion of the solute front.

For a continuous source under piston flow conditions, the solute front advances without dispersion and the velocity of the front (v_s) relative to the mean velocity of the fluid is $v_s = v/R$. With moderate dispersion and adequate transport length, the front assumes a sigmoidal shape, and v_s represents the position of the front defined by $C/C_0 = 0.5$. For a pulse source, v_s characterizes the velocity of the solute center of mass. In some instances, these simple relations may be adequate for estimating the solute front position or peak arrival. However, dispersion can play a critical role in determining the peak concentration and its duration. Evidence suggests that brief exposures to high concentrations may be exceptionally detrimental to human health in some instances.

Advection–Dispersion Equation

The advection–dispersion equation (ADE, or convection–dispersion equation, CDE) has long been used as a model for solute transport. It can be derived from the flux terms described above (here taken as one-dimentional in the x direction) and a solute mass balance. The flux of solute due to convection is $J_c = v\theta C$, while the diffusive–dispersive flux is $J_d = -D_d \partial C/\partial x$. Combining these two fluxes, we have $J = J_c + J_d = v\theta C - D_d \partial C/\partial x$. Consider a volume of soil (Figure 8) and steady one-dimensional flow. If the inflow of solute mass at $x = 0$ and the outflow at $x = L$ are not equal, then mass must accumulate in or be lost from the soil volume. Assuming for the moment that the solute can only be stored in the fluid, then either the volume of fluid at a constant concentration – given by $AL\theta$ – must change, or the solute concentration (C) in a fixed volume of fluid must change. The dimensions of the volume AL are constant and, when the flow is steady, the fluid content θ is constant. Therefore, only the concentration can change. The time rate of change in solute mass inside the volume is $AL\theta\, dC/dt$.

The difference in the inflow and outflow of mass per unit time can be written as $(\text{mass}_{out} - \text{mass}_{in})/\Delta t$. The flux J is the mass per unit area and time, so an equivalent expression is $A(J_{x=L} - J_{x=0})$. The spatial rate of change in solute flux is $\partial J/\partial x$, and the difference between the outflow and inflow is $A[J_{x=0} - (J_{x=0} + L\partial J/\partial x)]$. Equating this difference to the change in the stored mass gives:

$$AL\theta \frac{\partial C}{\partial t} = A\left[J_{x=0} - \left(J_{x=0} + L\frac{\partial J}{\partial x}\right)\right] \qquad [4]$$

which readily simplifies to:

$$\theta \frac{\partial C}{\partial t} = -\frac{\partial J}{\partial x} \qquad [5]$$

Incorporating our earlier expression for the solute flux J, we have:

$$\theta \frac{\partial C}{\partial t} = -\frac{\partial \left(v\theta C - D_d \frac{\partial C}{\partial x}\right)}{dx} \qquad [6]$$

Finally, with D_d, v, and θ constant, the ADE is:

$$\frac{\partial C}{\partial t} = -v\frac{\partial C}{\partial x} + \frac{D_d}{\theta}\frac{\partial^2 C}{\partial x^2} \qquad [7]$$

Eqn [7] is often written with D_d/θ replaced by an apparent diffusion–dispersion coefficient D.

The ADE can readily be modified to incorporate adsorption by accommodating solute storage on the soil solids in addition to storage in the fluid phase. Thus, the storage term becomes $AL\theta C + AL\rho_b C_s$ or, in terms of the solute concentration in the fluid and assuming a linear isotherm, $AL(\theta C + \rho_b K_d C)$. Then Eqn [6] becomes:

$$(\theta + \rho_b K_d)\frac{\partial C}{\partial t} = -\frac{\partial\left(v\theta C - D_d\frac{\partial C}{\partial x}\right)}{\partial x} \qquad [8]$$

Note that this formulation also assumes that the adsorption equilibrium is attained instantaneously. Alternative derivations that remove this restriction are available. Removing the constants from the differentials and dividing by θ yields:

$$R\frac{\partial C}{\partial t} = -v\frac{\partial C}{\partial x} + \frac{D_d}{\theta}\frac{\partial^2 C}{\partial x^2} \qquad [9]$$

where R is the retardation factor given by Eqn [3].

Other isotherms, including the Langmuir (which introduces an adsorption maximum) and the Freundlich (which introduces a nonlinear concentration dependence), have been used to describe adsorption during solute transport. Realistic transport simulations of reactive species must often rely on considerably more complex, coupled reaction–transport models.

Many chemical species are susceptible to production or decay in the soil. Zero-order decay proceeds at a rate that is independent of the concentration of the substance where $\partial C/\partial t$ is a constant. First-order production and decay proceed at rates that depend on the concentration and follow $\partial C/\partial t = kC$, where k is the rate constant (T^{-1}). Such potential sinks and sources, as well as physical sinks (e.g., root uptake) can readily be incorporated into the ADE.

Estimates of the parameters in the ADE can be obtained by moment analysis of solute breakthrough curves. A number of analytical solutions for the ADE are also available. Depending on the concentration detection mode (column effluent or 'resident' pore water) the following solutions have been recommended for a step change in concentration from C_0 to C at time $t = 0$:

$$\text{Flux mode}: \frac{C}{C_0} = \frac{1}{2}\text{erfc}\left[\frac{Rx - vt}{2(DRt)^{\frac{1}{2}}}\right] + \frac{1}{2}\exp\left(\frac{vx}{D}\right)\text{erfc}\left[\frac{Rx + vt}{2(DRt)^{\frac{1}{2}}}\right] \qquad [10]$$

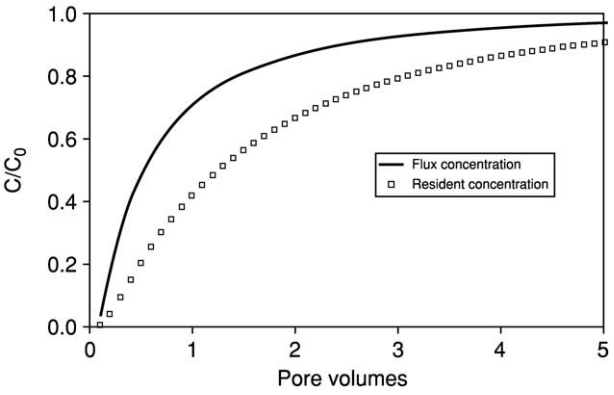


Figure 9 Effect of concentration detection mode on breakthrough curve at $Br = 1$.

$$\text{Resident mode}: \frac{C}{C_0} = \frac{1}{2}\text{erfc}\left[\frac{Rx - vt}{2(DRt)^{\frac{1}{2}}}\right] + \left(\frac{v^2 t}{\pi D r}\right)^{\frac{1}{2}}\exp\left(-\frac{(Rx - vt)^2}{4DRt}\right) - \frac{1}{2}\left(1 + \frac{vx}{D} + \frac{v^2 t}{DR}\right)\exp\left(\frac{vx}{D}\right)\text{erfc}\left[\frac{Rx + vt}{2(DRt)^{\frac{1}{2}}}\right] \qquad [11]$$

Here 'exp' and 'erfc' are the exponential function and complementary error function, respectively. The differences between the two solutions become appreciable in highly dispersive systems. Figure 9 shows the solutions at $Br = 1$. More complex and multidimensional formulations of the ADE generally require numerical solutions.

Two-Region Advection–Dispersion Equation

This equation, also referred to as the 'mobile–immobile water model' (MIM), is a special case of the ADE, in which the liquid phase is divided into two distinct (mobile and immobile) regions. Advective–dispersive solute flux occurs within the mobile water region, with solutes entering and leaving the immobile water region by diffusion. The ADE for solute transport within the mobile water region can then be written as:

$$\frac{\partial C_m}{\partial t} = D\frac{\partial^2 C_m}{\partial x^2} - v_m\frac{\partial C_m}{\partial x} - \left(\frac{\theta_{im}}{\theta_m}\right)\frac{\partial C_{im}}{\partial t} \qquad [12]$$

where the subscripts 'm' and 'im' refer to the mobile and immobile regions, respectively. The kinetic term

$\theta_{im}dC_{im}/dt$ is usually expressed in terms of a first-order rate constant and the concentration difference $C_m - C_{im}$. Eqn [12] can be solved analytically with or without making any assumptions about the spatial arrangement of solids and voids. In the first case, the mobile and immobile liquid regions are explicitly defined in terms of pore-space geometry. The majority of these models are for packed beds of uniform-sized, spherical soil aggregates; here θ_m is the macropore or inter-'aggregate' water content, and θ_{im} is the intra-'aggregate' water content. While it is possible to extend this approach to nonuniform and nonspherical aggregates, its predictive capability is limited. This is because the relationship between packed beds of sieved aggregates and the pore characteristics of undisturbed soil is far from clear. In the second case, the mobile–immobile water fraction is simply a fitting parameter.

Eqn [12] is commonly employed to parameterize asymmetric solute breakthrough curves with 'heavy tails.' Deviations from the Gaussian distribution associated with the ADE are often observed in breakthrough curves where more of the observed distribution is in the tails. Improvements in fit associated with the MIM as compared to the ADE can probably often be attributed to the presence of additional model parameters. Since it is rarely possible to measure the mobile and immobile water fractions independently, this parameter is usually estimated inversely. A two-site conceptualization of the soil, where a fraction of the adsorption sites are considered to have noninstantaneous adsorption, leads to an equivalent mathematical formulation.

Fractional Advection Dispersion and Continuous Time Random Walk Models

Recently, a fractional advection–dispersion equation (or FADE) and the continuous time random walk (CTRW) model have been proposed. These models appear to address the issues of scale-dependent dispersivity and heavy-tailed breakthrough curves well.

The FADE makes use of fractional calculus and is a generalization of the ADE that incorporates the family of Lévy distributions, of which the Gaussian distribution is a member. At the root of the FADE is the incorrect assumption in the derivation of the ADE above that the mass balance based on fluxes and fluid content applies in the infinitesimally small continuum limit. In fact it cannot, as the soil clearly divides into discrete pores and solids, and the averaged concepts of flux and fluid content lose meaning.

Instead of using a space-dependent dispersion coefficient to accommodate the apparent increase in D with scale, the FADE uses a constant dispersion coefficient and a non-Fickian fractional derivative for the diffusive–dispersive flux. The one-dimensional, symmetrical FADE can be written:

$$\frac{\partial C}{\partial t} = -v\frac{\partial C}{\partial x} + \frac{1}{2}D\left[\frac{\partial^{\gamma} C}{\partial x^{\gamma}} + \frac{\partial^{\gamma} C}{\partial(-x)^{\gamma}}\right] \qquad [13]$$

where γ is the fractional order of the derivative and D has dimensions $L^{\gamma}T$.

Figure 10 shows resident solute concentration data (measured with time-domain reflectometry probes) at different distances from the end of a single column. In this case, solute is being leached from the column. The data are fitted to both the ADE and the FADE. Clearly, the FADE is superior in its ability to fit the heavy tails of the breakthrough curves. Table 2 lists

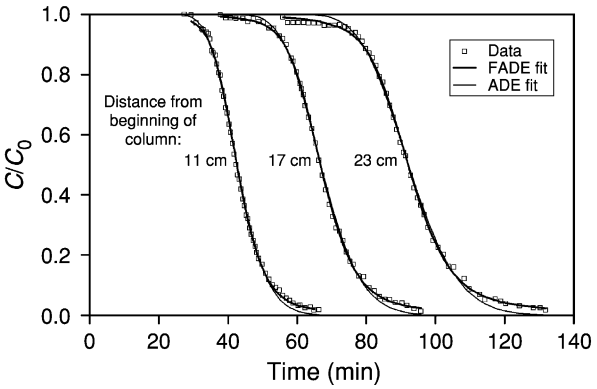


Figure 10 Comparison of fractional advection–dispersion equation (FADE) and ADE fits to data from an unsaturated soil column. (Adapted from Pachepsky Y, Benson DA, and Timlin D (2001) Transport of water and solutes in soils as in fractal porous media. In: Sparks D and Selim M (eds) *Physical and Chemical Processes of Water and Solute Transport/Retention in Soils*, pp. 51–77. Special Publication 56. Madison, WI: Soil Science Society of America.)

Table 2 Dispersion coefficients fitted to advection–dispersion equation (ADE) and fractional ADE (FADE) for breakthrough curves in Figure 10

	Dispersion coefficient	
Length (cm)	*ADE ($cm^2\,min^{-1}$)*	*FADE ($cm^{-1.6}\,min^{-1}$)*
11	0.035	0.030
17	0.038	0.029
23	0.042	0.028

Adapted from Pachepsky Y, Benson DA, and Timlin D (2001) Transport of water and solutes in soils as in fractal porous media. In: Sparks D and Selim M (eds) *Physical and Chemical Processes of Water and Solute Transport/Retention in Soils*, pp. 51–77. Special Publication 56. Madison, WI: Soil Science Society of America, with permission.

the dispersion coefficients for each of the models at each distance. Perhaps more important than the fitting of the tails is the fact that a single dispersion coefficient (approx. $0.03\,cm^{1.6}\,min^{-1}$) represents all of the curves in the case of the FADE, while the dispersion coefficients required to fit the ADE grow from approximately $0.035\,cm^2\,min^{-1}$ at 11 cm from the source to more than $0.04\,cm^2\,min^{-1}$ at 23 cm as the solute advances through the column. The FADE allows the parameter describing the changes with scale ($\gamma \approx 1.6$ in this case) to be separated from the transport parameter D, which ideally needs to be measured at one scale only.

Like the FADE, the CTRW model is founded on a conceptual picture in which solute particles move along various paths and encounter spatially varying velocities. The particle spatial transitions (direction and distance given by displacement vector **s**) in time t are represented by a joint probability density function $\Psi(\mathbf{s}, t)$. Estimation of this function is central to application of the CTRW model. The functional form $\Psi(\mathbf{s}, \sim t) \sim t^{-1-\beta}$ ($\beta > 0$) has been found to be of particular interest. β characterizes the nature and magnitude of the dispersive processes. $\beta \geq 2$ is reported to be equivalent to the ADE. β between 1 and 2 reflects moderate non-Fickian behaviour, while $0 < \beta < 1$ indicates strong 'anomalous' behaviour.

The ability of the CTRW model to fit observations is similar to that of the FADE for the results in Figure 10. Three parameters (β, C^*, and C_1) are involved. For breakthrough curves in time, available fitting routines return β, T^*, and r^*, which, for $1 < \beta < 2$, are related to C^* and C_1 as follows:

$$T^* = LC_1 \text{ and } r^* = C_1^*/\beta L^{-1+1/\beta}$$
$$C_1$$

where L is the distance from the source. Inverting these equations gives C^* and C_1, which can then be used to compute breakthrough curves at different locations. Thus C^* and C_1 should be constants for a 'stationary' porous medium (Table 3). Here there is systematic variation, but the precise nature of the medium and the significance of the magnitude of this variation are unknown.

Despite the promise of these models, their application in soil science remains almost completely unexplored currently.

Table 3 Parameters for continuous time random walk (CTRW) fitted to data in Figure 10. C^* and C_1 should be constant for solute transport in a 'stationary' porous medium

Length	*CTRW parameters*				
(cm)	T^*	β	r^*	C^*	C_1
11	44.1	1.68	0.11	1.92	4.01
17	68.5	1.72	0.092	2.09	4.03
23	94.2	1.73	0.084	2.23	4.10

List of Technical Nomenclature

α	Longitudinal dispersivity (L)
β	Characterizes the nature of magnitude of CTRW dispersive processes (–)
γ	Order of fractional derivative
ΔP	Pressure drop (MLT^{-2})
η	Dynamic viscosity (MLT^{-1})
θ	Volumetric fluid content (L^3L^{-3})
θ_{im}	Volumetric water content of the immobile region (L^3L^{-3})
θ_m	Volumetric water content of the mobile region (L^3L^{-3})
ρ_b	Soil bulk density (ML^{-3})
ϕ	Porosity (L^3L^{-3})
$\Psi(\mathbf{s}, t)$	Joint probability density function for CTRW (–)
A	Cross-sectional area of column and elemental volume (L^2)
Br	Brenner number
C	Solute concentration (ML^{-3})
C_0	Source solute concentration (ML^{-3})
C_{im}	Solute concentration in the immobile water region (ML^{-3})
C_m	Solute concentration in the mobile water region (ML^{-3})
C_s	Adsorbed concentration on soil (MM^{-1})
C^*, C_1, r^*, T^*	CTRW parameters
D	Apparent diffusion–dispersion coefficient (L^2T^{-1}) in ADE and ($L^\gamma T^{-1}$) in FADE
D^*	Dimensionless diffusion–dispersion coefficient (–)
D_b	Bulk diffusion coefficient (L^2T^{-1})
D_e	Effective diffusion coefficient (L^2T^{-1})
D_d	Lumped diffusion–dispersion coefficient (L^2T^{-1})
f_{oc}	Fraction of organic carbon in a soil

J	Solute flux ($ML^{-2}T^{-1}$)
J_c	Convective flux of solute ($ML^{-2}T^{-1}$)
J_d	Diffusive-dispersive flux of solute ($ML^{-2}T^{-1}$)
J_s	Diffusive flux of solute ($ML^{-2}T^{1}$)
k	First-order rate constant (T^{-1})
K_d	Soil and/or water partitioning or soil-water distribution coefficient (L^3M^{-1})
K_{oc}	Organic carbon partitioning coefficient (L^3M^{-1})
l	Characteristic length (L)
L	Length of column, elemental volume, and capillary tube (L)
Pe	Peclet number (–)
q	Fluid flux ($L^3L^{-2}T^{-1}$)
r	Capillary tube radius (L)
R	Retardation factor (–)
$\mathbf{s}$	Spatial transition vector in CTRW (L)
T	Pore volumes eluted (–)
t	Time (T)
v	Average fluid velocity (LT^{-1})
v_m	Average liquid velocity in the mobile region (LT^{-1})
v_{max}	Maximum velocity (LT^{-1})
v_s	Solute velocity (LT^{-1})
x	Spatial coordinate aligned with average fluid velocity (L)
y	Radial distance (L)

See also: **Cation Exchange**; **Chemical Equilibria**; **Colloid-Facilitated Sorption and Transport**; **Leaching Processes**; **Pesticides**; **Pollution:** Groundwater; **Remediation of Polluted Soils**; **Sorption-Desorption, Kinetics**; **Vadose Zone:** Hydrologic Processes; **Waste Disposal on Land:** Liquid; Municipal; **Water Potential**

Further Reading

Adler PM (1992) *Porous Media: Geometry and Transports*. Boston, MA: Butterworth-Heinemann.

Bear J (1972) *Dynamics of Flow in Porous Media*. New York: American Elsevier.

Bear J and Bachmat Y (1990) *Introduction to Modelling of Transport Phenomena in Porous Media*. Dodrecht, the Netherlands: Kluwer Academic.

Dullien FAL (1992) *Porous Media: Fluid Transport and Pore Structure*, 2nd edn. New York: Academic Press.

Elrick DE and Clothier BE (1990) *Solute Transport and leaching*. Agronomy Monograph 30, pp. 93–126. Madison, WI: American Society of Agronomy.

Fried JJ and Combarnous MA (1971) Dispersion in porous media. *Advances in Hydroscience* 7: 169–282.

Hillel D (1998) *Environmental Soil Physics*. New York: Academic Press.

Jury WA and Flühler H (1992) Transport of chemicals through soil: mechanisms, models, and field applications. *Advances in Agronomy* 47: 141–201.

Kutílek M and Nielsen DR (1994) *Soil Hydrology*. Reiskirchen, Germany: Catena-Verlag.

Leij FJ and Dane JH (1989) *A Review of Physical and Chemical Processes Pertaining to Solute Transport*. Agronomy and Soils Department Series No. 132, Alabama Agricultural Experiment Station. Auburn, AL: Auburn University Press.

Parker JC and van Genuchten MT (1984) *Determining Transport Parameters from Laboratory and Field Tracer Experiments*. Bulletin 84–3. Blacksburg, VA: Virginia Agricultural Experiment Station.

Sahimi M (1995) *Flow and Transport in Porous Media and Fractured Rock*. Weinheim, Germany: VCH-Verlag.

Selim HM and Sparks DL (eds) (2001) *Physical and Chemical Processes of Water and Solute Transport/Retention in Soils*, Special Publication 56. Madison, WI: Soil Science Society of America.

Toride N, Leij FJ, and van Genuchten MT (1995) *The CXTFIT Code for Estimating Transport Parameters From Laboratory or Field Tracer Experiments*. Research Report No. 137, US Salinity Laboratory. Riverside, CA: USDA–ARS.

van Genuchten MT and Wierenga PJ (1986) Solute dispersion coefficients and retardation factors. In: Klute A (ed.) *Methods of Soil Analysis*. 2nd edn, part 1, pp. 1025–1054. Madison, WI: American Society of Agronomy.

SORPTION

Contents
Metals
Organic Chemicals
Oxyanions

Metals

D L Sparks, University of Delaware, Newark, DE, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

The sorption (retention) of metals, including alkali (e.g., K), alkaline earth (e.g., Ca), and transition (e.g., Cd and Ni), on soil mineral and organic constituents is one of the most important processes in controlling the fate, transport, and bioavailability of metals in soil and water environments. This review discusses terminology for sorption processes, the role of surface functional groups on inorganic and organic soil components in sorption of metals, the types of surface complexes and products that metals can form in soils, characteristics of metal sorption in soils, and the importance of surface precipitation of metals such as Co, Ni, and Zn in sequestering metals in soils.

Terminology and Role of Sorption Processes in Soils

Adsorption, surface precipitation, and polymerization are all examples of sorption, a general term that is used when the retention mechanism at a surface is unknown. Adsorption can be defined as the accumulation of a substance or material at an interface between the solid surface and the bathing solution. Adsorption can include the removal of solute (a substance dissolved in a solvent) molecules from the solution, solvent (continuous phase of a solution, in which the solute is dissolved) from the solid surface, and attachment of the solute molecule to the surface. Adsorption does not include surface precipitation (the formation of a three-dimensional phase product on a surface), or polymerization (the formation of small multinuclear inorganic species such as dimers or trimers) processes. There are various metal sorption mechanisms that can occur at soil mineral surfaces involving both physical and chemical processes (Figure 1). It is useful to define a number of terms pertaining to retention (adsorption/sorption) of metal ions. The adsorbate is the material that accumulates at an interface, the solid surface on which the adsorbate accumulates is referred to as the adsorbent, and the ion in solution that has the potential of being adsorbed is the adsorptive. If the general term 'sorption' is used, the material that accumulates at the surface, the solid surface, and the ion in solution that can be sorbed are referred to as sorbate, sorbent, and sorptive, respectively.

Adsorption is one of the most important chemical processes in soils. It determines the quantity of plant nutrients, metals, pesticides, and other organic chemicals that are retained on soil surfaces and therefore is one of the primary processes that affects transport of nutrients and contaminants in soils. Adsorption also affects the electrostatic properties, e.g., coagulation and settling, of suspended particles and colloids. Both physical and chemical forces are involved in adsorption of solutes from solution. Physical forces include van der Waals forces (e.g., partitioning) and electrostatic outer-sphere complexes (e.g., ion exchange). Chemical forces result from short-range interactions that include inner-sphere complexation (which is defined below) involving a ligand exchange mechanism, covalent bonding, and hydrogen bonding.

Surface Functional Groups

Surface functional groups in soils play a significant role in adsorption processes. A surface functional group is a chemically reactive molecular unit attached at the boundary of a solid with the reactive groups of the unit exposed to the soil solution. Surface functional groups can be organic (e.g., carboxyl, carbonyl, phenolic) or inorganic molecular units. The major inorganic surface functional groups in soils are the siloxane surfaces associated with the plane of oxygen atoms bound to the silica tetrahedral layer of a phyllosilicate (clay mineral) and hydroxyl groups that are associated with the edges of inorganic minerals such as kaolinite, amorphous materials, and metal oxides, oxyhydroxides, and hydroxides.

Surface Complexation

When the interaction of a surface functional group with an ion or molecule present in the soil solution

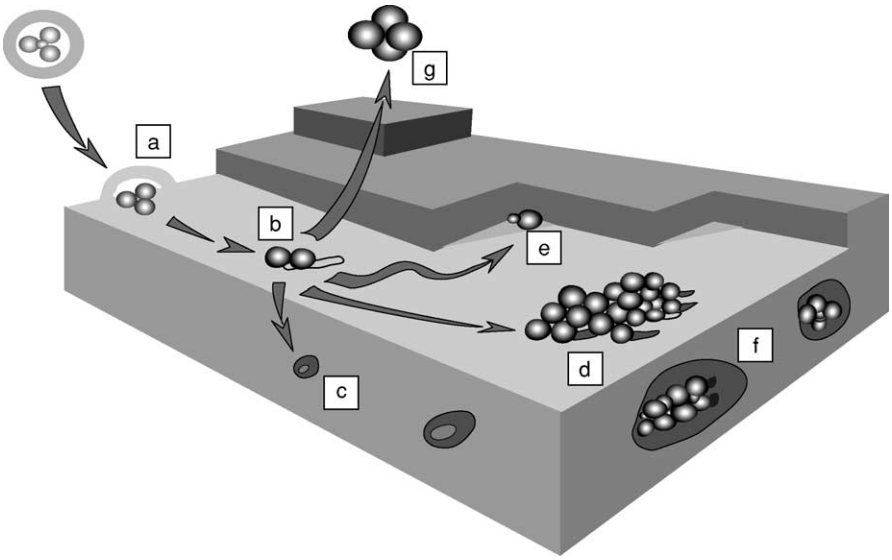


Figure 1 Various mechanisms of sorption of an ion at the mineral–water interface: (1) adsorption of an ion via formation of an outer-sphere complex (a); (2) loss of hydration water and formation of an inner-sphere complex (b); (3) lattice diffusion and isomorphic substitution within the mineral lattice (c); (4) and (5) rapid lateral diffusion and formation either of a surface polymer, (d), or adsorption on a ledge (which maximizes the number of bonds to the adatom) (e). Upon particle growth, surface polymers end up embedded in the lattice structure (f); finally, the adsorbed ion can diffuse back in solution, either as a result of dynamic equilibrium or as a product of surface redox reactions (g). Reproduced from Charlet L and Manceau A (1993) Structure, formation, and reactivity of hydrous oxide particles: insights from X-ray absorption spectroscopy. In: *Environmental Particles*, pp. 117–164. Boca Raton, FL: Lewis Publishers.

creates a stable molecular entity, it is called a surface complex. The overall reaction is referred to as surface complexation. There are two types of surface complexes that can form: outer-sphere and inner-sphere. Figure 2 shows surface complexes between metal cations and siloxane ditrigonal cavities on 2:1 clay minerals. Such complexes can also occur on the edges of clay minerals. If a water molecule is present between the surface functional group and the bound ion or molecule, the surface complex is termed outer-sphere. If there is not a water molecule present between the ion or molecule and the surface functional group to which it is bound, this is an inner-sphere complex. Inner-sphere complexes can be monodentate (metal is bonded to only one oxygen) and bidentate (metal is bonded to two oxygens) and mononuclear and binuclear.

Outer-sphere complexes involve electrostatic coulombic interactions and are thus weak compared to inner-sphere complexes in which the binding is covalent or ionic. Outer-sphere complexation is usually a rapid process that is reversible, and adsorption only occurs on surfaces that are of opposite charge to the adsorbate.

Inner-sphere complexation is usually slower than outer-sphere complexation; it is often irreversible. Inner-sphere complexation can increase, reduce, neutralize, or reverse the charge on the sorptive regardless of the original charge. Adsorption of ions via inner-sphere complexation can occur on a surface regardless of the original charge. It is important to remember that outer- and inner-sphere complexation can, and often does, occur simultaneously.

Ionic strength effects on metal sorption are often used as indirect evidence for whether an outer-sphere or inner-sphere complex forms. For example, if metal sorption is dependent on ionic strength (I), outer-sphere complexes are assumed to form, while if metal sorption is independent of I, inner-sphere complexes are assumed to be the predominant surface complex. However, such conclusions are based strictly on macroscopic investigations. To determine definitively the type of surface complex or product that metals form on soil surfaces, one must employ molecular-scale investigations.

Over the past decade or more, *in situ* (under natural conditions where water is present) molecular-scale analytical techniques such as X-ray absorption fine-structure spectroscopy (XAFS) and Fourier transform infrared spectroscopy (FTIR) have been employed to determine the type of metal surface complexes and products that form on soil surfaces. Based on these studies, one can conclude that alkali metals and alkaline earth metals primarily form outer-sphere complexes, while trace and heavy metals primarily form inner-sphere complexes. At higher

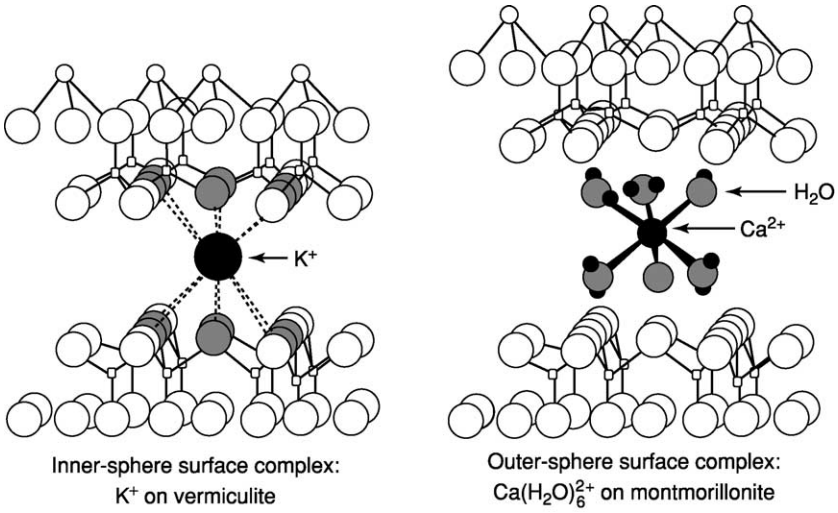


Figure 2 Examples of inner- and outer-sphere complexes formed between metal cations and siloxane ditrigonal cavities on 2:1 clay minerals. Reproduced with permission from Sposito G (1984) *The Surface Chemistry of Soils*. New York: Oxford University Press.

metal loadings and pHs, sorption of metals such as Co, Cr, Ni, and Zn on phyllosilicates and metal hydr(oxides) can result in the formation of surface precipitates. Thus, environmental factors such as pH, surface loading, ionic strength, type of sorbent, and time all affect the type of sorption complex or product. An example of this is shown for Pb sorption on montmorillonite over an I of 0.006–0.1 and a pH range of 4.48–6.77 (Table 1). Employing XAFS analysis, at a pH of 4.48 and an I of 0.006, outer-sphere complexation on basal planes in the interlayer regions of the montmorillonite predominated. At a pH of 6.77 and I of 0.1, inner-sphere complexation on edge sites of montmorillonite was most prominent, and at pH of 6.76, I of 0.006 and pH of 6.31, I of 0.1, both inner- and outer-sphere complexation occurred. These data are consistent with other findings that inner-sphere complexation is favored at higher pH and ionic strength. Clearly, there is a continuum of adsorption complexes that can exist in soils.

Sorption of Metal Cations

Sorption of metal cations is pH-dependent and is characterized by a narrow pH range where sorption increases to nearly 100%, traditionally known as an adsorption edge (Figure 3). The pH position of the adsorption edge for a particular metal cation is related to its hydrolysis or acid-base characteristics. In addition to pH, sorption of metals is dependent on sorptive concentration, surface coverage, and the type of the sorbent.

One can measure the relative affinity of a cation for a sorbent or the selectivity. The properties of the cation, sorbent, and the solvent all affect the selectivity. With monovalent alkali metal cations, electrostatic interactions predominate and the general order of selectivity is $Li^+ < Na^+ < K^+ < Rb^+ < Cs^+$. This order is related to the size of the hydrated radius. The ion in the above group with the smallest hydrated radius, Cs^+, can approach the surface the closest and be held the most tightly. However, on some hydrous oxides, the reverse order is often observed.

This has been particularly noted for some hydrous metal oxides. The reason for this selectivity is not well understood, but may be related to the effect of the solid on water that is present on the oxide surface or to variation in the solution matrix.

With divalent metal ions there is little consistency in the selectivity order. The type of surface and the pH both appear to have major effects on the selectivity sequence. Differences in the H^+/M^{n+} (where M^{n+} is a metal ion of valence n) could cause reversals in selectivity since the ion with the higher H^+/M^{n+} stoichiometry would be favored at higher pH. Divalent transition and heavy metal cations, both of which are often sorbed as inner-sphere complexes, are more strongly sorbed than alkaline earth cations.

Metal Surface Precipitation

As the amount of metal cation sorbed on a surface (surface coverage or loading which is affected by the pH at which sorption occurs) increases, sorption can proceed from mononuclear adsorption to surface precipitation. There are several thermodynamic reasons for surface precipitate formation: 1) the solid surface may lower the energy of nucleation by providing sterically similar sites; 2) the activity of the surface precipitate is <1; and 3) the solubility

Table 1 Effect of ionic strength (I) and pH on type of Pb adsorption complexes on montmorillonite

I ($mol\,l^{-1}$)	pH	Removal from solution (%)	Adsorbed Pb(II) ($mmol\,kg^{-1}$)	Primary adsorption complex[a]
0.1	6.77	86.7	171	Inner-sphere
0.1	6.31	71.2	140	Mixed
0.006	6.76	99.0	201	Mixed
0.006	6.40	98.5	200	Outer-sphere
0.006	5.83	98.0	199	Outer-sphere
0.006	4.48	96.8	197	Outer-sphere

[a]Based on results from X-ray absorption fine-structure spectroscopy data analysis.
Reproduced with permission from Strawn and Sparks (1999).

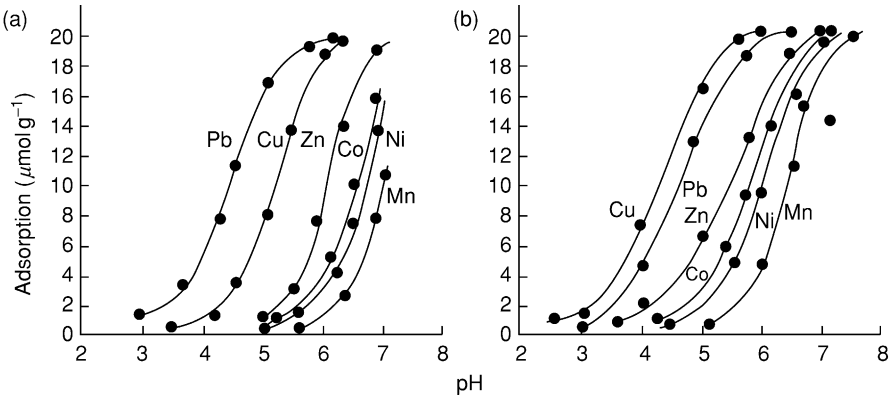


Figure 3 Sorption of a range of metals on (a) hematite and (b) goethite when they were added at a rate of 20 μmol g^{-1} of adsorbate. The values for the pK_1 for dissociation of the metals to give the monovalent MOH^+ ions are Pb, 7.71; Cu, 8; Zn, 8.96; Co, 9.65; Ni, 9.86; and Mn, 10.59. Reproduced with permission from McKenzie RM (1980) The adsorption of lead and other heavy metals on oxides of manganese and iron. *Australian Journal of Soil Research* 18: 61–73.

of the surface precipitate is lowered because the dielectric constant of the solution near the surface is less than that of the bulk solution. There are several types of surface precipitates. They can arise via polymeric metal complexes (dimers, trimers, etc.) that form on mineral surfaces and via the sorption of aqueous polymers. Homogeneous precipitates can form on a surface when the solution becomes saturated and the surface acts as a nucleation site. When adsorption attains monolayer coverage sorption continues on the newly created sites causing a precipitate on the surface. When the precipitate consists of chemical species derived from both the aqueous solution and dissolution of the mineral, it is referred to as a coprecipitate. The composition of the coprecipitate varies between that of the original solid and a pure precipitate of the sorbing metal. The ionic radius of the sorbing metal and sorbent ions must be similar for coprecipitates to form. Thus Co(II), Mn(II), Ni(II), and Zn(II) form coprecipitates on sorbents containing Al(III) and Si(IV) but not Pb(II) which is considerably larger (1.20 Å). Coprecipitate formation is most limited by the rate of mineral dissolution, rather than the lack of thermodynamic favorability. If the formation of a precipitate occurs under solution conditions that would, in the absence of a sorbent, be undersaturated with respect to any known solid phase, this is referred to as surface-induced precipitation.

Thus there is often a continuum between surface complexation (adsorption) and surface precipitation. This continuum depends on several factors: 1) ratio of the number of surface sites vs. the number of metal ions in solution; 2) the strength of the metal-oxide bond; and 3) the degree to which the bulk solution is undersaturated with respect to the metal hydroxide precipitate. At low surface coverages surface complexation (e.g., outer- and inner-sphere adsorption) tends to dominate. As surface average increases, nucleation occurs and results in the formation of distinct entities or aggregates on the surface. As surface loadings increase further, surface precipitation becomes the dominant mechanism (Figure 4).

Using an *in-situ* molecular-scale technique, X-ray absorption fine structure spectroscopy (XAFS), it has been shown by a number of scientists that multinuclear metal hydroxide complexes and surface precipitates of Co(II), Cr(III), Cu(II), Ni(II), Pb(II), and Zn(II) can form on metal oxides, phyllosilicates, soil clays, and soils. These metal hydroxide phases occur

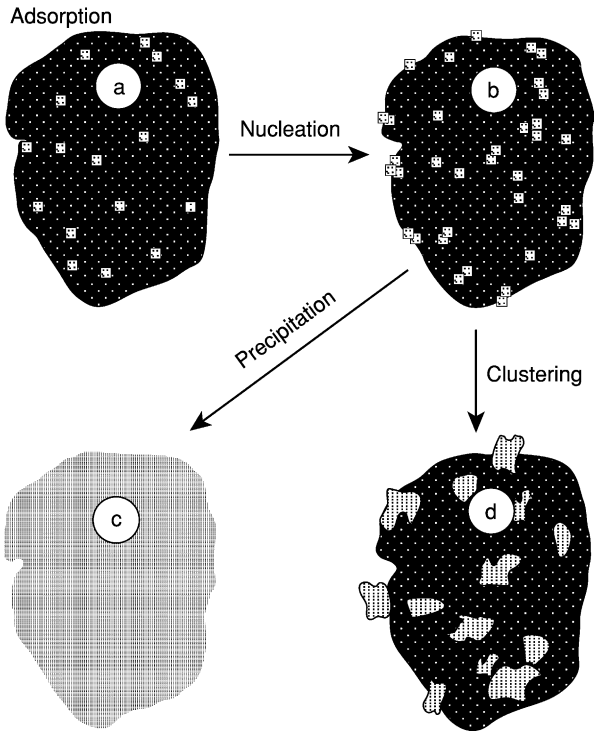


Figure 4 An illustration of metal ion sorption reactions on (hydr)oxide. (a) At low surface coverage, isolated site binding (adsorption) is the dominant sorption mechanism; (b) with increased metal loading. M-hydroxide nucleation begins; (c) further increases in metal loadings results in surface precipitation or (d) surface clusters. Reproduced with permission from Fendorf SE (1992) *Oxidation and Sorption Mechanisms of Hydrolysable Metal Ions on Oxide Surfaces*. PhD dissertation. Newark, DE: University of Delaware.

at metal loadings below a theoretical monolayer coverage and in a pH range well below the pH where the formation of metal hydroxide precipitates would be expected according to the thermodynamic solubility product.

Sorption of Co, Ni, and Zn, on an array of phyllosilicates and Al-oxide, can result in formation of mixed metal-Al hydroxide surface precipitates which appear to be coprecipitates. The precipitate phase shares structural features common to the hydrotalcite group of minerals and the layered double hydroxides (LDH) observed in catalyst synthesis. The LDH structure is built of stacked sheets of edge-sharing metal octahedra containing divalent and trivalent metal ions separated by anions between the interlayer spaces. The general structural formula can be expressed as $[Me^{2+}_{1-x}\,Me^{3+}_{x}(OH)_2]^{x+}\cdot(x/n)A^{-n}\cdot -mH_2O$, where for example, Me^{2+} could be Mg(II), Ni(II), Co(II), Zn(II), Mn(II), and Fe(II) and Me^{3+} is Al(III), Fe(III), and Cr(III). The LDH structure exhibits a net positive charge x per formula unit which is balanced by an equal negative charge from interlayer anions A^{n-} such as Cl^-, Br^-, I^-, NO_3^-, OH^-, ClO_4^-, and CO_3^{2-}; water molecules occupy the remaining interlayer space. The minerals takovite, $Ni_6\,Al_2(OH)_{16}\,CO_3\cdot H_2O$ and hydrotalcite, $Mg_6\,Al_2(OH)_{16}\,CO_3\cdot H_2O$ are among the most common natural mixed-cation hydroxide compounds containing Al.

The mechanism for the formation of metal hydroxide surface precipitates is not clearly understood. It is clear that the type of metal ion determines whether metal hydroxide surface precipitates form, and the type of surface precipitate formed, i.e., metal hydroxide or mixed metal hydroxide, is dependent on the sorbent type. The precipitation could be explained by the combination of several processes. First, the electric field of the mineral surface attracts metal ions, e.g., Ni, through adsorption, leading to a local supersaturation at the mineral–water interface. Second, the solid phase may act as a nucleation center for polyhydroxy species and catalyze the precipitation process. Third, the physical properties of water molecules adsorbed at the mineral surface are different from free water, causing a lower solubility of metal hydroxides at the mineral–water interface. With time Al, which is released by weathering of the mineral surface, slowly diffuses into the octahedral layer of the mineral and partially replaces the metal (e.g., Ni) in the octahedral sites. A Ni-Al (LDH) is formed, which is thermodynamically favored over α-Ni hydroxide.

The formation of metal hydroxide surface precipitates appears to be an important way to sequester metals. As the surface precipitates age, metal release is greatly reduced. Thus, the metals are less prone to leaching and being taken up by plants. This is due to silication of the interlayer of the LDH phases creating precursor phyllosilicate surface precipitates.

See also: **Colloid-Facilitated Sorption and Transport**; **Heavy Metals**; **Metal Oxides**; **Metals and Metalloids, Transformation by Microorganisms**

Further Reading

Barrow NJ (1985) Reactions of anions and cations with variable-charged soils. *Advances in Agronomy* 38: 183–230.

Charlet L and Manceau A (1993) Structure, formation, and reactivity of hydrous oxide particles: Insights from X-ray absorption spectroscopy. In: *Environmental Particles*, pp. 117–164. Boca Raton, FL: Lewis Publishers.

Fendorf SE (1992) *Oxidation and Sorption Mechanisms of Hydrolysable Metal Ions on Oxide Surfaces*. Ph.D. Dissertation. Newark, DE: University of Delaware.

McBride MB (1991) Processes of heavy and transition metal sorption by soil minerals. In: *Interactions at the Soil Colloid–Soil Solution Interface,* vol. 190, pp. 149–176. Dordrecht: Kluwer Academic Publishers.

McKenzie RM (1980) The adsorption of lead and other heavy metals on oxides of manganese and iron. *Australian Journal of Soil Research* 18: 61–73.

Scheidegger AM, Lamble GM, and Sparks DL (1997) Spectroscopic evidence for the formation of mixed-cation, hydroxide phases upon metal sorption on clays and aluminum oxides. *Journal of Colloid Interface Science* 186: 118–128.

Scheidegger AM and Sparks DL (1996) A critical assessment of sorption-desorption mechanisms at the soil mineral/water interface. *Soil Science* 161: 813–831.

Schindler PW (1981) Surface complexes at oxide-water interfaces. In: *Adsorption of Inorganics at Solid-Liquid Interfaces*, pp. 1–49. Ann Arbor, MI: Ann Arbor Science.

Sparks DL (2002) *Environmental Soil Chemistry*, 2nd Ed. San Diego, CA: Academic Press.

Sposito G (1984) *The Surface Chemistry of Soils*. New York: Oxford University Press.

Sposito G (1989) *The Chemistry of Soils*. New York: Oxford University Press.

Strawn DG and Sparks DL (1999) The use of XAFS to distinguish between inner- and outer-sphere lead adsorption complexes on montmorillonite. *Journal of Colloid Interface Science* 216: 257–269.

Strawn DG and Sparks DL (1999) Sorption kinetics of trace elements in soils and soil materials. In: Selim HM and Iskandar I (eds) *Fate and Transport of Heavy Metals in the Vadose Zone*, pp. 1–28. Chelsea, MI: Lewis Publishers.

Stumm W (1992) *Chemistry of the Solid–Water Interface*. New York: John Wiley & Sons.

Towle SN, Bargar JR, Brown GE Jr, and Parks GA (1997) Surface precipitation of Co(II) (aq) on Al_2O_3. *Journal of Colloid Interface Science* 187: 62–82.

Organic Chemicals

B Xing, University of Massachusetts, Amherst, MA, USA
J J Pignatello, Connecticut Agricultural Experimental Station, New Haven, CT, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Organic chemicals enter soils through the use of pesticides, application of sewage sludge, or the disposal or accidental release of petrochemicals or industrial wastes. Some have negative effects on the environment and public health. A key process in soils regulating the transport, bioavailability, and reactivity of organic compounds is sorption. 'Sorption' and the opposite term, 'desorption,' refer to the process of molecular exchange between the gas or solution phase and an immobile phase (the sorbent). The sorbed component is sometimes referred to as the sorbate (American) or sorptive (British).

Sorption may be categorized by the type of intermolecular attractive forces between sorbate and sorbent. Physisorption and ion exchange sorption involve relatively weak intermolecular forces, while chemisorption involves strong bonding of a covalent nature. Physisorption is, by far, the most common type encountered in the context of soil and environmental sciences.

The reactive mineral components in soils are silicate clays and oxyhydroxide minerals of Si, Fe, Al, Mg, Mn, etc. Soil organic matter (SOM) consists mainly of polymeric humic substances derived from decomposition of biological materials such as lignin and cutin. Other substances within the scope of SOM are carbonaceous materials such as soot and charcoal. SOM is usually the predominant sorbent unless the SOM content is extremely low – some estimates suggest less than 0.01% of organic carbon by weight – or unless the water content is low. Therefore, the focus here is on sorption by SOM. Emphasis is placed on sorption of nonionic compounds in water-saturated systems, because the majority of organic compounds, including pesticides, are uncharged within the range of pH usually encountered in soils (5.5–7.5), and because water is abundant, especially in the subsurface.

Physisorption of Nonionic Compounds

General Considerations

Forces involved Physisorption involves weak intermolecular forces such as London–van der Waals forces and hydrogen bonding. These forces act simultaneously in combinations permitted by the structures of the interacting species. London–van der Waals forces include dipole–dipole, dipole–induced dipole (attraction of a permanent dipole with the dipole it induces in its neighbor), and induced dipole–induced dipole (mutual attraction of momentary dipoles produced by the synchronization of electronic motion in interacting neighbors). Additionally, the interaction of a charge with a neutral molecule may involve charge–dipole and charge–induced-dipole forces. The strengths of London–van der Waals interactions depend on the molecular area of contact, separation distance (to the inverse sixth power), and applicable molecular properties of the interacting species such as dipole moment, ionization potential, and polarizability.

The most common type of hydrogen bond is between an acidic proton and a lone pair of electrons (i.e., $XH\cdots:Y$) which are located on O, N, or S atoms. This bond is highly oriented (XHY angle, $180^\circ \pm 15^\circ$), ranges in enthalpy from 12 to 35 kJ mol^{-1}, and involves a combination of dipole–dipole attraction and molecular orbital overlap. Other much weaker hydrogen bonds are known, such as the one formed between an acidic proton and the π system of an aromatic ring, but whether they are important in soil systems in competition with hydrogen bonds from water molecules is unclear.

Categories Physisorption is often broadly classified as absorption or adsorption. Absorption (often called partitioning) occurs when molecules penetrate the solid–fluid interface and are allowed to roam the internal molecular or atomic matrix of the sorbent. SOM is the only natural component of soils that is penetrable in this way, but polluted soils may contain additional absorptive phases such as solvents, oils, tars, etc.

Adsorption is a term used in the phenomenological sense to imply association at a solid–fluid interface. Mechanistically, however, adsorption may encompass several qualitatively different processes, including binding of molecules independently to discrete surface sites, partitioning of molecules into an ordered hydration phase near the surface, condensation of the sorbate into a liquid-like or solid-like state in small pores, and adsorption on the surfaces of water films associated with the particles.

Thermodynamic considerations Net sorption from water (eqn [1e]) can be written as a series of steps in a thermodynamic cycle:

$$\{x\} \rightleftharpoons \{\ldots\} + x_g \quad [1a]$$

$$\{\ldots\} \rightleftharpoons (\ ) \quad [1b]$$

$$\mathbf{M} \rightleftharpoons \mathbf{M}' \quad [1c]$$

$$x_g + \mathbf{M}' \rightleftharpoons x\mathbf{M}' \quad [1d]$$

$$\{x\} + \mathbf{M} \rightleftharpoons (\ ) + x\mathbf{M}' \quad [1e]$$

Equation [1a] represents removal of solute x from water to the gas phase, where {} symbolizes the solvation shell around x. Equation [1b] represents collapse of the solvation shell into the cavity vacated by the solute and reconfiguration of solvation-shell water molecules to the bulk water state symbolized by ( ). Equation [1c] is reorganization of the sorbent matrix $\mathbf{M}$ to some altered state $\mathbf{M}'$ that may be necessary to accommodate the sorbate; and eqn [1d] represents formation of bonds between sorbate and the altered sorbent.

Matrix reorganization may include displacement of water or other sorbed molecules into solution, or structural changes in the sorbent such as swelling and pore flexing in response to the adsorbate. Little is known about matrix changes and their effects on sorption thermodynamics.

When sorption is from the vapor state, eqns [1a] and [1b] do not apply, and net sorption is the sum of eqns [1c] and [1d]:

$$x_g + \mathbf{M} = x\mathbf{M}' \quad [1f]$$

The curve defining the equilibrium relationship between sorbed-phase concentration S and fluid-phase concentration C at constant temperature is the sorption isotherm. Various isotherm models have been used depending on the mechanism(s) involved.

It is convenient to define the sorption distribution ratio, K_d, as:

$$K_d = \frac{S}{C} \quad [2]$$

Fluid concentration is expressed in units of pressure or in moles per unit volume. The units of S depend on convention and the implied mechanism being invoked. Ideally, adsorption processes should be based on surface area (or pore volume if condensation is considered) and absorption on solid volume. However, sorbent dry weight is commonly used as a surrogate for these properties, since they are often not

accurately known. K_d is concentration-dependent unless the isotherm is linear.

K_d is related to the Gibbs free energy of sorption, ΔG^0_{sorp} by:

$$K_d = \exp\left(-\frac{\Delta G^0_{sorp}}{RT}\right) \quad [3]$$

where R is the gas constant and T is temperature. ΔG^0_{sorp} – which is concentration-dependent if K_d is concentration-dependent – encompasses all enthalpy and entropy changes that occur during sorption. This means that the position of equilibrium depends on (1) the difference in the strengths between all sorbate–sorbent and all solute–water interactions, and (2) the entropy changes associated with eqns [1a–1d]. There are virtually no examples where the contributions of these factors have been fully delineated.

Parts of molecules that carry no charge, are not strongly polarized, and lack functional groups that can form strong hydrogen bonds with water are said to be hydrophobic. The hydrophobic effect operates on such molecules when dissolved in aqueous solution and is due to entropy and enthalpy costs associated with the ordered alignment of water molecules near the hydrophobic regions of the solute (apparently, with –OH groups pointed toward the organic), and the consequent disruption of water–water hydrogen bonds. The hydrophobic effect reduces a compound's water solubility and thus contributes to the overall driving force for its sorption from water. For a series of compounds of related structure (e.g., aromatic hydrocarbons), K_d is found to correlate with the molecular hydrophobic surface area or with the *n*-octanol–water partition coefficient (K_{ow}), a commonly used parameter reflecting a compound's net hydrophobic character. This is true whether the sorbent domain is the polar surface of a mineral, or the comparatively apolar surface or interior of organic matter.

Physisorption of Nonionic Compounds to Mineral Solids

Minerals present mainly two kinds of environments for sorption: (1) the hydroxylated surface found on hydrous oxide minerals and the edges of aluminosilicate clays, and (2) the interlayers of layer silicate clays, which contain siloxane surfaces of variable charge density.

Hydroxylated surfaces have –OH groups protruding into solution from the topmost layer of lattice metal ions. Sorption of apolar compounds from dilute solution to nonmicroporous hydrous oxide surfaces is proportional to mineral surface area and is typically not very sensitive to surface charge, solution pH, and solution ionic strength (although high ionic strength may reduce water solubility and increase the tendency for sorption). It is also proportional to the hydrophobic surface area and to the K_{ow} of the sorbate, at least for closely related compounds. While data are sparse, sorption to hydroxylated surfaces tends to be endothermic and sorption isotherms tend to be linear to slightly nonlinear.

In the case of apolar compounds, there is good reason to believe that sorption to hydroxylated surfaces occurs by partitioning (i.e., absorption) of the sorbate into the semiordered multilayer of water under the influence of the surface, known as the vicinal water layer. This is depicted as 'B' in Figure 1a. Apolar molecules are thought to be unable to penetrate the closest one or two layers of water molecules

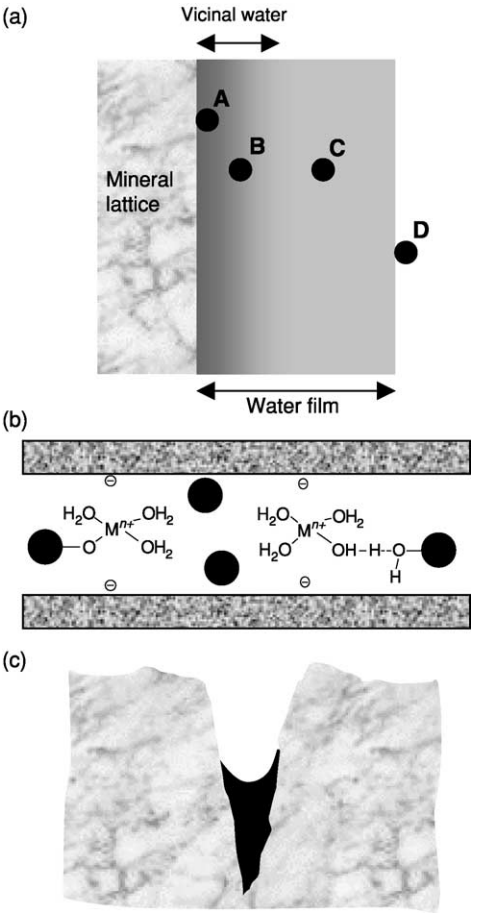


Figure 1 Depiction of different types of sorption in mineral particles: (a) interaction with discrete sites on the hydroxylated surface, A; partitioning in the vicinal water layer on the hydroxylated surface, B; dissolution in water films, C; and adsorption on the surface of water films, D; (b) sorption in a clay interlayer: left to right, direct coordination to interlayer metal ion; association with the uncharged siloxane surface; and hydrogen-bonding to a metal-coordinated water molecule; (c) condensation in a small pore.

and hence do not actually contact the surface. This is inferred from the observation that sorption of apolar organic vapors is strongly depressed by water vapor, and from the fact that per unit surface area of bare hydrous oxide surface heats of adsorption of organic molecules are much smaller than that of water. As the compound becomes more capable of hydrogen bonding to surface hydroxyl groups, it is better able to compete with strongly bound water molecules for surface sites ('A' in Figure 1a).

The siloxane surface existing on the faces of many layer silicate clays consists of oxygen atoms bridging underlying Si^{IV} atoms. The oxygens may bear permanent negative charge, depending on whether they lie near a point of isomorphic substitution in the underlying lattice. The charged sites are closely associated with metal or organic cations, and the surface in the vicinity of the charge is strongly hydrophilic (Figure 1b). Polar compounds can hydrogen-bond to metal ion-coordinated water molecules, or can coordinate directly with metal ions if they have an appropriate functional group. The neutral regions between charges are hydrophobic or only weakly hydrophilic; hence, hydrophobic molecules or parts of molecules may compete favorably with water for the bare surface in those regions. Sorption in clay interlayers is complicated by accessibility. The charge density and the exchangeable cation with its hydration sphere control the interlayer spacing and the size of the adsorption 'sites' available for molecules.

Under relatively dry conditions, multilayer adsorption, in which the sorbate condenses as a liquid-like phase on mineral surfaces, may also occur. Due to capillary forces, condensation progresses from the smallest to the largest of pores (or narrowest to widest part of a pore) as the partial pressure increases (Figure 1c).

Water has a strong negative effect on vapor adsorption to soils. Sorption of organic vapors decreases by an order of magnitude or more as the soil moisture content increases to approximately the wilting point, which corresponds to an average of 3–5 molecular layers of water (Figure 2). This inhibition is due to competition by water molecules for pore volume space and polar surface sites on inorganic surfaces. At moisture contents above the wilting point, sorption approaches what it would be in water-saturated soil.

When dealing with sorption of vapors to unsaturated soils, it is often necessary to take into account the association of organic sorbates with the water located in surface films and pores. This water is technically considered part of the immobile phase. Sorption may occur by a combination of aqueous-phase dissolution ('C' in Figure 1a) and vapor-phase adsorption at the gas–water interface ('D' in Figure 1a). The latter is especially important for compounds such as alkanes that are highly insoluble in water.

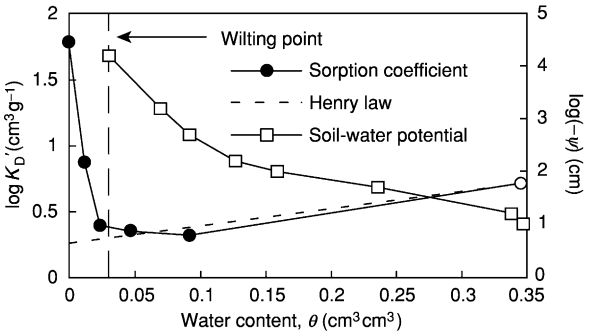


Figure 2 Toluene sorption coefficient (K_D') and soil-water potential (ψ) as functions of volumetric water content in Lundgaard soil (1.1% organic carbon). The Henry law line is the calculated sorption assuming dissolution into water and aqueous–solid partitioning only (i.e., zero vapor–solid sorption). (Reproduced with permission from Poulsen TG, Yamaguchi T, Moldrup P, de Jonge LW, and Rolston DE (2000) Predicting volatile organic vapor sorption from soil specific surface area and texture. *Journal of Environmental Quality* 29: 1642–1649.)

Physisorption of Nonionic Compounds to Natural Organic Matter

General considerations SOM is usually the dominant sorbent of neutral compounds in soils, because it presents a hydrophobic domain for hydrophobic molecules. Sorption isotherms in SOM and SOM-rich soils usually obey the semiempirical Freundlich equation:

$$S = K_F C^N \qquad [4]$$

where K_F and N are constants. The value of N is almost always less than or equal to 1, reflecting a constant $(N = 1)$-to-diminishing $(N < 1)$ tendency for sorption with increasing loading. The Freundlich relationship is only approximate and N may vary depending on the range of concentrations over which it is measured. It should be noted that all sorption isotherms tend toward linearity below a certain concentration, a result the Freundlich equation fails to predict.

Many researchers have assumed N is 1, or have taken N to be 1 for purposes of simplifying the sorption term in transport or bioavailability models. In that case, the isotherm is given simply by:

$$S = K_p C \qquad [5]$$

where K_p is referred to as the partition coefficient.

On the assumption that sorption occurs exclusively to the organic matter component of the soil, the distribution, Freundlich, or partition coefficient is

often normalized to the fraction of organic carbon (f_{oc}) in the sample; i.e.:

$$K_{oc} = K_d(\text{or } K_F \text{ or } K_p) \cdot f_{oc}^{-1} \quad [6]$$

Linear free energy relationships between K_{oc} and parameters that represent hydrophobic properties such as K_{ow} or saturated water solubility (C_{sat}) have been extensively tested, e.g.:

$$\log K_{oc} = a + b\log K_{ow} \quad [7]$$

$$\log K_{oc} = c - d\log C_{sat} \quad [8]$$

where a and b or c and d are fitting parameters. Such relationships work best for a set of closely related compounds in a given soil; the reason being that the enthalpy and entropy of eqns [1a–1d] depend on sorbate molecular structure, and because soil organic matter is not uniform from place to place. Other structure–property relationships have been studied based on, for example, parachor, connectivity index, or UNIFAC group contribution method. Solution properties such as pH and ionic strength have modest – and often contradictory – effects on sorption of nonionic compounds to SOM.

Natural organic materials in soils A discussion of sorption to soil natural organic matter must take into account the nature and abundance of all the different types of refractory components that make it up. Ordinarily the most abundant component is humic substances. Humic substances are highly altered, refractory macromolecules derived from biomass. They may also contain recognizable fragments of carbohydrates and proteins covalently attached to, or sequestered in, the humic solid. Humic substances are operationally defined as consisting of fulvic acid (soluble at all pH), humic acid (soluble in alkaline solutions), and humin (insoluble at all pH and typically strongly associated with inorganic particles). The soil water may also contain 'dissolved' humic substances – ranging from small, truly dissolved molecules to colloidal particles – that can sorb organic compounds, although somewhat less effectively than particulate humic substances. Kerogens are biogenic, macromolecular substances that have undergone low-temperature chemical and physical diagenesis over geologic time. Kerogens are abundant in soft coals and shale and may be found in soils as a result of reworking processes. Hard coals are biogenic macromolecular substances that have been subjected to high-temperature and -pressure alteration (catagenesis) over geologic time and are found in some soils. Black carbon is the carbonaceous by-product of biomass burning (charcoal) or fossil fuel combustion (soot). Black carbon may constitute a minute-to-small fraction of total organic C in soil.

Physisorption to humic substances and kerogen Humic substances and kerogens are considered to be random-network macromolecular solids; that is, they are composed of macromolecules of heterogeneous structure that are randomly aggregated to form a three-dimensional phase. This phase is hydrated to a degree depending on the type of humic substance and relative humidity, and may contain metal ions coordinated to charged groups, such as carboxylate. From fulvic acid to humic acid to humin to kerogen, there is a trend of increasing molecular weight, apolarity, C content, fused ring size, density, cross-linking, and glass-to-rubber transition temperature.

For such materials, an absorption rather than an adsorption model is better supported by the evidence and is widely accepted. External surface areas are too low – approx. $10^0\,m^2\,g^{-1}$ for humic acid particles based on N_2 as the probe – to account for their high uptake of organic chemicals. Furthermore, their surfaces are not hard like a mineral crystal, but diffuse and flexible due to solvation and charge repulsion of polar interfacial groups. Organic molecules reaching the interface pass through and enter interior regions, where they intermingle with the strands of the macromolecules. A similar paradigm for small-molecule sorption by synthetic organic polymers is well established. Sorption to humic substances appears to be slightly exothermic, although few data are available.

According to established polymer theory, amorphous macromolecular solids may exist in 'glassy' or 'rubbery' states, depending on temperature. These states differ in physical properties and sorbent behavior. The rubbery state is more expanded and its macromolecular segments have greater mobility. Sorption in rubbery solids is similar to liquid dissolution in that specific sorption sites have fleeting existence and sorbed molecules experience an average chemical environment. This results in a linear isotherm, like eqn [5], and noncompetitive sorption behavior when multiple solutes are present.

The glassy state is denser and more rigid. Also, because the glassy state is incompletely relaxed, it contains internal voids not filled by macromolecules. These voids (or 'holes') are of molecular dimensions and may accumulate sorbate. Hence, sorption in the glassy state exhibits dual-mode behavior. One mode is 'dissolution' in the bulk solid that is linear and noncompetitive, analogous to sorption in the rubbery state. The other mode is 'hole-filling' in the voids that is best modeled by a Langmuir-type equation. The dual-mode isotherm is given by:

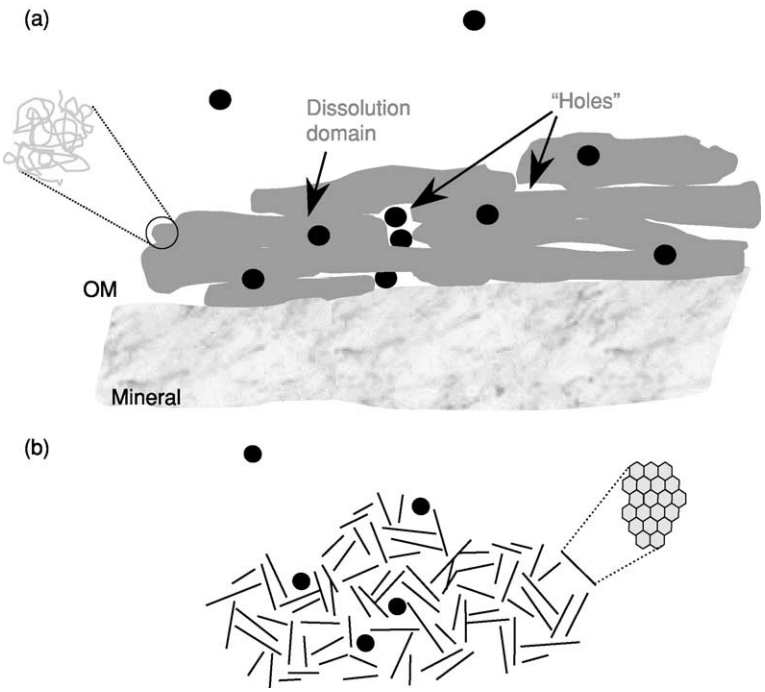


Figure 3 Sorption in soil organic matter (OM) fractions: (a) Macromolecular solids such as humic substances and kerogens: sorbate may undergo solid-phase dissolution in domains composed of macromolecules in a relatively fluid state and hole-filling in voids within the organic phase and between the organic phase and the mineral surface; (b) micrographitic substances such as black carbon: sorbate may adsorb in shallow micropores and some may penetrate into deep micropores between polyaromatic platelets.

$$S = K_D C + \frac{S_H^{max} K_H C}{1 + K_H C} \qquad [9]$$

where K_D is the dissolution domain partition coefficient, and S_H^{max} and K_H are the hole domain capacity and affinity parameters, respectively. Equation [9] predicts nonlinear and competitive sorption, both of which have been realized for many soils, humic acid reconstituted in particulate form, and kerogens.

In the context of SOM, holes may be regarded as semipermanent voids within the folds of individual macromolecules, between macromolecules, or between macromolecules and the mineral surface to which they are anchored. This is shown in **Figure 3a**. When eqn [9] is applied to heterogeneous natural materials: (a) the K_D term represents the sum of partition coefficients for all rubbery substances present and the dissolution domains of all glassy substances present; and (b) the Langmuir term represents a weighted average of holes of different sorption potential that may exist. The affinity for a given solute is greater in holes than in the dissolution domain.

As sorbate loading increases, the isotherm can become more complex. The glassy state may become softened ('plasticized') by the sorbate, causing it to become rubbery and the holes to disappear. At high loadings the isotherm in a rubbery solid begins to curve upward with respect to the solution concentration axis owing to the effects of swelling.

Natural organic matter in many samples appears to have glassy character. The glass transition temperature of some natural substances has been measured, including wet and dry humic acids (43–72°C), coals (>300°C), and wet lignin (60–90°C). Lignin is an amorphous, random polymer from the woody parts of plants that is regarded to be a major precursor of terrestrial humic substances. Whole-soil SOM may exhibit multiple glass transitions due to its heterogeneous nature. Synthetic polymers, humic acids, and coals have all been shown to have high internal surface area associated with the holes.

Nonlinearity and competitive effects tend to increase in the expected order of increasing glassy character of the solid:

$$\text{dissolved humic acid} < \text{particulate humic acid} < \text{whole-soil SOM} < \text{humin} < \text{kerogen}$$

Artificially cross-linking humic acid by flocculating it with polyvalent metal ions (e.g., Al^{3+}) increases nonlinearity and competition.

Figure 4 shows an example of sorption of a hydrophobic organic compound phenanthrene in SOM. It is evident that phenanthrene sorption is nonlinear with $N = 0.772$. It can also be seen that the dual-mode

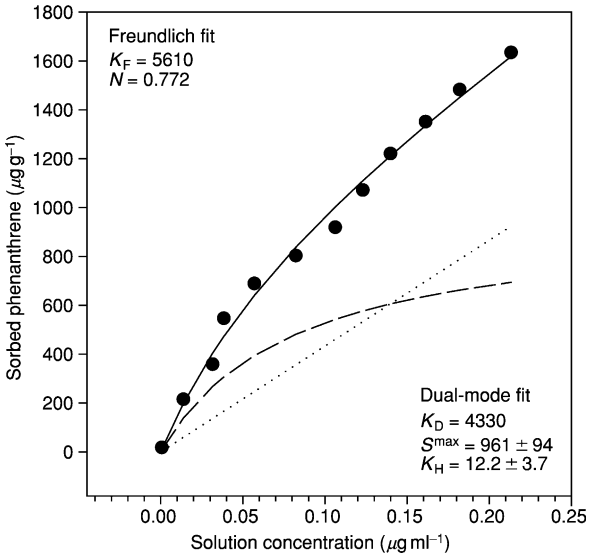


Figure 4 Phenanthrene sorption in soil with a high organic content. The fitting parameters of both Freundlich and dual-mode equations are shown. The solid line is the dual-mode fit, assuming only one type of Langmuir (hole) sites, and solid circles are actual data points; dotted and dashed lines are the contribution from partitioning and Langmuir (hole)-type sorption, respectively. The maximum surface sorption capacity calculated from surface area and molecular size is 286 $\mu g\, g^{-1}$. See text for details.

model fits the data well. Langmuir (hole) type of sorption dominates the overall uptake till the solution concentration reaches 0.14 $\mu g\, ml^{-1}$, then partitioning contributes more. The overall dissolution coefficient is 4330 $ml\, g^{-1}$ and Langmuir (hole) sorption capacity 961 $\mu g\, g^{-1}$. If the surface of this SOM sample (0.88 $m^2\, g^{-1}$) were rigid as an inorganic mineral, the maximal (external) surface coverage would be 286 $\mu g\, g^{-1}$ calculating from the molecular volume of phenanthrene as 182 $cm^3\, mol^{-1}$. Clearly, phenanthrene molecules have diffused into and sorbed within the SOM matrix, in contrast to surface adsorption.

SOM is heterogeneous and polydisperse in nature, and its characteristics can vary substantially as a function of age, source, diagenesis, and mode of transformation. Variations in SOM such as polarity and hydrophobicity strongly affect K_{oc}, up to several orders of magnitude. This is one reason why many different K_{oc}–K_{ow} free energy relationships are reported in the literature.

Aromatic and aliphatic moieties are the main structural components of SOM. Many investigations report positive correlations between the amount of HOC sorption and percentage of aromatic C of SOM. In addition, condensed aromatic domains are suggested to be the locations in SOM samples and isolated humic materials where nonlinear and competitive sorption occurs. Generally speaking, diagenetically altered organic matter samples are enriched in aromatic groups and show a higher degree of isotherm nonlinearity and competitive sorption, and a larger sorption capacity than geologically young samples. Organic matter in shale compared with recently deposited peat samples is an example in case. Therefore, HOC sorption is thought to be dominated by aromatic components of SOM rather than aliphatic components.

Recent research indicates that aliphatic components of SOM, particularly the paraffinic groups, can significantly sorb HOC. One example is that some soil humin samples often have a higher HOC sorption capacity than the humic acids extracted from the same soils, but these humic acids have higher aromatic content than the soil humins. It has been suggested that HOC is partitioned into the flexible, paraffinic carbon moieties to explain the high HOC sorption capacity. However, sorption of HOC by many soil humin samples is reported to be highly nonlinear and competitive, inconsistent with partitioning as the only sorption mechanism.

To account for both high sorption capacity and nonlinearity in soil humins, a new substructural conformation is proposed for the paraffinic carbons (Figure 5). Crystalline and amorphous aliphatic domains can coexist in SOM, which is supported by solid-state nuclear magnetic resonance (NMR) studies. A proposed conformation of semicrystalline polymer interphase is shown in Figure 5a. Sorption by crystalline regions is negligible due to their crystalline nature (inability for HOC molecules to diffuse). Upon interaction of the crystalline–amorphous complex with soil mineral surfaces, the first several molecular layers of the amorphous region may rearrange to take a more condensed form (Figure 5b). This condensed region can cause enhanced nonlinear sorption and both condensed and expanded amorphous aliphatic regions contribute to high HOC sorption. This proposed configuration is further supported by the experimental data reported in the early literature, where isotherm nonlinearity and competitive sorption of HOC by humic acids adsorbed on minerals are greater than unsorbed humic acids.

Physisorption to black carbon Black carbons are carbonaceous materials composed of single and stacked polyaromatic sheets in a highly disordered arrangement. The sheets may be functionalized along the edges with hydroxyl, carboxyl, keto, or other groups. The pore structure of black carbon depends on the source material and conditions of synthesis. From gas adsorption measurements it has high surface

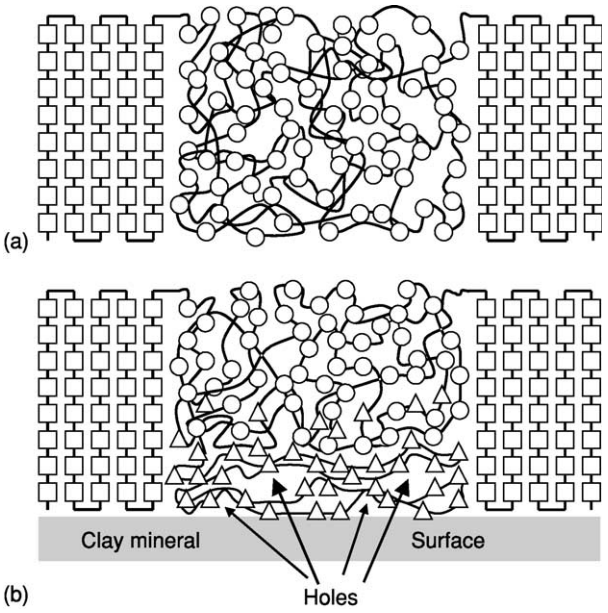


Figure 5 Schematic configuration showing the distribution of crystalline (squares) and amorphous (circles) aliphatic organic carbons (a). Interaction of the aliphatic carbons with a clay mineral surface rearranges part of the amorphous domain into a condensed configuration (triangles) having nanometer-sized holes (b). These holes are the locus of Langmuir adsorption sites, responsible for nonlinear and competitive sorption. (Reproduced with permission from Gunasekara AS and Xing B (2003) Sorption and desorption of naphthalene by soil organic matter: importance of aromatic and aliphatic components. *Journal of Environmental Quality* 32: 240–246.)

area (typically, more than $50\,m^2\,g^{-1}$), high porosity ($0.1–0.2\,cm^3\,g^{-1}$), and small pore size – i.e., below approximately 20 Å, with much porosity in the range 4–10 Å. Recent evidence shows that small apolar molecules can penetrate the interior of black carbon and cause swelling. Sorption isotherms of apolar compounds on prepared black carbons tend not to fit Freundlich or Langmuir models very well. For example, the isotherm of benzene in a wood charcoal is reverse S-shaped, with a linear region from 10^{-7} to 10^{-4} times its water solubility. While heats of sorption for black carbons have not been reported, sorption of apolar compounds from water to related carbonaceous sorbents such as graphite and activated carbon is exothermic.

A model that takes into account the existence of both macromolecular substances (humics plus kerogen) and black carbons in soils is clearly desirable. Models have been proposed of the form:

$$S = r_{HS} \cdot K_p C + r_{BC} \cdot f_{BC}(C) \qquad [10]$$

where r_{HS} and r_{BC} are the fractions of total organic carbon as humic substance and black carbon, respectively; S, K_p, and C are as defined above; and $f_{BC}(C)$ is a nonlinear function describing sorption to black carbon. In one example, $f_{BC}(C)$ is a Freundlich term like that in eqn [4]. In another, $f_{BC}(C)$ is a term that represents the potential to condense into liquid-like or solid-like form in pores that are assumed fixed in size and shape and similar to those that exist in synthesized activated carbons. Such models assume linear sorption in humic substances, which is not always true. Multicomponent models will be improved by advances in our understanding of the true nature of sorption to black carbon.

Kinetics of Physisorption

The rates of uptake and release from soil particles often control the transport and bioavailability of chemicals in soil systems. In well-mixed soil-water suspensions it is found experimentally that a large fraction of the sorbate is taken up or released within a few hours, but days, weeks, or even months may be required to reach complete equilibrium, depending on compound and soil. Moreover, exhaustive desorption at ordinary temperatures to air or water may reveal a minor fraction that seems permanently bound, yet is recoverable by solvent extraction or other vigorous means.

In general, soils contain structures that can hinder the diffusion of molecules to and from all available sites. These structures include pore networks in the interstices of particle aggregates as well as organic matter, which can be considered a highly viscous absorptive phase. Models based on pore diffusion and intraorganic matter diffusion being rate-limiting to sorption have been presented in the literature. Because most organic compounds have such a high affinity for the organic matter, it is tempting to regard intraorganic matter diffusion as rate-limiting unless the fraction of organic matter is very low. Generally, intraparticle sorption or desorption rate coefficients decrease with molecular size (or K_{ow} as a surrogate for size) and also appear to decrease with the introduction of hydrogen-bonding functional groups. This finding alone, however, does not distinguish pore diffusion and intraorganic matter diffusion. Fractions that become highly resistant to desorption have been generated in purely inorganic as well as purely organic reference materials. (Sorption or desorption in clay interlayers appears to be rapid, however, and does not generate resistant fractions.) Diffusion rates in soils are frequently independent of nominal particle radius, implying that the controlling diffusion length-scale is much smaller, perhaps as small as 10–100 nm.

Diffusion is highly dependent on the nature and geometry of the diffusing medium and the boundary conditions. Classic diffusion laws are difficult to

apply to soils owing to their heterogeneous nature in both composition and particle geometries. From studies on homogeneous inorganic reference materials, it is known that diffusion in fixed-pore networks containing mesopores (widths of 2–50 nm) or macropores (greater than 50 nm) is slowed by sorption to pore walls and by tortuosity. Tortuosity includes deviation from straight-line paths and pore noninterconnectedness. Diffusion in fixed micropore (less than 2 nm) networks, where pore openings are not much larger than the diffusant diameter, is additionally slowed by steric effects.

Intraorganic matter diffusion behavior depends on whether the solid is in the rubbery or glassy state. Diffusion in rubbery solids involves a more-or-less cooperative interchange between diffusant and polymer strands. The ideal rubbery sorbent is characterized by linear sorption isotherms, concentration-independent diffusion coefficients, and curves describing fractional mass uptake and release with time that are coincident. Diffusion in glassy sorbents generally shows the opposite. Molecules traversing the glassy solid experience a more condensed matrix and they encounter holes along the way where the diffusant can be momentarily detained. For these reasons, diffusion in glassy solids may be several orders of magnitude slower than in rubbery solids. The apparent diffusion coefficient in glassy solids is concentration-dependent because the availability of holes is concentration-dependent, and because the solid can undergo transition to a rubbery state at high concentrations. Diffusion coefficients of toluene, *n*-hexane, and acetone in pressed humic acid disks are in the range 10^{-8}–10^{-9} $cm^2\,s^{-1}$, about the same as in rubbery polymers and 3–4 orders of magnitude smaller than those in water. There is evidence to suggest that the diffusivity of a given compound follows the order: humic acid > humin > kerogen/coal. Diffusion within black carbons has received little attention so far.

To avoid difficulties inherent in the application of diffusion models to soils, many people have employed nonmechanistic rate models based on simpler mathematics. Recent ones include the two-site, multicompartment, and stochastic models. Although these models have met with some success, they usually suffer from an inability to translate to new systems. The two-site model, which is often used in soil-column elution studies, partitions the sorbate into instantaneous and kinetic compartments. The kinetic compartment is described by first-order terms for the forward and reverse steps. Overall the rate law has three parameters: fraction in the kinetic compartment, bulk solid equilibrium sorption coefficient, and rate constant for desorption from the kinetic compartment. The multicompartment model is popular in batch experiments in which molecules are instantly removed as they desorb by a stream of gas or by a polymeric sorbent included in the suspension. This model assumes that desorption occurs concurrently from 'fast,' 'slow,' and sometimes 'very slow' compartments, each described by a first-order term having two parameters: the rate constant and the fraction of chemical in that compartment. Stochastic models assume that the rate can be described with an array of equilibrium sorption coefficients and/or first-order rate constants that are distributed continuously according to a probability density function. Such models have three or two parameters, respectively.

Physisorption Hysteresis

Hysteresis in the sorption isotherm exists when the curve constructed by allowing the solid to take up the compound (the sorption branch) is different from the curve constructed in the reverse manner, by diluting the fluid and reequilibrating (the desorption branch). Typically, the desorption branch is shifted upward of the sorption branch, and may appear to extrapolate to a nonzero sorbed concentration at zero solution-phase concentration. Hysteresis has been observed in many studies. An example is given in Figure 6. Hysteresis poses a challenge to the elucidation of sorption mechanism and to using sorption models predictively.

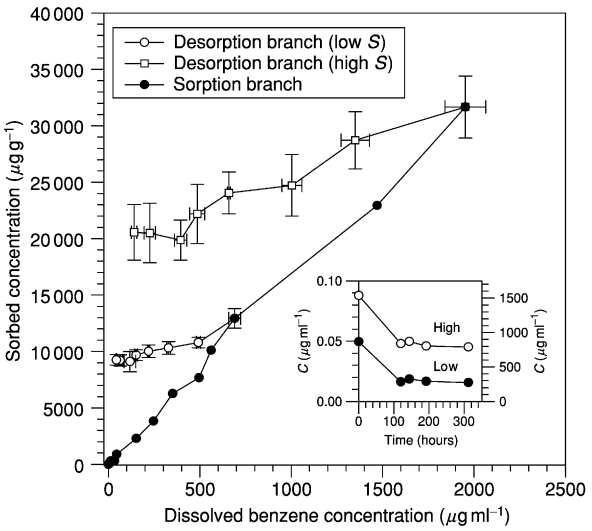


Figure 6 Hysteresis of benzene in a reference soil (Pahokee) from the International Humic Substances Society. Sorption was carried out in one step for 14 days. Desorption was carried out in consecutive steps (triplicate samples), each step for 14 days. Inset shows uptake rate curves (change in solution concentration) at two concentrations, indicating apparent equilibrium after approximately 4 days.

Hysteresis may be an artifact of the experimental technique. When sorbed concentrations are determined indirectly based on the amount taken up from solution, as is often the practice, artificial hysteresis may result from chemical or biological reactions or from dilution of competing substances such as cosolutes or colloidal sorbents in the desorption step. Failure to allow enough time for diffusion to all sites can give artificial hysteresis regardless of whether sorbed concentration is determined directly or indirectly. Nevertheless, in many cases hysteresis is observed even when those artifacts have been eliminated.

True hysteresis (also referred to as irreversible sorption) is caused by spontaneous mechanisms associated with irreversible pore-filling and -emptying. Since it would violate the laws of classical thermodynamics for both isotherm branches simultaneously to represent the reversible state, true hysteresis always indicates a nonequilibrium condition. Proposed explanations for true hysteresis fall into three categories: formation of metastable states of the sorbate, network percolation effects, and irreversible pore deformation. The first two are associated with capillary condensation of vapors in mesoporous solids. During uptake in a pore there can be a delay in the transition from the metastable multilayer film to the thermodynamic condensed plug state; evaporation from the meniscus of the condensate is less favorable than from the film surface. Hysteresis due to network percolation effects may arise in complex systems of pores, and it results when desorption is delayed because condensate in some pores does not have a continuous vapor channel to the surface. Capillary condensation hysteresis applies primarily to fixed-pore mesoporous materials.

Irreversible pore deformation has been identified in sorption of organic vapors or solutes by organic sorbents, including dry smectite clays, organosmectite clays, glassy polymers, and (recently) natural organic matter in soils and prepared charcoals. By this mechanism, the incoming molecule causes dilation of the pore due to the flexibility of the organic matrix, and the pore does not instantly relax when the molecule leaves. When this happens it may be surmised that some of the free energy of sorption goes into expanding the pores and the sorbate leaves from pores existing in a different thermodynamic state than the ones they entered. When desorption abruptly occurs, it is possible that some molecules get trapped in the collapsing organic matrix.

Ion-Exchange Sorption

Permanently charged and ionizable organic molecules in their ionic form can undergo 'ion-exchange' reactions with charged sites in soil particles from which existing ions are displaced. Cation exchange is far more common than anion exchange due to the prevalence of negative charges in soil particles. Examples of permanently charged compounds that can undergo cation exchange include tetraalkylammonium surfactants and cationic herbicides such as paraquat and diquat. Alkylamines, aromatic amines, and *N*-heterocyclic aromatic compounds are examples of ionizable compounds that can sorb via ion exchange, depending on the degree of their protonation, which, in turn, depends on soil pH.

Typically, cation exchange is nonlinear and subject to competition by other cations present. According to diffuse double-layer theory, cation exchange on mineral surfaces occurs by an accumulation of organic ions in the vicinal water film near the surface due to electrostatic force between the ion and the oppositely charged surface. However, sorption of organocations often exceeds the cation exchange capacity (CEC) of the solid in proportion to the apolar content of the molecule, such as alkyl chain length or aromatic ring content. This may be rationalized by a contribution to overall sorption from the hydrophobic effect operating on the apolar parts of the ion. In order to maintain electrical neutrality, organocations in excess of the CEC must carry along a counterion (e.g., Cl^-) to the interfacial region. Thus, sorption of organocations is best represented as the sum of true ion exchange and sorption of the ion pair:

$$R^+ + M^+ \cdots \equiv \rightleftharpoons M^+ + R^+ \cdots \equiv \qquad [11]$$

$$R^+ + X^- + \equiv \rightleftharpoons R^+X^- \cdots \equiv \qquad [12]$$

where R^+ represents the organocation, X^- a counter ion, M^+ a competing cation, and $\equiv$ the surface.

Ion-exchange reactions in SOM also occur. Organocations interact electrostatically with dissociated carboxyl or phenoxyl groups on humic molecules; hence, ion exchange in SOM is strongly pH-dependent. Ion exchange in SOM may be facilitated by physisorptive interactions of the noncharged parts of the molecule with humic groups other than the negative charge itself. Just like ion exchange on minerals, the hydrophobic effect may be expected to play some role in the sorption of organocations. Sorption of organoanions to SOM is typically weak, since positively charged sites on SOM are rare and the solid is negatively charged overall at ordinary pH. However, some organoanions – e.g., alkylbenzene sulfonate surfactants and the anion of pentachlorophenol – appear to be capable of sorbing to SOM, possibly as their ion pairs with Na^+ or other cations.

Chemisorption

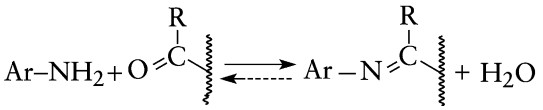
Chemisorption entails appreciable overlap of molecular orbitals between sorbate and sorbent, such as ligand–metal coordination bonds and covalent bonds. Molecules capable of chemisorption may also be capable of physisorption.

Ligand–metal bonds may be formed between carboxylate, phenolate, thiolate, sulfonate, or phosphate groups on the solute and metal ions – especially Fe^{3+} and Al^{3+} – on the surface of minerals, usually in exchange for hydroxyl or water ligands:

$$\equiv M{-}OH(-OH_2) + L^- \rightleftharpoons \equiv M{-}L + OH^-(H_2O) \quad [13]$$

where $\equiv$ represents the surface, M a metal ion, and L a ligand. Equation [13] is more favorable when L is bidentate; for example, salicylate, oxalate, catecholate, or *o*-phthalate and their derivatives. In that case the ligand can coordinate through both functional groups and to more than one metal ion. It is also possible for the metal ion to be attached to SOM through an appropriate functional group on the humic structure; in that case, an organic ligand (e.g., 2,4-dichlorophenoxyacetate (2,4-D anion)) may coordinate to the cation and thus be linked through cation-bridging to SOM. The energetics and kinetics of surface ligand exchange reactions are not well characterized. The intrinsic equilibrium constant seems to depend on structure in the same way as when the corresponding reaction takes place between the ligand and free metal ion in solution.

Covalent bonds between sorbate and humic substances are comparatively rare. Aromatic amines often generate nonextractable residues and are postulated to react, quasi-reversibly, with certain carbonyl groups on humic macromolecules to form imines:



[14]

The reactive carbonyl may be present in ketone, aldehyde, or quinone functional groups. The rate and extent of imine formation are currently unpredictable.

Other cases are known in which the organic solute is converted to reactive forms that then couple covalently with humic substances. Concern about potential remobilization of these so-called bound residues has spawned an active area of research. Coupling reactions are technically outside the definition of sorption, since the molecules undergoing coupling are transformed species, not the original ones, and since they are seldom reversible in the sense that the original chemical is liberated. Examples are found among pesticides, explosive compounds, phenols, aromatic amines, and polycyclic aromatic hydrocarbons (PAHs). Typically, the initial reaction is an enzyme- or mineral surface-catalyzed oxidation reaction, although reduction reactions have been reported. Rarely has the coupling reaction itself been assigned with certainty to a biotic or an abiotic origin, much less a specific mechanism. In most cases coupling is inferred by the inability to recover by solvent extraction a portion of an isotope when the compound is added in isotopically labeled form. Oxidative coupling of some compounds – for example the fungicide cyprodinil and the herbicide 2,4-dichlorophenoxyacetic acid – has been verified by nuclear magnetic resonance spectroscopy. In another example, trinitrotoluene was microbially reduced to amino toluenes that subsequently coupled to organic matter, most likely abiotically via eqn [14]. It has been proposed that aromatic amines are oxidized by manganese oxide minerals in soils to give free radicals that subsequently couple with humic substances.

See also: **Hydrocarbons**; **Organic Matter:** Interactions with Metals; **Pesticides**

Further Reading

Alexander M (1995) How toxic are toxic chemicals in soil? *Environmental Science and Technology* 29: 2713–2717.

Cheng HH (1990) *Pesticides in the Soil Environments: Processes, Impacts, and Modeling*. Madison, WI: SSSA.

Chiou CT (2002) *Partition and Adsorption of Organic Contaminants in Environmental Systems*. Englewood Cliffs, NJ: John Wiley.

Dec J and Bollag J-M (1997) Determination of covalent and noncovalent binding interactions between xenobiotic chemicals and soil. *Soil Science* 162: 858–874.

Gunasekara AS and Xing B (2003) Sorption and desorption of naphthalene by soil organic matter: importance of aromatic and aliphatic components. *Journal of Environmental Quality* 32: 240–246.

Israelachvili JN (1992) *Intermolecular and Surface Forces*, 2nd edn. London, UK: Academic Press.

Naidja A, Huang PM, and Bollag J-M (2000) Enzyme–clay interaction and their impact on transformations of natural and anthropogenic organic compounds in soil. *Journal of Environmental Quality* 29: 677–691.

Pignatello JJ (2000) The measurement and interpretation of sorption and desorption rates for organic compounds in soil media. *Advances in Agronomy* 69: 1–73.

Poulsen TG, Yamaguchi T, Moldrup P, de Jonge LW, and Rolston DE (2000) Predicting volatile organic vapor sorption from soil specific surface area and texture. *Journal of Environmental Quality* 29: 1642–1649.

Rouquerol F, Rouquerol J, and Sing K (1999) *Adsorption by Powders and Porous Solids*. San Diego, CA: Academic Press.

Schwarzenbach RP, Gschwend PM, and Imboden DM (2002) *Environmental Organic Chemistry*, 2nd edn. Englewood Cliffs, NJ: John Wiley.

Senesi N (1992) Binding mechanisms of pesticides to soil humic substances. *Science of the Total Environment* 123: 63–76.

Xing B and Pignatello JJ (1997) Dual-mode sorption of low-polarity compounds in glassy polyvinychloride and soil organic matter. *Environmental Science and Technology* 31: 792–799.

Oxyanions

C P Schulthess, University of Connecticut, Storrs, CT, USA
H Wijnja, Connecticut Agricultural Experiment Station, New Haven, CT, USA
W Yang, University of Connecticut, Storrs, CT, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

An oxyanion is a net negatively charged compound consisting of a cation at its center surrounded by oxygens. There are many different oxyanions present in soils. The most common ones of interest to soils include phosphates (PO_4^{3-}), arsenite (AsO_3^{3-}), arsenate (AsO_4^{3-}), selenite (SeO_3^{2-}), selenate (SeO_4^{2-}), chromate (CrO_4^{2-}), molybdate (MoO_4^{2-}), carbonate (CO_3^{2-}), sulfate (SO_4^{2-}), nitrate (NO_3^-) and perchlorate (ClO_4^-). Oxyanions are prevalent in nature, and most of these are essential nutrients for plants. Most all of them, however, are also toxic to plants and the environment when their concentrations rise above a certain critical level. For example, excessive amounts of some oxyanions, namely nitrates and phosphates, are directly responsible for the eutrification of our ponds and lakes.

There are, of course, numerous other oxyanions that are discussed in the environmental literature, particularly in connection to industrial and nuclear waste disposal, or spills and remediation. Industrial oxyanion wastes include many of those already mentioned above. Radioactive waste and environmental studies may include oxyanions such as pertechnetate (TcO_4^-). Since no stable isotopes of Tc exist (which are prepared in very small amounts in cyclotrons), and since it is rather dangerous to work with this radionuclide, a surrogate oxyanion is typically studied instead, namely perrhenate (ReO_4^-).

The fate of oxyanions in soils in the environment is strongly dependent on their sorption to soil mineral components. That is, an oxyanion that is adsorbed to an immobile solid phase will not be transported to a new location, because movement of oxyanions in the environment is through the liquid phase. Retention by solids results in a longer availability of oxyanions in a defined area. For example, leaching from the soil's upper horizons down toward groundwater does not occur as long as it is sorbed to the immobile solid phase. Lateral liquid movement to nearby streams is also minimized, as is its subsequent transport to larger bodies of water.

All this, of course, assumes that the solid phase that the oxyanion is sorbed to is itself immobile. Keep in mind, however, that various solids (such as colloidal-sized particles) are very small and can be translocated by the moving water. This translocation of mobile solids along with the ions sorbed on them can be either vertically through soil horizons, or horizontally with the groundwater flow, or as soil erosion moving horizontally over the soil surface.

An understanding of sorption of oxyanions to solid soil components is very important for an evaluation of the quality of the environment. If the oxyanions are retained weakly, then the moving water will easily remove them from the soil. If the retention of oxyanions by the immobile solid-phase is moderate in strength, then the moving water will remove the oxyanions from the area slowly. Conversely, if the oxyanions are retained strongly by immobile solids, then the moving water will remove them from the area very, very slowly. If the solid phases are small enough to be mobilized by the moving water, then attention to the turbidity of the liquid phase or the presence of surface soil erosion must be included in any evaluation of the environmental quality present. Once again, this is because the mobile solids typically will have sorbed contaminants on them.

Techniques for Detection and Quantification of Oxyanions in Solution

Quantitative studies of the adsorption of oxyanions on soils and minerals require determination of the oxyanion concentration remaining in the aqueous phase. That is, the amount adsorbed equals the amount of anion added to the solid-liquid suspension minus the amount remaining in the liquid phase at equilibria. Many different instruments can be used for the liquid-phase analysis, such as ion-selective electrodes, colorimetric methods, ion chromatography, atomic absorption, and inductively coupled plasma-atomic emission spectrometry. With some of these methods (namely, the first three mentioned), one must be particularly cautious to avoid problems attributed to speciation and complexation interferences. These complications can usually be avoided by measuring the liquid phase concentration with

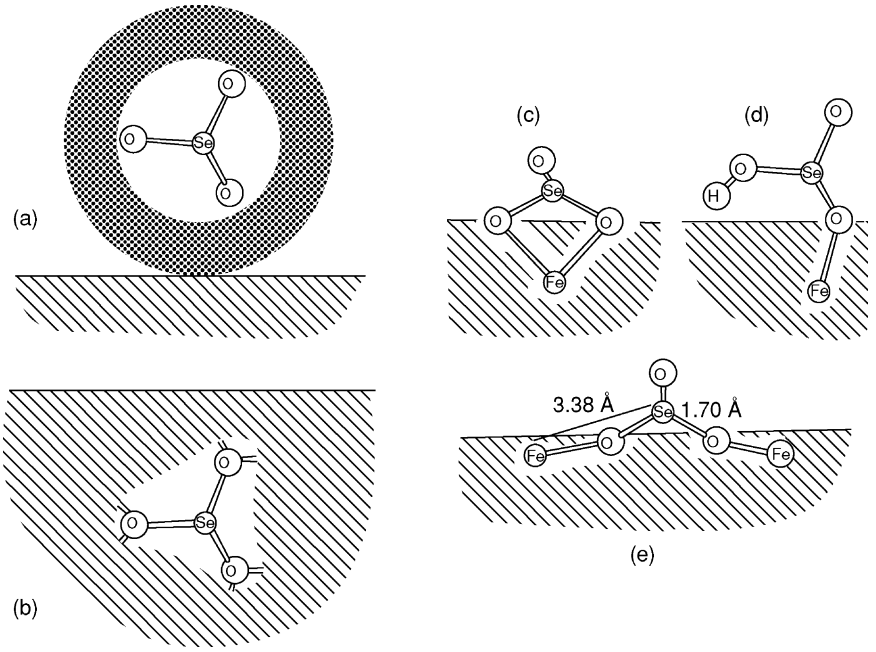


Figure 1 Schematic of several of the possible geometric arrangements of an oxyanion at the solid–water interface. Selenite is the adsorbate oxyanion and an iron oxide the adsorbent mineral. The area above each surface (indicated by solid horizontal lines) is water, while the solid mineral phases are below the solid horizontal lines. (Reprinted from Brown GE Jr, Parks GA, and Chisholm-Brause CJ (1989) *In-situ* X-ray absorption spectroscopy studies of ions at oxide–water interfaces. *Chimia* 43: 248–256.)

a high background ionic strength and at a specified pH condition. One does not generally quantify the amount of oxyanions adsorbed by direct measurement of the solid-phase concentration. This is because many of the solid-phase measuring methods require expensive instruments and/or the quantitative precision is not satisfactory.

Techniques for Investigation of Oxyanions Adsorbed at Mineral–Water Interfaces

Various surface-sensitive techniques have been used to obtain structural information of surfaces and surface species that are needed to identify adsorption mechanisms and reactions. An important aspect of the application of spectroscopic techniques to the investigation of adsorbed species is the experimental conditions under which the spectra of the sample are collected. Invasive or *ex situ* methods commonly require extensive sample dessication and vacuum, and therefore have the potential for modifying or destroying sorption complexes normally present at the solid–liquid interface in an aqueous environment. Conversely, noninvasive or *in situ* methods require little or no alteration of a sample from its natural state. Clearly, in order to resolve the structure of surface species present at a mineral–water interface of natural particles, *in situ* spectroscopy techniques are needed that can be applied in the presence of liquid water. Three examples of *in situ* techniques that have been used for the investigation of oxyanions on mineral–water interfaces are attenuated total reflectance–Fourier-transformed infrared (ATR-FTIR), Raman, and X-ray absorption spectroscopy (XAS). A very brief overview of these is given below.

The predominant application of spectroscopic analysis of adsorbed oxyanions at the solid–liquid interface is to understand the symmetry and binding mechanism of the anion on the surface. Quantitative measurements ("How many ions are adsorbed?") is not the primary focus, because, as was noted above, this is much more easily determined based on the mass balance of the elements added and recovered from the liquid phase. Spectroscopy is the only method, however, that can determine how the ions are held on the surface. Figure 1 shows five commonly encountered possibilities. Using selenite as the example oxyanion, it can be held as an outer-sphere (Figure 1a), where at least one molecule of water exists between the anion and the surface. It may be embedded in the internal regions of the mineral (Figure 1b) due to solid diffusion of the anion into the mineral or due to some coprecipitation of the anion with a cation onto the surface of the mineral. Figure 1c–e shows the anion held as an inner sphere, where the anion directly bonds with the mineral surface. An inner-sphere anion can be monodentate (Figure 1d) or bidentate

(Figure 1c, e). 'Monodentate' means that there is only one bond formed between the anion and the surface, whereas 'bidentate' means that there are two bonds formed between the anion and the surface. The bidentate inner-sphere bonds can be mononuclear (Figure 1c) or binuclear (Figure 1e), which refer to the minimum number of mineral cations that are needed to link the anion's oxygens to the mineral.

Vibrational Spectroscopies

Vibrational spectroscopy techniques such as FTIR and Raman spectroscopy probe the vibrational properties of molecules. Characteristic vibrational properties of oxyanions originate from *X*—O(H) stretching and bending vibrations. Each oxyanion has its characteristic pattern of vibrations that relates to its symmetry state. The symmetry, and consequently the vibration pattern, changes when an oxyanion complexes with other ions such as protons or metal cations. The changes in the number and position of the virbational bands provide information about the speciation and coordination state of the oxyanion. This behavior is the underlying principle of the determination of the types of surface species using vibrational spectroscopy.

Spectra of outer-sphere complexes are similar to the spectra of the free anion in solution. Differences in position and number of adsorption bands between the spectra of solution and adsorbed species indicate coordination with the surface, that is, inner-sphere complexation. Determination of the coordination state of an inner-sphere complex (mono- or bidentate and mono- or binuclear) is usually based on the comparison of spectra of adsorbed species with the literature data of spectra of coordinated compounds with known structures. Extensive infrared and Raman spectral data of inorganic and coordination compounds are available.

Attenuated total reflectance–Fourier-transformed infrared (ATR-FTIR) spectroscopy is an *in situ*, infrared sampling technique that can be used to provide information about surface species at the mineral–water interface. Note that, since water strongly adsorbs infrared, it was the introduction of ATR (and other similar detection modifications, such as cylindrical internal reflectance, CIR) that has given us the needed breakthrough to very effectively use FTIR in aqueous (i.e., natural) mediums. This technique has been applied successfully to investigations of various anions (PO_4^{3-}, $B(OH)_4^-$, SO_4^{2-}, CO_3^{2-}, NO_3^-) at metal oxide–solution interfaces.

The use of ATR-FTIR does not fully avoid water absorption of the infrared signal. The spectral window in metal oxide aqueous suspensions is limited by the strong adsorption by water and metal oxide lattice vibrations in the region below 1000 cm^{-1} and the strong adsorption by water in the region of 3000 – 4000 cm^{-1}. For example, the vibrations of the oxyanions of Se and As are expected below 1000 cm^{-1} and, consequently, are difficult to observe in these systems with the ATR-FTIR technique.

Diffuse reflectance FTIR (DRIFT) is an infrared sampling technique that requires air-drying of the sample. Usually this does not significantly affect the surface speciation and coordination of the adsorbed oxyanion because a very thin layer of water should remain on the mineral surface. Since there is less interference with the strong absorbing water, this sampling technique offers the possibility to investigate OH-stretching in the spectral range of 3400–3800 cm^{-1}. The OH-stretching bands can provide information about the effect of oxyanion adsorption on the surface OH groups of the mineral surface.

Raman spectroscopy can provide a window in the spectral region below 1000 cm^{-1} in the investigation of aqueous minerals and, accordingly, can offer complementary data to IR spectroscopy. One of the attractive aspects of Raman spectroscopy for its application to minerals in aqueous suspension is that Raman is transparent to water and, therefore, offers the possibility of *in situ* investigation of surface species. The recent advances in Raman instrumentation have brought new opportunities for applications to the investigation of surface species at mineral–water interfaces. Raman spectroscopy has been used to study the interfacial species on metal oxides of SeO_4^{2-}, SO_4^{2-}, AsO_4^{3-} and AsO_3^{3-}.

X-Ray Absorption Spectroscopy

X-ray absorption spectroscopy (XAS) is another techniques that can be applied to *in situ* studies of adsorbates at mineral–water interfaces. XAS is an element-specific, short-range probe, providing information about the local structural and compositional environment of an adsorbing atom. X-ray absorption near-edge structure (XANES) spectra provide information about the bonding and oxidation state of the adsorbing atom. Extended X-ray absorption fine structure (EXAFS) can provide information about the composition and structure in the vicinity of the element. The absence of back-scattering beyond the first shell indicates an outer-sphere complex. In the case of inner-sphere complexes, the number, distance, and type of back-scatterers provide information for identifying the type of inner-sphere complex (mono-versus bidentate). XAS has been used to identify the adsorption mechanisms of the oxyanions of Se, As, and Cr. The technique requires high-energy X-rays, which are only available at synchrotron facilities.

Sorption Strength of Oxyanions

The strength of the bond of oxyanions with mineral surfaces strongly affects the fate of the oxyanion in the environment. Weakly held anions leach easily and are lost from the upper soil horizons down toward the groundwater or are readily carried in the surface runoff to nearby streams. There are a few ways to determine the relative strength of these bonds. One common method used involves adsorption isotherms and another involves adsorption edges.

An adsorption isotherm analysis is based on measuring the amount of anion adsorbed by a solid phase as a function of the equilibrium concentration present in the liquid phase, as shown in Figure 2 for sulfate adsorption onto an aluminum oxide. The curvature of the isotherm, which is particularly visible at low aqueous concentrations, is a function of the adsorption strength. Qualitatively, the sharper the rate of increase and the sharper the amount of curvature observed in the adsorption isotherm, the stronger the adsorption strength of the compound with the surface. It is the curvature of the adsorption isotherm, not the adsorption maximum, that gives an indication of the adsorption strength.

Efforts to quantify the adsorption strength have been the focus of much debate in the literature for nearly a century. For example, if we assume a simple adsorption mechanism:

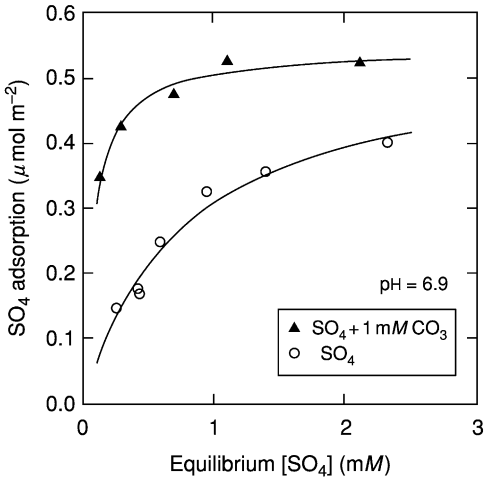


Figure 2 Adsorption isotherms of sulfate on an aluminum oxide (bayerite). Notice that the presence of carbonates increased the adsorption strength of sulfate, which is contrary to the competitive adsorption behavior typically observed when two adsorbing anions are in solution. (Reprinted from Wijnja H and Schulthess CP (2000) Interaction of carbonate and organic anions with sulfate and selenate adsorption on an aluminum oxide. *Soil Science Society of America Journal* 64: 898–908.)

$$\text{Solid Surface} + \text{Aqueous Anion} \rightleftharpoons \text{Adsorbed Anion} \qquad [1]$$

then the amount adsorbed can be predicted mathematically using data-fitting techniques on the Langmuir isotherm equation:

$$\text{Amount Adsorbed} = MKc/(1 + Kc) \qquad [2]$$

where M is maximum adsorption, which can be obtained from the isotherm at high aqueous concentrations; K is the Langmuir constant, which is equal to the ratio of P/R, with P the adsorbed anion concentration, R the (concentration of unoccupied surface sites) $\times$ (aqueous anion concentration); and c is the aqueous anion concentration at reaction equilibria. The adsorption strength is quantified here by the Langmuir constant, K. Much debate arises, however, about just how the adsorption mechanism should be modeled; that is, modeling the adsorption reaction by eqn [1] and other similar equations are not yet fully validated. Note that a model that does not fit the data is clearly disproved. However, a model that does fit the data is not automatically adequately validated.

One can also infer the relative strength of an adsorbing anion based on the pH value of the adsorption edge, which is shown in Figure 3. For a given mixture, the amount of aqueous anion adsorbed by the solid phase is strongly dependent on the equilibrium pH of the system. Working down the pH scale, there is no adsorption of anions at high pH values. At intermediate pH values, there is an abrupt increase in adsorption. The pH value measured at the midpoint of this abrupt adsorption increase is the pH value of the adsorption edge. At much lower pH values, there is a decrease in the amount of anions adsorbed. The entire bell-shaped curve is known as an adsorption envelope.

If we assume that the adsorption mechanism involves the coadsorption of hydroxyl anions (OH^-), for example:

$$\text{Surface—OH} + \text{Anion}^{n-}(\text{aq}) \rightleftharpoons \text{Surface—Anion}^{1-n} + \text{OH}^-(\text{aq}) \qquad [3]$$

then we can use the pH value of the adsorption edge to compare the relative adsorption strength of the anion. At equimolar concentrations, the anion that has a higher adsorption edge value is the anion that binds more strongly with the surface. Notice in eqn [3] that a strong surface–anion bond will require a higher concentration of OH^- anions to reverse the reaction. The midpoint of equilibria for eqn [3] occurs at the adsorption edge. Accordingly, if the

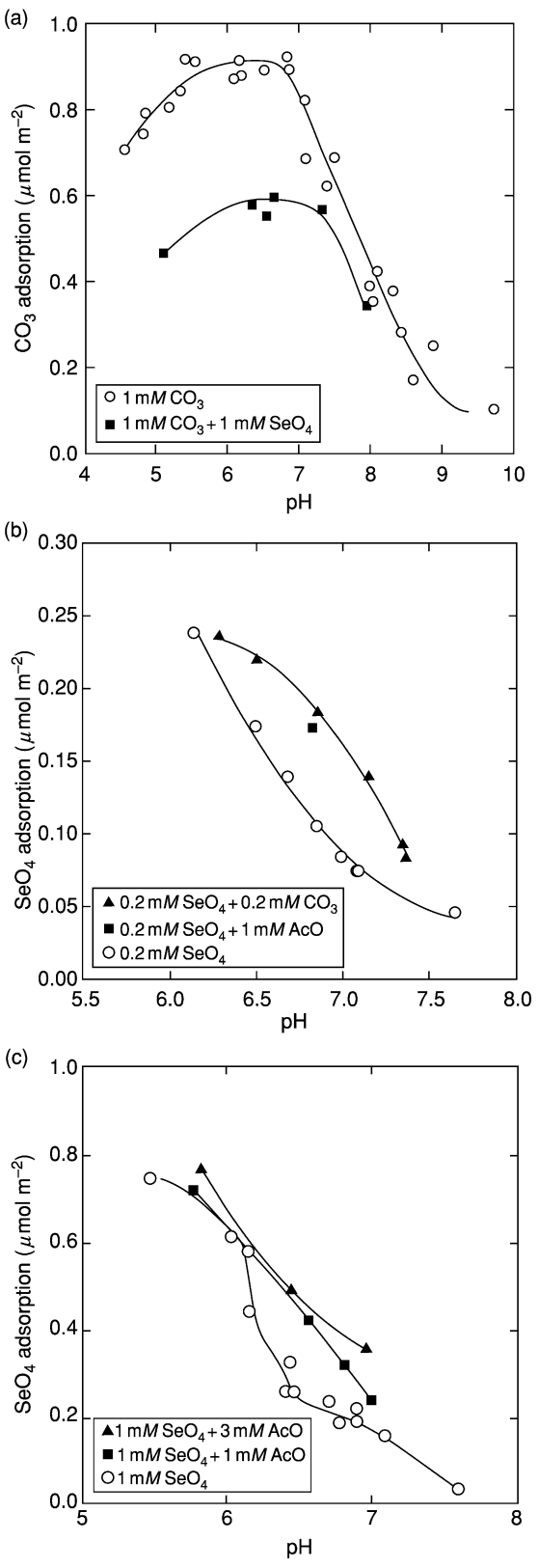


Figure 3 Adsorption envelopes (adsorption as a function of pH) for various oxyanions on goethite: (a) adsorption of carbonate. Notice the competitive behavior of selenate on carbonate adsorption; (b, c) adsorption of selenate in the presence of carbonate and acetate. Notice the enhancement by these oxyanions of selenate adsorption. (Reprinted from Wijnja H and Schulthess CP (2002) Effect of carbonate on the adsorption of selenate and sulfate on goethite. *Soil Science Society of America Journal* 66:1190–1197.)

adsorption of an anion is strong, then the quantity of OH^- anions needed to reverse the reaction will be high, and the pH value of the adsorption edge will be high also.

Efforts to quantify the adsorption strength based on the pH value of the adsorption edge are not yet fully resolved. The problem arises in that we are not definitively sure that eqn [3] or other similar equations fully describe the adsorption process. Consequently, the mathematical interpretation of the pH-dependent data is still discussed extensively in soil science research articles. We do know, however, that if the pH is kept constant, then eqn [3] will simplify to eqn [1], and equations similar to eqn [3] will simplify to equations similar to eqn [1]. We also know that anion adsorption does involve the release of hydroxyl anions. Note that it is difficult to distinguish hydroxyl release (which chemically removes protons from solution) from proton coadsorption (which physically removes protons from solution). Accordingly, essentially identical concepts can be presented using equations that illustrate proton coadsorption to the surface rather than hydroxyl release from the surface.

Using the tools outlined above, a general review of the published data suggests the following trend: the larger the negative charge present on the oxyanion, the stronger the affinity of the oxyanion toward the surface. Within each group, there is much uncertainty on which is more strongly adsorbed, and this is probably due to numerous reasons (such as, presence of surface contaminants). In general, the affinity of trivalent oxyanions follows the sequence: $AsO_3^{3-} > PO_4^{3-} \approx AsO_4^{3-}$. Divalent oxyanions adsorb more weakly, and in general follow the affinity sequence $SeO_3^{2-} > CrO_4^{2-} > CO_3^{2-} > SeO_4^{2-} \approx SO_4^{2-} > MoO_4^{2-}$. Monovalent oxyanions such as ReO_4^-, NO_3^- and ClO_4^- are very weakly adsorbed. Another, more useful observation is the close correlation of adsorption affinities of oxyanions with the stability constants (pK_a values) of their conjugate acids; noteworthy is the high adsorption affinity of $B(OH)_4^-$ by minerals. For these reasons, nitrates or perchlorates are sometimes used as background electrolytes in adsorption studies and are often assumed not to adsorb at all or to cause no significant impact on the adsorption process of the other compounds present.

Impact of Ions on Adsorption of Oxyanions

The presence of dissolved electrolytes in the mixture affects the adsorption of oxyanions. Adsorption studies, therefore, are pursued under constant ionic-strength conditions. These background electrolytes consist of dissolved cations and anions such as Na^+Cl^- or $Na^+NO_3^-$. It is critical to the adsorption study that the reactivity of these ions with the surface be minimal, preferably completely unreactive or at least constant in their reactivity. Sometimes, an aqueous phase complexation reaction occurs between the oxyanion being studied and the components of the background electrolyte. This will complicate the analysis and should be avoided as much as possible, preferably by using a different background electrolyte.

The presence of dissolved electrolytes affects the adsorption process in numerous ways. First, it raises the ionic strength of the solution. This tends to lower the electrostatic attraction of the adsorbate due to the lowering of surface potential or the masking of the surface charge. This impact on the adsorption process is not a serious concern as long as the ionic strength is kept constant throughout the study.

Second, the background anion can compete with the oxyanion for adsorption sites, which is very likely at high ionic-strength concentrations. Although the adsorption of the background anion (e.g., Cl^-) is typically extremely weak, it is rarely zero and its competitive impact increases significantly as its concentration in solution increases. This impact on the adsorption process is not serious as long as the concentration of the background ions are kept constant. It will, however, cause some complications to attempts to understand or quantify the adsorption strength of the oxyanion. When comparing adsorption isotherms or adsorption edges of adsorption envelopes, it is helpful if the oxyanions are compared at the same ionic strength and initial conditions. This impact is shown in Figure 3a at low pH, where the concentration of weakly competitive aqueous Cl^- anions has increased to the point of rendering it significantly competitive with the carbonate oxyanion. The source of the Cl^- anion at low pH is from the addition of acids (namely, HCl) needed to achieve the low-pH conditions.

Competition between anions is also shown in Figure 3a, where the presence of selenate significantly reduces the adsorption of carbonate. Competition between oxyanions is commonly observed. Competitive behavior should be expected due to the structural similarities of the oxyanions and the similarity of their adsorption sites. High Cl^- concentrations will also lower the amount adsorbed over a broad pH range due to competitive adsorption.

Third, the reactivity of the background anion (e.g., Cl^-) with the surface may be pH-dependent. Note that the first two problems mentioned above assume, for simplicity, that the reactivity of the background anion is pH-independent. However, this is perhaps one of the most serious problems that can arise in the study of oxyanion-adsorption processes. In these situations, the reactivity of the background anion must be monitored if the adsorption reactivity of the oxyanion with the solid surface is to be properly understood.

Fourth, some anions may be present in the mixture that actually enhance the adsorption reaction of the aqueous oxyanion with the solid surface. Note that this is a counterintuitive process because traditional views do not have an anion causing an increase in adsorption of another similarly charged ion. Figure 2 shows a sharp increase in adsorption of sulfate on an aluminum oxide in the presence of carbonate. Figure 3b and c shows that this enhancing effect is also pH-dependent. The enhancement of oxyanion adsorption, such as sulfate and selenate adsorption on aluminum oxide and iron oxide, has been observed in the presence of carbonates, acetate, and formate anions. At high concentrations, these anions seem to become competitive. Hence, an enhancing effect does not exclude the existence of a concurrent competitive effect. The enhancing effect is most pronounced at low concentration levels, where any possible competitive effect would also be at its lowest level (Figure 4). It is

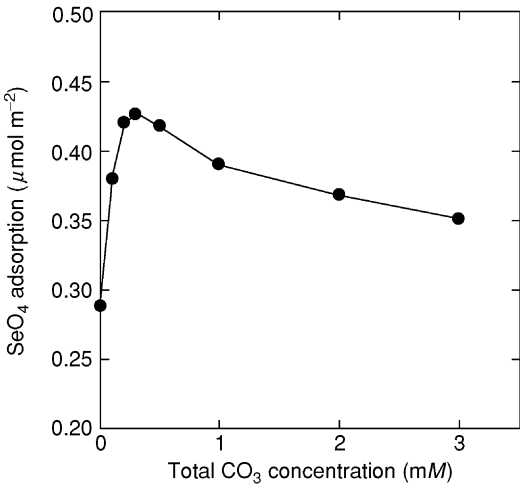


Figure 4 Adsorption of 1 m*M* selenate at pH 6.8 on an aluminum oxide (bayerite) as a function of the amount of carbonate added. Notice that the presence of carbonates increases the adsorption of selenate at low carbonate concentrations. (Reprinted from Wijnja H and Schulthess CP (2000) Interaction of carbonate and organic anions with sulfate and selenate adsorption on an aluminum oxide. *Soil Science Society of America Journal* 64: 898–908.)

not clear how prevalent the enhancing effect is in nature or in laboratory studies. We do know, however, that carbonates are nearly always present in aquatic systems. Their concentration in laboratory solutions varies widely depending on the pure water sources used or on the techniques used to minimize or control the carbonate concentrations. In other words, aqueous carbonate impurities can easily enter the sample preparations unbeknownst to the experimenters and cause a wide range in the variance of the data collected. This creates a serious problem for those that wish to understand and quantify the adsorption of oxyanions by solid surfaces in aqueous mediums.

Sorption Mechanisms of Oxyanions

The adsorption reaction of oxyanions at a mineral surface is described as a surface complexation reaction between the oxyanion and a surface functional group forming a surface complex. With respect to the types of surface complexes or adsorption mechanisms, there are commonly two major types distinguished: inner-sphere and outer-sphere complexes. An inner-sphere complex is formed via a ligand-exchange reaction with a surface-OH or surface-OH_2^+ group and is directly coordinated with the underlying surface metal ion. In a monodentate inner-sphere complex, one bond is formed with a surface group. In a bidentate complex, two bonds are formed, either with one surface metal ion (mononuclear) or with two surface metal ions (binuclear). In contrast, outer-sphere complexes form mainly by electrostatic interactions and at least one water molecule is interposed between the oxyanion and the mineral surface. Diffuse-ion swarm ions do not form a complex with a surface functional group, but instead are only bound through electrostatic interactions that neutralize the surface charge. Spectroscopic tools are essential for the development of detailed descriptions of the adsorbed oxyanion. A mechanistic description of the adsorption process requires spectroscopic information of the adsorbed species, plus information on the stoichiometry of the reaction and reaction rates, as well as a measure of the adsorption strength.

Oxyanions with a relatively high adsorption affinity such as PO_4^{3-}, AsO_4^{3-}, SeO_3^{2-}, and $B(OH)_4^-$ form predominantly inner-sphere complexes on metal oxide surfaces. Both monodentate and bidentate complexes have been found on metal oxides for several of these oxyanions. For example, it has been shown that the relative distribution of monodentate and bidentate inner-sphere complexes of AsO_4^{3-} and CrO_4^{2-} varies with the surface coverage.

The CO_3^{2-} anion forms monodentate inner-sphere complexes at metal oxide–water interfaces. Oxyanions with low affinity, such as NO_3^- and ClO_4^-, absorb mainly as outer-sphere and diffuse-swarm ions.

Several anions that have high or moderate sorption affinity, such as AsO_3^{3-}, SeO_4^{2-}, and SO_4^{2-}, show intermediate behavior, where both inner- and outer-sphere surface complexes coexist at metal oxide–water interfaces. For these moderately adsorbing anions, the relative distribution of inner- versus outer-sphere adsorption depends on pH, ionic strength, and type of mineral.

When an anion is held as an inner-sphere compound, then it has probably been preceded by an initial outer-sphere attraction to the mineral surface. The mechanism of forming an inner-sphere bond must involve the release of a molecule of water from the solid–liquid interface into the bulk solution. A net volume change occurs to the system whenever this process occurs, and this can be measured with a dilatometer. Based on measured volume changes, there is approximately one molecule of H_2O released into the bulk solution for every aqueous molecule of SO_4^{2-}, SeO_4^{2-}, or $H_2PO_4^-$ adsorbed onto an amorphous iron hydroxide at pH 4.5, which should be expected for inner-sphere adsorption of these oxyanions.

Anion adsorption is accompanied by proton uptake. The net proton uptake is the result of a ligand exchange reaction (OH release), surface protonation, oxyanion protonation on the surface, or a combination of these. Quantifying the number of protons coadsorbed (or OH released) by the surface per molecule of oxyanion adsorbed is an important component of our understanding of the adsorption process. This stoichiometry of the reaction directly affects how the surface reaction will be described, such as the 1:1 stoichiometry described in eqn [3]. This, in turn, also affects how the surface complexation model will be developed. The best way to obtain data on the adsorption or coadsorption of protons is by an acid–base titration method known as the backtitration technique.

Anion adsorption affects the electrostatic properties of mineral surfaces. The net surface charge of a mineral surface is the result of proton-transfer reactions and surface-complexation reactions. Electrophoretic mobility measurements can reveal the net surface charge changes due to oxyanion adsorption. In general, the adsorption of an oxyanion tends to lower the positive charge present on a surface. The positively charged surface does not necessarily become negatively charged following the adsorption of oxyanions. Instead, it tends to approach a net neutral surface charge. Taking these data in conjunction with the proton coadsorption data mentioned above, the adsorption of oxyanions and proton

coadsorption tends to form net neutral products on the mineral surfaces.

The adsorption of oxyanions generally reaches a maximum at around the pK_a value for the conjugate Bronsted acid, as shown in Figure 3a for carbonate adsorption ($pK_a = 6.3$). One can generally infer that the higher the value of the oxyanion's acidity constant (pK_a), the stronger will be its adsorption by a hydrophilic surface. This further reflects the importance of the coadsorption of protons on the adsorption mechanism of oxyanions. The mechanism of bringing the coadsorbed protons to the mineral surface does not necessarily follow a separate, independent path from that taken by the adsorbing oxyanion. The proton can be brought toward the mineral surface by the protonated oxyanions. The fate of the coadsorbed proton may be: (1) to remain complexed with the adsorbed oxyanion (hence, the oxyanion is held as a Bronsted acid on the mineral surface); (2) to remain complexed with the surface (hence, the oxyanion's negative charge is neutralized by a positively charged surface-OH_2^+ surface, which may or may not be directly below or adjacent to the adsorbing oxyanion); or (3) it may react with a surface-OH group and form an H_2O product that diffuses back to the bulk solution (this may also be described schematically as a hydroxyl exchange mechanism, where the hydroxyl is neutralized by free protons in the aqueous phase).

Site Preference and Site Reactivity

Surface hydroxyl groups constitute the complexation sites on metal oxide surfaces suspended in aqueous solutions. The surface groups behave as weak Bronsted acids or bases, capable of ionizing to yield a pH-dependent charged surface or serving as a ligand in surface complexation reactions. Analysis of the crystal structure of metal oxides indicates the existence of various different types of surface hydroxyl groups with different reactivities. Their reactivity depends, among other factors, on their coordination environment (singly, doubly, and triply metal-coordinated). The different crystal planes of a metal oxide mineral have a different composition with respect to the types of surface groups and densities of these groups.

With respect to complexation of oxyanions, the singly coordinated surface hydroxyls are believed to be most reactive. An FTIR study on the effect of As oxyanions on the various different types of surface groups has indicated that the singly coordinated surface groups are involved in the ligand exchange reaction to form inner-sphere complexes.

There are several things that the adsorbed oxyanion can do after it has adsorbed to the mineral surface. It can, of course, desorb from the surface and diffuse back to the aqueous bulk solution. Given the high energy of hydration of many of these oxyanions, this desorption step is very likely and is the basis for establishing the equilibrium condition where only a specific density of anions remain adsorbed to the surface for a given concentration of anions in solution.

The adsorbed oxyanion can also diffuse into the mineral through a process known as solid diffusion (product is shown in Figure 1b). Solid diffusion at ambient temperatures is a very slow process. However, the oxyanion may also became entrapped inside the mineral through the process of oxyanion-enhanced dissolution of the mineral followed by coprecipitation of the oxyanion–metal complex.

A final option for the adsorbed oxyanion is to migrate laterally, traveling horizontally on the mineral surface along a two-dimensional matrix. This is dependent on the strength of the bond of the oxyanion with the surface. The lateral, two-dimensional mobility of the oxyanion allows it to migrate toward a more stable adsorption site (more stable, that is, relative to its point of initial collision and reaction with the surface). Lateral mobility of adsorbed ions is known to occur on solid–gas interfaces, but data are missing that accurately quantify this mobility on solid–liquid interfaces.

See also: **Cation Exchange**

Further Reading

Fendorf S, Eick MJ, Grossl P, and Sparks DL (1997) Arsenate and chromate retention mechanisms on goethite. 1. Surface structure. *Environmental Science and Technology* 31: 315–320.

Nakamoto K (1997) *Infrared and Raman Spectra of Inorganic and Coordination Compounds*, 5th edn. New York: John Wiley & Sons, Inc.

Schulthess CP and Sparks DL (1986) Backtitration technique for proton isotherm modeling of oxide surfaces. *Soil Science Society of America Journal* 50: 1406–1411.

Schulthess CP and Sparks DL (1991) Equilibrium based modeling of chemical sorption on soils and soil constituents. *Advances in Soil Science* 16: 121–163.

Schulthess CP and Hu Z (2001) Impact of carbonate on the adsorption of selenate and sulfate on goethite. *Soil Science Society of America Journal* 66: 1190–1197.

Sun XH and Doner HE (1996) An investigation of arsenate and arsenite bonding structures on goethite by FTIR. *Soil Science* 161: 865–872.

Yamaguchi NU, Okazaki M, and Hashitani T (1999) Volume changes due to SO_4^{2-}, SeO_4^{2-}, and $H_2PO_4^-$ adsorption on amorphous iron (III) hydroxide in an aqueous suspension. *Journal of Colloid and Interface Science* 209: 386–391.

SORPTION–DESORPTION, KINETICS

D L Sparks, University of Delaware, Newark, DE, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

In the past two decades, as concerns and interests about soil and water quality have increased, soil and environmental chemists, environmental and chemical engineers, and geochemists have increasingly realized that sorption/desorption reactions in soil and other subsurface environments are time-dependent. Thus, to predict accurately the fate, mobility, speciation, and bioavailability of environmentally important plant nutrients, trace elements, radionuclides, and organic chemicals in soils, one must understand the kinetics and mechanisms of the reactions. This article will provide an overview of the topic of kinetics of sorption/desorption on soils with applications to environmentally important reactions.

Rate-Limiting Steps and Time Scales for Sorption Processes

A number of transport and chemical reaction processes can affect the rate of soil chemical reactions. The slowest of these will limit the rate of a particular reaction. The actual chemical reaction (CR) at the surface, for example, adsorption, is usually very rapid and not rate-limiting. Transport processes (Figure 1) include: (1) transport in the solution phase, which is rapid, and in the laboratory, can be eliminated by rapid mixing; (2) transport across a liquid film at the particle/liquid interface (film diffusion (FD)); (3) transport in liquid-filled macropores (>2 nm), all of which are nonactivated diffusion processes and occur in mobile regions; (4) diffusion of a sorbate along pore wall surfaces (surface diffusion); (5) diffusion of sorbate occluded in micropores (<2 nm) (pore diffusion); and (6) diffusion processes in the bulk of the solid, all of which are activated diffusion processes. Pore and surface diffusion can be referred to as interparticle diffusion while diffusion in the solid is intraparticle diffusion.

The energy of activation value (E_a), which can be obtained by employing the Arrhenius equation ($\ln k = \ln A_f - E_a/RT$ where k is the rate constant, A_f is the frequency factor, R is the universal gas constant, and T is absolute temperature), is a useful kinetic parameter to assess rate-limiting steps. Some E_a values for various reactions or processes are shown in Table 1. Low E_a values usually indicate diffusion-controlled transport and physical adsorption processes, whereas higher E_a values would indicate chemical reaction or surface-controlled processes.

Sorption reactions can range over a wide time scale – from seconds to years. The type of soil component can drastically affect the reaction rate. For example, sorption reactions are often more rapid on clay minerals such as kaolinite and smectites than on vermiculitic and micaceous minerals. This is in large part due to the availability of sites for sorption. For example, kaolinite has readily available planar external sites and smectites have primarily internal sites that are also quite available for retention of sorbates. Thus, sorption reactions on these soil constituents are often quite rapid, even occurring on time scales of seconds and milliseconds.

On the other hand, vermiculite and micas have multiple sites for retention of metals and organics, including planar, edge, and interlayer sites, with some of the latter sites being partially to totally collapsed. Consequently, sorption and desorption reactions on these sites can be slow, tortuous, and mass transfer-controlled. Often, an apparent equilibrium may not be reached even after several days or weeks. Thus, with vermiculite and mica, sorption can involve two to three different reaction rates: high rates on external sites, intermediate rates on edge sites, and low rates on interlayer sites.

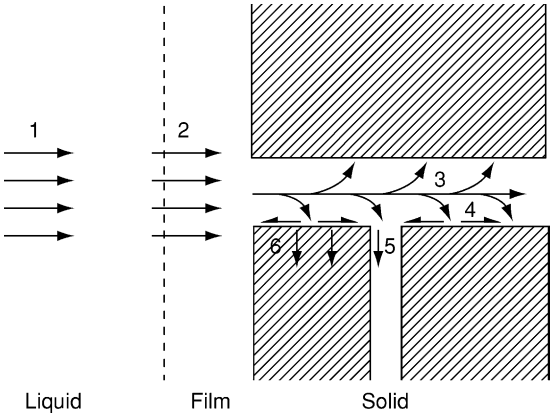


Figure 1 Transport processes in solid–liquid soil reactions – nonactivated processes: (1) transport in soil solution; (2) transport across a liquid film at the solid–liquid inerface; (3) transport in a liquid-filled macropore; activated processes; (4) diffusion of a sorbate at the surface of the solid; (5) diffusion of a sorbate occluded in a micropore; (6) diffusion in the bulk of the solid. Reproduced with permission from Aharoni C and Sparks DL (1991) *Soil Science Society of America Special Publication* 27.

Table 1 Energy of activation (E_a) values for various reactions or processes

Reaction of process	*Typical range of E_a values (kJ mol^{-1})*
Physical adsorption	8 to 25
Aqueous diffusion	<21
Pore diffusion	20 to 40
Cellular and life-related reactions	21 to 84
Mineral dissolution or precipitation	34 to 151
Mineral dissolution via surface reaction control	42 to 84
Polymer diffusion	60 to 70
Ion exchange	>84
Isotopic exchange in solution	75 to 201
Solid-state diffusion in minerals at low temperatures	84 to 503

Reproduced with permission from Lasaga AC and Kirkpatrick RJ (eds) (1981) *Kinetics of Geochemical Processes.* Washington, DC: Mineralogical Society of America; and Langmuir D (1997) *Aqueous Environmental Geochemistry.* Englewood Cliffs, NJ: Prentice Hall.

Metal sorption reactions on oxides, hydroxides, and humic substances depend on the type of surface and metal being studied, but the CR appears to be rapid. For example, CR rates of metals and oxyanions on goethite occur on millisecond time scales. Half-times for divalent Pb, Cu, and Zn sorption on peat range from 5 to 15 s. Heavy metal sorption on oxide and clay minerals increases with longer residence times (contact time between metal and sorbent). The mechanisms for these lower reaction rates are not well understood, but have been ascribed to diffusion phenomena, sites of lower reactivity, and surface nucleation/precipitation.

Sorption/desorption of metals, oxyanions, radionuclides, and organic chemicals on soils can be very slow, and may demonstrate a residence time effect, which has been attributed to diffusion into micropores of inorganic minerals and into humic substances, retention on sites of varying reactivity, and surface nucleation/precipitation.

Sorption/Desorption Kinetics of Metals

The kinetics of heavy-metal sorption on soil components and soils is typically characterized by a biphasic process in which sorption is initially rapid followed by slow reactions (Figure 2). The rapid step, which occurs over milliseconds to hours, can be ascribed to CR and film diffusion processes. During this rapid reaction process, a large portion of the sorption may occur. For example, in Figure 2a one sees that ~90% of the total Ni sorbed on kaolinite and pyrophyllite occurred within the first 24 h. For Pb sorption on a Matapeake soil, 78% of the total Pb sorption

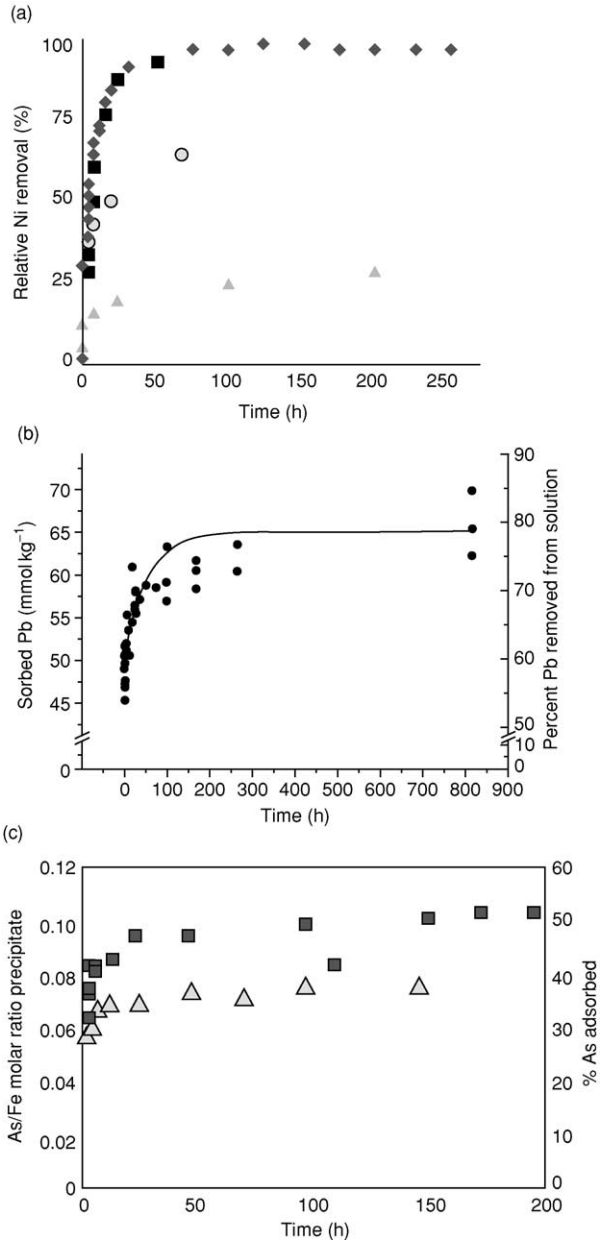


Figure 2 Kinetics of metal and oxyanion sorption on soil minerals and soil. (a) Kinetics of Ni sorption (%) on pyrophyllite (◆); kaolinite (■); gibbsite (▲); and montmorillonite (●) from a 3 mmol l^{-1} Ni solution, an ionic strength $I = 0.1$ mol l^{-1} $NaNO_3$, and a pH of 7.5. Reproduced with permission from Scheidegger AM, Lamble GM, and Sparks DL (1997) Spectroscopic evidence for the formation of mixed-cation, hydroxide phases upon metal sorption on clays and aluminum hydroxides. *Journal of Colloid and Inter. Science* 186: 118–128. (b) kinetics of sorption on a Matapeake soil from a 12.25 mmol l^{-1} Pb solution, an ionic strength $I = 0.05$ mol l^{-1}, and a pH of 5.5. Reproduced with permission from Strawn DG and Sparks D (2000) Effects of soil organic matter on the kinetics and mechanisms of Pb(II) sorption and desorption in the soil. *Soil Science Society of America Journal* 64: 144–156; (c) kinetics of As(V) sorption on ferrihydrite at pH 8.0 and 9.0. As(V) Total 1×10^{-4} mol l^{-1}; Fe(III) Total 5×10^{-4} mol l^{-1}; ■ Adsorption at pH 8.0; △ Adsorption at pH 9.0. Reproduced with permission from Juller CC, Davis JA, and Waychunas GA (1993) Surface chemistry of ferrihydride: part 2, kinetics of arsenate adsorption and co-precipitation. *Geochimica Cosmochimica Acta* 57: 2271–2282.

occurred in 8 min. Following the initial fast reaction, slow sorption continued, but only about 1% additional Pb was sorbed after 800 h (Figure 2b). Figure 2c shows a biphasic reaction for As(V) sorption on ferrihydrite. Within 5 min, a majority of the total sorption had occurred. Slow sorption continued for at least 192 h.

However, in some cases, the magnitude of sorption can greatly increase with longer reaction times. For example, it has been found that metal (e.g., Cd, Ni, and Zn) sorption on porous Fe-oxides such as goethite can significantly increase as time increases. These observations have been ascribed to a three-step mechanism: (1) adsorption of metals on external surfaces; (2) solid-state diffusion of metals from external to internal sites; and (3) metal binding and fixation at positions inside the goethite particle.

The amount of contact time between metals and soil sorbents (residence time) can dramatically affect the degree of desorption, depending on the metal and sorbent. Examples of this are shown in Figure 3 and Table 2. In Figure 3 the effect of residence time on Pb^{2+} and Co^{2+} desorption from hydrous Fe-oxide (HFO) was studied. With Pb, between pH 3 and 5.5 there was a minor effect of residence time (over 21 weeks) on Pb^{2+} desorption, with only minor hysteresis occurring (hysteresis varied from <2% difference between sorption and desorption to ~10%). At pH 2.5, Pb^{2+} desorption was complete within a 16-h period and was not affected by residence time (Figure 3a). In a soil where 2.1% soil organic matter (SOM) was present, residence time had little effect on the amount of Pb desorbed, but marked hysteresis was observed at all residence times (Table 2). This could be ascribed to the strong metal–soil complexes that occur and perhaps to diffusion processes.

With Co^{2+}, extensive hysteresis was observed over a 16-week residence time (Figure 3b) and the hysteresis increased with residence time. After a 16-week residence time, 53% of the Co^{2+} was not desorbed, and even at pH 2.5, hysteresis was observed. The extent of Co reversibility with residence time was attributed to Co incorporation into a recrystallizing solid by isomorphic substitution and not to micropore diffusion. Similar residence time effects on Co desorption have been observed with both Fe- and Mn-oxides and with soil clays.

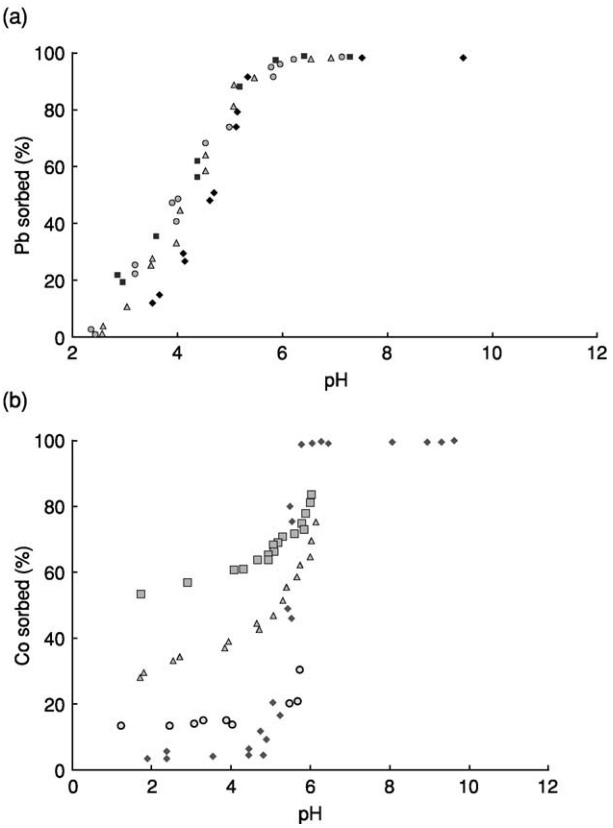


Figure 3 Fractional sorption–desorption of (a) Pb^{2+} (adsorption pH edge ◆ aging time = 0 weeks; desorption ■ aging time = 2 weeks; △ aging time = 12 weeks; ● aging time = 21 weeks) and (b) Co^{2+} (adsorption pH edge ◆ aging time = 0 weeks; desorption ○ aging time = 2 weeks; △ aging time = 9 weeks; ■ aging time = 16 weeks) to hydrous Fe(III) as a function of pH and HFO-Pb^{2+} (a) and HFO-Co^{2} (b) aging time. From Ainsworth CC, Pilou JL, Gassman PL, and Van Der Sluys WG (1994) Cobalt, cadmium, and lead sorption to hydrous iron oxide: residence time effect. *Soil Science Society of America Journal* 58: 1615–1623.

Table 2 Effect of residence time on Pb desorption from a Matapeake soil

Residence time (days)	*Sorbed Pb (mmol kg^{-1})*	*Desorbed Pb (mmol kg^{-1})*	*Percentage Pb desorbed (mmol kg^{-1})*
1	54.9	27.9	50.8
10	60.0	28.7	47.0
32	66.1	30.5	46.1

Reproduced with permission from Strawn DG and Sparks DL (2000) Effects of soil organic matter on the kinetics and mechanisms of Pb(II) sorption and desorption in soils. *Soil Science Society of America Journal* 64: 144–156.

Kinetics of Metal Hydroxide Surface Precipitation/Dissolution

In addition to diffusion processes, the formation of metal hydroxide surface precipitates and subsequent residence time effects on soil sorbents can greatly affect metal release and hysteresis. Metal hydroxide precipitates can form on phyllosilicates, metal oxides, and soils. It has generally been thought that the kinetics of formation of surface precipitates was slow. However, recent studies have shown that metal hydroxide precipitates can form on time scales

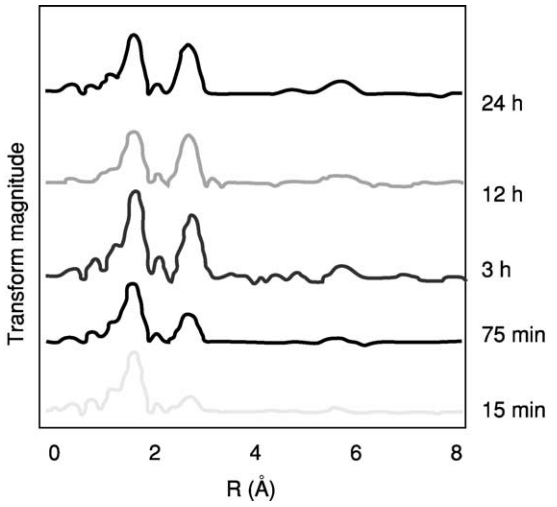


Figure 4 Radial structure functions (derived from X-ray absorption fine structure (XAFS) analyses) for Ni sorption on pyrophyllite for reaction times up to 24 h, demonstrating the appearance and growth of the second shell (peak at approximately 2.8 Å) contributions due to surface precipitation and growth of a mixed Ni-Al hydroxide phase. Reproduced with permission from Scheidegger AM, Strawn DG, Lamble GM, and Sparks DL (1998) The kinetics of mixed Ni-Al hydroxide formation on clay and aluminum oxide minerals: a time-resolved XAFS study. *Geochimica Cosmochimica Acta* 62: 2233–2245.

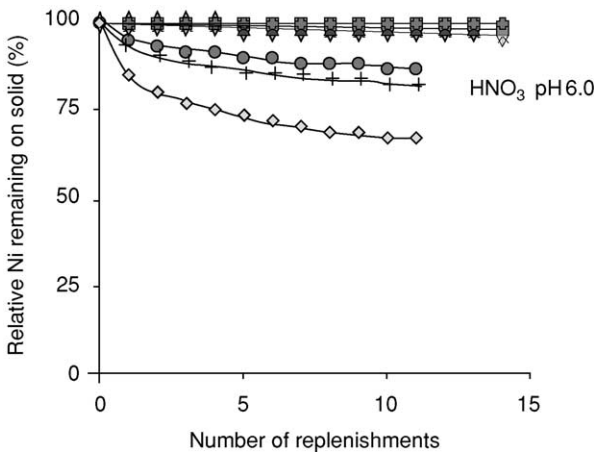


Figure 5 Dissolution of Ni from surface precipitates formed on pyrophyllite at residence times of 1 h to 2 years. The figure shows the relative amount of Ni^{2+} remaining on the pyrophyllite surface following extraction for a 24-h period (each replenishment represents a 24-h extraction) with HNO_3 at pH 6.0. ✥ 2 year; ■ 1 year; △ 6 months; × 3 months; ✥ 1 month; ● 24 h; + 12 h; ◇ 1 h. Reproduced with permission from Scheckel KG and Sparks DL (2001) Dissolution kinetics of Ni surface precipitates on clay mineral and oxide surfaces. *Soil Science Society of America Journal* 65: 719–728.

of minutes. In Figure 4 one sees that mixed Ni-Al hydroxide precipitates formed on pyrophyllite within 15 min, and they grew in intensity as time increased. Similar results have been observed with other soil components and with soils.

The formation and subsequent 'aging' of the metal hydroxide surface precipitate can have a significant effect on metal release. In Figure 5 one sees that as residence time ('aging') increases from 1 h to 2 years, Ni release from pyrophyllite, as a percentage of total Ni sorption, decreased from 23 to ~0%, when HNO_3 (pH 6.0) was employed as a dissolution agent for 14 days. This enhanced stability is due to the transformation of the metal-Al hydroxide precipitates to a metal-Al phyllosilicate precursor phase as residence time increases. This transformation occurs via a number of steps (Figure 6). There is diffusion of Si originating from weathering of the sorbent into the interlayer space of the layered double hydroxide precipitate (LDH), replacing the anions such as NO_3^-. Polymerization and condensation of the interlayer Si slowly transform the LDH into a precursor metal-Al phyllosilicate. The metal stabilization that occurs in surface precipitates on Al-free sorbents (e.g., talc) may be due to Ostwald ripening, resulting in increased crystallization. Thus, with time, metal (e.g., Co, Ni, and Zn) sorption on soil minerals often results in a continuum of processes from adsorption to precipitation to solid-phase transformation.

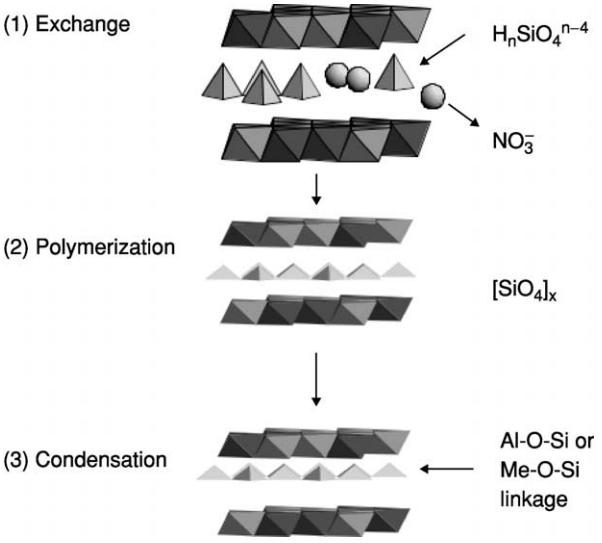


Figure 6 Hypothetical reaction process illustrating the transformation of an initially precipitated Ni-Al LDH to a phyllosilicate like phase during aging. The initial step involves the exchange of dissolved silica for nitrate within the layered double hydroxide (LDH) interlayer followed by polymerization and condensation of silica on to the octahedral Ni-Al layer. The resultant solid possesses structural features common to 1:1 and 2:1 phyllosilicates. Reproduced with permission from Ford RG, Scheinost AC, and Sparks DL (2001) Frontiers in metal sorption/precipitation mechanisms on soil mineral surfaces. *Advances in Agronomy* 74: 41–62.

The formation of metal surface precipitates could be an important mechanism for sequestering metals in soils such that they are less mobile and bioavailable.

Such products must be considered when modeling the fate and mobility of metals such as Co^{2+}, Mn^{2+}, Ni^{2+}, and Zn^{2+} in soil and water environments.

Kinetics of Organic Contaminant Sorption/Desorption

Similarly to metals and oxyanions, many investigations have shown that sorption/desorption of organic contaminants is characterized by a rapid, reversible stage followed by a much slower, nonreversible stage or biphasic kinetics. The rapid phase has been ascribed to retention of the organic chemical in a labile form that is easily desorbed. The labile form of the chemical is also available for microbial attack. However, the much slower reaction phase involves the entrapment of the chemical in a nonlabile form that is difficult to desorb and is resistant to biodegradation. This slower sorption/desorption reaction has been ascribed to diffusion of the chemical into micropores of organic matter and inorganic soil components.

Recent theories have explained the slow diffusion of the organic chemical in SOM by considering SOM as a combination of 'rubbery' and 'glassy' polymers. The rubbery-like phases are characterized by an expanded, flexible, and highly solvated structure with pores of subnanometer dimensions (holes). Sorption in the rubbery phase results in linear, noncompetitive, and reversible behavior. The glassy phases have pores that are of subnanometer size and sorption in this phase is characterized by nonlinearity and is competitive.

The above theories, relating slow diffusion into organic matter to diffusion in polymers, are partially validated in recent studies measuring E_a values for organic chemical sorption. E_a values of 60–70 kJ mol^{-1} have been determined for some chlorobenzenes, polychlorinated biphenyls (PCBs) and polycyclic aromatic hydrocarbons (PAHs) which are in the range for diffusion in polymers (Table 1). These values are much higher than those for pore diffusion (20–40 kJ mol^{-1}), suggesting that intraorganic matter diffusion may be a more important mechanism for slow organic chemical sorption than interparticle pore diffusion.

An example of the biphasic kinetics that is observed for many organic chemical reactions in soils/sediments is shown in Figure 7. In this study 55% of the labile PCBs was desorbed from sediments in a 24-h period, while little of the remaining 45% nonlabile fraction was desorbed in 170 h (Figure 7a). Over another 1-year period about 50% of the remaining nonlabile fraction desorbed (Figure 7b).

Similarly to some metals, it has been demonstrated that with increased residence time the nonlabile portion

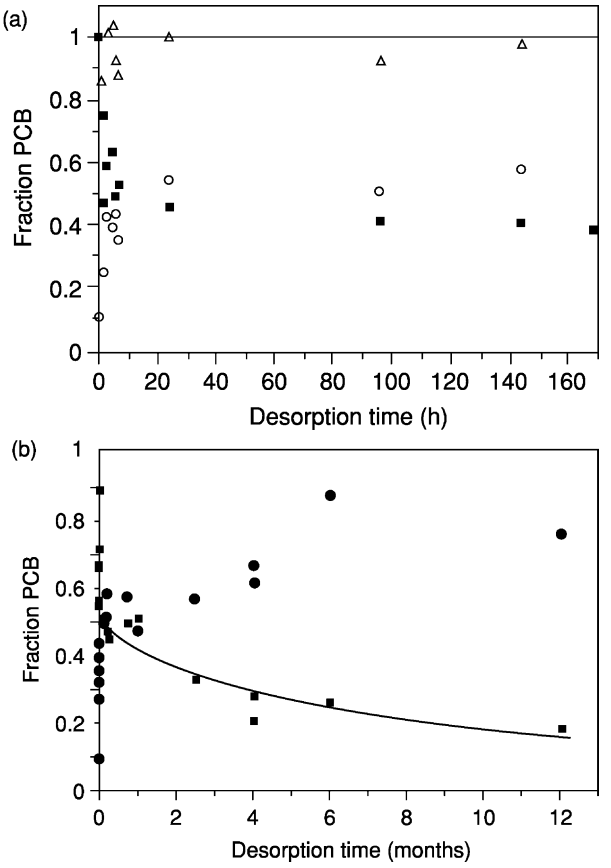


Figure 7 (a) Short-term polychlorinated biphenyl (PCB) desorption in hours (h) from Hudson river sediment contaminated with 25 mg kg^{-1} PCB. Distribution of the PCB between sediment (■) and XAD-4 resin (○) is shown, as well as the overall mass balance (△). The resin acts as a sink to retain the PCB that is desorbed. (b) Long-term PCB desorption in months from Hudson river sediment contaminated with 25 mg kg^{-1} PCB. Distribution of the PCB between the sediment (■) and XAD-4 resin (●) is shown. The line represents a nonlinear regression of the data by the two-box (site) model. Reproduced with permission from Carroll KM, Harkness MR, Bracco AA, and Balcarcel RB (1994) Application of a permeant/polymer diffusional model to the desorption of polychlorinated biphenyls from Hudson river sediments. *Environmental Science Technology* 28: 253–258.

of the organic chemical in the soil/sediment becomes more resistant to release and is less available for microbial degradation. Soils reacted in the laboratory over a short period of time often show more rapid organic chemical release than field soils that have been contaminated for many years. The difference in release has been ascribed to greater diffusion into micropores of clay minerals and humic components that occurred at longer times.

One of the implications of these results is that while many transport and degradation models for organic contaminants in soils and water assume that the sorption process is an equilibrium process, the above

studies clearly show that kinetic reactions must be considered when making predictions about the mobility and fate of organic chemicals. Moreover, K_d values that are determined based on a 24-h equilibration period and which are commonly used in fate and risk assessment models can be inaccurate since 24-h K_d values often overestimate the amount of organic chemical in the solution phase.

The finding that many organic chemicals are quite persistent in the soil environment has both good and bad features. The beneficial aspect is that the organic chemicals are less mobile and may not be readily transported in groundwater supplies. The negative aspect is that their persistence and inaccessibility to microbes may make decontamination more difficult, particularly if *in situ* remediation techniques such as biodegradation are employed.

Conclusions and Future Directions

This article has discussed various basic and applied aspects of kinetics of sorption/desorption processes on soils and soil components including: rate-limiting steps and time scales for sorption processes, sorption/desorption kinetics of metals, kinetics of metal hydroxide surface precipitation/dissolution, and kinetics of organic contaminant sorption/desorption.

While great strides have been made in our understanding of kinetics and mechanisms of sorption processes on soils, a number of areas need further investigation and will be undoubtedly frontiers in the field. These include: the role of microbes on sorption kinetics and mechanisms at the soil/mineral water interface; the coupling of macroscopic studies with molecular-scale tools to study sorption reactions in real time; development of more predictive kinetic models to describe sorption processes on multireactive, heterogeneous soils; long-term sorption and, particularly, desorption rate studies; and increased knowledge of nucleation/precipitation phenomena at the soil/water interface and their impact on nutrient/contaminant mobility and bioavailability in the soil environment.

See also: **Cation Exchange**; **Sorption:** Metals; Oxyanions

Further Reading

Aharoni C and Sparks DL (1991) Kinetics of soil chemical reactions: a theoretical treatment. In: Sparks DL and Suarez DL (eds) *Rates of Soil Chemical Processes*. SSSA special publication no. 27, pp. 1–18. Madison, WI: Soil Science Society of America.

Ainsworth CC, Pilou JL, Gassman PL, and Van Der Sluys WG (1994) Cobalt, cadmium, and lead sorption to hydrous iron oxide: residence time effect. *Soil Science Society of America Journal* 58: 1615–1623.

Alexander M (2000) Aging, bioavailability, and overestimation of risk from environmental pollutants. *Environmental Science Technology* 34: 4259–4265.

Bruemmer GW, Gerth J, and Tiller KG (1988) Reaction kinetics of the adsorption and desorption of nickel, zinc and cadmium by goethite: I. Adsorption and diffusion of metals. *Journal of Soil Science* 39: 37–52.

Pignatello JJ (2000) The measurement and interpretation of sorption and desorption rates for organic compounds in soil media. *Advances in Agronomy* 69: 1–73.

Sparks DL (1989) *Kinetics of Soil Chemical Processes*. San Diego, CA: Academic Press.

Sparks DL (1998) Kinetics of sorption/release reactions on natural particles. In: Huang PM, Senesi N, and Buffie J (eds) *Structure and Surface Reactions of Soil Particles*, pp. 413–448. New York: John Wiley.

Sparks DL (1999) Kinetics and mechanisms of chemical reactions at the soil mineral/water interface. In: Sparks DL (ed.) *Soil Physical Chemistry*, 2nd edn, pp. 135–191. Boca Raton, FL: CRC Press.

Sparks DL and Suarez DL (eds) (1991) *Rates of Soil Chemical Processes*. SSSA special publication no. 27. Madison, WI: Soil Science Society of America.

Sposito G (1994) *Chemical Equilibria and Kinetics in Soils*. New York: John Wiley.

SPATIAL PATTERNS

J H Görres and J A Amador, University of Rhode Island, Kingston, RI, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Biological processes in the soil and associated properties are highly variable at different spatial scales. The sources of spatial variability in soils include both natural and anthropogenic processes that may act at microscopic, aggregate, field, landscape, or regional scale. Together these factors create complex mosaics of environments affecting biological processes at different scales. Identifying the sources of spatial heterogeneity and quantifying the heterogeneity of biological processes in soils are challenging. However, once understood, spatial heterogeneity may offer management options that affect ecosystem productivity, biodiversity, and soil and environmental quality.

Background

Soil biological processes provide crucial feedback between belowground decomposition and aboveground processes such as productivity, herbivory, and diversity. Soil organisms recycle much of primary productivity in terrestrial ecosystems, contributing to sustained ecosystem growth. At the ecosystem level, the interaction between decomposers and producers is well understood. However, high spatial heterogeneity of soil biological processes may obscure ecologic relationships at a finer scale. For this reason, the origins of spatial variability in soil biological processes have themselves become an important topic of research.

Soil biological processes and their spatial variability have gained practical environmental and economic recognition. For example: bioremediation relies on soil microorganisms that can metabolize organic pollutants; precision agriculture attempts to reduce the amount of agrochemicals applied by variable applications that depend on the spatial distribution of nutrients in a field; and riparian wetland buffers promote denitrification, which reduces nitrate loading to streams, lakes, and estuaries. Understanding soil spatial variability is essential to successfully managing soil biological processes to improve environmental quality.

Organisms are at the root of many soil processes, influencing the development of physical and chemical soil properties. The close relationship between soil organisms and their habitat confounds a clear distinction between biological and physicochemical soil properties. For example, soil porosity is a physical soil property, but biogenic mechanisms such as aggregation and bioturbation contribute to its development. Components of the nitrogen cycle affect soil pH. Conversely, soil biological processes are affected by physicochemical properties such as water content and pH. Here soil biological processes are defined as those that are clearly associated with the distribution of organisms, such as microbial biomass and functional diversity, and their metabolic activity, such as decomposition and nutrient mineralization.

A common characteristic of soil biological processes is their high spatial variability. Depending on sampling intensity, coefficients of variation range from ten to several hundred percent. Spatial heterogeneity of soil biological processes is linked to the patchiness of the distribution of population and community level processes, and to pedological factors. It is clear that their spatial variability derives from a superimposition of processes that occur at multiple scales. Spatial variability has been categorized into variations at microsite (approx. 10^{-3}–10^{-1} m), field (10^{0}–10^{2} m), landscape (10^{1}–10^{2} m), and regional scale (greater than 10^{3}–10^{5} m). A case could also be made for adding a continental and global scale in which climatic and vegetational factors of ecoregions determine soil biological processes.

Regardless of scale, spatial variation can be attributed to a combination of random and deterministic variations. Random variations are those in which soil properties or processes are spatially independent, i.e., the property value at one point is not related to values at any other point. Deterministic variations of a property are characterized by spatial dependence. Spatial dependence develops because of processes that connect points in space, such as migration of organisms, diffusion of chemicals, or the integrative effect of plant roots. In environments favoring patchy distributions of organisms, such as soils, spatial dependence includes within-patch and between-patch components. Statistical techniques exist to characterize the properties of both random and deterministic variations.

Sources of Variability

Pedological Factors of Soil Biological Variability

Factors influencing soil formation lay out a coarse, multiscale spatial pattern of factors that influence soil formation. Pedologists distinguish five factors, four

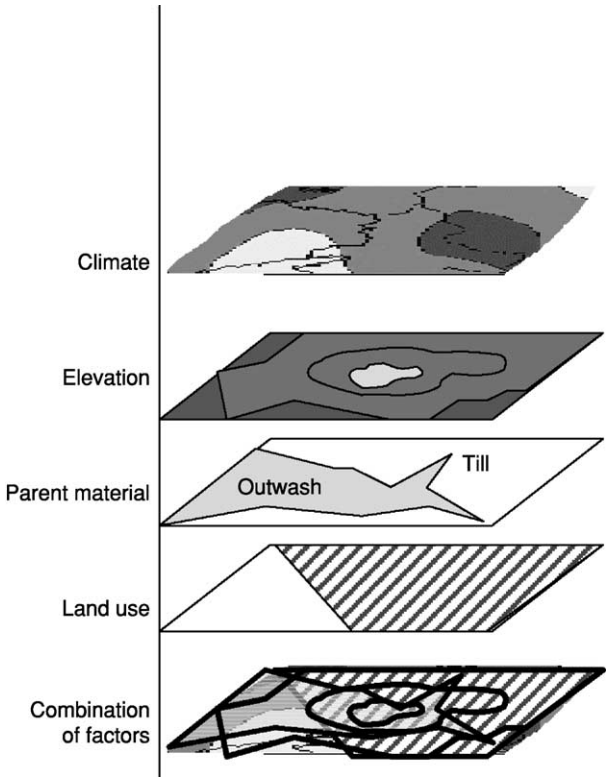


Figure 1 Superposition of intrinsic and extrinsic properties. Climate, elevation, parent material, and land use combine to form a complex soil-property pattern that affects soil biological processes.

of which are distributed at different spatial scales: parent material, climate, topography, organisms, and time. Parent material and topography are regarded as landscape-scale factors. Climate has an overarching function affecting soil biology at all scales. For example, soil moisture, an important factor of biological activity in soils, is affected by topography, i.e., landscape position, elevation, and aspect, but also by climate. At finer scales, the action of organisms may be affected by resource distribution, microclimates, and microtopography.

The superimposition of parent material, climate, and topography (Figure 1) within a region establishes a coarse pattern, which is modified by organisms to form soils, creating multiscale distributions of biological soil properties. For example, nematode distributions are affected by large-scale processes such as texture and climate, and small-scale, organism-related processes such as life-history traits. Denitrification also displays multiscale variability. Microsite characteristics such as localized anaerobiosis at the center of soil aggregates and the availability of detrital resources to denitrifying bacteria are essential for denitrification. Yet anaerobic conditions crucial to denitrification are also controlled by landscape-scale factors such as proximity of groundwater to the soil surface.

In managed ecosystems, additional factors impact the spatial heterogeneity of soil properties. Soil variability is introduced by intrinsic and extrinsic factors. Intrinsic factors are those that act in the absence of soil management and these may coincide with pedological factors of soil formation. Extrinsic factors are associated with land management. In silvicultural, agricultural, and urban ecosystems, management tools such as artificial drainage, tillage, fertilization, planting, and control of pH are superimposed on intrinsic soil processes, adding to the variability of soil biological properties and processes. For example, enzymatic activity has been shown to be high within a crop row and low in the spaces between rows. Non-point source pollution, acid deposition, and land-use patterns may add to soil variability at the landscape, regional, and continental scale.

Spatial Scales Related to Organisms

Biological processes occur at scales ranging from molecular (e.g., enzymes, extracellular polysaccharides, etc.) and organismal, to population and metapopulation. Thus, organisms create connections between points in space in several ways: their bodies occupy space; their motion spans distances between points in space; diffusion and bulk flow transport metabolic waste products (e.g., mineralized nutrients, gases); the presence and activities of larger organisms such as earthworms or plants may create spatially extended spheres of opportunity for other soil dwellers such as nematodes and microorganisms; small organisms may move by phoresis on, or in, larger motile organisms. These processes create spatial correlations that are expressed at different scales.

Whether spatial correlations due to body size are detected depends on the size of the soil sample taken. For example, the soil sample volume, also known as the sample support, is often many times greater than the dimensions of microscopic soil inhabitants, obscuring spatial correlation introduced by the body size of organisms. The longest linear dimensions of soil organisms range from less than a micrometer (microorganisms) to several meters (plants). Even within a group of organisms such as the soil fauna, sizes vary over several orders of magnitude (Table 1). The effect of plants on the spatial variability of mineral nitrogen can thus be detected, using typical soil cores (e.g., 5 cm in diameter and 10 cm long). However, the typical soil core is several orders of magnitude larger than the space occupied by a single nematode. The spatial effects of organisms many times smaller than a typical core, e.g., of a nematode, cannot be resolved

Table 1 Order of magnitude of body diameter ranges for selected soil animals in relation to pore-size classes

Pore-size range	*Faunal-size class*	*Organism*	*Order of magnitude of organism diameter (μm)*
Mesopores 1–100 μm	Microfauna	Ciliates	10^0–10^1
		Amebae	10^0–10^1
		Nematodes	10^1
	Mesofauna	Collembola	10^2–10^3
		Mites	10^2–10^3
Macropores >100 μm		Insects	
	Macrofauna	Spiders	
		Earthworms	10^3–10^4
		Millipedes	10^3–10^4

at that sampling scale. Finer-scale sample support has been used (e.g., $2.5 \times 2.5 \times 5$ mm), particularly in the investigation of genetic structure of microbial communities.

Spatial dependence related to populations and metapopulations may be more evident because these occupy more space. Indeed, the distribution of soil organisms is often clumped, introducing characteristic spatial correlations that are related to life-history traits and migration patterns. For example, entomopathogenic nematodes release their young in insect larvae. Juvenile nematodes spread from these larval contagions, creating spatial patterns that at least initially show strong aggregation around the point of contagion.

Because organisms, even when they are of dissimilar scale, are connected via trophic or symbiotic interactions, the spatial distribution of one organism can affect the distribution of another population. The rhizosphere is defined as soil attached to plant roots. It is the focus of microbial and faunal activity due to rhizodeposits, enriching the food resources of microorganisms. Similarly, the soil surrounding earthworm burrows, known as the drilosphere, has elevated abundances of soil organisms because of detrital resources translocated from the surface. As a result of the elevated biological activity and resource enrichment, carbon and nitrogen mineralization rates are enhanced.

The activities of organisms alter soil properties beyond their immediate neighborhood, thus affecting other organisms. For example, consecutive microbial transformations, e.g., ammonification followed by nitrification, may occur in different locations because molecules move from the transformation sites by diffusion or by bulk movement of water. The diffusion of substrates may link points in space, with different microorganisms acting locally. The relationship of processes at these points may not be immediately apparent, because they occur in different locations and at different times and are affected by microbial populations with different functions.

Migration of organisms introduces another scale of spatial dependence. Migration distances of microorganisms, micro-, and mesofauna are limited by the distribution of soil air and water, as well as resource distribution and transport (phoresis) by other organisms. Aquatic microfauna, soil animals that live in water-filled pores (e.g., nematodes and protozoa), and their aerial counterparts, soil animals that live in the air-filled pores (e.g., mites, Collembola), depend on connected water-filled or air-filled spaces, respectively. Unless macrofauna extend the dispersal distances, microfauna and mesofauna migrate distances up to several decimeters.

Resource distribution directs motion of chemotactic soil animals such as entomopathogenic and bacterivorous soil nematodes; but even in the absence of chemotactic adaptations, resource distribution affects the distribution of consumers. Thus, the range of motion of an organism depends on a complex set of factors such as soil type, water status, distribution of resources, and whether the organism has particular traits for finding its prey or grazing substrates. In addition, many organisms and microbiological transformations are sensitive to abiotic constraints such as pH, texture, moisture, substrate, and O_2 availability. For example, C mineralization depends on water content and quality, temperature, organic matter content, bulk density, and texture. Nematode abundance depends on similar factors. The actual scale of potential motion may thus be secondary to other factors.

Some studies aim at characterizing spatial distributions at an even smaller scale. Such studies may be conducted to investigate the distribution of organisms or soil microbial processes at aggregate scale, or genetic diversity at subaggregate scales. To investigate the potential for denitrification in unsaturated soils, microelectrodes have been used to map the distribution of oxygen availability in individual soil peds. Microbial diversity measured using genetic techniques has been conducted at the subaggregate scale ($2.5 \times 2.5 \times 5$ mm). Studies conducted at these scales are important for the elucidation of mechanisms which contribute to biodiversity, e.g., the probability of genetic exchange during bacterial encounters, or to a mechanistic understanding of biogeochemical processes.

Quantification of Spatial Properties

First-Order Statistics and Spatial Independence

In soil science, both first- and second-order statistics are used to quantify spatial variability. Typical measures of first-order statistics are mean and standard

deviation. Because of the patchiness of biological properties, distributions are skewed and the coefficient of variation is often high. For this reason, log-normal transformations or nonparametric techniques are used for inferential analysis of data. Coefficients of variation (CV) of greater than 150% have been reported for abundance of soil fauna. For enzymatic activity, CVs can be as low as 30% for phosphatase and protease, but as high as 130% for urease. CVs of several hundred percent have been reported for microbial transformation rates.

Mean and variance are used beyond inferential statistics. For example, the Taylor power law relates sample mean, μ, and variance, σ:

$$\sigma^2 = a\mu^{\mathrm{b}} \qquad [1]$$

where b is a measure of aggregation or 'clumpiness' of the distribution, and a is a scale factor. The Taylor power law is often invariant for a given population over a wide range of spatial scales. Tests of the invariance of the law might include rhizosphere scale, field scale, and farm scale. Invariance of the law has been interpreted to mean that the distribution of the population depends only on the dispersal traits of the organism, not on spatial constraints. Its scale invariance can be helpful, because, once a and b are known at one scale, the constants can be applied at another scale. The invariant properties of the Taylor power law are useful in the design of efficient sampling schemes for soil organisms and in the control of plant-parasitic nematodes.

Second-Order Statistics and Spatial Dependence

First-order estimates are appropriate when soil properties at all sample points are spatially independent. However, the variance of most properties has both random (spatially independent) and deterministic (spatially dependent) components. Total sample variance, σ^2_{T}, can be expressed as:

$$\sigma^2_{\mathrm{T}} = \sigma^2_{\mathrm{R}} + \sigma^2_{\mathrm{D}} \qquad [2]$$

where the variances associated with random and deterministic variations are σ^2_{R} and σ^2_{D}, respectively. Deterministic variance includes components such as trends, periodicities, and characteristic patch sizes. Organism size and its range of motion as well as physicochemical scales may contribute to the deterministic variations of biological soil properties. Care has to be taken when using the paradigm expressed in Eqn [2]. The perception of deterministic and random variations in soil may depend on sampling or investigative scale, when multiple scales are contributing to spatial variations. Increasing the scale of an investigation may change the ratio of deterministic to random contributions.

The assumption – inherent in first-order statistics – that samples taken from different locations are independent of each other is not valid at all scales. Second-order analyses are more appropriate when spatial dependence in the data is suspected. In soil science, the main tool to parameterize spatial dependence is semivariance analysis. For this analysis, data needs to satisfy the stationarity requirements which prohibits trends. For example, denitrification increases along a hydrosequence from dry, well-drained soils to poorly drained soils. Hence, semivariance analysis of denitrification along the entire hydrosequence would violate the stationarity requirement. Trends are easily detected and can be removed from the data prior to analysis.

Semivariance, $\gamma(h)$, measures how much of the variance of a spatial distribution can be explained by the relationship between points that are a given distance, h, apart:

$$\gamma(h) = 1/(2N(h))\Sigma(x_{\mathrm{i}} - x_{\mathrm{j}})^2 \qquad [3]$$

where x_{i} and x_{j} are all point pairs that are a distance h apart, and N is the number of pairs $(x_{\mathrm{i}}, x_{\mathrm{j}})$ separated by distance h. A semivariogram is constructed in two steps. First, the covariances of point pairs separated by particular lag distances, h, are computed and plotted against the lag distance (Figure 2). Then, a regression model is fitted to the covariance data to estimate three parameters: the range lag (R), the nugget variance (Co), and the spatial variance (C) (Figure 2). Co is given by the y intercept of the regression model. The nugget value estimates the compound variance associated with sampling too few points, analytical errors, and random variations. The sum of nugget and spatial structure variance is called the sill variance, $C + \mathrm{Co}$. The sill variance estimates variance at lags beyond the range lag, R, where sample pairs become spatially independent. The range lag, R, can be interpreted as the characteristic scale of a process.

The ratio $C/(C + \mathrm{Co})$ is called spatial structure and expresses how much of the total variance is attributable to spatial dependence. If the nugget is equal to the sill variance, the data have no spatial structure; it is not spatially dependent. Conversely, if the nugget is zero, spatial structure is high and spatial correlation between points is high.

Figure 3 shows three examples of variograms relating to microbial biomass N (Figure 3a), soil respiration (Figure 3b), and predaceous nematode (Figure 3c) data collected from an old field in southern New England, USA. The three variograms have distinctly different shapes. The variogram for microbial

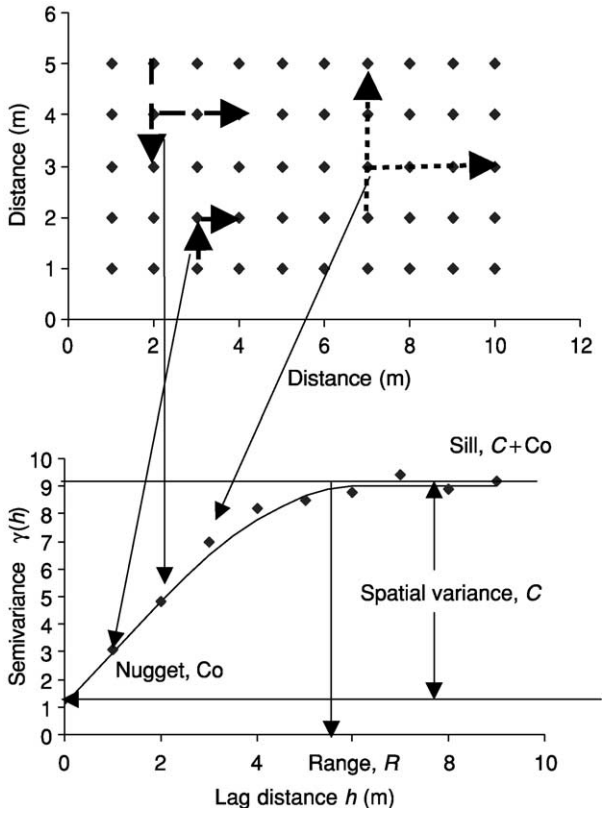


Figure 2 Examples of point pairs separated by distances of 1 m (thin arrows), 2 m (dashed arrows), 3 m (dotted arrows). Semivariance values are computed using Eqn [3] based on the entirety of point pairs for each lag distance, *h*. Nugget variance is the semivariance, Co, at zero predicted by a model line. Sill variance is the sum of nugget, Co, and spatial variance, *C*.

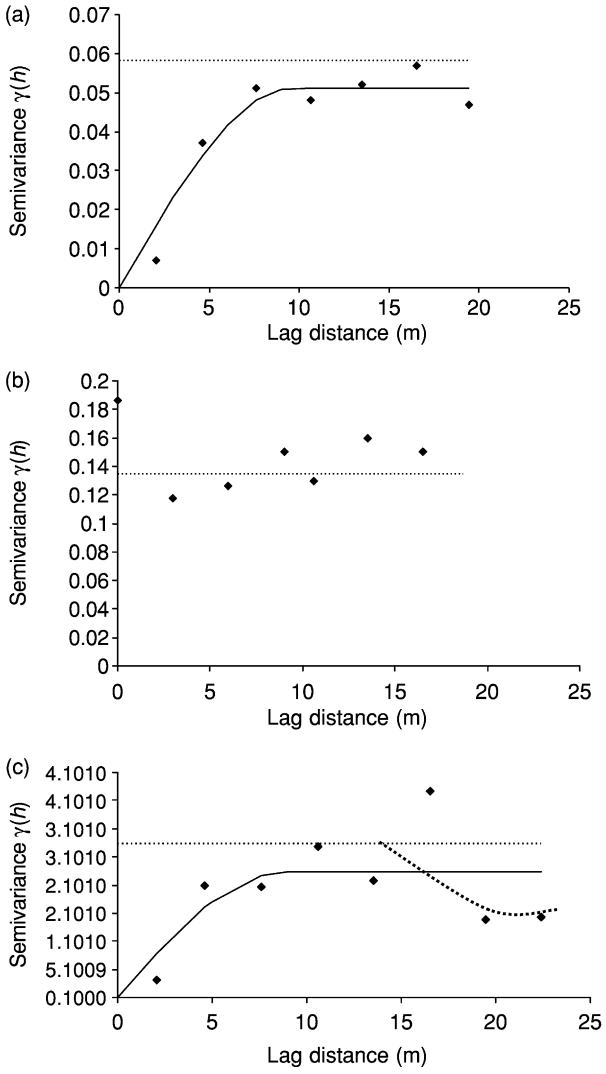


Figure 3 Typical semivariogram shapes of biological properties. Rhomboid, covariance at lag *h*; continuous line, fitted spherical model; dashed line, sample variance; dotted line, hole effect. (a) Biomass N; (b) soil respiration; (c) predaceous nematodes.

biomass N displays strong spatial dependence from 0 to 8 m, expressed by the gradual increase in semivariance from zero to the sill variance (Figure 3a). In contrast, the semivariogram of soil respiration has no spatial structure. Semivariance values are scattered about the sample variance, indicating that the nugget and the sill variances are identical and that there is no spatial correlation of respiration for the sample distance and the sample support (Figure 3b). The semivariogram for predaceous nematodes has an additional feature (Figure 3c): after the semivariance reaches the sill, its value drops again. This is known as the 'hole effect.' Semivariance holes are caused by strongly repetitive or patchy patterns. Both range and hole lag distances are characteristic scales of the pattern.

Sampling to Characterize Spatial Variability

Sampling schemes employed to collect soil samples for first- and second-order statistics are subject to specific sampling rules. Under the assumption that samples are independent of each other, random sampling can be employed, and the design is guided by the central-limit theorem and considerations of accuracy and confidence.

If spatial dependence is assumed, sampling schemes need to fulfill an additional criterion. In theory, sampling ought to comply with the Nyquist theorem to avoid misrepresentation of patterns, which is also called aliasing. The Nyquist theorem states that a pattern must be sampled at a frequency of greater than twice the wavelength of the shortest variation that can be resolved. Sampling design needs to trade off the physical dimension, or sample support, of each sample and the number of samples collected. If the sampling design does not comply with the Nyquist

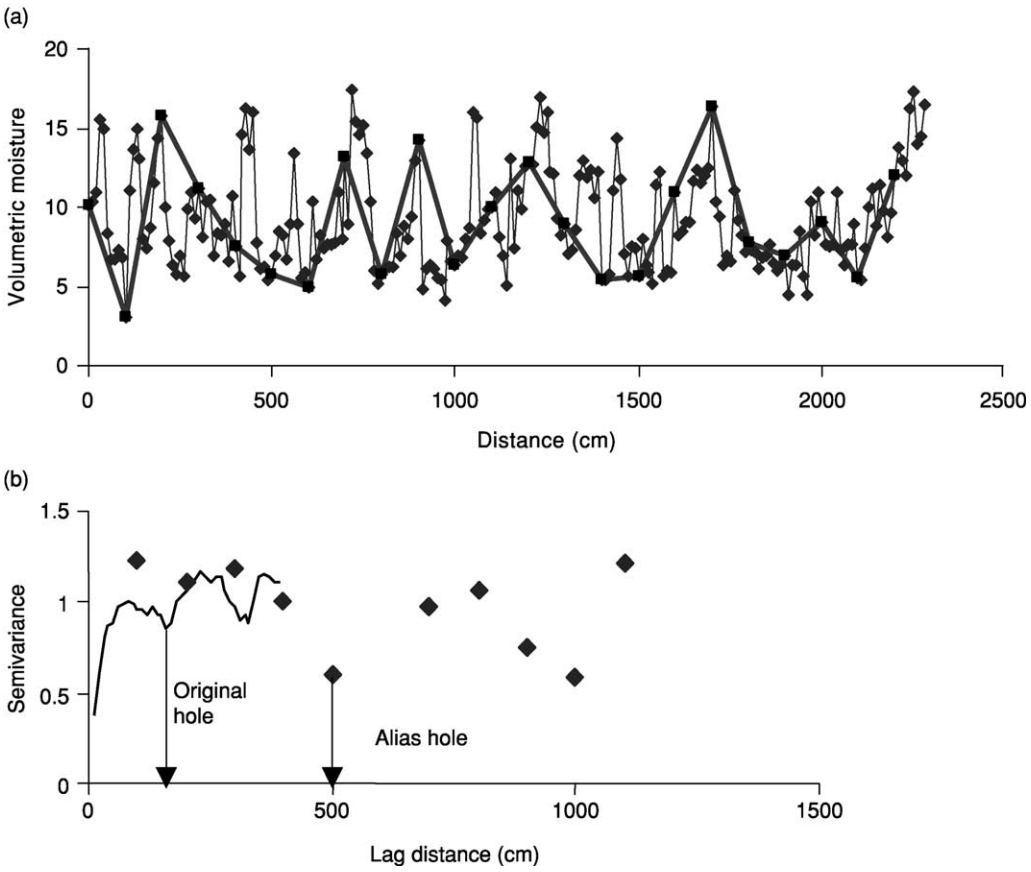


Figure 4 Sampling at an insufficient rate may result in aliasing, i.e., the misrepresentation of a fine-grain regular pattern by a coarser variation: (a) Moisture data of field transect with a ridge-and-furrow pattern of 1.5 m. Thin line with rhomboids, sampled at intervals of 10 cm (entire data set); thick line, sampled at 100-cm intervals (subsample); (b) semivariance: (continuous line) semivariogram of transect data sampled at 10 cm; rhomboid semivariance values for transect sampled every 100 cm. The 'hole effect' occurs at 140 cm when the entire data set is analyzed; it occurs at 500 cm when a transect is sampled every 100 cm.

theorem the field pattern may be misrepresented by an alias pattern, which usually overestimates the wavelength of the shortest variation. Aliasing may affect mean, variance, and scale estimates.

An example of aliasing is shown in Figure 4. Volumetric soil moisture was determined using time-domain reflectometry at 220 points along a field transect, oriented normal to the plow direction (Figure 4a, thin line). Sampling points were 10 cm apart. The associated semivariogram displays a hole at 1.4 m, roughly the distance between plough ridges (Figure 4b). A subsample was selected from the data sets, with point separations of 1 m (Figure 4a, thick line). This sample does not satisfy the Nyquist theorem (1.4 samples per small-scale variation, rather than the more-than-two samples required to satisfy the theorem). As a result, a long-wavelength variation is erroneously detected which finds expression in the semivariogram as a dilation of the hole lag distance. The larger variation seen in the subsample is called an 'alias' of the original data.

Nested sampling schemes are used to give consideration to the multiscale nature of soil variability. Nested sampling distributes sampling at a primary coarse scale with a finer-scale grid being applied locally at the coarse grid, yielding data at two dissimilar scales.

Examples of nested sampling schemes are shown in Figure 5. The nested, stratified random sampling scheme shown in Figure 5a was used to compare the spatial dependence of nematode distributions in soils from a forest and an old field. Plots (17×17 m) were divided into 25 subplots. For each subplot at least one sample was taken from a random location. In five randomly selected subplots, nine regularly spaced samples were taken from a 30×30 cm area surrounding the random location (Figure 5a). The range over which C mineralization was spatially correlated was dominated by nest-scale variations in the autumn, and by the large-scale variations for winter, spring, and summer sampling. Range values for nematode distributions and moisture indicate spatial dependence

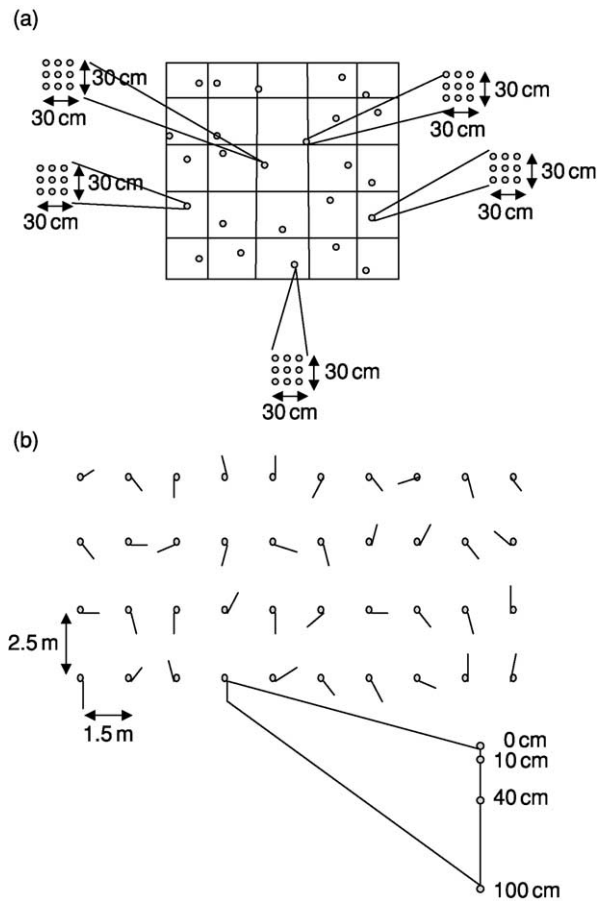


Figure 5 Examples of nested-sampling designs: (a) Nested sampling used for quantifying spatial dependence of nematode, CO_2, and soil-moisture distribution. (Adapted from Görres JH, Dichiaro MJ, Lyons JB, and Amador JA (1998) Spatial and temporal patterns of soil biological activity in a forest and an old field. *Soil Biology and Biochemistry* 30: 219–230 with permission.) (b) Nested sampling design to quantify spatial dependence of NO_3 availability. (Adapted from Gross KL, Pregitzer KS, and Burton AJ (1995) Changing scales of nitrogen availability in early and late successional plant communities. *Journal of Ecology* 83: 357–367, with permission.)

at a larger scale than for CO_2 for most sampling dates (Table 2).

The nested sampling scheme in Figure 5b was used to detect multiscale variations in soil nutrient availability. A primary grid of a field consisted of 40 points spaced at 2.5-m spacings in the N-S direction and at 1.25 m in the E-W direction. A 1-m-long transect originated at each grid point. Each transect was oriented at random and sampled at distances of 0, 10, 40, and 100 cm along the transect, giving the pattern in Figure 5b.

Nested sampling schemes are useful in discerning variability at different scales, an important step in elucidating spatial variability. Usually, only two scale components are investigated, and it is difficult

Table 2 Range values determined from semivariogram analysis of C mineralization, soil moisture, and total nematode abundance distributions in an old field and a forest soil using nested sampling

Semivariogram range						
	Total nematodes (m)		*C Mineralization (m)*		*Soil moisture (m)*	
Date	*Old field*	*Forest*	*Old field*	*Forest*	*Old field*	*Forest*
5 Jun 1994	2.9	1.0	0.9	5.6	14.0	4.9
4 Nov 1994	2.8	10.3	0.3	0.2	10.0	>20.0
20 Jan 1995	3.2	1.1	1.4	3.5	2.8	2.6
25 Mar 1995	6.0	2.6	0.9	0.9	6.7	1.3

to assess whether these are the only levels of nesting present. Detecting several levels of a spatial hierarchy may not be practicable because of the cost of collecting and analyzing many samples, a prerequisite in many spatial analyses.

Models of Spatial Variability

Linear Statistical Models

Statistical descriptions of spatial distributions in themselves do not explain the cause of variability. Attempts to attribute biological effects to physical, chemical, and biological factors with linear models are necessarily incomplete. Typically, variance of biological distributions is explained to approximately 50% even when a multitude of factors are used in the model. The difference between measured variance of a soil biological property and a multivariate model may be caused by analytical error, nonlinear effects of population dynamics, dissimilar scales between the variables, and the absence of important factors from the model.

Physical Mechanistic Process Models

Decision-making in environmental management is sometimes guided by mechanistic models describing processes with chemical, physical, or biological mechanisms. Mechanistic models are used to estimate the contamination of groundwater and surface water by nutrients and pesticides. They require information about soil physical structure, decomposition, and mineralization rates at the scales of interest.

Most mechanistic models simulate soil biological processes at a point. A point may represent a landscape unit, field, or forest stand. Usually, biological processes are treated as black boxes, while physical concepts in the model have strong mechanistic components. Some models of physical mechanisms

use multiple data layers as spatial input. The data required for the analysis are frequently found in and extracted from environmental data bases such as geographic information systems (GISs). Typically, simulations are run for every point independently, but connections between points can be made by including the contribution of groundwater or surface-water flow to simulated contaminant inputs.

Biological Mechanistic Process Models

Life-history traits and motility affect spatial population characteristics and can often be described by random processes such as the Neyman–Scott point process, which assumes that offspring disperses randomly from contagions, e.g., mother or parasitized organisms, which are themselves randomly distributed. The net effect of a Neyman–Scott process is a clumped distribution. Different levels of aggregation are shown in Figure 6. A simple random process (Figure 6a) distributes organisms almost evenly across an area. Any clumping here is a result of chance. In the double random Neyman–Scott process, the appearance of the distribution depends on the initial distribution pattern of the contagions and the distance that offspring could have traveled at the time of sampling. Both Figure 6b and c show Neyman–Scott processes, with dispersion distances of offspring of 7.5 m and 15 m, respectively.

An example of a Neyman–Scott process is that of endoparasitic nematodes, i.e., those that live within a root. Many endoparasites lay all their eggs in a single location, from where nematode larvae migrate, resulting in a clumped distribution. Aggregations arising from doubly random processes can have distinct spatial characteristics that increase spatial structure measured by semivariogram analysis even though they were caused by random processes.

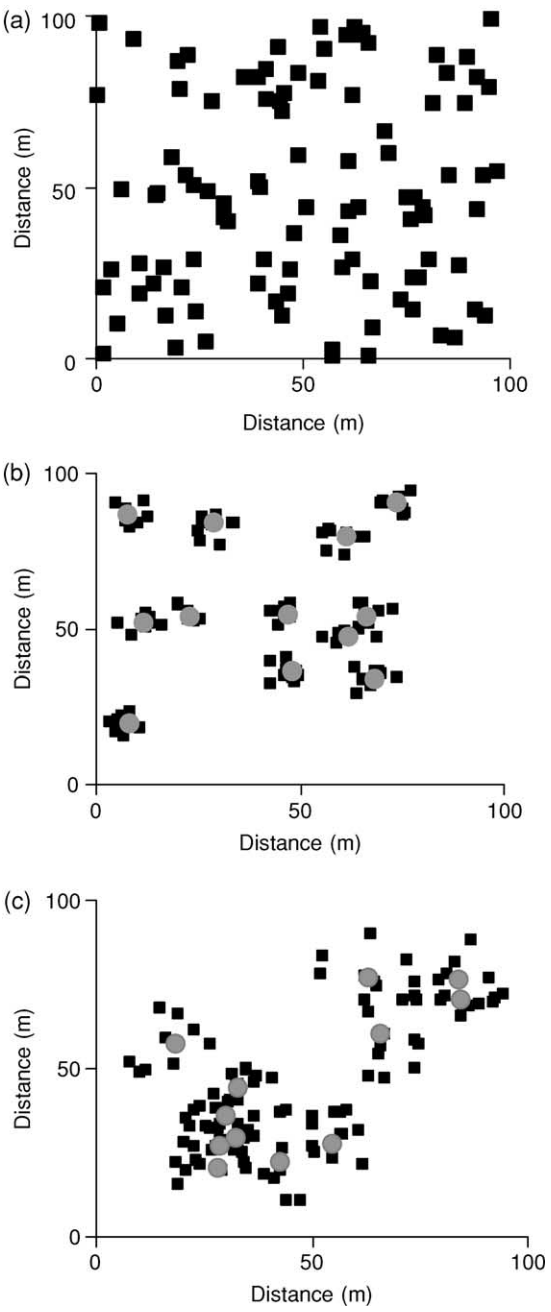


Figure 6 Distributions with different aggregation levels: (a) Distribution simulated with a single random process; (b) strongly aggregated distribution, simulated using the Neyman–Scott process with offspring dispersed randomly within a circle of 7.5 m radius around randomly distributed parent; (c) aggregated, distribution simulated using the Neyman–Scott process with offspring dispersed randomly within a circle of 15 m radius around randomly distributed parent.

Applications of Spatial Analysis

Environmental scientists and managers have become more aware of soil variability at different scales. Managers may use spatial information at landscape or field scale. Many data sets which are used to manage the environment are stored as spatially referenced information within GISs that have become important databases and analysis tools supporting decision-making and research. For instance, mapping of critical loads for soils has supported strategies to reduce soil acidification in Europe. The critical-loads approach requires several data layers pertaining to spatially variable soil biological processes. For nitrogen deposition, critical loads use spatial information of denitrification, plant uptake, and immobilization. Critical-load maps are essential to estimating where depositions of nitrogen and sulfur exceed the ability of the soil to buffer acidification and thus to taking steps to target specifically reductions in acid emissions.

Against the backdrop of the large amount of information available in spatially referenced form,

first- and second-order statistics parameters of soil variability allow managers to estimate uncertainty associated with mean property estimation. In turn, this enables the manager to assess the risk inherent in any management action informed by measures of central tendency. Farm managers using precision agriculture or integrated pest management use the spatial distribution of soil properties explicitly to dose agrochemical inputs to minimize environmental impacts while increasing farm productivity.

At smaller scales, information on the spatial variability of soil and vegetation properties can guide experimental or survey designs to support mechanistic studies of ecosystem processes. In this way, the power of inferential tests can be increased. In fundamental research addressing questions of soil biodiversity, knowledge of the spatial structure of small habitat units may help estimate the frequency or the chance of genetic exchange between microorganisms. Likewise, knowledge of how habitats of soil-dwelling predators and prey or of grazers and forage overlap can help estimate the flow of energy and matter through soil food chains.

Differences in space may be seen as gradients that drive ecologic processes such as the flow of materials in an ecosystem. As such the application of spatial analysis may both further knowledge of ecosystem functioning and improve environmental quality.

Summary

Soil biological processes are spatially variable. Many interacting factors together explain this variability at multiple scales. However, often spatial statistical descriptions of heterogeneity cannot explain the variation. Biological and physicochemical models are used to improve understanding of soil variability. Knowledge of the distribution of soil biological processes at point, field, and landscape scale can support management of environmental quality and ecosystem productivity as well as fundamental research.

See also: **Geographical Information Systems**; **Scaling:** Physical Properties and Processes; Transport Processes; **Spatial Variation, Soil Properties**; **Statistics in Soil Science**

Further Reading

Amador JA, Wang Y, Savin MC, and Görres JH (2000) Fine-scale spatial variability of physical and biological soil properties in Kingston, Rhode Island. *Geoderma* 98: 83–94.

Bonmati M, Ceccanti N, and Nanniperi P (1991) Spatial variability of phosphatase, urease, protesase, organic carbon and total nitrogen in soil. *Soil Biology and Biogeochemistry* 23: 391–396.

Bull KR, Brown MJ, Dyke H *et al.* (1995) Critical loads for nitrogen deposition for Great Britain. *Water, Air and Soil Pollution* 85: 2527–2532.

Burrough PA (1983) Multiscale sources of spatial variability in soil. I. The application of fractal concepts to nested levels of soil variation. *Journal of Soil Science* 34: 577–597.

Ferris H, Mullens TA, and Foord KE (1990) Stability and characteristics of spatial description parameters for nematode populations. *Journal of Nematology* 22: 427–439.

Groffman PM and Tiedje JM (1989) Denitrification in north temperate forest soils: relationship between denitrification and environmental factors at the landscape scale. *Soil Biology and Biogeochemistry* 21: 621–626.

Gross KL, Pregitzer KS, and Burton AJ (1995) Changing scales of nitrogen availability in early and late successional plant communities. *Journal of Ecology* 83: 357–367.

Isaaks EH and Srivastava RM (1989) *Applied Geostatistics*. Oxford, UK: Oxford University Press.

Kampichler C (1999) Fractal concepts in studies of soil fauna. *Geoderma* 88: 283–300.

McBratney AB and Pringle MJ (1999) Estimating average and proportional variograms of soil properties and their potential use in precision agriculture. *Precision Agriculture* 1: 125–152.

Parkin TB (1993) Spatial variability of microbial processes in soil – a review. *Journal of Environmental Quality* 22: 409–417.

Robertson GP and Freckman DW (1995) The spatial distribution of nematode trophic groups across a cultivated ecosystem. *Ecology* 76: 1425–1432.

Smith JL, Halvorson JJ, and Bolton H, Jr (1994) Spatial relationship of soil microbial biomass and C and N mineralization in a semi-arid shrub-steppe ecosystem. *Soil Biology and Biogeochemistry* 26: 1151–1159.

Vogel J, Normand P, Thioulouse J, Nesme X, and Grundmann GL (2003) Relationship between spatial and genetic distance on *Agrobacterium* spp. in 1 cubic centimeter of soil. *Applied and Environmental Microbiology* 69: 1482–1487.

Webster R and Oliver MA (2001) *Geostatistics for Environmental Scientists*. Chichester, UK: John Wiley.

Wirth SJ (2001) Regional-scale analysis of soil microbial biomass and soil basal CO_2-respiration in Northeastern Germany. In: Stott DE, Mohtar RH, and Steinhardt GC (eds) *Sustaining the Global Farm*, pp. 486–493. 10th International Soil Conservation Organization Meeting, 24–29 May 1999. West Lafayette, IN: Purdue University and the USDA–ARS National Soil Erosion Laboratory.

ENCYCLOPEDIA OF SOILS IN THE ENVIRONMENT

EDITED BY
DANIEL HILLEL
CYNTHIA ROSENZWEIG
DAVID POWLSON
KATE SCOW
MICHAIL SINGER
DONALD SPARKS

VOLUME FOUR

ENCYCLOPEDIA OF SOILS IN THE ENVIRONMENT

FOUR-VOLUME SET

by Daniel Hillel (Editor-in-Chief)

Hardcover: 2200 pages
Publisher: Academic Press; 1 edition (November 8, 2004)
Language: English
ISBN-10: 0123485304
ISBN-13: 978-0123485304

Book Description

More than ever before, a compelling need exists for an encyclopedic resource about soil the rich mix of mineral particles, organic matter, gases, and soluble compounds that foster both plant and animal growth. Civilization depends more on the soil as human populations continue to grow and increasing demands are placed upon available resources.
The Encyclopedia of Soils in the Environmentis a comprehensive and integrated consideration of a topic of vital importance to human societies in the past, present, and future.

This important work encompasses the present knowledge of the world's variegated soils, their origins, properties, classification, and roles in the biosphere. A team of outstanding, international contributors has written over 250 entries that cover a broad range of issues facing today's soil scientists, ecologists, and environmental scientists.
This four-volume set features thorough articles that survey specific aspects of soil biology, ecology, chemistry and physics. Rounding out the encyclopedia's excellent coverage, contributions cover cross-disciplinary subjects, such as the history of soil utilization for agricultural and engineering purposes and soils in relation to the remediation of pollution and the mitigation of global climate change.

This comprehensive, yet accessible source is a valuable addition to the library of scientists, researchers, students, and policy makers involved in soil science, ecology, and environmental science.

Also available online via ScienceDirect featuring extensive browsing, searching, and internal cross-referencing between articles in the work, plus dynamic linking to journal articles and abstract databases, making navigation flexible and easy. For more information, pricing options and availability visit www.info.sciencedirect.com.

* A distinguished international group of editors and contributors
* Well-organized encyclopedic format providing concise, readable entries, easy searches, and thorough cross-references
* Abundant visual resources — photographs, figures, tables, and graphs — in every entry
* Complete up-to-date coverage of many important topics — essential information for scientists, students and professionals alike

EDITOR-IN-CHIEF

Daniel Hillel
Columbia University
New York, NY
USA

EDITORS

Jerry L Hatfield
National Soil Tilth Laboratory
Ames, IA
USA

Kate M Scow
University of California
Davis, CA
USA

David S Powlson
Rothamsted Research
Harpenden
UK

Michael J Singer
University of California
Davis, CA
USA

Cynthia Rosenzweig
NASA Goddard Institute for Space Studies
New York, NY
USA

Donald L Sparks
University of Delaware
Newark, DE
USA

EDITORIAL ADVISORY BOARD

R Bardgett
Lancaster University
Lancaster
UK

J L Boettinger
Utah State University
Logan, UT
USA

G Gee
Pacific Northwest National Laboratory
Richland, WA
USA

R Keren
The Volcani Center
Bet Dagan
Israel

J Kimble
USDA Natural Resources Conservation Service
Lincoln, NE
USA

M B Kirkham
Kansas State University
Manhattan, KS
USA

M Kutilek
Prague
Czech Republic

D Martens
Southwest Watershed Research Center
Tucson, AZ
USA

K Mengel
Justus Leibig University
Giessen
Germany

K Reichardt
Center for Nuclear Energy in Agriculture
Piracicaba
Brazil

K Ritz
Cranfield University
Silsoe
UK

R Schulin
Swiss Federal Institute of Technology Zurich
Schlieren
Switzerland

N Senesi
Università di Bari
Bari
Italy

J T Sims
University of Delaware
Newark, DE
USA

K Smith
University of Edinburgh
Edinburgh
UK

R L Tate
Rutgers University
New Brunswick, NJ
USA

N van Breemen
Wageningen Agricultural University
Wageningen
The Netherlands

W H van Riemsdijk
Department of Soil Quality
Wageningen
The Netherlands

FOREWORD

The *Encyclopedia of Soils in the Environment* is a vitally important scientific publication and an equally important contribution to global public policy. The *Encyclopedia* brings together a remarkable range of cutting-edge scientific knowledge on all aspects of soil science, as well as the links of soils and soil science to environmental management, food production, biodiversity, climate change, and many other areas of significant concern. Even more than that, the *Encyclopedia* will immediately become an indispensable resource for policy makers, analysts, and students who are focusing on one of the greatest challenges of the 21st century. With 6.3 billion people, our planet is already straining to feed the world's population, and is failing to do so reliably in many parts of the world. The numbers of chronically poor in the world have been stuck at some 800 million in recent years, despite long-standing international goals and commitments to reduce that number by several hundred million. Yet the challenge of food production will intensify in coming decades, as the human population is projected to rise to around 9 billion by mid-century, with the increased population concentrated in parts of the world already suffering from widespread chronic under-nourishment.

Unless the best science is brought to these problems, the situation is likely to deteriorate sharply. Food production systems are already under stress, for reasons often related directly to soils management. In Africa, crop yields are disastrously low and falling in many places due to the rampant depletion of soil nutrients. This situation needs urgent reversal, through increasing use of agro-forestry techniques (e.g. inter-cropping cereals with leguminous nitrogen-fixing trees) and increasing the efficient applications of chemical fertilizers. In other impoverished, as well as rich, parts of the planet, decades of intensive agriculture under irrigation have led to salinization, water-logging, eutrophication of major water bodies, dangerous declines of biodiversity and other forms of environmental degradation. These enormous strains are coupled with the continuing pressures of tropical deforestation and the lack of new promising regions for expanding crop cultivation to meet the needs of growing populations. Finally, there looms the prospect of anthropogenic climate change. Global warming and associated complex and poorly understood shifts in precipitation extremes and other climate variables all threaten the world's natural ecosystems and food production systems in profound yet still imperfectly understood ways. The risks of gradual or abrupt climate change are coupled with the risks of drastic perturbations to regional and global food supplies.

The *Encyclopedia* offers state-of-the-art contributions on each of these challenges, as well as links to entries on the fundamental biophysical processes that underpin the relevant phenomena. The world-scale and world-class collaboration that stands behind this unique project signifies its importance for the world community. It is an honor and privilege for me to introduce this path-breaking endeavor.

Jeffrey D Sachs
Director
The Earth Institute at Columbia University
Quetelet Professor of Sustainable Development
Columbia University, New York, USA

PREFACE

The term 'soil' refers to the weathered and fragmented outer layer of our planet's land surfaces. Formed initially through the physical disintegration and chemical alteration of rocks and minerals by physical and biogeochemical processes, soil is influenced by the activity and accumulated residues of a myriad of diverse forms of life. As it occurs in different geologic and climatic domains, soil is an exceedingly variegated body with a wide range of attributes.

Considering the height of the atmosphere, the thickness of the earth's rock mantle, and the depth of the ocean, one observes that soil is an amazingly thin body – typically not much more than one meter thick and often less than that. Yet it is the crucible of terrestrial life, within which biological productivity is generated and sustained. It acts like a composite living entity, a home to a community of innumerable microscopic and macroscopic plants and animals. A mere fistful of soil typically contains billions of microorganisms, which perform vital interactive biochemical functions. Another intrinsic attribute of the soil is its sponge-like porosity and its enormous internal surface area. That same fistful of soil may actually consist of several hectares of active surface, upon which physicochemical processes take place continuously.

Realizing humanity's utter dependence on the soil, ancient peoples, who lived in greater intimacy with nature than many of us today, actually revered the soil. It was not only their source of livelihood, but also the material from which they built their homes and that they learned to shape, heat, and fuse into household vessels and writing tablets (ceramic, made of clayey soil, being the first synthetic material in the history of technology). In the Bible, the name assigned to the first human was Adam, derived from 'adama,' meaning soil. The name given to that first earthling's mate was Hava (Eve, in transliteration), meaning 'living' or 'life-giving.' Together, therefore, Adam and Eve signified quite literally 'Soil and Life.'

The same powerful metaphor is echoed in the Latin name for the human species – Homo, derived from humus, the material of the soil. Hence, the adjective 'human' also implies 'of the soil.' Other ancient cultures evoked equally powerful associations. To the Greeks, the earth was a manifestation of Gaea, the maternal goddess who, impregnated by Uranus (god of the sky), gave birth to all the gods of the Greek pantheon.

Our civilization depends on the soil more crucially than ever, because our numbers have grown while available soil resources have diminished and deteriorated. Paradoxically, however, even as our dependence on the soil has increased, most of us have become physically and emotionally detached from it. Many of the people in the so-called 'developed' countries spend their lives in the artificial environment of a city, insulated from direct exposure to nature, and some children may now assume as a matter of course that food originates in supermarkets.

Detachment has bred ignorance, and out of ignorance has come the delusion that our civilization has risen above nature and has set itself free of its constraints. Agriculture and food security, erosion and salination, degradation of natural ecosystems, depletion and pollution of surface waters and aquifers, and decimation of biodiversity – all of these processes, which involve the soil directly or indirectly – have become abstractions to many people. The very language we use betrays disdain for that common material underfoot, often referred to as 'dirt.' Some fastidious parents prohibit their children from playing in the mud and hurry to wash their 'soiled' hands when the children nonetheless obey an innate instinct to do so. Thus soil is devalued and treated

as unclean though it is the terrestrial realm's principal medium of purification, wherein wastes are decomposed and nature's productivity is continually rejuvenated.

Scientists who observe soil closely see it in effect as a seething foundry in which matter and energy are in constant flux. Radiant energy from the sun streams onto the field and cascades through the soil and the plants growing in it. Heat is exchanged, water percolates through the soil's intricate passages, plant roots extract water and transmit it to their leaves, which transpire it back to the atmosphere. Leaves absorb carbon dioxide from the air and synthesize it with soil-derived water to form the primary compounds of life. Oxygen emitted by the leaves makes the air breathable for animals, which consume and in turn fertilize plants.

Soil is thus a self-regulating bio-physio-chemical factory, processing its own materials, water, and solar energy. It also determines the fate of rainfall and snowfall reaching the ground surface – whether the water thus received will flow over the land as runoff, or seep downward to the subterranean reservoir called groundwater, which in turn maintains the steady flow of springs and streams. With its finite capacity to absorb and store moisture, and to release it gradually, the soil regulates all of these phenomena. Without the soil as a buffer, rain falling over the continents would run off entirely, producing violent floods rather than sustained river flow.

Soil naturally acts as a living filter, in which pathogens and toxins that might otherwise accumulate to foul the terrestrial environment are rendered harmless. Since time immemorial, humans and other animals have been dying of all manner of disease and have then been buried in the soil, yet no major disease is transmitted by it. The term *antibiotic* was coined by soil microbiologists who, as a consequence of their studies of soil bacteria and actinomycetes, discovered streptomycin (an important cure for tuberculosis and other infections). Ion exchange, a useful process of water purification, also was discovered by soil scientists studying the passage of solutes through beds of clay.

However unique in form and function, soil is not an isolated body. It is, rather, a central link in the larger chain of interconnected domains and processes comprising the terrestrial environment. The soil interacts both with the overlying atmosphere and the underlying strata, as well as with surface and underground bodies of water. Especially important is the interrelation between the soil and the climate. In addition to its function of regulating the cycle of water, it also regulates energy exchange and surface temperature.

When virgin land is cleared of vegetation and turned into a cultivated field, the native biomass above the ground is often burned and the organic matter within the soil tends to decompose. These processes release carbon dioxide into the atmosphere, thus contributing to the earth's greenhouse effect and to global warming. On the other hand, the opposite act of reforestation and soil enrichment with organic matter, such as can be achieved by means of conservation management, may serve to absorb carbon dioxide from the atmosphere. To an extent, the soil's capacity to store carbon can thus help to mitigate the greenhouse effect.

Thousands of years are required for nature to create life-giving soil out of sterile bedrock. In only a few decades, however, unknowing or uncaring humans can destroy that wondrous work of nature. In various circumstances, mismanaged soils may be subject to erosion (the sediments of which tend to clog streambeds, estuaries, lakes, and coastal waters), to leaching of nutrients with attendant loss of fertility and eutrophication of water bodies, to waterlogging and impaired aeration, or to an excessive accumulation of salts that may cause a once-productive soil to become entirely sterile. Such processes of soil degradation, sometimes called 'desertification,' already affect large areas of land.

We cannot manage effectively and sustainably that which we do not know and thoroughly understand. That is why the tasks of developing and disseminating sound knowledge of the soil and its complex processes have assumed growing urgency and importance. The global environmental crisis has created a compelling need for a concentrated, concise, and definitive source of information – accessible to students, scientists, practitioners, and the general public – about the soil in all its manifestations – in nature and in relation to the life of humans.

Daniel Hillel
Editor-in-Chief
May 2004

INTRODUCTION

The *Encyclopedia of Soils in the Environment* contains nearly 300 articles, written by the world's leading authorities. Pedologists, biologists, ecologists, earth scientists, hydrologists, climatologists, geographers, and representatives from many other disciplines have contributed to this work. Each of the articles separately, and all of them in sequence and combination, serve to summarize and encapsulate our present knowledge of the world's variegated soils, their natural functions, and their importance to humans.

Concise articles surveying specific aspects of soils (soil genesis, soil chemistry and mineralogy, soil physics and hydrology, and soil biology) are complemented by articles covering transdisciplinary aspects, such as the role of soils in ecology, the history of soil utilization for agricultural and engineering purposes, the development of soil science as a discipline, and the potential or actual contributions of soils to the generation, as well as to the mitigation, of pollution and of global climate change.

This comprehensive reference encompasses both the fundamental and the applied aspects of soil science, interfacing in general with the physical sciences and life sciences and more specifically with the earth sciences and environmental sciences.

The *Encyclopedia of Soils in the Environment* manifests the expanding scope of modern soil science, from its early sectarian focus on the utilitarian attributes of soils in agriculture and engineering, to a wider and much more inclusive view of the soil as a central link in the continuous chain of processes constituting the dynamic environment as a whole. Thus it both details and integrates a set of topics that have always been of vital importance to human societies and that are certain to be even more so in the future.

Daniel Hillel
Editor-in-Chief
May 2004

CONTENTS

Contents are given as follows: CHAPTER NAME Author(s) Page number

VOLUME 1

A

ACID RAIN AND SOIL ACIDIFICATION L Blake 1

ACIDITY N Bolan, D Curtin and D Adriano 11

AERATION D E Rolston 17

AGGREGATION

Microbial Aspects S D Frey 22

Physical Aspects J R Nimmo 28

AGROFORESTRY P K R Nair 35

AIR PHASE see AERATION; DIFFUSION

ALBEDO see ENERGY BALANCE; RADIATION BALANCE

ALLOPHANE AND IMOGOLITE see AMORPHOUS MATERIALS

ALLUVIUM AND ALLUVIAL SOILS J L Boettinger 45

ALUMINUM SPECIATION D R Parker 50

AMMONIA D E Kissel and M L Cabrera 56

AMORPHOUS MATERIALS J Harsh 64

ANAEROBIC SOILS P W Inglett, K R Reddy and R Corstanje 72

ANION EXCHANGE see CATION EXCHANGE

APPLICATIONS OF SOILS DATA P J Lawrence 78

ARCHAEA J E T McLain 88

ARCHEOLOGY IN RELATION TO SOILS J A Homburg 95

B

BACTERIA

Plant Growth-Promoting Y Bashan and L E de-Bashan 103

Soil L J Halverson 115

BACTERIOPHAGE M Radosevich, K E Williamson and K E Wommack 122

BIOCONTROL OF SOIL-BORNE PLANT DISEASES C E Pankhurst and J M Lynch 129

BIODIVERSITY D H Wall 136

BUFFERING CAPACITY B R James 142

BULK DENSITY see POROSITY AND PORE-SIZE DISTRIBUTION

C

CALCIUM AND MAGNESIUM IN SOILS N Bolan, P Loganathan and S Saggar 149

CAPILLARITY D Or and M Tuller 155

CARBON CYCLE IN SOILS

Dynamics and Management C W Rice 164

Formation and Decomposition C A Cambardella 170

CARBON EMISSIONS AND SEQUESTRATION K Paustian 175

CATION EXCHANGE L M McDonald, V P Evangelou and M A Chappell 180

CHEMICAL EQUILIBRIA A P Schwab 189

CHEMICAL SPECIATION MODELS see SURFACE COMPLEXATION MODELING

CHERNOZEMS see GRASSLAND SOILS

CHILDS, ERNEST CARR E G Youngs 195

CIVILIZATION, ROLE OF SOILS D Hillel 199

CLASSIFICATION OF LAND USE see LAND-USE CLASSIFICATION

CLASSIFICATION OF SOILS R W Arnold 204

CLASSIFICATION SYSTEMS

Australian R W Fitzpatrick 211

FAO F O Nachtergaele 216

Russian, Background and Principles M Gerasimova 223

Russian, Evolution and Examples D Konyushkov 227

USA D J Brown 235

CLAY MINERALS D G Schulze 246

CLIMATE CHANGE IMPACTS P Bullock 254

CLIMATE MODELS, ROLE OF SOIL P Smith 262

COLD-REGION SOILS C-L Ping 268

COLLOID-FACILITATED SORPTION AND TRANSPORT R Kretzschmar 276

COMPACTION J J H van den Akker and B Soane 285

COMPOST T L Richard 294

CONDITIONERS R E Sojka, J A Entry and W J Orts 301

CONSERVATION see EROSION: Water-Induced; Wind-Induced; SUSTAINABLE SOIL AND LAND MANAGEMENT; TERRACES AND TERRACING

CONSERVATION TILLAGE M R Carter 306

COVER CROPS L Edwards and J Burney 311

CROP ROTATIONS C A Francis 318

CROP WATER REQUIREMENTS L S Pereira and I Alves 322

CROP-RESIDUE MANAGEMENT D C Reicosky and A R Wilts 334

CRUSTS

Biological J Belnap 339

Structural R L Baumhardt and R C Schwartz 347

CULTIVATION AND TILLAGE M R Carter and E McKyes 356

D

DARCY'S LAW D Swartzendruber 363

DEGRADATION C J Ritsema, G W J van Lynden, V G Jetten and S M de Jong 370

DENITRIFICATION D A Martens 378

DESERTIFICATION D Hillel and C Rosenzweig 382

DIFFUSION T Addiscott and P Leeds-Harrison 389

DISINFESTATION A Gamliel and J Katan 394

DISPERSION see FLOCCULATION AND DISPERSION

DISSOLUTION PROCESSES, KINETICS K G Scheckel and C A Impellitteri 400

DRAINAGE, SURFACE AND SUBSURFACE N R Fausey 409

DRYLAND FARMING G A Peterson 414

E

EARTHWORMS see FAUNA

EDAPHOLOGY A L Ulery 419

ELECTRON PARAMAGNETIC RESONANCE see ELECTRON-SPIN RESONANCE SPECTROSCOPY

ELECTRON-SPIN RESONANCE SPECTROSCOPY N Senesi and G S Senesi 426

ELECTROSTATIC DOUBLE-LAYER see CATION EXCHANGE

ENERGY BALANCE M Fuchs 438

ENVIRONMENTAL MONITORING P J Loveland and P H Bellamy 441

ENZYMES IN SOILS R P Dick and E Kandeler 448

EROSION

Irrigation-Induced G A Lehrsch, D L Bjorneberg and R E Sojka 456

Water-Induced J E Gilley 463

Wind-Induced T M Zobeck and R S Van Pelt 470

ESSENTIAL ELEMENTS E A Kirkby 478

EUTROPHICATION A J Gold and J T Sims 486

EVAPORATION OF WATER FROM BARE SOIL C W Boast and F W Simmons 494

EVAPOTRANSPIRATION G Stanhill 502

F

FACTORS OF SOIL FORMATION

Biota A H Jahren 507

Climate O C Spaargaren and J A Deckers 512

Human Impacts J Sandor, C L Burras and M Thompson 520

Parent Material K R Olson 532

Time E F Kelly and C M Yonker 536

FAUNA T Winsome 539

VOLUME 2

FERTIGATION U Kafkafi and S Kant 1

FERTILITY J L Havlin 10

FERTILIZERS AND FERTILIZATION H W Scherer 20

FIELD CAPACITY see WATER CYCLE

FLOCCULATION AND DISPERSION I Shainberg and G J Levy 27

FLUORESCENCE SPECTROSCOPY N Senesi and V D'Orazio 35

FOLIAR APPLICATIONS OF NUTRIENTS M Tagliavini and M Toselli 53

FOOD-WEB INTERACTIONS P C de Ruiter and J C Moore 59

FORENSIC APPLICATIONS W F Rowe 67

FOREST SOILS J R Boyle 73

FOURIER TRANSFORM INFRARED SPECTROSCOPY D Peak 80

FRACTAL ANALYSIS Y Pachepsky and J W Crawford 85

FREEZING AND THAWING

Cycles B Sharratt 98

Processes G N Flerchinger, G A Lehrsch and D K McCool 104

FUNGI K Ritz 110

G

GEOGRAPHICAL INFORMATION SYSTEMS J Bo¨hner, T Selige and R Ko¨the 121

GERMINATION AND SEEDLING ESTABLISHMENT A Hadas 130

GLOBAL WARMING see CARBON EMISSIONS AND SEQUESTRATION; CLIMATE CHANGE IMPACTS; GREENHOUSE GAS EMISSIONS

GRASSLAND SOILS J A Mason and C W Zanner 138

GREEN MANURING see COVER CROPS

GREENHOUSE GAS EMISSIONS K A Smith 145

GROUNDWATER AND AQUIFERS Y Bachmat 153

GROUNDWATER POLLUTION see POLLUTION: Groundwater

H

HEAT AND MOISTURE TRANSPORT R Horton and A Globus 169

HEAT CAPACITY see THERMAL PROPERTIES AND PROCESSES

HEAT FLOW see THERMAL PROPERTIES AND PROCESSES

HEAVY METALS D C Adriano, N S Bolan, J Vangronsveld and W W Wenzel 175

HILGARD, EUGENE WOLDEMAR R Amundson 182

HOOGHOUDT, SYMEN BAREND P A C Raats and R R van der Ploeg 188

HUMIFICATION T C Balser 195

HYDRAULIC PROPERTIES, TEMPERATURE EFFECTS S A Grant 207

HYDRIC SOILS G W Hurt 212

HYDROCARBONS P Kostecki, R Morrison and J Dragun 217

HYDRODYNAMIC DISPERSION see SOLUTE TRANSPORT

HYDRODYNAMICS IN SOILS T P A Ferre´ and A W Warrick 227

HYSTERESIS J H Dane and R J Lenhard 231

I

IMMISCIBLE FLUIDS R J Lenhard, J H Dane and M Oostrom 239

INCEPTISOLS A Palmer 248

INDUSTRIAL POLLUTION see POLLUTION: Industrial

INFILTRATION T P A Ferre´ and A W Warrick 254

INFRARED SPECTROSCOPY see FOURIER TRANSFORM INFRARED SPECTROSCOPY

IRON NUTRITION K Mengel and H Kosegarten 260

IRRIGATION

Environmental Effects S Topcu and C Kirda 267

Methods D L Bjorneberg and R E Sojka 273

ISOTOPES IN SOIL AND PLANT INVESTIGATIONS K Reichardt and O O S Bacchi 280

ISOTROPY AND ANISOTROPY T-C J Yeh, P Wierenga, R Khaleel and R J Glass 285

J

JENNY, HANS R Amundson 293

K

KELLOGG, CHARLES J D Helms 301

KINETIC MODELS P M Jardine 307

KIRKHAM, DON D R Nielsen and R R van der Ploeg 315

L

LAMINAR AND TURBULENT FLOW see HYDRODYNAMICS IN SOILS

LANDFILLS see WASTE DISPOSAL ON LAND: Municipal

LAND-USE CLASSIFICATION J A LaGro Jr 321

LAWES, JOHN BENNET AND GILBERT, JOSEPH HENRY A E Johnston 328

LEACHING PROCESSES B E Clothier and S Green 336

LIEBIG, JUSTUS VON R R van der Ploeg, W Bo¨hm and M B Kirkham 343

LIMING E J Kamprath and T J Smyth 350

LIPMAN, JACOB G. J C F Tedrow 358

LOESS A J Busacca and M R Sweeney 364

LOWDERMILK, WALTER CLAY J D Helms 373

LYSIMETRY T A Howell 379

M

MACRONUTRIENTS C W Wood, J F Adams and B H Wood 387

MACROPORES AND MACROPORE FLOW, KINEMATIC WAVE APPROACH P F Germann 393

MAGNESIUM IN SOILS see CALCIUM AND MAGNESIUM IN SOILS

MANURE MANAGEMENT J T Sims and R O Maguire 402

MARBUT, CURTIS FLETCHER J P Tandarich 410

MATRIC POTENTIAL see HYDRODYNAMICS IN SOILS; WATER POTENTIAL; WATER RETENTION AND CHARACTERISTIC CURVE

MEDITERRANEAN SOILS J Torrent 418

METAL OXIDES A C Scheinost 428

METALS AND METALLOIDS, TRANSFORMATION BY MICROORGANISMS S M Glasauer, T J Beveridge, E P Burford, F A Harper and G M Gadd 438

METALS, HEAVY see HEAVY METALS

MICROBIAL PROCESSES

Environmental Factors P G Hartel 448

CommunityAnaly sis C H Nakatsu 455

Kinetics N S Panikov 463

MICRONUTRIENTS L M Shuman 479

MINERAL–ORGANIC–MICROBIAL INTERACTIONS P M Huang, M C Wang and M K Wang 486

MINERALS, PRIMARY P M Huang and M K Wang 500

MINERALS, SECONDARY see CLAY MINERALS

MINIMUM TILLAGE see CONSERVATION TILLAGE

MISCIBLE DISPLACEMENT see SOLUTE TRANSPORT

MORPHOLOGY P R Owens and E M Rutledge 511

MULCHES C L Acharya, K M Hati and K K Bandyopadhyay 521

MYCORRHIZAL FUNGI L M Egerton-Warburton, J I Querejeta, M F Allen and S L Finkelman 533

VOLUME 3

N

NEMATODES D A Neher and T O Powers 1

NEUTRON SCATTERING M J Fayer and G W Gee 6

NITROGEN IN SOILS

- Cycle M S Coyne and W W Frye 13
- Nitrates D S Powlson and T M Addiscott 21
- Nitrification J I Prosser 31
- Plant Uptake A Hodge 39
- Symbiotic Fixation J I Sprent 46

NITROGEN FERTILIZERS see FERTILIZERS AND FERTILIZATION

NUCLEAR WASTE DISPOSAL G W Gee, P D Meyer and A L Ward 56

NUTRIENT AVAILABILITY N K Fageria and V C Baligar 63

NUTRIENT MANAGEMENT G D Binford 71

O

ORGANIC FARMING C A Francis 77

ORGANIC MATTER

- Principles and Processes M Schnitzer 85
- Genesis and Formation K M Haider and G Guggenberger 93
- Interactions with Metals N Senesi and E Loffredo 101

ORGANIC RESIDUES, DECOMPOSITION A J Franzluebbers 112

ORGANIC SOILS D L Mokma 118

OVERLAND FLOW T S Steenhuis, L Agnew, P Ge´rard-Marchant and M T Walter 130

OXIDATION–REDUCTION OF CONTAMINANTS C J Matocha 133

P

PADDY SOILS C Witt and S M Haefele 141

PARENT MATERIAL see PEDOLOGY: Basic Principles; FACTORS OF SOIL FORMATION: Parent Material

PEDOLOGY

- Basic Principles M J Singer 151
- Dynamic F C Ugolini 156

PEDOMETRICS I O A Odeh and A B McBratney 166

PENMAN, HOWARD LATIMER J L Monteith 176

PENMAN–MONTEITH EQUATION R Allen 180

PERCOLATION see HYDRODYNAMICS IN SOILS

PERMAFROST see POLAR SOILS

PERMEABILITY see HYDRODYNAMICS IN SOILS

PERSISTENT ORGANIC POLLUTANTS (POPS) see POLLUTANTS: Persistent Organic (POPs)

PESTICIDES R H Bromilow 188

PETROLEUM see HYDROCARBONS

pH N Bolan and K Kandaswamy 196

PHOSPHORUS IN SOILS

Overview J T Sims and P A Vadas 202

Biological Interactions M D Mullen 210

PHYTOTOXIC SUBSTANCES IN SOILS M Qadir, S Schubert and D Steffens 216

PLANT–SOIL–WATER RELATIONS R A Feddes and J C van Dam 222

PLANT–WATER RELATIONS C Gimenez, M Gallardo and R B Thompson 231

POISEUILLE'S LAW see HYDRODYNAMICS IN SOILS

POLAR SOILS J C F Tedrow 239

POLLUTANTS

Biodegradation P B Hatzinger and J W Kelsey 250

Effects on Microorganisms M E Fuller 258

Persistent Organic (POPs) D Johnson 264

POLLUTION

Groundwater H Rubin 271

Industrial S P McGrath 282

POLYMERS AND MICROORGANISMS M C Rillig 287

POORLY CRYSTALLINE ALLUMINOSILICATES see AMORPHOUS MATERIALS

POROSITY AND PORE-SIZE DISTRIBUTION J R Nimmo 295

POTASSIUM IN SOILS P M Huang, J M Zhou, J C Xie and M K Wang 303

PRECIPITATION, WATERSHED ANALYSIS J V Bonta 314

PRECIPITATION–DISSOLUTION PROCESSES W P Robarge 322

PRECISION AGRICULTURE see SITE-SPECIFIC SOIL MANAGEMENT

PREFERENTIAL FLOW see UNSTABLE FLOW; MACROPORES AND MACROPORE FLOW, KINEMATIC WAVE APPROACH

PRODUCTIVITY D L Karlen 330

PROFILE see MORPHOLOGY

PROTOZOA W Foissner 336

Q

QUALITY OF SOIL B J Wienhold, G E Varvel and J W Doran 349

R

RADIATION BALANCE J L Hatfield, T J Sauer and J H Prueger 355

RADIONUCLIDES see ISOTOPES IN SOIL AND PLANT INVESTIGATIONS

RAINFED FARMING see DRYLAND FARMING

RANGE MANAGEMENT G L Anderson 360

RECYCLING OF ORGANIC WASTES see POLLUTANTS: Biodegradation

REDISTRIBUTION see WATER CYCLE

REDOX POTENTIAL R D DeLaune and K R Reddy 366

REDOX REACTIONS, KINETICS P S Nico and S Fendorf 372

REMEDIATION OF POLLUTED SOILS E Lombi and R E Hamon 379

REMOTE SENSING

Organic Matter D K Morris, C J Johannsen, S M Brouder and G C Steinhardt 385

Soil Moisture T J Jackson 392

RHIZOSPHERE A C Kennedy and L Z de Luna 399

RICHARDS, LORENZO A. W R Gardner 407

ROOT ARCHITECTURE AND GROWTH L E Jackson 411

ROOT EXUDATES AND MICROORGANISMS B-J Koo, D C Adriano, N S Bolan and C D Barton 421

S

SALINATION PROCESSES I Shainberg and G J Levy 429

SALINITY

Management D Hillel 435

Physical Effects D Russo 442

SALT BALANCE OF SOILS see SALINATION PROCESSES

SALT-AFFECTED SOILS, RECLAMATION R Keren 454

SAND DUNES H Tsoar 462

SATURATED AND UNSATURATED FLOW see HYDRODYNAMICS IN SOILS; VADOSE ZONE: Hydrologic Processes

SCALING

Physical Properties and Processes G Sposito 472

Transport Processes R P Ewing 477

SEPTIC SYSTEMS R L Lavigne 485

SHIFTING CULTIVATION R Lal 488

SITE-SPECIFIC SOIL MANAGEMENT C J Johannsen and P G Carter 497

SLASH AND BURN AGRICULTURE see SHIFTING CULTIVATION

SLUDGE see WASTE DISPOSAL ON LAND: Liquid; Municipal

SODIC SOILS G J Levy and I Shainberg 504

SOIL–PLANT–ATMOSPHERE CONTINUUM J M Norman and M C Anderson 513

SOLUTE TRANSPORT M C Sukop and E Perfect 521

SORPTION

Metals D L Sparks 532

Organic Chemicals B Xing and J J Pignatello 537

Oxyanions C P Schulthess, H Wijnja and W Yang 548

SORPTION–DESORPTION, KINETICS D L Sparks 556

SPATIAL PATTERNS J H Go¨rres and J A Amador 562

VOLUME 4

SPATIAL VARIATION, SOIL PROPERTIES R Webster 1

SPECIFIC SURFACE AREA K D Pennell 13

STATISTICS IN SOIL SCIENCE R Webster 19

STERILIZATION see DISINFESTATION

STOCHASTIC ANALYSIS OF SOIL PROCESSES D Russo 29

STRESS–STRAIN AND SOIL STRENGTH S K Upadhyaya 38

STRUCTURE V A Snyder and M A Va´zquez 54

SUBSOILING R L Raper 69

SULFUR IN SOILS

- Overview M A Tabatabai 76
- Biological Transformations S D Siciliano and J J Germida 85
- Nutrition M A Tabatabai 91

SURFACE COMPLEXATION MODELING S Goldberg 97

SUSTAINABLE SOIL AND LAND MANAGEMENT J L Berc 108

SWELLING AND SHRINKING D Smiles and P A C Raats 115

T

TEMPERATE REGION SOILS E A Nater 125

TEMPERATURE REGIME see THERMAL PROPERTIES AND PROCESSES

TENSIOMETRY T K Tokunaga 131

TERMITES see FAUNA

TERRA ROSSA see MEDITERRANEAN SOILS

TERRACES AND TERRACING G R Foster 135

TESTING OF SOILS A P Mallarino 143

TEXTURE G W Gee 149

THERMAL PROPERTIES AND PROCESSES D Hillel 156

THERMODYNAMICS OF SOIL WATER P H Groenevelt 163

TILLAGE see CONSERVATION TILLAGE; CULTIVATION AND TILLAGE; ZONE TILLAGE

TILTH D L Karlen 168

TIME-DOMAIN REFLECTOMETRY G C Topp and T P A Ferre´ 174

TROPICAL SOILS

Arid and Semiarid H C Monger, J J Martinez-Rios and S A Khresat 182

Humid Tropical S W Buol 187

U

UNSTABLE FLOW T S Steenhuis, J-Y Parlange, Y-J Kim, D A DiCarlo, J S Selker, P A Nektarios, D A Barry and F Stagnitti 197

URBAN SOILS J L Morel, C Schwartz, L Florentin and C de Kimpe 202

V

VADOSE ZONE

Hydrologic Processes J W Hopmans and M Th van Genuchten 209

Microbial Ecology P A Holden and N Fierer 216

VIRUSES see BACTERIOPHAGE

VOLCANIC SOILS G Uehara 225

W

WAKSMAN, SELMAN A. H B Woodruff 233

WASTE DISPOSAL ON LAND

Liquid C P Gerba 238

Municipal D A C Manning 247

WATER AVAILABILITY see PLANT–SOIL–WATER RELATIONS

WATER CONTENT AND POTENTIAL, MEASUREMENT G S Campbell and C S Campbell 253

WATER CYCLE D K Cassel and B B Thapa 258

WATER EROSION see EROSION: Water-Induced

WATER HARVESTING D Hillel 264

WATER MANAGEMENT see CROP WATER REQUIREMENTS

WATER POTENTIAL D Or, M Tuller and J M Wraith 270

WATER REQUIREMENTS see CROP WATER REQUIREMENTS

WATER RETENTION AND CHARACTERISTIC CURVE M Tuller and D Or 278

WATER TABLE see GROUNDWATER AND AQUIFERS

WATER, PROPERTIES D Hillel 290

WATER-REPELLENT SOILS J Letey 301

WATERSHED MANAGEMENT M D Tomer 306

WATER-USE EFFICIENCY M B Kirkham 315

WEED MANAGEMENT D D Buhler 323

WETLANDS, NATURALLY OCCURRING E K Hartig 328

WIDTSOE, JOHN A. AND GARDNER, WILLARD G S Campbell and W H Gardner 335

WIND EROSION see EROSION: Wind-Induced

WINDBREAKS AND SHELTERBELTS E S Takle 340

WOMEN IN SOIL SCIENCE (USA) M J Levin 345

WORLD SOIL MAP H Eswaran and P F Reich 352

Z

ZERO-CHARGE POINTS J Chorover 367

ZONE TILLAGE J L Hatfield and A T Jeffries 373

Table of Contents, Volume 4

S (cont.) **1**

Spatial Variation, Soil Properties **1**

Introduction 1

Soil Classification 1

Sampling and Estimation 1

The Geostatistical Approach 4

Random Variables and Random Functions 4

Stationarity 4

Intrinsic Variation and the Variogram 4

Estimating the Variogram 5

Models for Variograms 5

Spherical 7

Exponential 7

Models with reverse curvature at the origin 7

Unbounded models 8

Anisotropy 8

Combining Trend and Random Fluctuation 9

Combining Classification with Geostatistics 9

Coregionalization - Simultaneous Variation in Two or More Variables 9

Modeling the Coregionalization 10

Example 11

Spatial Prediction - Kriging 11

Further Reading 13

Specific Surface Area **13**

Introduction 13

Direct Physical Measurement 13

Adsorption from Solution 15

Adsorption from the Gas Phase 16

Retention of Polar Liquids 17

Selection of Surface Area Measurement Technique 18

List of Technical Nomenclature 18

Further Reading 19

Statistics in Soil Science **19**

Why Statistics? 19

Population, Units, and Samples 20

Replication and Randomization 20

Descriptive Statistics 21

The Mean 21

Characteristics of Variation 21

Histogram 21

Variance 21

Estimation variance, standard error, and confidence 21

Coefficient of Variation 22

Additivity of variances 23

Statistical Significance 23

Transformations 24

Analysis of Variance 25

Fixed Effects, Random Effects, and Intraclass Correlation 26

Covariance, Correlation, and Regression 26

Covariance 26

Correlation 26

Spearman rank correlation 27

Regression 27

Further Reading 28

Stochastic Analysis of Soil Processes 29
Introduction 29
Modeling Flow and Transport in Unsaturated, Heterogeneous Soils 29
Stochastic Analysis of the Flow 29
Stochastic Analysis of the Transport 33
Summary 35
List of Technical Nomenclature 36
Further Reading 37
Stress-Strain and Soil Strength 38
Concept of Stress 38
Principal Stress 39
Effective Stress 39
Concept of Strain 40
Principal Strain 41
Void Ratio and Volumetric Strain 41
Constitutive Laws 41
Material Behavior under Load 43
Nonlinear Elastic Behavior 43
Variable-Moduli Models 44
Duncan and Chang model 45
Elastoplastic Behavior of Soil 46
Yield Criteria 46
Hardening Cap 47
Postyield Behavior 47
Plastic potential and flow rule 47
Hardening law 48
Critical-State Soil Mechanics 48
Modified Cam-Clay Constitutive Relationship 48
NSDL-AU Model 50
Composite Soil-Strength Parameters 51
Cone Penetrometer 51
Measurement of Soil Sinkage and Shear 51
Some Recent Developments 53
Further Reading 53
Structure 54
Introduction 54
Hierarchical Levels of Soil Structure and Bonding Mechanisms 55
Microaggregates less than 2 mu m in Diameter 55
Microaggregates 2-20 mu m in Diameter 57
Microaggregates 20-250 mu m in Diameter 57
Macroaggregates (greater than 250 mu m Diameter) 57
General properties of macroaggregates and their dynamic nature in soil management systems 57
Earthworm casts 60
Role of Wetting and Drying on Soil Structure Development 60
Characterization of Soil Structure Based on Visual Assessment 61
Visual Inspection of Soil Aggregates 61
Image Analysis 63
Mathematical Models of Soil Structure 63
Aggregate Size Distributions 63
Pore Size Distributions 63
Scaling Models of Soil Structure 63
Fractal scaling relations between hierarchical levels in a given soil 64
Miller scaling relations between different soils 64
Structure of Tilled Agricultural Soils 65
Initial Soil Conditions Produced by Tillage: Soil Tilth 66

Post-tillage Soil Structural Transformations 66
Surface crusting of tilled soils 66
Fracture and plastic deformation of aggregates during wetting and drying cycles 67
Long-Term Effects of Tillage 68
The Notion of Structural Quality and the Nonlimiting Water Range 68
Summary 68
Further Reading 68
Subsoiling 69
Introduction 69
Measurement of Subsoiling 69
Benefits of Subsoiling 69
Subsoiler Design 70
Force Required for Subsoiling 71
Management Practices 72
When to Subsoil 72
Maintaining Surface-Residue Coverage 74
Subsoiling in Irrigated Fields 74
Subsoiling in Perennial Crops 74
Considerations Before Subsoiling 75
Summary 76
List of Technical Nomenclature 76
Further Reading 76
Sulfur in Soils 76
Overview 76
Introduction 76
Carbon-Nitrogen-Phosphorus-Sulfur Relationships 77
Sources of Sulfur in soils 77
Minerals Sources 77
Fertilizers Sources 77
Atmospheric Sources 77
Chemical Nature of Organic Sulfur in Soils 78
Inorganic Sulfur in Soils 78
Fate of Inorganic Sulfate in Soils 78
Leaching losses 78
Sulfate Adsorption by Soils 80
Mechanisms of Sulfate Adsorption by Soils 81
Coordination with hydrous oxides 81
Exchange on edges of silicate clays 81
Molecular adsorption 82
Sulfur Transformations in Soils 82
Mineralization 82
Sources of mineralizable S 82
Role of arylsulfatase in S mineralization 82
Pattern of sulfate release 82
Factors affecting sulfur mineralization 82
Oxidation of Elemental S in Soils 83
Reduction of Sulfate in Waterlogged Soils 84
Volatilization of S Compounds from Soils 84
Further Reading 85
Biological Transformations 85
Introduction 85
The Global Sulfur Cycle 85
The Biological Availability of Sulfur 86
Mineralization 87
Assimilation 87

Sulfur Oxidation 88
Disproportionation of Sulfur 89
Reduction of Sulfur 89
Sulfur Transformations and Environmental Quality 89
Conclusion 90
Further Reading 90
Nutrition 91
Introduction 91
Sulfur Requirements of Crops 91
Functions of Sulfur in Plants 91
Sulfur in Soils 92
Sources of Mineralizable Sulfur in soils 93
Organic Nitrogen and Sulfur Relationship in Soils 93
Sulfur Availability Indexes 93
Sulfur Requirement of Plants 94
Sulfur Metabolism in Plants 94
Sulfur-Containing Materials Added to Soils 95
Fertilizers 95
Sulfur in the Atmosphere 96
Sulfur in Precipitation 96
Further Reading 96
Surface Complexation Modeling 97
Introduction 97
Description of Models 97
Surface Configuration of the Solid-Solution Interface 98
Surface Complexation Reactions 99
Equilibrium Constants for Surface Complexation 99
Mass and Charge Balances 100
Charge-Potential Relationships 101
Obtaining Values of Adjustable Parameters 101
Surface Site Density 101
Capacitances 101
Surface Complexation Constants 101
Applications to Ion Adsorption on Natural Samples 102
Constant Capacitance Model 102
Diffuse-Layer Model 104
Triple-Layer Model 105
One-pK Model 106
List of Technical Nomenclature 107
Further Reading 107
Sustainable Soil and Land Management 108
Introduction 108
The Birth of Soil Conservation in US Agriculture 108
Emerging Agricultural Technology Shifts National Conservation Priorities 109
Recognition of New Environmental Issues 109
Defining Sustainable Development 110
Policy Aspects 110
Sustainable and Organic Agricultural Systems 111
Inventories and Assessment 112
Soil Loss 112
Water Quality 113
Loss of Farmland 114
Further Reading 115
Swelling and Shrinking 115
Introduction 115
Swelling Soils 116

History....116
Elements of Theory....117
Material Balance in Swelling Systems....117
Flux Laws in Swelling Systems....118
Water Potential in a Swelling Soil....118
Potential Gradient....119
Darcy Law and Hydraulic Conductivity....119
Flow Equation....119
Overview....120
Scale of Discourse....121
Material Balance and Coordinates....121
Water Flow....121
Multidimensional Flow....121
Measurement....121
Water Retention and Hydraulic Conductivity Functions in Unsaturated Swelling Soils....122
Summary....122
List of Technical Nomenclature....122
Further Reading....122
T....125
Temperate Region Soils....125
Introduction....125
State Factors....125
Climate....125
Parent Materials....126
Organisms....127
Topography....128
Age....129
Effects of Humans....129
Classification....130
List of Technical Nomenclature....130
Further Reading....131
Tensiometry....131
Introduction....131
Equilibrium....132
Response Time....132
Range of Applications....133
List of Technical Nomenclature....134
Further Reading....134
Terraces and terracing....135
Benefits and Limitations....135
Runoff and Erosion on Hillslopes without Terraces....135
Types of Terraces....135
Gradient Terraces....135
Construction....135
Spacing....136
Deposition and sediment delivery....137
Outlet channels....138
Moisture conservation....138
Parallel-Impoundment, Underground-Outlet Terraces....138
Overland-flow interceptors....139
Impoundments....139
Water-conveyance system....139
Bench Terraces....139
Flat benches with vertical backslopes....139
Outward-sloping benches....140

Inward-sloping benches 140
Naturally formed benches 141
Ridge 141
Summary 141
Further Reading 143
Testing of Soils 143
Introduction 143
Soil-Test Extractants and Methodologies 144
Phosphorus Soil Tests 144
Nitrogen Soil Tests 144
Calibration of Soil Tests for Crop Production 145
Calibration of Soil Tests for Environmental Purposes 146
Quality of Soil Testing 147
Laboratory Quality Control 147
Soil Sampling for Soil Testing 147
Use of Soil Tests to Determine Nutrient Loading Rates 148
Summary 148
Further Reading 149
Texture 149
Introduction 149
Soil Particle Size: Measurements and Classification 149
Size Measurements 150
Fine Soil Measurements 150
Size and Textural Classification Schemes 151
Textural Classification 152
Universal classification (engineering) scheme 152
USDA classification 152
Use of Texture Data for Estimating Hydraulic Properties 154
Summary 155
Further Reading 155
Thermal Properties and Processes 156
Introduction 156
Modes of Energy Transfer 156
Energy Balance for a Bare Soil 157
Conduction of Heat in Soil 157
Volumetric Heat Capacity of Soils 158
Thermal Conductivity of Soils 158
Simultaneous Transport of Heat and Moisture 159
Thermal Regime of Soil Profiles 161
Further Reading 162
Thermodynamics of Soil Water 163
Introduction 163
Classical (Equilibrium) Thermodynamics of Soil Water 163
The Tensiometer (and the Pressure Membrane Apparatus) 164
The Maxwell Relations and Their Use in the Hydrostatics of Swelling Soils 165
Nonequilibrium Thermodynamics of Soil Water 165
Modern Development 167
Further Reading 168
Tilth 168
Introduction 168
History of Tilth 169
Soil Organic Matter and Tilth 169
Soil-Management Effects on Tilth 170
Tillage Effects on Soil Tilth 170
Extended Crop Rotations 171

Cover Crops 171
Evaluating Soil Tilth 172
Tilth Indices 172
Summary 173
Further Reading 174
Time-Domain Reflectometry 174
Introduction 174
Measurement of Soil Water Content Using TDR 175
Instrumentation and Wave-Guides 178
Impact of Salinity on Water Content Measurement 179
Measurement of Bulk Electrical Conductivity Using TDR 179
Impact of Water Content on Pore-Water Salinity Measurement 180
Conclusion 180
List of Technical Nomenclature 181
Further Reading 181
Tropical Soils 182
Arid and Semiarid 182
Introduction 182
Climatic Controls 182
Processes of Soil Formation in Arid and Semiarid Soils 183
Factors of Soil Formation in Arid and Semiarid Soils 184
Properties of Arid Soils 185
Properties of Semiarid Soils 185
Human Land Use of Arid and Semiarid Soils 186
Further Reading 187
Humid Tropical 187
Introduction 187
Humid Tropical Setting 188
Soil-Temperature Regimes of the Tropics 188
Soil-Moisture Regimes of the Tropics 188
Chemical and Mineralogical Composition of Soils 190
Mineral and Chemical Grouping of Soils 191
Reworked Sediments of Low Fertility 191
Acid Igneous Rock 193
Limestone and Other Base-Rich Rock 193
Volcanic Material 193
Major River Basins 193
Floodplains 193
Organic Deposits 194
Human Utilization 194
Further Reading 194
U 197
Unstable Flow 197
Introduction 197
Conditions for Unstable Flow 197
Finger Diameter and Structure 197
Further Reading 201
Urban soils 202
Introduction 202
Definition 202
Use of Soils in Urban Areas 202
Evolution of Soils in Urban Areas 203
Composition and Heterogeneity of Urban Soils 203
Soil Contamination in Urban Areas 205
Garden soils 205
Functions of Urban Soils 206

Pedology and Archeology....207
Management of Urban Soils....207
Further Reading....208
V....209
Vadose zone....209
Hydrologic Processes....209
Introduction....209
Physical Processes....209
Chemical Processes....211
Biological Processes....213
Scale Issues....214
Opportunities and Challenges....215
Further Reading....215
Microbial Ecology....216
Introduction....216
Definition of the Vadose Zone....216
Abundance and Distribution of Microbes in the Vadose Zone....217
Quantity Along a Depth Gradient....217
Distribution at the Meso- and Microscales....218
Diversity of Vadose Microorganisms....218
Activity of Vadose-Zone Microbes....219
Physical and Chemical Characteristics Affecting Vadose-Zone Microbial Ecology....219
Modeling Vadose-Zone Microbial Processes....221
Applications of Vadose-Zone Microbial Ecology....223
Pollutant Remediation....223
The Vadose Zone as an Analog for Other Extreme Habitats....223
Further Reading....224
Volcanic Soils....225
Introduction....225
Rock Weathering....225
Soil Formation....225
Vertisols....227
Oxisols....227
Ultisols....227
Andisols....229
Human Interactions with Volcanic Soils....229
List of Technical Nomenclature....232
Further Reading....232
W....233
Waksman, Selman A....233
Further Reading....238
Waste Disposal on Land....238
Liquid....238
Introduction....238
Types of Land-Treatment Processes....239
Slow-Rate Process....239
Rapid-Infiltration Process....240
Overland-Flow Process....241
Treatment Mechanisms....241
Filtration....241
Adsorption and Precipitation....241
Chemical Reactions....242
Volatilization....242
Biological Mechanisms....242
Fate of Specific Contaminants....242

Metals and Trace Elements 242
Organic Compounds 243
Disinfection By-products 244
Endocrine Disruptors 244
Pharmaceuticals 244
Microorganisms 244
Further Reading 246
Municipal 247
Introduction 247
MSW Composition and Degradation 247
Microbial Reactions and Waste Degradation 249
Mineralogical Reactions Within Waste 250
Summary 252
List of Technical Nomenclature 252
Further Reading 252
Water Content and Potential, Measurement 253
Introduction 253
Measurement of Water Content 253
Gravimetric Methods 253
Dielectric Properties 254
Thermal Properties 254
Nuclear Methods 254
Measurement of Water Potential 255
Solid-Phase Sensors 255
Heat-Dissipation Matric-Potential Sensors 255
Electrical Resistance Matric Potential Sensors 256
Dielectric Matric-Potential Sensors 256
Filter-Paper Method 256
Tensiometer 256
Vapor-Pressure Methods 257
Water-Potential Units 257
Further Reading 257
Water Cycle 258
Introduction 258
Overview 258
Water Cycle Processes 259
Environmental and Economic Implications of Water-Cycle Processes 261
Land-Management Impacts 262
Summary 263
Further Reading 264
Water Harvesting 264
Definition 264
Surface Runoff 264
Runoff Control and Utilization 265
Ancient Methods 266
Water Spreading 266
Runoff Inducement 266
Modern Methods 268
Further Reading 270
Water Potential 270
Introduction 270
The 'Total' Soil Water Potential and its Components 270
Pim(h) 272
Pie(h) 272
Pis(h) 272
Pia(h) 272

Measurement of Potential Components 273
Water Potential 273
Pressure Potential 274
Matric Potential 275
Osmotic Potential 277
Further Reading 277
Water Retention and Characteristic Curve 278
Introduction 278
The Matric Potential 278
The Bundle of Cylindrical Capillaries Model 278
Liquid Retention and Pore Shape 279
Modeling SWC 281
Empirical SWC Models 281
Fractal Representation of the Soil Pore Space and the SWC 283
Physically Based Models for the SWC 283
Lattice Boltzmann Approach 284
Pore Network Models and the SWC 286
Hysteresis of the SWC 286
Measurement of SWC Relationships 287
Pressure-Plate Apparatus and Pressure-Flow Cells 288
Paired Sensors for Field SWC Measurement 288
Further Reading 288
Water, Properties 290
Introduction 290
Molecular Structure 290
Hydrogen Bonding 291
States of Water 291
Ionization and pH 292
Solvent Properties of Water 293
Osmotic Pressure 294
Solubility of Gases 295
Adsorption of Water on Solid Surfaces 295
Vapor Pressure 295
Surface Tension 296
Curvature of Water Surfaces and Hydrostatic Pressure 297
Contact Angle of Water on Solid Surfaces 298
The Phenomenon of Capillarity 298
Density and Compressibility 299
Dynamic and Kinematic Viscosity 300
Further Reading 300
Water-Repellent Soils 301
Characterizing the Degree of Water Repellency 301
Water-Entry Pressure Head 302
Infiltration Rate 302
Mitigating Water Repellency 304
Summary 305
List of Technical Nomenclature 305
Further Reading 305
Watershed Management 306
Introduction 306
An Interdisciplinary Task 306
Hydrology and Streamflow Variation 306
Stormflow and Floods 307
Snow Hydrology 308
Drought and Low Streamflow 308
Water Quality 308

Land-Use Impacts....309
Agricultural Lands....309
Forest Lands....309
Grazing Lands....311
Mined Land....311
Urban Areas....311
Riparian Areas and Wetlands....311
Approaches, Challenges, and Tools for Watershed Management....312
List of Technical Nomenclature....314
Further Reading....315
Water-Use Efficiency....315
Introduction....315
History....316
Factors that Influence Water-Use Efficiency....316
Soil Factors....316
Soil-water content....316
Method of irrigation....316
Plant Factors....317
Species adaptation....317
Plant breeding....317
Cultural Factors....318
Planting patterns....318
Seed quality....318
Weeds....318
Disease and insect pests....318
Tillage....319
Rotations....319
Fertilization....319
Climate....319
Measurement of Water-Use Efficiency....320
Possibilities of Increasing Water-Use Efficiency....321
Further Reading....322
Weed Management....323
Introduction....323
Impacts of Tillage on Weed Management....323
General Effects....323
Control Options and Efficacy....323
Biology and Ecology of Weed Responses to Tillage Systems....324
Classification of Weed Species....324
Summer Annual Species....324
Weed Seed Bank and Seedling Biology....325
Winter Annual and Biennial Species....326
Perennial Species....326
Summary....327
List of Technical Nomenclature....327
Further Reading....328
Wetlands, Naturally Occurring....328
Introduction....328
Wetland Definitions....329
Wetland Classification....329
Estuarine Wetlands....329
Palustrine Wetlands....329
Hydric Soils....330
Soil Chemistry....331
Nitrogen....331
Iron and Manganese....333

Sulfur 333
Carbon 333
Policies, Regulations, and Protection 334
Further Reading 334
Widtsoe, John A. and Gardner, Willard 335
Further Reading 339
Windbreaks and Shelterbelts 340
Introduction 340
Use of Windbreaks and Shelterbelts in Soil Management 340
Influence of Shelters on Aerodynamics and Microclimate 340
Wind Speed Reduction 340
Turbulence Fields 341
Pressure Fields 341
Surface Fluxes of Heat and Moisture 341
Soil Moisture 342
Meteorologic Mechanisms that Move Gases in Soil 343
Type 1: Large-Scale Pressure Changes 343
Type 2: Traveling Atmospheric Pressure Waves 343
Type 3: Static Horizontal Pressure Gradient in Soil (Standing Wave) 344
Type 4: Fluctuations on the Standing Wave 344
Type 5: Turbulence 344
Type 6: Venturi Effects 344
Type 7: Rainfall 344
Type 8: Phase Change of Liquid to Vapor 344
Aesthetic and Recreational Value 344
List of Technical Nomenclature 345
Further Reading 345
Women in Soil Science (USA) 345
Introduction 345
The Pioneers (1895-1965) 345
Foundations: Building on the Pioneers (1959-1975) 347
In the Classroom, in the Field, and in the Laboratory (1970-1990) 349
1990 and Ahead 351
Acknowledgment 352
Further Reading 352
World Soil Map 352
Introduction 352
Soil Classification 353
Global Distribution 356
Gelisols 356
Histosols 356
Andisols 357
Spodosols 360
Oxisols 360
Vertisols 361
Aridisols 361
Ultisols 362
Mollisols 363
Alfisols 363
Inceptisols 363
Entisols 364
Summary 364
Further Reading 364
Z 367
Zero-Charge Points 367
Introduction 367

Components of Surface Charge 367
Surface Charge Balance.....369
Points of Zero Charge.....369
Point of Zero Charge.....369
Point of Zero Net Proton Charge.....371
Point of Zero Net Charge 371
List of Technical Nomenclature 372
Further Reading.....373
Zone Tillage 373
Further Reading.....375

SPATIAL VARIATION, SOIL PROPERTIES

R Webster, Rothamsted Research, Harpenden, UK

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

The soil mantles the land more or less continuously, except where there is bare rock and ice, and in a way so complex that no description of it can follow every detail. Further, our knowledge of the soil's properties beneath the surface is fragmentary because it derives from observations on small volumes of material widely separated from one another. Any representation of the whole mantle involves simplification and inference or prediction as to what the soil is like between sampling points with the uncertainty that they entail.

Research in the last 40 years has provided us with quantitative descriptions based on samples. Two main approaches may be discerned. In the first the soil is divided into discrete classes (strata) which are sampled to give estimates of mean values and variances using classical statistics. The other sees soil as a suite of continuous variables and seeks to describe the continuity in terms of spatial dependence and specifically uses geostatistics. The two approaches are not mutually exclusive and they can be combined.

Soil Classification

Peasant cultivators and farmers have for centuries recognized different kinds of soil, and they have divided their land where the soil changes so that they can manage it. In other words, they have classified the soil and land spatially. More formal classification of soil has its roots in nineteenth-century biological taxonomy and practice in geological survey. Finite circumscribed regions are divided into parcels by boundaries, which are sharp lines across which the soil changes in some sense. For any one region, the outcome is a map, technically a choropleth map, showing the region tessellated into spatial classes, which constitute a general-purpose classification. The map may purport to show the classes of some predefined scheme of classification; alternatively the boundaries on it may be drawn where the soil changes more than elsewhere and between which the soil is relatively homogeneous. There are thousands of examples.

The map has usually been accompanied by a text describing each of the classes displayed, with data on individual soil properties from representative sites. The spatial variation for any one soil property thus appears as a stepped function, as in Figure 1a of a transect across a region. Variation within the classes may be acknowledged, but it is not evident. The reality is more like Figure 1b, which is the same transect but now with all the data from sampling at 10-m intervals shown, and for which there is a summary in Table 1. Some of the boundaries can still be recognized where there are large jumps in the data, but others are not so obvious.

By the 1960s, taxonomists were putting numerical limits on the discriminating criteria for consistency. This helped to codify description. It did nothing, however, to quantify the variation in properties that could not be assessed readily in the field; and it was unhelpful to the map-maker who wished to place boundaries where there were maxima in the rate of change in the landscape. Description needed a formal statistical basis, a need first recognized by civil engineers in the 1960s.

Sampling and Estimation

In the classical approach, a soil property, z, takes values at an infinity of points, $\mathbf{x}_i = \{x_{i1}\, x_{i2}\}$, $i = 1, 2, \ldots, \infty$, in a region $\mathcal{R}$. These values, $z(\mathbf{x}_i)$, comprise the population, which has a mean, μ, and variance, here denoted as σ_i^2 signifying the total variance in $\mathcal{R}$. The region is divided into K spatial strata or classes, $\mathcal{R}_k$, $k = 1, 2, \ldots, K$, which are mutually exclusive and exhaustive and which are what the map displays; each has its own mean and variance, denoted μ_k and σ_k^2, respectively. The region is then sampled, and the property at the N sampling points is measured to give data, $z(\mathbf{x}_1), z(\mathbf{x}_2), \ldots, z(\mathbf{x}_N)$, of which n_k belong in class $\mathcal{R}_k$.

If sampling is unbiased, then the mean for the kth class is estimated simply by:

$$\hat{\mu}_k = \bar{z}_k = \frac{1}{n_k}\sum_{i=1}^{n_k} z(\mathbf{x}_i) \quad \text{for} \quad \mathbf{x}_i \in \mathcal{R}_k \qquad [1]$$

The variance within $\mathcal{R}_k$ is estimated from the same sample by

$$\hat{\sigma}_k^2 = s_k^2 = \frac{1}{n_k - 1}\sum_{i=1}^{n_k} \{z(\mathbf{x}_i) - \bar{z}_k\}^2 \quad \text{for} \quad \mathbf{x}_i \in \mathcal{R}_k \quad [2]$$

In the classical approach, the soil map plus the class means and variances summarize the available information on the spatial variation of z in $\mathcal{R}$. The analysis may be elaborated by adding skewness coefficients, computed from the third moments about the means

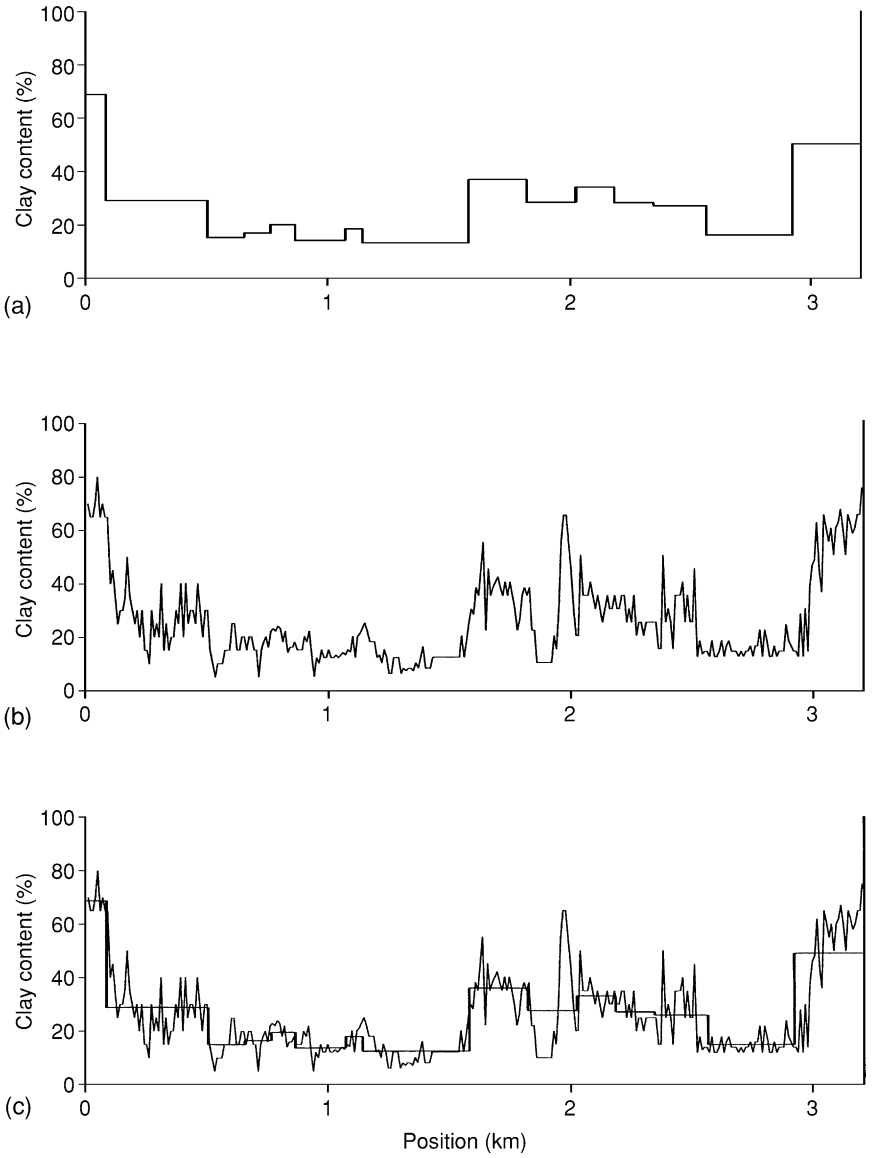


Figure 1 Variation of clay content along a transect: (a) as represented by a classification; (b) actual values measured in topsoil at 10-m intervals; and (c) superimposition of the classification on the reality.

Table 1 Summary of 321 observations of clay content (recorded as percentage by weight) in topsoil and subsoil

	Topsoil	*Subsoil*
Mean (%)	25.6	39.0
Median (%)	20.0	36.0
Variance $(\%)^2$	255.49	936.81
Standard deviation (%)	16.0	30.6
Skewness	1.2	0.2

for the classes, and perhaps higher-order moments. The data might be transformed so that they approximate a normal distribution and so that the means and variances are sufficient summaries.

More importantly, the statistics can be embraced in a single summary. When a soil surveyor subdivides a particular region to display the spatial distribution of the soil, he or she usually tries to create classes of the same categorical level, for example, all soil series or all soil families. Ideally the variances within these are equal, i.e., there is a common within-class variance:

$$\sigma_{\mathrm{W}}^2 = \sigma_k^2 \quad \text{for all} \quad k \qquad [3]$$

The differences between classes can be represented by the between-class variance, σ_{B}^2. This is essentially the variance among the means of the classes, and can be estimated from the data as follows. A quantity B is calculated as the sum of the squares of the differences

Table 2 One-way analysis of variance

Source	*Degrees of freedom*	*Mean square*	*Parameter estimated*
Between classes	$K-1$	$\frac{1}{K-1}\sum_{k=1}^{K} n_k(\bar{z}_k - \bar{\bar{z}})^2$	$n^*\sigma_B^2 + \sigma_W^2$
Within classes	$N-K$	$\frac{1}{N-K}\sum_{k=1}^{K}\sum_{i=1}^{n_k}\{z(\mathbf{x}_{ik}) - \bar{z}_k\}^2$	σ_W^2
Total	$N-1$	$\frac{1}{N-1}\sum_{i=1}^{N}\{z(\mathbf{x}_i) - \bar{\bar{z}}\}^2$	σ_T^2

between the class means and the mean of all the data, $\bar{z}_k - \bar{\bar{z}}$, the latter being the equivalent of $\hat{\mu}$:

$$B = \frac{1}{K-1}\sum_{k=1}^{K} n_k(\bar{z}_k - \bar{\bar{z}})^2 \quad [4]$$

If all the classes are sampled equally, so that $n = n_k$ for all k, then s_B^2 is computed from B simply by:

$$s_B^2 = (B - s_W^2)/n \quad [5]$$

If the n_k are not equal, n in Eqn [5] is replaced by:

$$n^* = \frac{1}{K-1}\left(N - \frac{\sum_{k=1}^{K} n_k^2}{N}\right) \quad [6]$$

and

$$s_B^2 = (B - s_W^2)/n^* \quad [7]$$

The whole can be set out in a one-way analysis of variance, as in Table 2.

For the whole population in $\mathcal{R}$, the variances σ_W^2 and σ_B^2 sum to the total variance, σ_T^2:

$$\sigma_T^2 = \sigma_W^2 + \sigma_B^2 \quad [8]$$

The ratios of these variances describe the relative effects of the spatial classification. One ratio is the intraclass correlation:

$$\rho_i = \frac{\sigma_B^2}{\sigma_W^2 + \sigma_B^2} \quad [9]$$

estimated by:

$$r_i = \frac{s_B^2}{\sigma_W^2 + \sigma_B^2} = \frac{B - s_W^2}{B + (n^* - 1)s_W^2} \quad [10]$$

The last expression enables the intraclass correlation to be calculated directly from the table of analysis of variance. The term derives its name from the fact that it expresses the 'correlation' among individuals within the same class.

The theoretical maximum of ρ_i is 1 when every class is uniform ($\sigma_W^2 = 0$). In practice there is always some variation within the classes, and so $\rho_i < 1$. Its theoretical minimum is zero when all the μ_k are equal, so that $\sigma_B^2 = 0$. Its estimate, r_i, and the estimate s_B^2 are often negative. The usual cause is sampling fluctuation, where the differences between means are small in relation to the variation within the classes, and one can take negative values of r_i as estimates of $\rho_i = 0$.

Another ratio expressing the effectiveness of the classification is simply s_W^2/s_T^2, sometimes called the 'relative variance.' Its complement, $1 - s_W^2/s_T^2$, can be regarded as the proportion of the variance accounted for by the classification, and in this respect it is like the coefficient of determination, R^2, in regression analysis. Like the intraclass correlation, it varies between 1 (uniformity within classes) and 0 (no differences between them), and for large N and K the two have very similar values.

This analysis is applied to the classification of the clay content of the topsoil at Sandford. Table 3 lists the means and variances of the 15 classes, and Table 4 summarizes the analysis of variance.

Table 3 Means and variances of clay content in the topsoil, recorded as percentage by weight, for 15 classes

Class	*Mean*	*Variance*
1	68.8	26.79
2	28.8	110.74
3	15.0	32.14
4	16.5	21.07
5	19.6	12.93
6	13.7	12.33
7	17.9	17.14
8	12.4	21.96
9	35.9	54.69
10	27.6	380.36
11	33.1	42.92
12	27.2	23.23
13	26.0	121.85
14	15.0	9.71
15	49.1	341.78

Table 4 Analysis of variance of clay content of topsoil

Source	*Degrees of freedom*	*Mean square*	*F-ratio*
Between classes	14	3786.67	40.3
Within classes	306	93.94	
Total	320	255.5	

The within-class component of variance (in Table 4) is $s^2_W = 93.94$, and the between-class component, s^2_B, is 176.1. This leads to an intraclass correlation for the classification as 0.65. The complement of the relative variance is 0.63.

There are substantial differences between the variances within the individual classes, so it is something of a liberty to treat them as estimates of the same quantity to arrive at sensible values for r_i and $1 - s^2_W/s^2_T$.

The Geostatistical Approach

The artificiality of imposing boundaries between classes to describe variation that is patently continuous worried quantitatively minded soil scientists, and soil physicists in a particular, for many years. A practicable alternative eluded them, however. They toyed with polynomials, but any such function would have to be of a very high order and could have no generality. The variation was too complex, perhaps chaotic, as Figure 1b shows. Such variation looks as though it might be random. It was this last idea that provided the breakthrough: if the variation appears random then why not treat it as if it were random? This is the basis of modern geostatistics and its approach to describing soil variation.

Random Variables and Random Functions

As in the classical approach, a region $\mathcal{R}$ is regarded as comprising an infinite number of points $\mathbf{x}_i$, $i = 1, 2, \ldots, \infty$. Whereas in the classical approach the values of z at these points constitute the population, in the geostatistical approach this population is assumed to be just one realization of a random process or random function that could generate any number of such populations. At each place $\mathbf{x}$ the soil property is a random variable, $Z(\mathbf{x})$ – notice the capital 'Z' – of many values. For a continuous variable such as hydraulic conductivity or pH, this number is infinite, and the whole process may be regarded as a doubly infinite superpopulation. The random variable at $\mathbf{x}$ has a distribution with a mean and variance and higher-order moments, and the actual value there, $z(\mathbf{x})$, is just one drawn at random from that distribution.

In these circumstances the quantitative description of the variation involves estimating the characteristics of what are assumed to be the underlying random processes. The characteristics include the means and variances, and perhaps higher-order moments, but most importantly the spatial covariances.

The spatial covariance between the variables at any two places $\mathbf{x}_1$ and $\mathbf{x}_2$ is given by:

$$C(\mathbf{x}_1, \mathbf{x}_2) = \mathrm{E}\left[\{Z(\mathbf{x}_1) - \mu(\mathbf{x}_1)\}\{Z(\mathbf{x}_2) - \mu(\mathbf{x}_2)\}\right] \quad [11]$$

where $\mu(\mathbf{x}_1)$ and $\mu(\mathbf{x}_2)$ are the means at $\mathbf{x}_1$ and $\mathbf{x}_2$, and E denotes the expected value. In practice $C(\mathbf{x}_1, \mathbf{x}_2)$ cannot be estimated, because there is only ever the one realization, and to overcome this apparent impasse assumptions of stationarity must be invoked.

Stationarity

Starting with the first moment, we assume that the mean, $\mu = \mathrm{E}[Z(\mathbf{x})]$, is constant for all $\mathbf{x}$, and so $\mu(\mathbf{x}_1)$ and $\mu(\mathbf{x}_2)$ can be replaced by the single value μ, which is estimated by repetitive sampling.

Next, when $\mathbf{x}_1$ and $\mathbf{x}_2$ coincide, Eqn [11] defines the variance, $\sigma^2 = \mathrm{E}[\{Z(\mathbf{x}) - \mu\}^2]$. This is assumed to be finite and, like the mean, to be the same everywhere. Equation [11] is then generalized so that it applies to any pair of points $\mathbf{x}_i$ and $\mathbf{x}_j$ separated by a vector, or lag $\mathbf{h} = \mathbf{x}_i - \mathbf{x}_j$, so that:

$$\begin{aligned} C(\mathbf{x}_i, \mathbf{x}_j) &= \mathrm{E}\left[\{Z(\mathbf{x}_i) - \mu\}\{Z(\mathbf{x}_j) - \mu\}\right] \\ &= \mathrm{E}\left[\{Z(\mathbf{x})\}\{Z(\mathbf{x} + \mathbf{h})\} - \mu^2\right] \\ &= C(\mathbf{h}) \end{aligned} \quad [12]$$

and this is also constant for any given $\mathbf{h}$. This constancy of the mean and variance and of a covariance that depends only on separation and not on absolute position constitutes second-order stationarity.

Equation [12] shows that the covariance is a function of the lag and only of the lag; it describes quantitatively the dependence between values of Z with changing lag. It is readily converted to the dimensionless autocorrelation by:

$$\rho(\mathbf{h}) = C(\mathbf{h})/C(\mathbf{0}) \quad [13]$$

where $C(\mathbf{0}) = \sigma^2$ is the covariance at lag $\mathbf{0}$.

Intrinsic Variation and the Variogram

In many instances the assumption of constant mean throughout a region is untenable, and if the mean changes the variance will appear to increase indefinitely with increasing area. The covariance cannot be defined then, because there is no value for μ to insert in Eqn [12]. Faced with this situation, geostatisticians consider the differences from place to place, and their squares, as follows. For small lag distances, the expected differences are zero:

$$\mathrm{E}\left[Z(\mathbf{x}) - Z(\mathbf{x} + \mathbf{h})\right] = 0 \quad [14]$$

and the expected squared differences define the variances for those lags:

$$\begin{aligned} \mathrm{E}\left[\{Z(\mathbf{x}) - Z(\mathbf{x} + \mathbf{h})\}^2\right] &= \mathrm{var}\left[Z(\mathbf{x}) - Z(\mathbf{x} + \mathbf{h})\right] \\ &= 2\gamma(\mathbf{h}) \end{aligned} \quad [15]$$

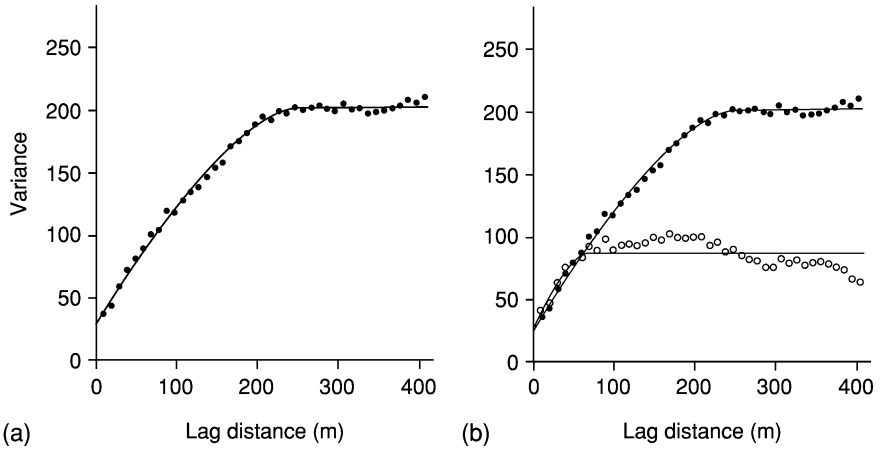


Figure 2 Experimental variograms of clay content of topsoil and the spherical models fitted to them: (a) global variogram with black circles for the experimental semivariances; and (b) the global variogram plus the within-class variogram with open circles for the experimental values.

Eqn [15] gives the variance of the difference at lag $\mathbf{h}$; and the quantity $\gamma(\mathbf{h})$, known as the semivariance, is the variance per point. Equations [14] and [15] constitute the intrinsic hypothesis of geostatistics. Like the covariance, the semivariance depends only on the lag and not on the absolute positions $\mathbf{x}$ and $\mathbf{x}+\mathbf{h}$. As a function, $\gamma(\mathbf{h})$ is the variogram, often still called the 'semivariogram.'

If the process $Z(\mathbf{x})$ is second-order stationary then the semivariance and the covariance are equivalent:

$$\begin{aligned}\gamma(\mathbf{h}) &= C(\mathbf{0}) - C(\mathbf{h}) \\ &= \sigma^2\{1 - \rho(\mathbf{h})\}\end{aligned} \qquad [16]$$

If it is intrinsic only, the covariance does not exist, but the semivariance remains valid, and it is this validity in a wide range of circumstances that makes the variogram so useful in summarizing spatial variation.

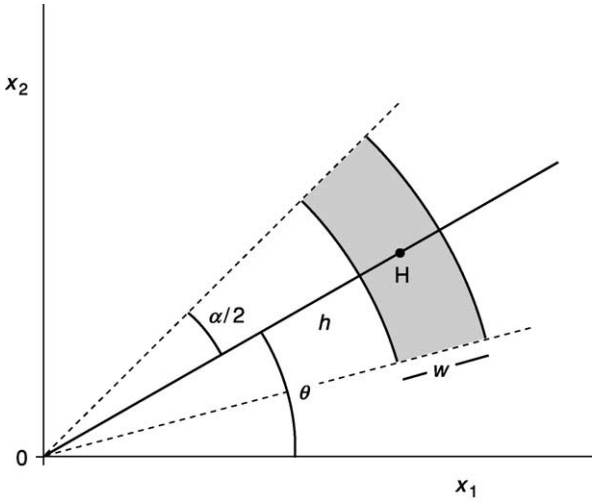


Figure 3 Discretization of the lag in two dimensions for irregularly scattered data. All separations within the gray sector are assigned to lag distance h and direction θ.

Estimating the Variogram

Semivariances are readily estimated from data, $z(\mathbf{x}_1)$, $z(\mathbf{x}_2), \ldots$, by the method of moments:

$$\hat{\gamma}(\mathbf{h}) = \frac{1}{2m(\mathbf{h})}\sum_{i=1}^{m(\mathbf{h})}\{z(\mathbf{x}_i) - z(\mathbf{x}_i+\mathbf{h})\}^2 \qquad [17]$$

in which $m(\mathbf{h})$ is the number of paired comparisons at lag $\mathbf{h}$. By changing $\mathbf{h}$ we obtain a sample or experimental variogram, which can be displayed as a graph of $\hat{\gamma}$ against $\mathbf{h}$. Figure 2a is an example in which the experimental semivariances for the data in Figure 1b are plotted as points against the lag distance, $h = |\mathbf{h}|$, for the one-dimensional transect.

The values of $\mathbf{h}$ define discrete points on the variogram, and so sampling is best planned with regular intervals along a line in one dimension or on a grid in two or three. Otherwise the actual separations have to be placed into 'bins,' with limits in separating distance and also in direction if there is more than one dimension (Figure 3).

Models for Variograms

The underlying variogram, Eqn [15], is a continuous function in as many dimensions as the variable $Z(\mathbf{x})$. The experimental variogram estimates it at a set of points with more or less error and point-to-point fluctuation arising from the sampling. To obtain a variogram to describe the spatial variation in $\mathcal{R}$, a plausible function is fitted to the experimental values. The usual approach is to fit the simplest model that makes sense.

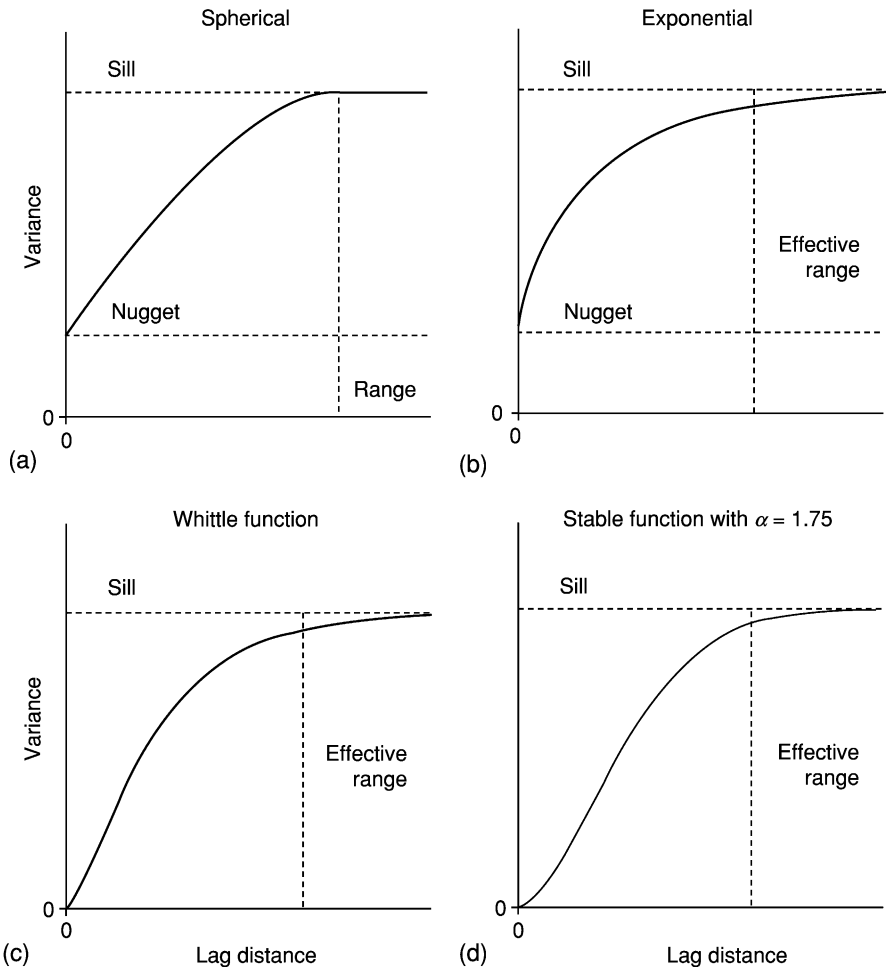


Figure 4 Four kinds of bounded variogram.

Figure 4a shows the principal features of many, if not most, experimental variograms. They are as follows:

1. The variance increases from near the ordinate with increasing lag distance;
2. The variance reaches a maximum at which it remains thereafter;
3. Any simple smooth line or surface placed through the points and projected to the ordinate cuts the ordinate at some value greater than zero.

The model must also be mathematically acceptable in that it cannot give rise to 'negative variances' when random variables are combined. Let $z(\mathbf{x}_i)$, $i = 1, 2, \ldots, n$, be a realization of the random variable $Z(\mathbf{x})$ with covariance function $C(\mathbf{h})$ and variogram $\gamma(\mathbf{h})$, and consider the linear sum:

$$y = \sum_{i=1}^{n} \lambda_i z(\mathbf{x}_i) \qquad [18]$$

where the λ_i are any arbitrary weights. The variable Y from which y derives is also a random variable, and its variance is given by:

$$\text{var}[Y] = \sum_{i=1}^{n} \sum_{j=1}^{n} \lambda_i \lambda_j C(\mathbf{x}_i - \mathbf{x}_j) \qquad [19]$$

where $C(\mathbf{x}_i - \mathbf{x}_j)$ is the covariance of Z between $\mathbf{x}_i$ and $\mathbf{x}_j$. This variance must be positive or zero; it may not be negative. If $Z(\mathbf{x})$ is intrinsic only then the covariances do not exist, and we must rewrite [19] as:

$$\text{var}[Y] = C(\mathbf{0}) \sum_{i=1}^{n} \lambda_i \sum_{j=1}^{n} \lambda_j - \sum_{i=1}^{n} \sum_{j=1}^{n} \lambda_i \lambda_j \gamma(\mathbf{x}_i - \mathbf{x}_j) \qquad [20]$$

where $\gamma(\mathbf{x}_i - \mathbf{x}_j)$ is the semivariance of Z between $\mathbf{x}_i$ and $\mathbf{x}_j$. The first term on the right-hand side of this equation is eliminated if the weights sum to zero, so that:

$$\text{var}[Y] = -\sum_{i=1}^{n}\sum_{j=1}^{n}\lambda_i\lambda_j\gamma(\mathbf{x}_i - \mathbf{x}_j) \qquad [21]$$

This too must guarantee non-negative variances, and only functions that do that are admissible. They are said to be 'conditional negative semidefinite,' the condition being that the weights in Eqn [21] sum to zero.

There are only a few families of simple functions that satisfy the above criteria. They can be divided into those that are bounded and those that are not. In the first group are the popular spherical and exponential models. Their formulae in their isotropic forms, i.e., for $h = |\mathbf{h}|$, are as follows.

Spherical The spherical function has the following equation:

$$\begin{aligned}\gamma(h) &= c_0 + c\left\{\frac{3}{2}\left(\frac{h}{a}\right) - \frac{1}{2}\left(\frac{h}{a}\right)^3\right\} \quad \text{for} \quad 0 < h \le a \\ &= c \quad \text{for} \quad h > a \\ &= 0 \quad \text{for} \quad h = 0\end{aligned} \qquad [22]$$

Here $\gamma(h)$ is the semivariance at lag h, and c is the *a priori* variance of the autocorrelated process. The quantity c_0 is the intercept on the ordinate and is known as the 'nugget variance,' a term derived from gold-mining. The combined $c_0 + c$ is known as the 'sill' of the model, and c is the sill of the correlated variance. These quantities are illustrated in **Figure 4a**, and **Figure 2a** shows the function fitted to the experimental variogram of clay content, **Figure 1b**. The values of the parameters, c_0, c, and a, are listed in **Table 5**.

The function has a distance parameter, a; this is its range, also know as its 'correlation range.' It marks the limit of spatial dependence; values at places closer to one another than a are more or less correlated, whereas those further apart are not. It implies that all the variance in $\mathcal{R}$ is encountered within that distance, and in this sense it corresponds to the concept of the representative elementary volume (REV). The spherical function gets its name from the formula for the volume of two intersecting spheres, which are of diameter a.

Table 5 Parameters of spherical models, Eqn [22], fitted to the experimental global and within-class variograms of clay in topsoil and subsoil

		Nugget, c_0	*Sill*, c	*Range*, a (m)
Topsoil	Global	28.2	172.8	265
	Within-class	28.1	59.3	70.9
Subsoil	Global	117.5	551.4	191
	Within-class	108.4	257.2	77.1
Subsoil × *topsoil*	Global	−1.1	210.8	229.6

The semivariance at lag zero is itself zero, and for continuous processes such as most physical properties of the soil, $\gamma(h)$ should increase gradually as h increases from zero. In practice, there are usually insufficient estimates of $\gamma(h)$ near the ordinate to fit a model through the origin, and therefore the conservative approach (described above) is taken. The nugget variance is therefore best regarded as embodying variation within the shortest sampling interval plus any measurement error.

Other functions with the same general form and finite ranges are the bounded linear function (valid in one dimension only), the circular (valid in one and two dimensions, but not in three), and the pentaspherical.

Exponential The equation for the exponential function is:

$$\gamma(h) = c_0 + c\left\{1 - \exp\left(-\frac{h}{r}\right)\right\} \qquad [23]$$

in which c_0 and c have the same meanings as before, but now with a distance parameter, r. The exponential model approaches its sill asymptotically and has no definite range therefore. A working range is often taken as $a' = 3r$, at which point the function has reached 95% of c. This model is shown in **Figure 4b**.

Models with reverse curvature at the origin Some variograms appear to approach the origin with decreasing gradients. These may be represented by the general equation:

$$\gamma(h) = c\left\{1 - \exp\left(-\frac{h^\alpha}{r^\alpha}\right)\right\} \qquad [24]$$

in which $0 < \alpha \le 2$. If $\alpha = 2$ we have the Gaussian function. This is at the limit of acceptability and gives rise to unstable prediction. It is best replaced by stable models with $\alpha < 2$; **Figure 4d** is an example. Another recommended function to describe such variation is the Whittle elementary correlation:

$$\gamma(h) = c\left\{1 - \frac{h}{r}K_1\left(\frac{h}{r}\right)\right\} \qquad [25]$$

in which r is again a distance parameter, and K_1 is the modified Bessel function of the second kind (**Figure 4c**). It has the added attraction in that it derives theoretically from diffusion in two dimensions. To all can be added a nugget variance, c_0, if desired.

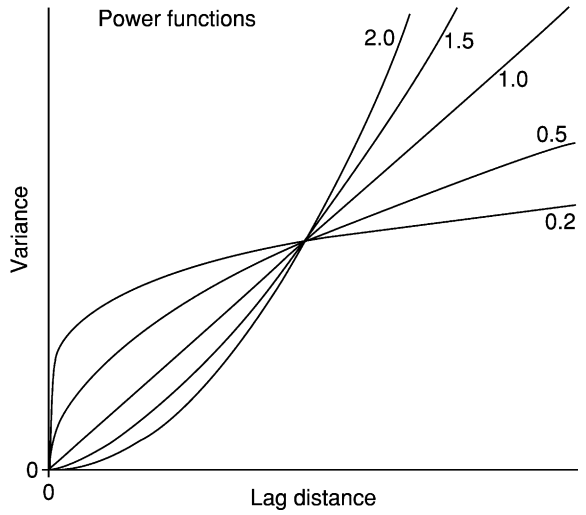


Figure 5 Four valid power functions, Eqn [26] with exponents, α, 0.2, 0.5, 1.0, and 1.5, plus the inadmissible limiting function with $\alpha = 2$.

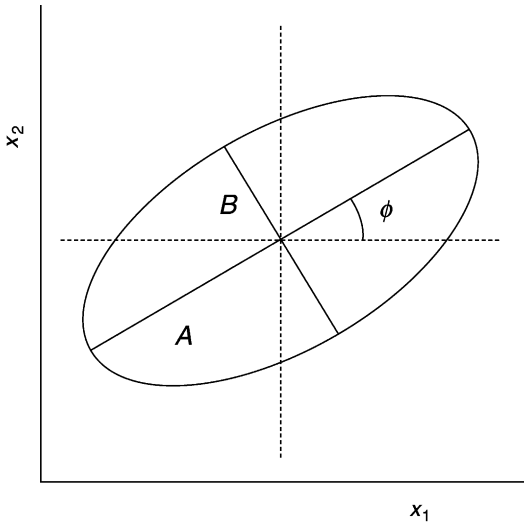


Figure 6 Ellipse showing parameters of anisotropy.

Unbounded models Variograms of processes that are intrinsic but not second-order stationary increase without bound as the lag distance increases. These can usually be fitted by power functions, for which the general equation including a nugget is:

$$\gamma(h) = c_0 + wh^{\alpha} \qquad [26]$$

The parameter w describes the intensity of the process, and the exponent, which must lie strictly between 0 and 2 (these limits are excluded), describes the curvature. If $\alpha < 1$ the curve is convex upward; if it is 1 we have a straight line; and if $\alpha > 1$ the curve is concave upward. The curve with $\alpha = 2$ is a parabola and describes a smoothly continuous process that is not random. Figure 5 shows the curves for several values of α.

Anisotropy The variogram of a two-dimensional process is itself two-dimensional, and if the process is anisotropic so is its variogram, which is then a function of both distance h and direction θ. In the simplest cases, the anisotropy is geometric, meaning that it can be made isotropic by a linear transformation of the coordinates. The transformation is defined by reference to an ellipse:

$$\Omega(\theta) = \sqrt{A^2\cos^2(\theta - \phi) + B^2\sin(\theta - \phi)} \qquad [27]$$

where A and B are the long and short diameters, respectively, of the ellipse, and ϕ is its orientation, i.e., the direction of the long axis (Figure 6). Equation [27] is embodied into the models as follows: For the bounded models, Ω replaces the distance parameter

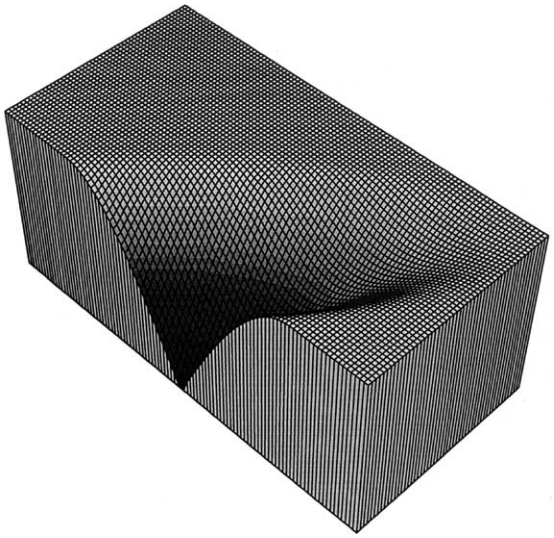


Figure 7 Perspective diagram of the variogram surface of an anisotropic spherical function.

of the isotropic variogram. So, for example, in the exponential:

$$\gamma(h,\theta) = c_0 + c\left\{1 - \exp\left(-\frac{h}{\Omega(\theta)}\right)\right\} \qquad [28]$$

and in the linear function:

$$\gamma(h,\theta) = c_0 + \Omega(\theta)h^{\alpha} \qquad [29]$$

in which $\alpha = 1$. Figure 7 shows an anisotropic spherical function. Notice how the range of the model changes with changing direction.

Combining Trend and Random Fluctuation

The above functions describe processes that are entirely random though correlated. We can represent the processes by the general model:

$$Z(\mathbf{x}) = \mu_V + \epsilon(\mathbf{x}) \quad [30]$$

in which μ_V is the mean, i.e., constant, in some neighbourhood V, and $\epsilon(\mathbf{x})$ is the autocorrelated variance as defined in Eqn [15]. It often happens that such models are unacceptable, either because there is an evident long-range trend across a region or because over short distances the variation appears smooth. In these circumstances, μ_V cannot be treated as constant but must be replaced by a deterministic term, say $u(\mathbf{x})$, that depends on the position $\mathbf{x}$. The model becomes:

$$Z(\mathbf{x}) = u(\mathbf{x}) + \epsilon(\mathbf{x}) \quad [31]$$

If $u(\mathbf{x})$ can describe the variation over the whole of $\mathcal{R}$ it is called 'trend.' If it is local only then is it known as 'drift.' In either event it is usually represented by a low-order polynomial, so Eqn [31] becomes:

$$Z(\mathbf{x}) = \sum_{j=0}^{J} a_j f_j(\mathbf{x}) + \epsilon(\mathbf{x}) \quad [32]$$

in which the a_j are unknown coefficients and the $f_j(\mathbf{x})$ known functions of our choosing.

It is fairly easy, even if somewhat arbitrary, to separate any long-range trend from the short-range, apparently random fluctuation and to estimate the parameters of the two components separately. It is not at all easy to do it where there is short-range drift. In these circumstances it involves a full structural analysis, effectively a process of trial and error.

Combining Classification with Geostatistics

In some instances neither a classification nor a variogram alone can serve to represent spatial variation in soil properties. The choropleth map implies abrupt changes, whereas the variogram is based on a model of random but continuous fluctuation. If there appears to be both kinds of variation then the two approaches may be combined. By recognizing the class boundaries, that is, by combining the information in **Figure 1a** and **1b**, and analyzing the variance (**Table 4**), residuals can be obtained from the class means. Their variance is the residual mean square, and a portion of this is likely to be autocorrelated and have its own variogram. **Figure 2b** shows by the circles an example of a within-class variogram obtained by superimposing the classification on the data. The curve through the points is again that of a spherical model with parameter values as given in **Table 5**.

The variogram of the residuals differs from the variogram of the original data in two important respects:

1. The sill of the fitted model is less by an amount approximately equal to the between-class variance, as expected;
2. The range of the model is much less. This is because the class-to-class variation, which evidently dominated the variation over the whole transect, has been removed to leave only the short-range correlation.

Coregionalization – Simultaneous Variation in Two or More Variables

Any two variables, say z_u and z_v, may be correlated and in particular linearly correlated. That relation is conventionally expressed by the product–moment correlation coefficient:

$$\rho = \frac{\mathrm{cov}[uv]}{\sqrt{\mathrm{var}[u] \times \mathrm{var}[v]}} \quad [33]$$

i.e., the covariance of z_u and z_v divided by the product of their standard deviations.

The two spatial random variables, $Z_u(\mathbf{x})$ and $Z_v(\mathbf{x})$, may also be spatially intercorrelated in that each is spatially correlated both with itself, i.e., autocorrelated, and with the other. The two variables are then said to be cross-correlated. In these circumstances, the two variables have autovariograms, one each, as defined by Eqn [15] and for present purposes denoted $\gamma_{uu}(\mathbf{h})$ and $\gamma_{vv}(\mathbf{h})$. They also have a cross-variogram, $\gamma_{uv}(\mathbf{h})$, defined by:

$$\gamma_{uv}(\mathbf{h}) = \frac{1}{2}\mathrm{E}\left[\{Z_u(\mathbf{x}) - Z_u(\mathbf{x}+\mathbf{h})\}\{Z_v(\mathbf{x}) - Z_v(\mathbf{x}+\mathbf{h})\}\right] \quad [34]$$

If both variables are second-order stationary with means μ_v and μ_v then will both have covariance functions, C_{uu} and C_{vv} as defined in Eqn [12], and a cross-covariance:

$$C_{uv}(\mathbf{h}) = \mathrm{E}[\{Z_u(\mathbf{x}) - \mu_u\}\{Z_v(\mathbf{x}+\mathbf{h}) - \mu_v\}] \quad [35]$$

There is also a cross-correlation coefficient, ρ_{uv}, given by:

$$\rho_{uv}(\mathbf{h}) = \frac{C_{uv}}{\sqrt{C_{uu}(0)C_{vv}(0)}} \quad [36]$$

This is effectively the extension of the Pearson product–moment correlation coefficient of Eqn [33] into the spatial domain, and when $\mathbf{h} = 0$ it is the Pearson coefficient.

The cross-covariance is in general not symmetric, i.e.:

$$\mathrm{E}[\{Z_u(\mathbf{x}) - \mu_u\}\{Z_v(\mathbf{x}+\mathbf{h}) - \mu_v\}] \neq \mathrm{E}[\{Z_v(\mathbf{x}) - \mu_v\}\{Z_v(\mathbf{x}+\mathbf{h}) - \mu_u\}] \quad [37]$$

In words, the cross-covariance between $Z_u(\mathbf{x})$ and $Z_v(\mathbf{x})$ in one direction is different from that in the other, or, expressed another way:

$$C_{uv}(\mathbf{h}) \neq C_{uv}(-\mathbf{h}) \quad \text{or equivalently} \quad C_{uv}(\mathbf{h}) \neq C_{vu}(\mathbf{h}) \quad [38]$$

since:

$$C_{uv}(\mathbf{h}) = C_{vu}(-\mathbf{h})$$

Asymmetry can be envisaged between two soil properties at different depths on a slope as a result of creep or solifluction. The subsoil would tend to lag behind the topsoil. Similarly, irrigation by flooding always from the same end of a field might distribute salts differentially in the direction of flow, but asymmetric covariances have not been reported in the literature as of 2003, as far as I know.

The cross-variogram and the cross-covariance function (if it exists) are related by:

$$\gamma_{uv}(\mathbf{h}) = C_{uv}(\mathbf{0}) - \frac{1}{2}\{C_{uv}(\mathbf{h}) + C_{uv}(-\mathbf{h})\} \quad [39]$$

This quantity contains both $C_{uv}(\mathbf{h})$ and $C_{uv}(-\mathbf{h})$ and in consequence loses any information on asymmetry; it is an even function, i.e., symmetric:

$$\gamma_{uv}(\mathbf{h}) = \gamma_{vu}(\mathbf{h}) \quad \text{for all} \quad \mathbf{h}$$

Cross-semivariances can be estimated in a way similar to that of the autosemivariances by:

$$\hat{\gamma}_{uv}(\mathbf{h}) = \frac{1}{2m(\mathbf{h})}\sum_{i=1}^{m(\mathbf{h})}\{z_u(\mathbf{x}_i) - z_u(\mathbf{x}_i+\mathbf{h})\}\{z_v(\mathbf{x}_i) - z_v(\mathbf{x}_i+\mathbf{h})\} \quad [40]$$

and the sample cross-variogram is formed by simple incrementation of $\mathbf{h}$. There is an equivalent formula for computing the cross-covariances. Notice that there must be numerous places where both z_u and z_v have been measured.

Modeling the Coregionalization

The cross-variogram can be modeled in the same way as the autovariogram, and the same restricted set of functions is available. There is one additional constraint. Any linear combination of the variables is itself a regionalized variable, and its variance cannot be negative. This is assured by adopting the linear model of coregionalization. In it the variable $Z_u(\mathbf{x})$ is assumed to be the sum of independent (orthogonal) random variables, $Y_j^k(\mathbf{x})$:

$$Z_u(\mathbf{x}) = \sum_{k=1}^{K}\sum_{j=1}^{k} a_{uj}^k Y_j^k(\mathbf{x}) + \mu_u \quad [41]$$

in which the superscript k is an index, not a power. There is a similar assumption for $Z_v(\mathbf{x})$. If the assumptions hold then the pair of variables has a cross-variogram:

$$\gamma_{uv}(\mathbf{h}) = \sum_{k=1}^{K}\sum_{j=1}^{k} a_{uj}^k a_{vj}^k g^k(\mathbf{h}) \quad [42]$$

The products in the second summation can be replaced by b_{uv}^k to give

$$\gamma_{uv}(\mathbf{h}) = \sum_{k=1}^{K} b_{uv}^k g^k(\mathbf{h}) \quad [43]$$

The quantities are the variances and covariances, e.g., nugget and sill variances, for the independent components of a spherical model. For two variables there are the three nugget variances, b_{uu}^1, b_{vv}^1, and b_{uv}^1, and similarly three for the sills of the correlated variances. The coefficients $b_{uv}^k = b_{vu}^k$ for all k, and, for each k, the matrix of coefficients:

$$\begin{bmatrix} b_{uu}^k & b_{uv}^k \\ b_{vu}^k & b_{vv}^k \end{bmatrix}$$

must be positive definite. Since the matrix is symmetric, it is sufficient that $b_{uu}^k \geq 0$ and $b_{vv}^k \geq 0$ and that its determinant is positive or zero:

$$\left|b_{uv}^k\right| = \left|b_{vu}^k\right| \leq \sqrt{b_{uu}^k b_{vv}^k} \quad [44]$$

This is Schwarz's inequality.

Any number of regionalized variables may be embodied in the linear model of coregionalization. If there are V of them the full matrix of coefficients, $[b_{ij}]$, will be of order V, and its determinant and all its principal minors must be positive or zero.

Schwarz's inequality has the following consequences:

1. Every basic structure, $g^k(\mathbf{h})$, present in the cross-variogram must also appear in the two autovariograms, i.e., $b_{uu}^k \neq 0$ and $b_{vv}^k \neq 0$ if $b_{vv}^k \neq 0$. As a corollary, if a basic structure $g^k(\mathbf{h})$ is absent from either autovariogram it may not be included in the cross-variogram;

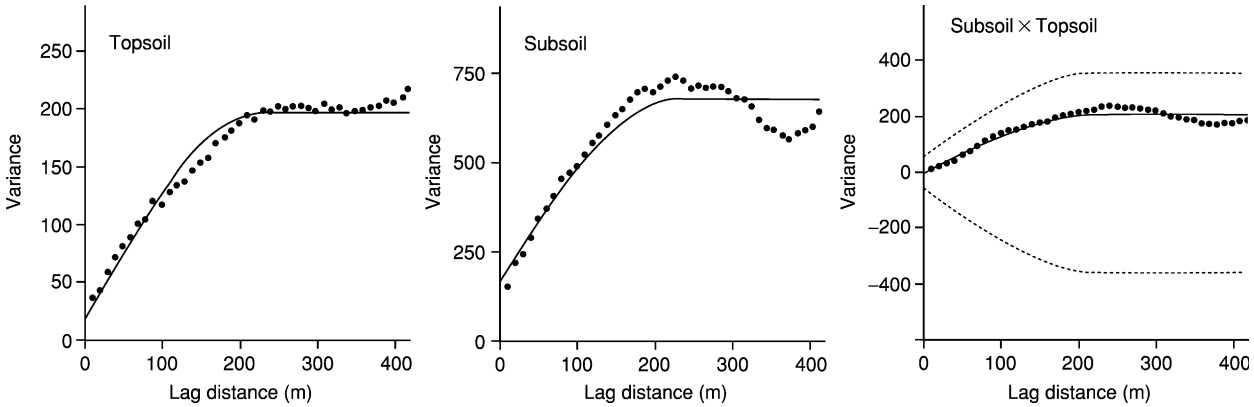


Figure 8 Auto- and cross-variograms of clay content in topsoil and subsoil with the linear model of coregionalization fitted. The dashed lines depict the hull of perfect correlation.

2. Structures may be present in the autovariograms without their appearing in the cross-variogram, i.e., b^k_{uv} may be zero when $b^k_{uu} > 0$ and $b^k_{vv} > 0$.

Parameters of the linear model of the coregionalization with the above constraints can be fitted by iteration. The distance parameters are usually first approximated by fitting models independently to the experimental variograms, and good compromise values are chosen from these. Then with those values fixed the values of the b^k_{uv} are found to minimize the sums of the squares of the residuals, subject to the condition that the solution guarantees non-negative variances, i.e., is conditionally negative semi-definite (CNSD). The validity of the resulting model may be checked by plotting it on a graph of the experimental cross-semivariances plus the limiting values that would hold if the correlation between the variables were perfect. These limits constitute the hull of perfect correlation, which is obtained from the coefficients b^k_{uu} and b^k_{vv} by:

$$\text{hull}\,[\gamma(\mathbf{h})] = \pm \sum_{k=1}^{K} \sqrt{b^k_{uu} b^k_{vv}}\, g^k(\mathbf{h})$$

The line should fit close to the experimental values for the model. It must also fall within the hull to be acceptable. If it lies close to the hull, the cross-correlation is strong; if, in contrast, it is far from the bounds, then the cross-correlation is weak.

Example The clay content of the subsoil was recorded at the same sampling points as those for topsoil in Figure 1b and, in combination with the topsoil data, illustrate the coregionalization. Table 1 contains a summary that includes the Pearson correlation coefficient, which is 0.59, indicating a modest correlation overall. The experimental autovariograms for the two depths together with their cross-variogram are shown in Figure 8 by the black circles. Spherical models can be fitted to all of them, and their parameters are listed in Table 5. Their ranges vary between 191 and 265 m, with a mean of 228 m. If this value is fixed for the coregionalization model the nugget and sill component components of the model are obtained as described above (Table 6). They are somewhat different from those fitted independently, but only somewhat. The final result is shown in Figure 8 by the solid lines through the plotted points. The model evidently fits well.

Table 6 Fitted nugget variance and sill variances of the correlated structure, i.e., the coefficients b^k_{uv} of Eqn [43], of the model of coregionalization of clay in topsoil and subsoil

		Topsoil	*Subsoil*
Nugget variance	Topsoil	17.841	
	Subsoil	−0.799	167.986
Sill variance	Topsoil	178.659	
	Subsoil	211.436	509.723

The dashed lines on the graph for subsoil × topsoil define the hull of perfect correlation. The cross-variogram falls within the hull, as it should, and the moderate distance it keeps from the upper bound is a measure of moderate cross-correlation over the lag distances computed.

Spatial Prediction – Kriging

The variogram and covariance functions are not only elegant mathematical descriptions of the real world of the soil, they are crucial for local estimation, or spatial prediction as it might better be called, by kriging.

'Kriging' is a general term for processes of weighted averaging of data to provide unbiased local estimates of unknown values of a variable with minimum

variance. It is named after D.G. Krige, who developed it for estimating the gold content of ore bodies in South Africa. In the simplest case, for a place $\mathbf{x}_0$ where the mean is unknown (the usual situation), the estimate is formed as:

$$\hat{Z}(\mathbf{x}_0) = \sum_{i=1}^{N} \lambda_i z(\mathbf{x}_i) \qquad [45]$$

where the $z(\mathbf{x}_i), i = 1, 2, \ldots, N$ are sample data at places $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N$, and the λ_i are weights. The weights sum to 1 to assure unbiasedness, i.e., $\sum_{i=1}^{N} \lambda_i = 1$, and the variance of the estimate is given by:

$$\begin{aligned}\sigma^2(\mathbf{x}_0) &= \mathrm{E}\left[\left\{\hat{Z}(\mathbf{x}_0) - z(\mathbf{x}_0)\right\}^2\right] \\ &= 2\sum_{i=1}^{N} \lambda_i \gamma(\mathbf{x}_i, \mathbf{x}_0) \\ &\quad - \sum_{i=1}^{N}\sum_{j=1}^{N} \lambda_i \lambda_j \gamma(\mathbf{x}_i, \mathbf{x}_j) \qquad [46]\end{aligned}$$

Here $\gamma(\mathbf{x}_i, \mathbf{x}_j)$ is the semivariance between the data points $\mathbf{x}_i$ and $\mathbf{x}_j$, and $\gamma(\mathbf{x}_i, \mathbf{x}_0)$ is the semivariance between the data point $\mathbf{x}_i$ and the target point $\mathbf{x}_0$.

This variance is minimized by solution of the $N + 1$ equations:

$$\begin{aligned}\sum_{i=1}^{N} \lambda_i \gamma(\mathbf{x}_i, \mathbf{x}_j) + \psi(\mathbf{x}_0) &= \gamma(\mathbf{x}_i, \mathbf{x}_0) \,\forall\, j \\ \sum_{i=1}^{N} &= 1 \qquad [47]\end{aligned}$$

in which the $\psi(\mathbf{x}_0)$ is a Lagrange multiplier introduced for the minimization. The solution yields the optimum weights, and these are inserted into Eqn [45] to give the required estimate at $\mathbf{x}_0$.

The minimized variance is also obtained from the solution as

$$\sigma^2_{\mathrm{OK}}(\mathbf{x}_0) = \sum_{i=1}^{N} \lambda_i \gamma(\mathbf{x}_i, \mathbf{x}_0) + \psi(\mathbf{x}_0) \qquad [48]$$

This particular form of the technique is ordinary punctual kriging – 'ordinary' because it is the most used, and 'punctual' because the estimates are for points of the same size and shape, i.e., the same supports, as the bodies of soil or other material on which the measurements were made. It is readily generalized for estimating blocks, B, larger than the supports of the data. Eqn [45] holds for the averaging, though with B replacing $\mathbf{x}_0$. In Eqn [47] the individual semivariances on the right-hand sides are replaced by the means of the semivariances between the data points and the target block, B, $\bar{\gamma}(\mathbf{x}_i, B)$. Finally, the kriging variance is given by:

$$\sigma^2_{\mathrm{OK}}(B) = \sum_{i=1}^{N} \lambda_i \bar{\gamma}(\mathbf{x}_i, B) + \psi(B) - \bar{\gamma}(B, B) \qquad [49]$$

where $\bar{\gamma}(B, B)$ is the mean semivariance within B, i.e., the within-block variance.

When the kriging equations are solved, it usually turns out that only the few points nearest to the target point or block carry any appreciable weight; the weights of the others are so close to zero that they can be ignored. Kriging is thus a local weighted average. It has two other intuitively attractive features:

1. Where data points are clustered the weight of the cluster is divided among its members so that the individual weights are small compared with those of isolated points;
2. Where data points lie approximately in a line between the target and more distant points they screen the latter, which tend as a result to have virtually no weight, however close they are to the target.

This also has practical implications. The kriging systems, Eqn [47], need never be large; they are swiftly solved, and instabilities with matrix inversion are rare.

It is now evident why the variogram, or the equivalent covariance function, is so important; the kriging systems need values drawn from it. As above, these must not give rise to negative kriging variances, and so a valid function must be fitted to the experimental variogram.

Ordinary kriging will serve in some 90% of cases; it is the 'work horse' of practical geostatistics. It requires the fewest assumptions and the least knowledge. Kriging has been much used to map soil properties, including concentrations of plant nutrients, salinity, trace element contents, and nematode infestation. The kriging systems are solved at close intervals on a grid, from which isarithms, 'contours,' of the estimates can be drawn by other graphics programs. This has led to the application of kriging in land reclamation and precision agriculture. The kriging variances can be mapped similarly; patches of large variance coincide with sparse sampling, and so the maps can show where denser sampling is necessary or desirable to achieve more reliable estimates.

Kriging can be elaborated to embody other knowledge. Universal kriging takes into account known or estimated trend in the target variable. The underlying model is that of Eqn [31], usually with a simple

low-order polynomial for the trend, as in Eqn [32]. Another way of dealing with trend is to regard the target variable as an intrinsic random function of order k, $k > 0$ (IRF-k), and working with generalized covariances. Measurements on related subsidiary variables can be combined with those of the target variable by co-kriging, in which the semivariances are drawn from the model of coregionalization, Eqn [43], with all combinations of u and v in the model. Kriging with external drift embodies knowledge of the trend in related subsidiary variables and is an extension of universal kriging. Indicator kriging and disjunctive kriging form linear combinations of nonlinear transforms of data to estimate the probabilities of variables' exceeding specified threshold values, and these techniques are of potential value in statutory controls of pollution and restoration of contaminated land.

See also: **Spatial Patterns**; **Statistics in Soil Science**

Further Reading

Chilès J-P and Delfiner P (1999) *Geostatistics: Modeling Spatial Uncertainty*. New York: John Wiley.

Goovaerts P (1997) *Geostatistics for Natural Resources Evaluation*. New York: Oxford University Press.

Olea RA (1999) *Geostatistics for Engineers and Earth Scientists*. Boston, MA: Kluwer Academic Publishers.

Webster R (1985) Quantitative Spatial Analysis of Soil in the Field. *Advances in Soil Science* 3: 1–70.

Webster R and Oliver MA (2001) *Geostatistics for Environmental Scientists*. Chichester, UK: John Wiley.

SPECIFIC SURFACE AREA

K D Pennell, School of Civil & Environmental Engineering, Atlanta, GA, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

The surfaces of soil particles play a critical role in many processes, including chemical reactions, contaminant adsorption, colloid filtration, and water imbibition and drainage. The specific surface area of soils and soil constituents can range from less than $0.1\,m^2\,g^{-1}$ or $1 \times 10^2\,m^2\,kg^{-1}$ up to $800\,m^2\,g^{-1}$ or $8 \times 10^5\,m^2\,kg^{-1}$. Soils consisting primarily of sands (i.e., particle diameters of 0.05–2.0 mm or 5×10^{-5} m to 2×10^{-3} m) possess relatively small specific surface areas, usually less than $0.5\,m^2\,g^{-1}$ or $5 \times 10^2\,m^2\,kg^{-1}$. In contrast, soils containing appreciable amounts of clay minerals and organic matter tend to have much larger specific surface areas (Table 1). Quantification of specific surface area may involve direct physical measurement of particle size and shape, adsorption of probe molecules from either the gas or aqueous phases, or the retention of polar liquids. The adsorption of nitrogen (N_2) gas, in conjunction with the Brenauer–Emmett–Teller (BET) equation, is the most common method of surface area determination. However, it is widely recognized that N_2 does not access the interlayer surfaces of expandable clay minerals upon drying. To overcome this limitation, the retention of polar compounds such as ethylene glycol monoethyl ether (EMGE) has been utilized to measure the total specific surface area (i.e., internal + external) of soils and expandable clay minerals. However, specific surface area values obtained by these two methods (i.e., N_2/BET and EGME) may be similar or divergent, depending upon sample composition and pretreatments. Therefore, it is important to recognize the dependence of specific surface area data on the measurement technique, and to select one or more methods that are appropriate for the system of interest.

Direct Physical Measurement

Direct physical measurement of specific surface area typically involves the use of light or electron microscopy to determine the shape and dimensions of individual soil particles. Such observations are often supplemented with X-ray diffraction measurements to assess crystallographic structure and interlayer spacing of clay minerals. Provided that a characteristic particle shape and size can be determined, the specific surface area can be obtained from mass–volume relationships. For example, the specific surface area (A_s) of a spherical particle may be calculated in the following manner:

$$A_s = \frac{4\pi r^2}{\rho_s V_s} = \frac{4\pi r^2}{\rho_s \frac{4}{3}\pi r^3} = \frac{3}{\rho_s r} \qquad [1]$$

where r is the radius of the solid particle, ρ_s is the density of the solid, and V_s is the volume of the solid. Using this approach, the specific surface area of quartz sand with a particle diameter of 1.0 mm

Table 1 Comparison of specific surface area values obtained by N_2 gas adsorption and ethylene glycol (EG) or ethylene glycol monoethyl ether (EGME) retention

Sample	Organic carbon content ($g\,kg^{-1}$)	Specific surface area ($\times 10^3\ m^2\ kg^{-1}$)	
		N_2/BET method	EG/EGME method
Kaolinite (KGa-1)[a]	0.0	10.05	16.0
Montmorillonite (SWy-1)[a]	0.0	31.82	662.0
Montmorillonite (SAz-1)[a]	0.0	97.42	820.0
Wyoming bentonite[b]	0.0	65.0	372.0
Lula aquifer sand[c]	0.1	7.7	10.5
Boston silt[b]	26.6	28.6	46.0
Webster soil[d]	33.2	8.2	168.4
Ashurst soil[b]	45.5	6.3	25.8
Houghton muck[d]	445.7	0.8	162.9

[a]van Olphen H and Fripiat JJ (1979) *Data Handbook for Clay Minerals and Other Non-metallic Minerals*, pp. 203–211. New York: Pergamon Press.
[b]Call F (1957) The mechanism of sorption of ethylene dibromide on moist soils. *Journal of the Science of Food and Agriculture* 8: 630–639.
[c]Rhue RD, Rao PSC, and Smith RE (1988) Vapor-phase adsorption of alkylbenzenes and water on soils and clays. *Chemosphere* 17: 727–741.
[d]Pennell KD, Boyd SA, and Abriola LM (1995) Surface area of soil organic matter reexamined. *Soil Science Society of America Journal* 59: 1012–1018.
BET, Brunauer–Emmett–Teller.

or 1.0×10^{-3} m and a particle density of $2.65\,g\,cm^{-3}$ or $2.65 \times 10^3\,kg\,m^{-3}$ would be $2.26 \times 10^{-3}\,m^2\,g^{-1}$ or $2.26\,m^2\,kg^{-1}$. In practice, specific surface area values of sands determined by direct observation are often several orders-of-magnitude smaller than measured values due to presence of nonspherical particles, surface roughness, and fine particles (Table 1).

An analogous mass–volume approach can be used to estimate the specific surface area of clay minerals, provided that the structural formula and unit cell dimensions are known. For example, consider a montmorillonite with a nominal structural formula of $K_{0.66}Si_{8.0}(Al_{3.34}Mg_{0.66})O_{20}(OH)_4$ and unit cell dimensions of $a = 0.5$ nm or 5×10^{-10} m, $b = 0.9$ nm or 9×10^{-10} m, and $c = 9.5$ nm or 9.5×10^{-10} m. Assuming that the particle density is approximately $2.8\,g\,cm^{-3}$ or $2.8 \times 10^3\,kg\,m^{-3}$, and that the edge area (i.e., c-dimension) is negligible compared with the area of the basal surfaces (i.e., a- and b-dimensions), the specific surface area of the montmorillonite can be estimated in the following manner:

$$A_s = \frac{2ab}{\rho_s V_s}$$

$$= \frac{2(5.0 \times 10^{-10}\,m)(9.0 \times 10^{-10}\,m)}{(2.8 \times 10^3 kg\,m^{-3})(5.0 \times 10^{-10}\,m)(9.0 \times 10^{-10}\,m)(9.5 \times 10^{-10}\,m)}$$

$$= 7.52 \times 10^5 m^2\,kg^{-1} \quad [2]$$

Using a slight variation of eqn [2], the specific surface area of the montmorillonite can also be estimated from the molecular weight of the unit cell and Avogadro's number (N_A):

$$A_s = \frac{(2ab)(N_A)}{MW_{mont}}$$

$$= \frac{2(5.0 \times 10^{-10}\,m)(9.0 \times 10^{-10}\,m)(6.022 \times 10^{23}\,mol^{-1})}{7.447 \times 10^{-1}\,kg\,mol^{-1}}$$

$$= 7.52 \times 10^5 m^2\,kg^{-1} \quad [3]$$

The relationships shown in eqns [2] and [3] are applicable to clay minerals existing as flat, plate-like structures with a thickness corresponding to that of the unit cell. When water is removed from expandable clay minerals, the interlayers collapse. The resulting clay particle will then consist of several unit cells stacked on top of one another. If the number of unit cells contained in a collapsed montmorillonite particle is 10, the resulting surface area would be approximately $75\,m^2\,g^{-1}$ or $7.5 \times 10^4\,m^2\,kg^{-1}$, which is consistent with specific surface area values reported for dry montmorillonite samples based on N_2/BET analysis (Table 1).

The mass–volume approaches described above are generally limited in applicability to clean sands or pure clay mineral samples. Many soil constituents, including metal oxides and organic matter, exist as irregular or poorly defined amorphous structures which are virtually impossible to characterize. Furthermore, the specific surface area of natural soils cannot be treated as a strictly additive property due to surface coatings and mineral-organic matter associations. Except in a few limited cases (e.g., clean sands), the specific surface area of a whole soil should not be estimated using a summation procedure based on the surface area contributions of individual soil constituents.

Adsorption from Solution

The adsorption of dissolved molecules from solution can be used to estimate the specific surface area of a solid, provided that the resulting adsorption isotherm exhibits a limiting or maximum value. Such adsorption isotherms are classified as Type I, and can be described by the Langmuir adsorption model. The Langmuir model can be derived by assuming that, once equilibrium is reached, the rate of solute adsorption on to open surface sites is equal to the rate of solute desorption from occupied surface sites. Several assumptions are inherent to the Langmuir model, including: (1) adsorption is localized or site-specific; (2) no interactions occur between adsorbed molecules; (3) the energy of adsorption is constant for all adsorption sites; (4) the adsorption capacity of the solid is limited; and (5) the maximum adsorption capacity corresponds to monolayer coverage. Although the model was originally developed to describe the adsorption of gases on solids, for solid–liquid systems the Langmuir equation may be written as:

$$C_s = \frac{C_{s,max}\,\beta\,C_a}{1+\beta\,C_a} \qquad [4]$$

where C_s is the solid-phase concentration of solute at equilibrium, $C_{s,max}$ is the maximum solid-phase concentration, β is the ratio of the adsorption and desorption rates, and C_a is the aqueous-phase concentration of solute at equilibrium. A series of Langmuir isotherms is presented in **Figure 1** to illustrate the effect of increasing the value of β from 0.01 to $0.25\,l\,mg^{-1}$ with $C_{s,max}$ fixed at $1.0\,g\,kg^{-1}$. As the value of β increases, the rate at which the adsorption isotherm approaches the maximum sorption capacity of the solid ($C_{s,max}$) increases. However, the shape or steepness of the isotherm has no bearing on the maximum sorption capacity, which is used to calculate specific surface area.

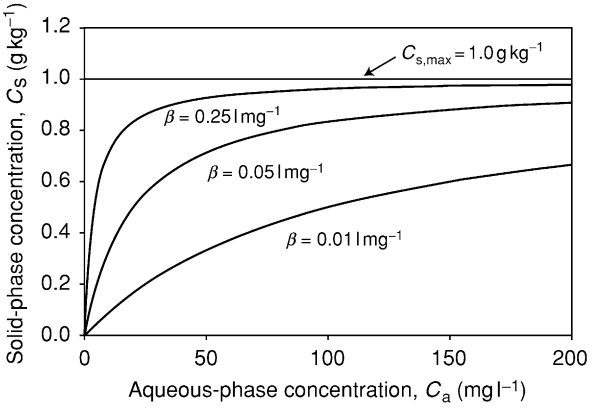


Figure 1 Effect of changes in the value of the β parameter (0.025, 0.05, and $0.01\,l\,mg^{-1}$) on Langmuir adsorption isotherms when $C_{s,max} = 1.0\,g\,kg^{-1}$.

In practice, the parameters β and $C_{s,max}$ can be obtained by directly fitting eqn [4] to experimental adsorption isotherm data (C_s versus C_a) using a nonlinear, least-squares regression procedure. Alternatively, eqn [4] can be rearranged to yield the linear form of the Langmuir equation:

$$\frac{C_w}{C_s} = \frac{1}{\beta\,C_{s,max}} + \frac{C_w}{C_{s,max}} \qquad [5]$$

Here, C_w/C_s (y-axis) is plotted against C_w (x-axis) and a linear regression procedure is then used to obtain a slope equal to $1/C_{s,max}$ and an intercept value equal to $1/\beta C_{s,max}$. With minor manipulation, the desired parameters are obtained as follows: $C_{s,max} = 1/\text{slope}$ and $\beta = (1/C_{s,max})(1/\text{intercept})$.

Several organic molecules have been used in conjunction with the Langmuir equation to determine specific surface area. In the past, organic dyes such as methylene blue were utilized because their concentration in solution could be determined by colorimetric analysis. More recently, cationic surfactants exhibiting visible or ultraviolet (UV) light absorbance have been employed for surface area determination. The cationic surfactant most widely used for this purpose is cetyl pyridinium bromide (CBP), which has a strong absorbance peak at wavelength of 259 nm or 2.59×10^{-7} m. On most mineral surfaces, adsorbed CPB molecules form a bilayer (i.e., double layer), yielding an effective molecular area (A_m) of $0.27\,nm^2$ or $2.7 \times 10^{-19}\,m^2$. If the measured value of $C_{s,max}$ for a nonexpanding clay mineral sample was $1.0\,g$ CPB kg^{-1} solid, the specific surface area would be calculated as:

$$\begin{aligned} A_s &= \frac{(C_{s,max})(N_A)(A_m)}{MW_{CBP}} \\ &= \frac{(1.0\,g\,kg^{-1})(6.22\times10^{23}\,mol^{-1})(2.7\times10^{-19}\,m^2)}{384.45\,g\,mol^{-1}} \\ &= 4.23\times10^{2}\,m^2\,kg^{-1} \end{aligned} \qquad [6]$$

Iron and aluminum oxides possess relatively low surface charge densities, and as a result CPB may not form complete bilayers on these surfaces. Therefore, soil samples are often treated to remove oxides prior to surface area analysis by CPB adsorption. In the case of expandable clay minerals such as montmorillonite, the adsorbed CPB bilayer on the interlayer surfaces is shared and therefore yields an effective molecular area of $0.54\,nm^2$ or $5.4 \times 10^{-19}\,m^2$. In addition, the external surface area must be obtained independently using the N_2/BET method in order to compute the contribution of internal surfaces to the overall adsorption of CPB.

Adsorption from the Gas Phase

The adsorption of gases is frequently used in conjunction with the BET equation to measure the specific surface area of soils and soil constituents. Gas adsorption on dry solid surfaces typically conforms to a Type II isotherm, characterized by the formation of multiple layers of adsorbed molecules. Derivation of the BET equation is based on the Langmuir model, modified to account for multilayer formation. The underlying assumptions of the BET equation are: (1) the heat of adsorption for the first layer is constant; (2) the heat of adsorption for the second and all succeeding layers is constant and equal to heat of condensation; (3) adsorption and desorption can only occur from exposed layers; and (4) the assumptions of the Langmuir model apply to each layer. Although originally derived on a molar basis, the BET equation can be expressed in terms of the solid-phase concentration (C_s) as:

$$C_s = \frac{(C_{s,mon})[(\alpha\ P/P_0)/(1-P/P_0)]\left[1-(n+1)(P/P_0)^n+n(P/P_0)^{n+1}\right]}{1+(\alpha-1)(P/P_0)-\alpha(P/P_0)^{n+1}} \quad [7]$$

where $C_{s,mon}$ is the solid-phase concentration at monolayer coverage, P is the vapor pressure, P_0 is the saturated vapor pressure, and n is the total number of adsorbed layers. The dimensionless parameter α is related to the heat of adsorption and is defined as:

$$\alpha = e^{[(Q_a-Q_c)/RT]} \quad [8]$$

where Q_a is the heat of adsorption on the exposed surface, Q_c is the heat of condensation of the liquid adsorbate, R is the ideal gas constant, and T is temperature. The BET eqn [7] reduces to the Langmuir eqn [4] when number of adsorbed layers is limited to one ($n=1$). If the number of adsorbed layers approaches infinity ($n=\infty$), eqn [7] reduces to the simplified form of BET equation commonly used for surface area determination:

$$C_s = \frac{(C_{s,mon})(\alpha)(P/P_0)}{(1-P/P_0)[1-P/P_0+\alpha(P/P_0)]} \quad [9]$$

A series of BET isotherms is shown in Figure 2 for α equal to 1, 10, and 100, with $C_{s,mon}$ fixed at $1\,g\,kg^{-1}$ and n equal to infinity. As the value of α increases, the inflection point or 'knee' of the BET adsorption isotherm becomes more evident. The inflection in the BET adsorption isotherm corresponds approximately to the point of monolayer coverage. Below the inflection point, gas adsorption occurs on exposed surfaces. The gradual increase in adsorption above the inflection point corresponds to multilayer formation on the surface, while the steep asymptotic rise in adsorption at relative vapor pressures approaching unity ($P/P_0=1$) corresponds to liquid condensation.

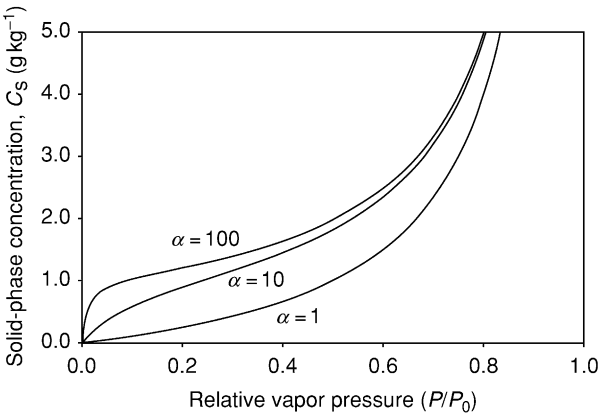


Figure 2 Effect of changes in the value of the α parameter (1, 10, and 100) on Brunauer–Emmett–Teller adsorption isotherms when $C_{s,man} = 1\ g\ kg^{-1}$ and $n =$ infinity.

The most common gas used for BET surface area analysis is nitrogen (N_2), although noble gases such as krypton (Kr) and argon (Ar) are used occasionally for solids possessing very small surface areas (less than approx. $0.1\,m^2\,g^{-1}$ or $1.0\times10^2\,m^2\,kg^{-1}$). Several volatile organic compounds (VOCs), including benzene, *p*-xylene, EGME, and water vapor have also been used for BET surface area analysis. Regardless of the gas or vapor used, the configuration of the adsorbed molecules on the surface must be known or estimated. The most common approach used to estimate the cross-sectional area of an adsorbed molecule (A_m) is related to the liquid density (ρ_l):

$$A_m = 1.091\left(\frac{MW}{\rho_l N_A}\right)^{2/3} \quad [10]$$

The coefficient of 1.091 in eqn [10] is based on the assumption of an ideal hexagonal packing of adsorbed molecules on the surface. For N_2, the value of A_m obtained using eqn [10] is $0.162\,nm^2$ or $1.62\times10^{-19}\,m^2$.

The adsorption of gases on solids can be measured experimentally using several methods. The most common method is based on the change in vapor pressure following the introduction of gas into a small glass bulb containing a dry soil sample. To obtain the adsorption isotherm, the volume of gas adsorbed is computed for each incremental gas dosage based on the change in vapor pressure at equilibrium. Upon reaching the saturated vapor pressure ($P/P_0=1$), the process may be reversed (the vapor pressure is incrementally reduced) to obtain a desorption isotherm, from which pore size analysis

can be performed. Several automated instruments based on this principle are available from commercial vendors. In general, automated surface area instruments report gas adsorption data as the volume of gas adsorbed per gram of soil (e.g., $1\,kg^{-1}$) at standard temperature and pressure (STP). Volumetric gas adsorption data are converted to a mass basis (grams per kilogram) using the molar volume of an ideal gas at STP ($22.414\,l\,mol^{-1}$) and molecular weight of the gas (e.g., $N_2 = 28.02\,g\,mol^{-1}$). The second experimental approach used to measure gas or vapor adsorption is based on the continuous introduction of gas stream at constant vapor pressure. Once equilibrium is attained, the soil sample is either weighed or extracted to determine the amount of adsorbed gas. The vapor pressure of the gas flow stream is then incrementally increased over the desired vapor pressure range to obtain an adsorption isotherm.

To obtain values for the two unknown parameters in the BET equation, α and $C_{s,mon}$, eqn [9] can be fit directly to the experimental adsorption data using a nonlinear, least-squares regression procedure. Alternatively, the experimental data can be expressed using the linear form of the BET equation:

$$\frac{P/P_0}{C_s(1-P/P_0)} = \frac{1}{C_{s,mon}\ \alpha} + \frac{(\alpha-1)P/P_0}{C_{s,mon}\ \alpha} \qquad [11]$$

Here, $(P/P_0)/[C_s(1 - P/P_0)]$ (y-axis) is plotted against P/P_0 (x-axis), and a linear regression procedure is used to obtain values for the slope, which is equal to $(\alpha - 1)/(\alpha\ C_{s,mon})$ and the intercept, which is equal to $1/(\alpha\ C_{s,mon})$. The two fitting parameters can then be obtained as follows: $\alpha = [(\text{slope})(1/\text{intercept}) + 1]$; and $C_{s,mon} = (1/\alpha)(1/\text{intercept})$. As a general rule, the value of α should be greater than 20 and the amount of sample should yield a total surface area between 40 and 120 m^2. In addition, it is often recommended that eqn [11] be applied to adsorption data over a relative vapor pressures range of 0.05–0.35. The specific surface area is then computed using the fitted value of $C_{s,mon}$ obtained from the experimental adsorption data and the cross-sectional area of the adsorbed molecule (A_m) from eqn [10]. For example, if $C_{s,mon}$ was determined to be $0.5\,g\,kg^{-1}$, the N_2/BET specific surface area (A_s) would be calculated in the following manner:

$$A_s = \frac{(C_{s,mon})(N_A)(A_m)}{MW_{N_2}}$$
$$= \frac{(0.5\,g\,kg^{-1})(6.022\times10^{23}\,mol^{-1})(1.62\times10^{-19}\,m^2)}{28.02\,g\,mol^{-1}}$$
$$= 1.74\times10^{3}\,m^2\,kg^{-1} \qquad [12]$$

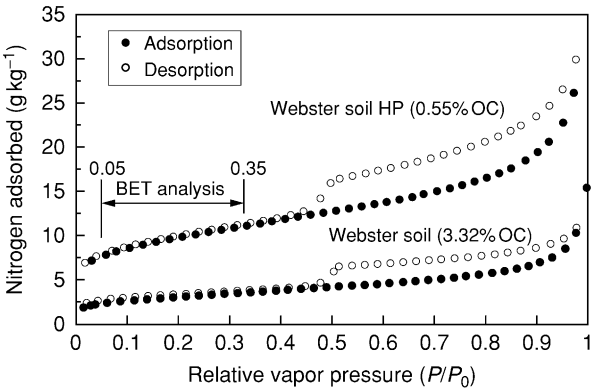


Figure 3 Adsorption and desorption of N_2 on hydrogen peroxide (HP)-treated and untreated Webster soil.

Treatment and preparation of soil samples prior to N_2 adsorption can strongly influence the measured specific surface area. For example, the removal of soil organic matter will often lead to increased specific surface area values; this has been attributed to the exposure of mineral surfaces covered by organic matter and the division of mineral particles held together by organic matter bridging. Nitrogen adsorption and desorption isotherms for hydrogen peroxide (HP)-treated and untreated Webster soil are shown in Figure 3. The resulting surface area values for the untreated and HP-treated samples were 0.79 and $7.38\,m^2\,g^{-1}$ or 7.9×10^2 and $7.38\times10^3\,m^2\,kg^{-1}$, respectively.

As noted previously, water must be removed from soil samples prior to the measurement of gas adsorption. The drying or dehydration process is known to collapse the interlayer space of expandable clay minerals, which are then not accessible to inert gases such as N_2. In addition, electron microscopy suggests that air-drying of soil samples results in the collapse and shrinkage of soil humic acid, whereas freeze-drying maintains a complex structural network characteristic of soil organic matter. Thus, freeze-drying of soils to remove water prior to N_2/BET analysis will tend to result in larger surface area values, which may be more representative of natural conditions.

Retention of Polar Liquids

The use of polar liquids such as ethylene glycol (EG) and EGME for surface area determination was based on the need to develop a relatively simple experimental technique that could be used to quantify the total surface area (i.e., internal + external) of expandable clay minerals. Due to the attractive forces between polar molecules and exchangeable cations, EG and

EGME are able to penetrate the interlayer space of expandable clay minerals. In practice, a dry soil or clay sample is placed in a vacuum desiccator containing EG or EGME as a free liquid or mixed with calcium chloride to maintain a constant vapor pressure of EG or EGME. Liquid EG or EGME is then added as drops to the solid sample until complete wetting is achieved. A suction of approximately 13.332 Pa or 0.1 mmHg is applied, and the sample is weighed over time until a constant weight is obtained. The specific surface area of the solid is then calculated as follows:

$$A_s = \frac{W_{EG/EGME}}{(W_{OD})(\text{EG/EGME conversion factor})} \quad [13]$$

where $W_{EG/EGME}$ is the weight of EG or EGME retained by the soil at the applied suction and W_{OD} is the oven-dry weight of the solid. The method is predicated on the assumption that EG and EGME are retained on solid surfaces at monolayer coverage. The mass-surface area conversion factors for EG and EGME are $3.1 \times 10^{-7}\,\text{kg}\,\text{m}^{-2}$ and $2.86 \times 10^{-7}\,\text{kg}\,\text{m}^{-2}$, respectively, and were derived from retention measurements performed on reference clay minerals with known unit cell dimensions and structural formula. It is usually recommended that the specific surface area of a reference expandable clay mineral, such as Wyoming montmorillonite (SWy-1), be measured to confirm the value of the conversion factor and to ensure that experimental procedure is operating properly. Although EG was used during the initial development of the retention method, more recent studies have employed EGME due to the shorter times required to attain a constant retention value.

Based on the success of the EG/EGME retention method for determining the total surface area of expandable clay minerals, the technique has subsequently been applied to natural soils to obtain a measure of total specific surface area. However, use of the EG/EGME retention method is complicated by the fact that polar molecules tend to form multilayers in the vicinity of cation exchange sites prior to complete monolayer coverage, retention may be influenced by the species of exchangeable cation, and the EG/EGME may partition into soil organic matter. As a result, the retention of EG or EGME by natural soils represents a measure of the uptake capacity of soils for a polar adsorbate, rather than a strict measure of total specific surface area. Despite this shortcoming, the EG/EGME retention method is an appropriate method for measuring the total specific surface area of pure clay minerals and soils that have been treated to remove organic carbon.

Selection of Surface Area Measurement Technique

Although the methods used to determine experimentally the specific surface area of soils and soil constituents are relatively well established, interpretation and appropriate use of the resulting data can be problematic. Complications arise primarily from two factors: (1) sample pretreatments including drying, organic-matter removal, and metal-oxide removal can markedly alter measured values of specific surface area; and (2) the use of different probe molecules and measurement techniques can result in similar or very divergent specific surface area values for the same soil. For example, soils that contain only trace amounts of organic matter and no expandable clay minerals will yield relatively consistent specific surface area values regardless of the method (see kaolinite and Lula aquifer sand in Table 1). In addition, the external and internal specific surface area of pure expandable clay mineral samples can be determined using a combination of N_2/BET (external surface area) and EG/EGME retention (total surface area; see SAz-1 and SWy-1 montmorillonites in Table 1). However, for natural soils containing appreciable amounts of organic matter and expandable clay minerals, the selection of an appropriate measurement technique, as well as the interpretation of specific surface area data, is far more difficult. This is particularly relevant if the intent is to quantify the total or internal surface area of a soil available under natural conditions using the EG/EGME retention method. Due to the potential for partitioning or dissolution of polar molecules into soil organic matter and interactions with exchangeable cations, the EG/EGME retention method is not suitable for most natural soils. In effect, EG/EGME retention provides an indication of the capacity of the soil to take up polar molecules, in contrast to a measure of total specific surface area. Despite the known limitations of N_2/BET surface area analysis, this method provides a relatively simple and reproducible technique for assessing the specific surface area of natural soils and soil constituents.

List of Technical Nomenclature

α	BET isotherm parameter
β	Langmuir isotherm parameter
ρ_l	Liquid density ($\text{g}\,\text{cm}^{-3}$)
ρ_s	Solid density ($\text{g}\,\text{cm}^{-3}$)
A_m	Area of adsorbed molecule (m^2)
A_s	Specific surface area ($\text{m}^2\,\text{kg}^{-1}$)

C_a	Aqueous-phase concentration ($mg\,l^{-1}$)
C_s	Solid-phase concentration ($kg\,kg^{-1}$)
$C_{s,max}$	Maximum solid-phase concentration ($kg\,kg^{-1}$)
$C_{s,mon}$	Monolayer solid-phase concentration ($kg\,kg^{-1}$)
N_A	Avogadro's number ($molecules\,mol^{-1}$)
n	Number of adsorbed layers
P	Vapor pressure (Pa)
P_0	Saturated vapor pressure (Pa)
P/P_0	Relative vapor pressure
Q_A	Heat of adsorption ($J\,mol^{-1}$)
Q_C	Heat of condensation ($J\,mol^{-1}$)
R	Ideal gas constant ($J\,K^{-1}\,mol^{-1}$)
r	Radius of particle (m)
V_s	Volume of solid (m^3)
W_{OD}	Weight of oven-dry solid (kg)

See also: **Cation Exchange**; **Clay Minerals**; **Organic Matter:** Principles and Processes; **Sorption:** Metals; Organic Chemicals

Further Reading

Adamson AW (1990) *Physical Chemistry of Surfaces*, 5th edn, pp. 591–681. New York: John Wiley.

Bower CA and Gschwend FB (1952) Ethylene glycol retention by soils as a measure of surface area and interlayer swelling. *Soil Science Society of America Proceedings* 16: 342–345.

Brunauer S, Emmett PH, and Teller E (1938) Adsorption of gases in multimolecular layers. *Journal of the American Chemical Society* 60: 309–319.

Carter DL, Mortland MM, and Kemper KD (1986) Specific surface. In: Klute A (ed.) *Methods of Soil Analysis, Part 1, Physical and Mineralogical Methods*, 2nd edn, pp. 413–423. Monograph No. 9. Madison, WI: American Society of Agronomy.

Greenland DJ and Mott CJB (1978) Surfaces of soil particles. In: Greenland DJ and Hayes MHB (eds) *The Chemistry of Soil Constituents*, pp. 321–353. New York: John Wiley.

Gregg SJ and Sing KSW (1982) *Adsorption, Surface Area and Porosity*, 2nd edn. New York: Academic Press.

Lowell S and Shields JE (1991) *Powder Surface Area and Porosity*, 3rd edn. London, UK: Champan & Hall.

McClellan AL and Harnsberger HF (1967) Cross-sectional areas of molecules adsorbed on solid surfaces. *Journal of Colloid and Interface Science* 23: 577–599.

Pennell KD (2003) Specific surface area. Topp GC and Dane JH (eds) *Methods of Soil Analysis*, part 4, *Physical Methods*, sect. 2.5, pp. 295–315. Madison, WI: American Society of Agronomy.

Pennell KD, Boyd SA, and Abriola LM (1995) Surface area of soil organic matter reexamined. *Soil Science Society of America Journal* 59: 1012–1018.

Sposito G (1984) *The Chemistry of Soils*, pp. 23–35. Oxford, UK: Oxford University Press.

van Olphen H and Fripiat JJ (1979) *Data Handbook for Clay Minerals and Other Non-Metallic Minerals*, pp. 203–211. New York: Pergamon Press.

STATISTICS IN SOIL SCIENCE

R Webster, Rothamsted Research, Harpenden, UK

© 2005, Elsevier Ltd. All Rights Reserved.

Why Statistics?

Data on soil embody variation from many sources. Much of the variation is natural – from place to place at all scales from the global to the infinitessimal, and from time to time as the soil responds to weather, plant growth, and processes in the rhizosphere engendered by that growth. Farmers have added to the variation by their enclosing, reclaiming, clearing, and fertilizing of the land, though within fields they have removed some variation by cultivation and drainage. Further sources of variation are mineral extraction and subsequent reclamation, dumping of waste, and pollution of many kinds. Investigators design experiments and surveys in such a way that they can estimate the variation from particular sources such as imposed treatments or strata in a population and the differences between them. Variation in data also arises from the way observations are made; from the people who make the observations, from the imprecision instruments, from imperfect

laboratory technique, and from sampling fluctuation. One may like to think that the contributions from the laboratory are negligible, though 'ring' tests have often revealed them not to be so. In general, however, sampling fluctuation in the field, arising from the spatial variation there, is much larger.

Any one measurement of a soil property is influenced by contributions from at least some of these sources. It cannot be taken as an absolutely accurate representation of the truth therefore; rather it must be seen in relation to the probable error.

Statistics are needed in soil science to estimate and express this error and to apportion it to the different sources. In this way sense (signal) can be separated from meaningless or uninteresting variation (noise) in comparisons between classes of soil, in expressing relations, in assessing the effects of treatments in experiments, and in prediction. The statistical repertoire is huge, so here is presented the fairly elementary techniques that soil scientists most often need.

The techniques can be divided into two groups, namely, descriptive and analytical. They also have two fairly distinct fields of application: survey and experiment. In the first, investigators observe and record the soil as it is on samples, and descriptive techniques tend to dominate. In the second, they deliberately control some of the variation so that they can assess the effects of changes in one or a few factors that are of specific interest by analysis.

Population, Units, and Samples

The soil is regarded for statistical purposes as a population comprising elements or units. The units are the bodies of soil on which measurements are made. They are more or less arbitrary and determined largely by convenience and practicality. They may be individual cores of soil, pits, or pedons; each may be the volume of soil occupied by the roots of a single plant or that deformed under the wheel of a tractor; they may be pots in a greenhouse experiment, samples in lysimeters, plots in a field experiment, or whole fields or farms. The variation among them depends very much on the size of the units; the larger they are the more variation they encompass within them and the less there is between them to be revealed in data. The size, shape, and orientation of the units, known as the 'support' in survey, must be defined and adhered to throughout an investigation. The population is then all such units in a specified region or falling within some other definition for the purposes of the investigation. It is an operational definition, often known as the 'target population.' In a more restricted sense, the population and the units comprising it may be the values of a particular soil property in the defined supports.

Populations in surveys are typically very large, in many instances infinite, or hypothetical, and in some they are poorly defined. Measurement is feasible only on small subsets, i.e., on samples, and these subsets must properly represent their populations for the measurements to apply to the larger populations. The units in an experiment, in contrast, are typically a few dozen, though ideally the experimental material should be representative of some larger population.

Replication and Randomization

Estimating the mean in a population and its associated error from measurements on a sample requires both replication and randomization. The need for replication is evident: a single value can contain no information on variation. Randomization is needed to avoid bias. At its simplest it means choosing units such that all have the same chance of being selected.

In a survey this takes the form of simple random sampling. To apply it requires (1) that each unit can be identified uniquely, and (2) a rule for the selection. Simple random sampling, as in Figure 1a, tends to be inefficient in that it takes no account of anything known about the population beforehand, such as its spatial dependence, a soil map of the region, the relations between soil and vegetation or physiography, or the farming history. To take advantage of such knowledge, the population is first stratified either into small grid cells, as in Figure 1b, or by type of soil, vegetation, physiography, or farming, as in Figure 1c.

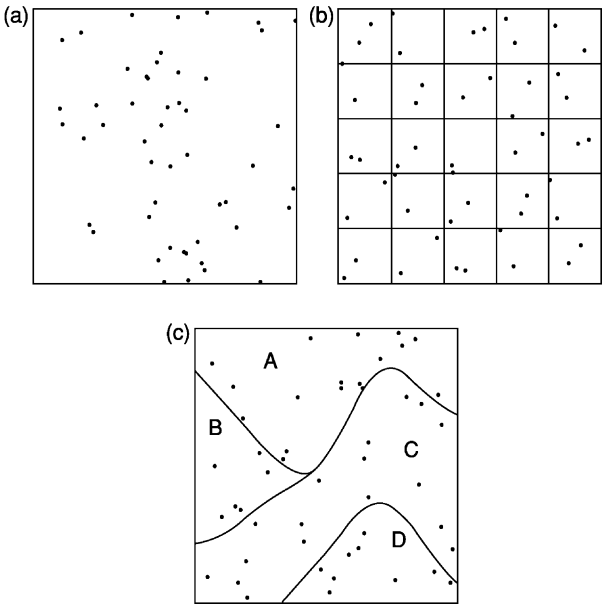


Figure 1 Probabilistic sampling of a region: (a) simple random sampling; (b) stratified random sampling with the region divided into 25 square cells (strata) and two points per cell; (c) stratified random sampling with four mapped classes (A–D) of soil as strata.

(a)

W	O	F	W	D
F	D	W	F	O
W	O	W	O	D
D	F	F	D	O

(b)

W	D	F	W	F
F	O	D	F	O
O	W	W	O	D
D	F	O	D	W

Figure 2 Layout of a randomized field experiment with four treatments; O (control), D (dung), W (industrial waste), and F (NPK fertilizer), and five-fold replication. (a) Completely randomized; (b) with the replicates arranged in five blocks.

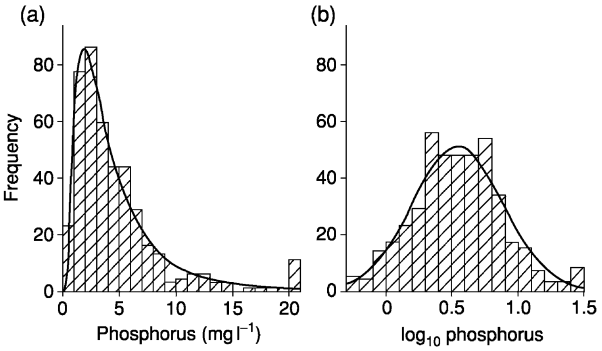


Figure 3 Histograms of 434 data on available phosphorus (a) in milligrams per liter and (b) transformed to common logarithms. The curve in b is that of the fitted normal distribution, and that in a shows the lognormal distribution.

The soil is then sampled at random within each stratum independently; this is stratified random sampling, and it enables the variation due to the strata to be distinguished from variation within them.

In experiments treatments are allocated to the units at random. The units may be arranged completely randomly, as in Figure 2a. In the field and greenhouse, they are usually grouped into blocks such that each block contains one unit of each treatment, as in Figure 2b; long-range variation then appears in the variation among blocks. There are many more elaborate designs.

Descriptive Statistics

The Mean

In almost all investigations, mean values are of prime interest. Provided sampling has been properly randomized, the mean of a set of N data, denoted z_1, $z_2, \ldots, z_N$,

$$\bar{z} = \frac{1}{N}\sum_{i=1}^{N} z_i \qquad [1]$$

estimates without bias the mean, μ, in the population from which the sample is drawn. How well it does so depends on the variation it embodies.

Characteristics of Variation

Histogram The variation in a set of measurements, if there are sufficient of them, can be displayed in a histogram. The scale of measurement is divided into segments of equal width, or 'bins,' the values falling in each bin are counted, and bars of height proportional to the counts are drawn. Figure 3 is an example; it summarizes graphically the way in which the frequency is distributed over the range of the data.

Variance Variation is best expressed quantitatively as variance. For a set of N data, it is the average squared difference between the observations and their mean:

$$S^2 = \frac{1}{N}\sum_{i=1}^{N}(z_i - \bar{z})^2 \qquad [2]$$

Its square root, S, is the standard deviation, which is often preferred because it is in the same units as the measurements and is therefore more intelligible.

Ideally this variance should estimate the variance of the larger population, of which the N observations are a sample. This variance of a population is by definition:

$$\sigma^2 = \mathrm{E}\left[(z-\mu)^2\right] \qquad [3]$$

where μ is mean of z in the population and E denotes expectation. Equation [2] gives a biased result, however: S^2 is a biased estimator, because $\bar{z}$ in the equation is itself more or less in error as the estimate of μ. To remove the bias, N must be replaced by $N-1$ in the denominator:

$$s^2 = \frac{1}{N-1}\sum_{i=1}^{N}(z_i - \bar{z})^2 \qquad [4]$$

The result, s^2, is now unbiased, provided the sampling was unbiased in the first place.

Estimation variance, standard error, and confidence Both S and s measure the dispersion in the observations; neither expresses the reliability of the estimate of μ. The reliance one can place in an estimate of the mean can be expressed in terms of the expected squared deviation of it from the true mean, i.e., $\mathrm{E}[(\bar{z}-\mu)^2]$. Its estimate is derived simply from s^2 by:

$$s^2(\bar{z}) = s^2/N \qquad [5]$$

This is the estimation variance, and its square root is the standard error, which is the standard deviation of means of samples of size N. The larger the sample is, the smaller this error is, other things being equal, and the more confidence there can be had in the estimate. So, the standard deviation, which describes the variation in a sample, is different from the standard error, which expresses the confidence associated with a mean. Standard errors accompany means in tables compiled from replicated measurements, and they can be shown by error bars on graphs.

Equation [5] gives the estimation variance for a simple random sample of size N. If the population has been divided into strata, then s^2, the population variance on the right-hand side of the equation, can be replaced by s^2_W, the variance within strata. The latter is generally less, often much less, than s^2, and so stratified sampling is more precise than simple random sampling by the factor $s^2_{\text{random}}(\bar{z})/s^2_{\text{stratified}}(\bar{z})$. It also means that one can achieve the same precision with a smaller sample, and in this sense stratified sampling is more efficient. This efficiency can be expressed as $N_{\text{random}}/N_{\text{stratified}}$. The within-stratum variance, s^2_W, can be estimated by the analysis of variance.

Standard errors can be converted to confidence intervals for known kinds of distribution. If the data are drawn from a normal (Gaussian) distribution and there are many of them, then an interval of $2s/\sqrt{N}$ centered on $\bar{z}$ spans a symmetric confidence interval of approximately 68%. Wider confidence intervals are readily calculated by multiplying by factors, η, such as in Table 1.

When data are few ($N < 30$) the above factors need to be replaced by Student's t for the number of degrees of freedom, f. The lower and upper symmetric confidence limits about a mean $\bar{z}$ are:

$$\bar{z} - t_f s/\sqrt{N} \quad \text{and} \quad \bar{z} + t_f s/\sqrt{N} \qquad [6]$$

where t_f is Student's t for f degrees of freedom. Values of t are available in tables and in most statistical computer packages. For $30 < N \leq 60$, t converges to η, and for $N > 60$ the values in Table 1 can be used.

There are formulae for calculating the confidence limits for other theoretical distributions. In many instances, however, it is easier to transform data to approximate a normal distribution and subsequently analyze the transformed values.

Table 1

Confidence (%)	80	90	95	99
η	1.28	1.64	1.96	2.58

Coefficient of Variation

The coefficient of variation (CV) is the standard deviation divided by the mean, i.e., $s/\bar{z}$. It is often multiplied by 100 and quoted as a percentage. The merit of the CV is that it expresses variation as pure numbers independent of the scales of measurement. It enables investigators and those reading their reports to appreciate quickly the degree of variation present and to compare one region with another and one experiment with another. The CV should not be used to compare variation in different variables, especially ones having different dimensions.

The CV is sensible only for variables measured on scales with an absolute zero. Otherwise the arbitrary choice of the zero affects it. Examples for which it should not be used are soil temperature in degrees Celsius (arbitrary zero at 273 K), color hue (which is approximately circular), and soil acidity on the pH scale (arbitrary zero equivalent to $-\log_{10}[H^+]$, with H^+ expressed in moles per liter).

For some soil properties, physics sets limits on the utility of the CV. For example, the minimum bulk density of the soil is determined by the physical structures that keep particles apart. Particles must touch one another, otherwise the soil collapses. For most mineral soils on dry land, a working minimum bulk density is approximately 1 g cm^{-3}. At the other end of the scale, the bulk density cannot exceed the average density of the mineral particles, approximately 2.7 g cm^{-3}. So the CV of bulk density is fairly tightly constrained.

The sensible use of the coefficient of variation for comparing two variables y and z relies on the assumption that they are the same apart from some multiplying factor, b, thus:

$$y = bz \qquad [7]$$

Then the mean of y is $\bar{y} = b\bar{z}$, its variance $s^2_y = b^2 s^2_z$, and its standard deviation is $s_y = bs_z$. From there, simple arithmetic shows that their CVs are the same. This principle offers a means of comparing variation by taking logarithms of the observations. Equation [7] becomes:

$$\log y = \log b + \log z \qquad [8]$$

The logarithm $\log b$ is a constant, and so the variances, $s^2_{\log y}$ and $s^2_{\log z}$, are equal, as are their standard deviations. The resulting measure of variation is therefore independent of the original scale of measurement.

The measure can be used to compare variation in two groups of observations. Consider again soil acidity. To compare the variation in acidity of a class A with that in another, class B, we treat the hydrogen ion concentration as the original variable, transform

it to pH, and compute the variances of pH. Whichever has the larger variance is the more variable, regardless of the mean. Further, we can make a formal significance test by computing $F = s^2_{\text{logy}}/s^2_{\text{logz}}$ and compare the result with F for N_y-1 and N_z-1 degrees of freedom.

Additivity of variances Variances are additive; those from two or more independent sources in an investigation sum to the total in the data. Their square roots, the corresponding standard deviations, are not. To obtain an average variation on the original scale of measurement from several sets of data, the arithmetic mean of their variances is computed, weighted as appropriate by the numbers of degrees of freedom, and then the square root of it is taken to give an 'average' or pooled standard deviation. This, divided by the mean, gives an average CV. More generally, the additive nature of variances confers great flexibility in analysis, enabling investigators to distinguish variation from two or more sources and estimate their components according to the design, as by the analysis of variance.

Statistical Significance

Significance in a statistical context means distinguishing a signal or the effects of some imposed treatment, or detecting differences between strata against a background of 'noise.' It is a matter of separating the variance due to the signal, treatments, or strata from that from other sources that are of no interest. The question being addressed is as follows: Given the magnitudes of the several components of variance, is the signal so strong or are the differences observed so large that they are highly unlikely to have occurred by chance? If the answer is 'yes,' then the result is said to be significant.

A significance test is prefaced by a hypothesis. This is usually that there is no real difference between populations or treatments and that any differences among the means of observations are due to sampling fluctuation. That is the 'null hypothesis,' often designated H_0, and the test is designed to reject it (not confirm it). The alternative, that there is a difference, is denoted either H_1 or H_A.

To judge, for example, whether two means differ, one computes from the sampling error, the probability, P, of obtaining the observed difference if the true means were identical, assuming that the form of the distribution is known. If P is small (conventionally <0.05), the null hypothesis is rejected, and the difference is judged significant. If P is large, then the null hypothesis is likely to be correct, but we have no measure of the probability that the two means are indeed identical; instead we take the view that we have too-little evidence to conclude that the difference observed applies to the population from which the sample has been drawn.

Mistaken conclusions can still be drawn as the result of significance tests. Mistakes can be of two kinds, denoted type I and type II. The first occurs when the null hypothesis is rejected on the basis of the sample evidence, i.e., a difference is declared significant when the populations are not different. The second is when the null hypothesis is accepted, i.e., there is insufficient evidence for a difference when the populations do differ.

The likelihood of drawing wrong conclusions can be diminished by increasing the sensitivity of the test, and that depends on the precision with which the means have been estimated, i.e., on their estimation variances or on the estimation variance of their difference. The latter is given by:

$$s^2_{\text{diff}} = \frac{s^2_{\text{W}}}{n_1} + \frac{s^2_{\text{W}}}{n_2} \qquad [9]$$

where n_1 and n_2 are the numbers of observations from which the means in classes 1 and 2 derive, and s^2_{W} is the variance within the classes, assumed to be common. If $n_1 = n_2$ then the variance of the difference is simply twice the estimation variance of the individual means.

Equation [9] shows two features. One is that the larger the variance is within the populations, the larger the variance between the means is and the less likely can a difference be established as significant. The second is that larger samples result in smaller variances and hence more sensitive comparisons. If the samples are large enough any soil can be established as different from almost any other for whatever property of interest.

The significance test is valuable in preventing false claims on inadequate evidence. Thus, a result might be summarized as:

> The mean measured pH of the topsoil was 5.7 compared with 6.7 in the subsoil; but because the samples were small (or because the variances were large) the difference was not statistically significant.

However, when a difference is deemed statistically significant because the null hypothesis is rejected, that does not mean that it is important or physically or biologically meaningful. For example, the difference between an observed mean pH of 5.7 in the topsoil and 5.9 in the subsoil would be of little consequence, whatever the probability of rejecting the null hypothesis. Also, while an investigator might regard a difference as significant only if $P \leq 0.05$, a reader may be willing to recognize one for which $0.05 \leq P < 0.1$

or, more stringently, only if $P \leq 0.01$. It is to some extent a matter of personal choice, and if the standard errors are reported then readers can reach their own judgments.

Note finally that the null hypothesis is highly implausible when horizons and different types of soil are being compared; they are different.

Transformations

It is often desirable to transform data to their square roots, or logarithms, or by other more elaborate functions. One reason for doing so is to obtain a new variate that approximates some known distribution, preferably normal (Gaussian) so that the usual parametric tests of significance can be applied.

The most serious departure from normality usually encountered with soil data is skewness, i.e., asymmetry, as in Figure 3a. The normal distribution is symmetric, its mean is at its center (its mode), and the mean of the data estimates this central value without ambiguity. The mean of data from a skewed distribution does not estimate the mode, nor does the median (the central value in the data). The meaning of the statistics can be uncertain therefore. A second feature of skewed data is that the variances of subsets depend on their means. If the data are positively skewed (again the usual situation) then the variances increase with increasing mean. This is undesirable when making comparisons. Third, estimation is 'inefficient' where data are skewed; that is, the errors are greater than they need be or, put another way, more data are required to achieve a given precision than would be if the distribution were normal. Transforming data to approximate normality overcomes these disadvantages. We achieve symmetry and hence remove ambiguity concerning the center. We stabilize the variances, and we make estimation efficient. The second of these is perhaps the most important.

No real data are exactly normal; all deviate more or less from normality. We have therefore to decide whether to transform them. This is best done by judicious exploration of the data aided by graphic display.

If a histogram looks symmetrical it can have superimposed on it a normal curve computed from the mean and variance of the data. The normal curve has the formula:

$$y = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{(z-\mu)^2}{2\sigma^2}\right\} \qquad [10]$$

where y is the probability density. Scale it to fit the histogram by multiplying by the number of observations and by the width of the bins. If the curve fits well then there is no need to transform the data.

If the histogram is skewed then the skewness coefficient, γ_1, can be computed in addition to the mean and variance. This dimensionless quantity can be obtained via the third moment of the data about their mean:

$$\gamma_1 = \frac{1}{NS^3} \sum_{i=1}^{N} (z_i - \bar{z})^3 \qquad [11]$$

A symmetric histogram has $\gamma_1 = 0$. Values of γ_1 greater than 0 signify positive skewness, i.e., long upper tails to the distribution and a mean that exceeds the median, which is common. Negative values of γ_1 signify negative skewness and are unusual.

If γ_1 is positive and less than 0.5 then there should be no need to transform the data. If $0.5 < \gamma_1 \leq 1$ then it might be desirable to convert the data to their square roots; and if $\gamma_1 > 1$ a transformation to logarithms is likely to give approximate normality.

The following example illustrates the situation. The data, which are summarized in Table 2, are 433 measurements of available phosphorus, P, in the topsoil. Their skewness coefficient is 3.95, i.e., they are strongly positively skewed, and this is apparent in their histogram (Figure 3a). Transforming to logarithms makes the histogram (Figure 3b) more-nearly symmetric, and, as the skewness in the logarithms is now only 0.34, the transformation seems satisfactory. Further, the normal curve appears to fit well.

Figure 4 shows how the transformation stabilizes the variances. In Figure 4a the variances of subsets of 44 from the full set of data on phosphorus are plotted against the means. Evidently, the variance increases strongly with increasing mean. Converting the data to their logarithms produces a result in which there is virtually no relation (Figure 4b).

These simple functions change only the general form of the distribution; they do not change the detail.

Table 2 Summary statistics of 433 values of available phosphorus measured in a survey of topsoil. Values of χ^2, with 18 df are added for the hypothesis that the data or their logarithms are from a normal distribution

	Scale	
Statistic	*P (mg l^{-1})*	*Log$_{10}$P*
Mean	4.86	0.546
Variance	26.52	0.1142
Standard deviation	5.15	0.338
Skewness	3.95	0.23
χ^2_{df18}	368.2	26.7

Normalizing the detail requires a more elaborate, normal score transform.

Analysis of Variance

The analysis of variance is at once one of the most powerful and elegant techniques in statistics. Its basis is that variances are additive and that the total variance in a population is the sum of the variances contributed by two or more sources. Working from the design of an investigation, it analyzes the data by separating the contributions from those sources and estimating the variances in them. Designs vary from the simple to highly complex, but all embody the same principle.

Here, two of the simplest designs are considered, such as appear in Figure 2. An investigator wants to know how manuring improves crop yield in the field. He or she has several (k) treatments, e.g., nothing (O), dung (D), industrial waste (W), and a complete artificial (NPK) fertilizer (F). The researcher replicates each m times by assigning them completely at random to plots of equal size. Figure 2a shows how the experiment might be layed out with $m = 5$ replicates. The investigator applies the treatments, grows the crop, and measures the yield at harvest, designated z.

The total variance in the yields in the experiment, s_T^2, comprises variance from two sources, namely that between the treatments, s_B^2, and that within them, s_W^2, and $s_T^2 = s_B^2 + s_W^2$. The total variance is estimated by Eqn [4]. The variance within any one treatment is estimated by the same formula but for only those data in that treatment. Pooling estimates for all treatments gives s_W^2. To complete the analysis, a quantity B can also be computed:

$$B = \frac{1}{k-1}\sum_{i=1}^{k} n_i(\bar{z}_i - \bar{\bar{z}})^2 \qquad [12]$$

where n_i is the number of plots in the ith treatment, $\bar{z}_i$ is the mean of the ith treatment, and $\bar{\bar{z}}$ is the general mean of the data. The computations are set out in Table 3. Finally, the analysis leads to a test of significance. The ratio $F = B/W$ is computed, the distribution of which has been worked out and tabulated for degrees of freedom $k-1$ in the numerator and $N-k$ in the denominator. If F exceeds the tabulated value at probability $P = 0.05$ (or $P = 0.01$ or $P - 0.001$, according to choice), the treatments are judged to have produced significant differences.

In the simple experiment illustrated in Figure 2a, all n_i are equal to 5, so that n_i can be replaced by $n = 5$, and $N = mk = 5 \times 4 = 20$. Things do not always go as planned, however, and if some of the plot yields are lost then the n_i can vary from treatment to treatment; and the more general formulae in Table 3 will take care of that.

The soil might vary across the experiment systematically, so that there is trend, or in an apparently random way at a coarse scale. This variation could swell the residual variation and so mask that due to the treatments. It can be taken into account by blocking. The m replicates are now arranged in m blocks such that in each block every treatment appears once and once only. Figure 2b shows an example in which five blocks are laid out side by side. The analysis follows the same procedure as in Table 3 except that there is an additional line for the blocks in which the sum of squares is that of deviations from the block means (Table 4). The residual sum of squares is diminished by this quantity, and, although the residual degrees of freedom are also diminished, the residual mean square, i.e., the residual variance, is usually less, and the experiment more sensitive therefore.

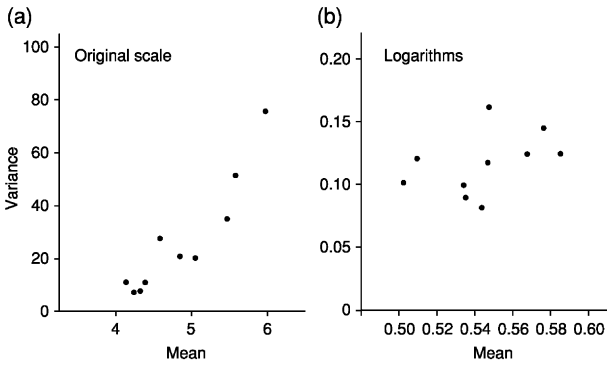


Figure 4 Graphs of variance against mean for 10 subsets of 44 phosphorus data (a) on the original scale in milligrams per liter and (b) after transformation to common logarithms.

Table 3 Table for the analysis of variance for a completely randomized design with a single classification

Source	*Degrees of freedom*	*Sums of squares*	*Mean square*	*F*
Treatments	$k-1$	$\sum_{i=1}^{k} n_i(\bar{z}_i - \bar{\bar{z}})^2$	$\frac{1}{k-1}\sum_{i=1}^{k} n_i(\bar{z}_i - \bar{\bar{z}})^2 = B$	B/W
Residual	$N-k$	$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(z_{ij} - \bar{z}_i)^2$	$\frac{1}{N-k}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(z_{ij} - \bar{z}_i)^2 = W$	
Total	$N-1$	$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(z_{ij} - \bar{\bar{z}})^2$	$\frac{1}{N-1}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(z_{ij} - \bar{\bar{z}})^2 = T$	

Table 4 Table for the analysis of variance for a balanced randomized block design with a single set of treatments

Source	*Degrees of freedom*	*Sums of squares*	*Mean square*	*F*
Treatments	$k-1$	$\sum_{i=1}^{k} n_i(\bar{z}_i-\bar{\bar{z}})^2$	$\frac{1}{k-1}\sum_{i=1}^{k} n_i(\bar{z}_i-\bar{\bar{z}})^2 = B$	B/W
Blocks	$m-1$	$\sum_{l=1}^{m} n_l(\bar{z}_l-\bar{\bar{z}})^2$	$\frac{1}{m-1}\sum_{l=1}^{m} n_l(\bar{z}_l-\bar{\bar{z}})^2 = M$	
Residual	$(k-1)\times(m-1)$	$T(N-1)-B(k-1)-M(m-1)$	$\frac{T(N-1)-B(k-1)-M(m-1)}{(k-1)(m-1)} = W$	
Total	$N-1$	$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(z_{ij}-\bar{\bar{z}})^2$	$\frac{1}{N-1}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(z_{ij}-\bar{\bar{z}})^2 = T$	

In soil survey, different classes of soil replace treatments. In the simplest cases, each class is sampled at random, and the n_i are rarely equal, either because it is difficult to obtain equal representation or because sampling is deliberately in proportion to area so as to maintain a fairly constant density. Effectively the classes are weighted in proportion to the areas they cover. $F = B/W$ can still be computed, and tested for significance, but, as above, this is less interesting than the differences between the means.

The variance between treatments or classes, s_B^2 can be obtained from B. The latter combines variation both from between treatments or classes and within them:

$$B = ns_B^2 + s_W^2 \qquad [13]$$

if $n_i = n$ for all i. Rearranging then gives:

$$s_B^2 = (B - s_W^2)/n \qquad [14]$$

If the n_i are unequal then n is replaced by n^*, given by:

$$n^* = \frac{1}{k-1}\left(N - \frac{\sum_{i=1}^{k} n_i^2}{N}\right) \qquad [15]$$

and:

$$s_B^2 = (B - s_W^2)/n^* \qquad [16]$$

Fixed Effects, Random Effects, and Intraclass Correlation

The value s_B^2 obtained as above estimates $\sum_{i=1}^{k}(\mu_i - \mu)^2/(k-1)$, where μ_i is the expected value of treatment or class i, and μ is the expected value in the whole population. In designed experiments, in which the effects are fixed by the experimenter, s_B^2 is of little interest. In soil survey, however, where it is often a matter of chance which classes are actually sampled, the differences $\mu_i - \mu$ are subject to random fluctuation. In this event, s_B^2 estimates the variance, σ_B^2, among a larger population of means and is termed a 'component of variance,' which is of considerable interest.

The between-class variance expressed as a proportion of the total variance, $\sigma_B^2 + \sigma_W^2$, is the intraclass correlation, ρ_i:

$$\rho_i = \frac{\sigma_B^2}{\sigma_B^2 + \sigma_W^2} \qquad [17]$$

which is estimated from the analysis of variance table by:

$$r_i = \frac{s_B^2}{s_B^2 + s_W^2} = \frac{B - W}{B + (n^* - 1)W} \qquad [18]$$

The intraclass correlation has a theoretical maximum of 1 when every class is uniform. In practice there is always some variation within classes, and so ρ_i never attains 1. The minimum of ρ_i is zero, when $s_W^2 = 0$. The calculated estimate of ρ_i can be negative, because $B < W$, and is usually best explained by sampling fluctuation.

Covariance, Correlation, and Regression

The relations between two variables can be expressed by correlation and regression.

Covariance

The covariance of a pair of variables, y and z, is estimated from data in a way analogous to the estimation of the variance, Eqn [4], by:

$$\hat{c}_{y,z} = \frac{1}{N-1}\sum_{i=1}^{N}\{y_i - \bar{y}\}\{z_i - \bar{z}\} \qquad [19]$$

where $\bar{y}$ and $\bar{z}$ are the means of y and z, respectively. Covariance is not easy to envisage, especially if y and z have different dimensions. The relation may be standardized by converting it to correlation, below.

Correlation

The correlation between two variables, strictly the linear correlation, or the product–moment coefficient of linear correlation, is a dimensionless quantity, usually denoted by ρ for the population parameter. Its estimate, r, is obtained from the covariance by:

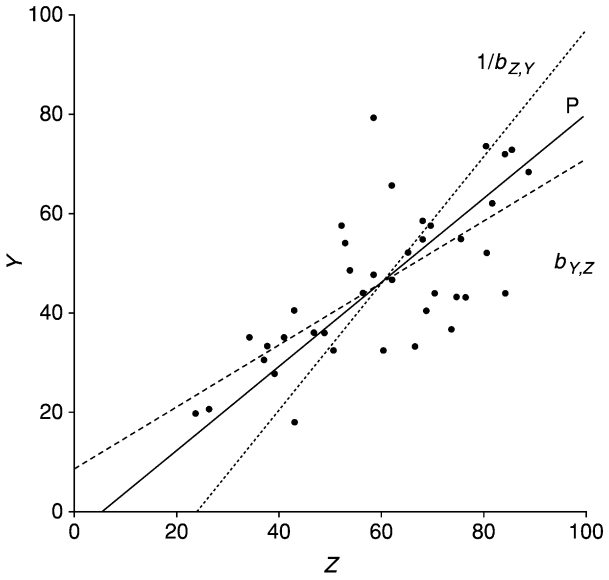


Figure 5 Scatter graph showing the relations between two variables Y and Z for which the correlation coefficient, r, is 0.699. The symbols are the observed values, the solid line, labeled 'P,' is the principal axis, with gradient 0.844, equivalent to an angle of 40.1° to the horizontal. The dashed line shows the regression of Y on Z, with gradient $b_{Y,Z} = 0.621$; and the dotted line shows that of Z on Y, for which $b_{Y,Z} = 0.788$, giving it a gradient of approximately $1/b_{Z,Y} = 1.269$.

$$r_{y,z} = \frac{\hat{c}_{y,z}}{\sqrt{s_y^2 s_z^2}} \qquad [20]$$

where s_y^2 and s_z^2 are the estimated variances of y and z. The coefficient was proposed by Karl Pearson, and for that reason it is often called the 'Pearson correlation coefficient.'

The coefficient is effectively a standardized version of the covariance. It measures the extent to which the data, when plotted as one variable against the other in a scatter graph, depart from a straight line (see Figure 5). It may vary between +1, signifying perfect positive correlation, and −1, for perfect negative correlation. Intermediate values indicate departures from the straight line, as in Figure 5, for which $r = 0.699$. In general, positive values of r indicate the tendency of y and z to increase together, whereas negative values arise when y increases as z decreases. A value of zero represents no linear relation.

Notice that the statistic refers specifically to linear correlation. The relation between two variables might be curved; the absolute value of r would then be necessarily less than 1 regardless of any scatter about the curve.

When the data, y_i, z_i, $i = 1, 2, \ldots, N$, are from a sample, then $c_{y,z}$ and $r_{y,z}$ estimate corresponding population parameters, $\mathrm{cov}_{y,z}$ and $\rho_{y,z}$. If the data can be assumed to be drawn from a bivariate normal distribution then one can test r for significance. One computes Student's t with $N - 2$ degrees of freedom:

$$t_{N-2} = \frac{r\sqrt{N-2}}{\sqrt{1-r^2}} \qquad [21]$$

The probability of this value's occurring on the null hypothesis that $\rho = 0$ can then be computed or looked up in a table of t.

Spearman rank correlation Where the distributions of the underlying variables are far from normal, the Pearson coefficient can be replaced by the Spearman rank correlation coefficient, usually denoted r_S. The values of each variable are ranked from smallest to largest and given new values 1, 2, ..., N. The correlation coefficient is then computed by applying eqns [19] and [20]. Alternatively, one may take the differences, d_i, $i = 1, 2, \ldots, N$, between the ranks and compute:

$$r_S = 1 - \frac{6\sum_{i=1}^{N} d_i^2}{N(N^2-1)}. \qquad [22]$$

Many soil variables observed in the field, such as grade of structure and frequency of mottles, are recorded as rankings rather than measured. In these circumstances the correlations between them can be expressed by the Spearman coefficient, whereas the Pearson coefficient would be inappropriate. Also, in these circumstances, tied ranks in any large set of data are inevitable, and Eqn [22] must be elaborated. The coefficient can be calculated in various ways, but from Eqns [19], [20], and [22] can be derived:

$$r_s = \frac{\sum_{i=1}^{N}(y_i-\bar{y})^2 - \sum_{i=1}^{N} T_{i,y} + \sum_{i=1}^{N}(z_i-\bar{z})^2 - \sum_{i=1}^{N} T_{i,z} - \sum_{i=1}^{N} d_i^2}{2\sqrt{\left\{(y_i-\bar{y})^2 - \sum_{i=1}^{N} T_{i,y}\right\}\left\{(z_i-\bar{z})^2 - \sum_{i=1}^{N} T_{i,z}\right\}}} \qquad [23]$$

in which:

$$T_i = \frac{t_i(t_i^2-1)}{12} \qquad [24]$$

where t_i is the number of observations tied at rank i.

For small samples, r_S is somewhat less sensitive than the Pearson coefficient in that larger values are necessary to establish statistical significance.

Regression

Regression treats the relation between two variables in a somewhat different way by designating one of them, y, as depending on the other, z, represented by the equation:

$$y = \beta_0 + \beta_1 z \quad [25]$$

The underlying rationale is often physical. For example, How is the soil's strength changed by additions of gypsum? Known amounts of gypsum can be added to the soil, the resultant changes in strength measured, and from the data how much on average the strength is changed by each increment in gypsum added can be estimated. The Gauss linear model can be adopted for this purpose:

$$y_i = \beta_0 + \beta_1 z_i + \epsilon_i \quad [26]$$

where y_i is the value of the random dependent variable, Y, here strength, in unit i, z_i is that of the independent variable, added gypsum, and ϵ_i is random error term that is uniformly and independently distributed with variance σ_ϵ^2. The quantities β_0 and β_1 are parameters of the model and are estimated as follows:

$$\hat{\beta}_1 = \frac{\hat{c}_{y,z}}{s_z^2} \quad [27]$$

and:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{z} \quad [28]$$

Equation [27] gives the rate at which strength changes in response to increments in gypsum. The quantity $\hat{\beta}_0$ in Eqn [28] is the intercept at $y = 0$ and is likely to be of subsidiary interest. Together $\hat{\beta}_0$ and $\hat{\beta}_1$ may be inserted into Eqn [25] for predicting unknown values of Y if we know those of z:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 z \quad [29]$$

The procedure minimizes the sum of the squares of the differences between the measured values y_i, $i = 1, 2, \ldots, N$, and those expected from Eqn [25], $\hat{y}_i$:

$$s_{Y,z}^2 = \frac{1}{N-2} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 \quad [30]$$

A somewhat different situation is common in soil survey. In it we sample the soil without knowing in advance what values we shall obtain for the variables of interest. For example, we may measure both cation exchange capacity, CEC, and clay content, and we may regard CEC as depending in a physical sense on clay content. The data on both variables now contain random components, and for this reason we designate them with the capital letters, Y and Z, respectively. We may express the relation as the regression of CEC on clay and estimate the parameters in the same way as for the Gauss linear model. In doing so we assign all the error to CEC and minimize the sum of squares, $s_{Y,Z}^2$, as in Eqn [30]. The purpose is now prediction, i.e., prediction of CEC, knowing the clay content. We could equally well compute the regression of clay on CEC. The roles of Y and Z are reversed, and we minimize:

$$s_{Z,Y}^2 = \frac{1}{N-2} \sum_{i=1}^{N} (z_i - \hat{z}_i)^2 \quad [31]$$

where:

$$\hat{z} = \hat{\beta}_0' + \hat{\beta}_1' y \quad [32]$$

The primes attached to $\hat{\beta}_0'$ and $\hat{\beta}_1'$ signify that these quantities refer to the regression of Z on Y, and that they are different from those for the regression of Y on Z. In other words, the line defined by Eqn [29] differs from that defined by [32], as **Figure 5** shows. The correct line to choose depends on which variable is to be predicted. To predict y from z, Eqn [29] is used; to predict z from y, Eqn [32] is used.

Further Reading

Cochran WG (1977) *Sampling Techniques*, 3rd edn. New York: John Wiley.

Draper NR and Smith H (1981) *Applied Regression Analysis*, 2nd edn. New York: John Wiley.

Mardia KV, Kent JT, and Bibby JM (1979) *Multivariate Analysis*. London, UK: Academic Press.

Snedecor GW and Cochran WG (1989) *Statistical Methods*, 8th edn. Ames, IA: Iowa State University Press.

Sokal RR and Rohlf FJ (1981) *Biometry*, 2nd edn. San Francisco, CA: WH Freeman.

Townend J (2002) *Practical Statistics for Environmental and Biological Scientists*. Chichester, UK: John Wiley.

Webster R (1997) Regression and functional relations. *European Journal of Soil Science* 48: 557–566.

Webster R (2001) Statistics in soil research and their presentation. *European Journal of Soil Science* 52: 331–340.

Webster R and Oliver MA (1990) *Statistical Methods in Soil and Land Resource Survey*. Oxford, UK: Oxford University Press.

Sterilization *See* **Disinfestation**

STOCHASTIC ANALYSIS OF SOIL PROCESSES

D Russo, Volcani Center, Bet Dagan, Israel

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Quantitative field-scale descriptions of water flow and chemical transport in the unsaturated (vadose) zone are essential for improving the basic understanding of the transport process in near-surface geological environments, and for providing predictive tools that, in turn, will be used to predict the future spread of pollutants in these environments. One of the distinctive features of a natural formation at the field scale is the spatial heterogeneity of its properties that affect flow and transport. This spatial heterogeneity is generally irregular; it occurs on a scale beyond the scope of laboratory samples and has a distinct effect on the spatial distribution of solutes, which results from transport through the formation. A fundamental question is that of how to develop predictive models that may incorporate the impact of field-scale spatial variability of the formation properties on vadose-zone flow and transport.

Modeling Flow and Transport in Unsaturated, Heterogeneous Soils

The starting point in modeling soil water flow is the uncertainty in the soil properties that affect flow and transport (because of their inherently erratic nature and the paucity of measurements). This uncertainty generally precludes the use of the traditional, deterministic modeling approach for the prediction of flow and transport on the field scale. In an alternative approach, uncertainty is set in a mathematical framework by modeling the relevant soil properties as random space functions (RSFs). As a consequence, the flow and transport equations are of a stochastic nature, and the dependent variables are also RSFs. The aim of the stochastic approach, therefore, is to evaluate the statistical moments of the variables of interest, given the statistical moments of the formation properties. In general, this is a formidable task and usually its scope is restricted to finding the first two moments. To simplify matters, it is common to treat each of the soil properties, as well as the various flow-controlled attributes, denoted by $p'(\mathbf{x})$, where $\mathbf{x}=(x_1,x_2,x_3)$ is the spatial coordinate vector, as stationary, characterized at second order by a spatially invariant mean, $P=<p'(\mathbf{x})>$, and a covariance, $C_{pp}(\mathbf{x}',\mathbf{x}'')=<p(\mathbf{x}')p(\mathbf{x}'')>$, in which $p(\mathbf{x})=p'(\mathbf{x})-P$, that depends on the separation vector $\boldsymbol{\xi}=\mathbf{x}'-\mathbf{x}''$, and not on $\mathbf{x}'$ and $\mathbf{x}''$ individually.

Under unsaturated flow conditions, the analysis is further complicated by the fact that the relevant flow parameters – hydraulic conductivity, K, and water capacity, C, which depend on formation properties – depend also on flow-controlled attributes in a highly nonlinear fashion. Consequently, under variably saturated conditions, the evaluation of the effects on flow and transport of spatial variations in the formation properties is extremely complex; it requires several simplifying assumptions regarding the constitutive relationships for unsaturated flow, the flow regime, and the spatial structure of the formation heterogeneity.

The focus here is on transport of conservative, nonreactive, nonvolatile solutes, under steady flow and in the absence of plant roots. Furthermore, the emphasis is on formations, for which the spatial distribution of their properties can be modeled by a unimodal distribution with a spatial correlation structure characterized by a covariance with a single, finite-length scale. It should be emphasized, however, that the theoretical framework described here can serve as a basis for more complicated situations, including transient water flow, transport of reactive solutes, and multiple-length scale, heterogeneous formations.

Quantification of solute transport in partially saturated, heterogeneous porous formations may be accomplished with a two-stage approach, which combines a stochastic continuum description of the steady-state flow with a general Lagrangian description of the motion of an indivisible particle of a passive solute that is carried by a steady-state flow. The first stage involves relating the statistical moments of the probability density function (PDF) of the velocity to those of the properties of the formation, while the second stage involves relating the statistical moments of the particle displacement PDF to those of the velocity PDF.

Stochastic Analysis of the Flow

A first-order perturbation approach is used here in order to obtain the first two statistical moments of the PDF of flow-controlled attributes, i.e., the pressure head, ψ, the log-unsaturated conductivity, $\log K$, the water flux vector, $\mathbf{q}'$, the water saturation, Θ, and, concurrently, those of the Eulerian velocity vector, $\mathbf{u}'$, for given statistics of the formation properties. The following assumptions are made:

1. The heterogeneous formation has a three-dimensional structure with axisymmetric anisotropy, and the flow domain is variably saturated and unbounded;
2. The local steady-state unsaturated flow obeys the Darcy law and continuity, which, if local isotropy is assumed, reads:

$$q_i'(\mathbf{x}) = -K(\psi,\mathbf{x})\frac{\partial\varphi(\mathbf{x})}{\partial x_i}$$
$$\frac{\partial q_i'(\mathbf{x})}{\partial x_i} = 0 \quad i = 1,2,3 \qquad [1]$$

where q_i' $(i=1,2,3)$ is the water flux vector, $\varphi = -x_1 - \psi$ is the hydraulic head, x_1 is directed vertically downward, and the Einstein summation convention is used in eqn [1] and elsewhere;
3. The local relationships between K and ψ (considered here as a positive quantity) and Θ and ψ are nonhysteretic and are given by the Gardner–Russo model, i.e.,

$$K(\psi,\mathbf{x}) = K_s(\mathbf{x})\exp[-\alpha(\mathbf{x})\psi] \qquad [2a]$$

$$\Theta(\psi,\mathbf{x}) = \left\{\exp\left[-\frac{1}{2}\alpha(\mathbf{x})\psi\right]\left[1+\frac{1}{2}\alpha(\mathbf{x})\psi\right]\right\}^{2/(m'+2)} \qquad [2b]$$

where K_s is saturated conductivity, $\alpha = \lambda^{-1}$ is a parameter of the formation, λ is the macroscopic capillary length scale and m′ is a parameter, selected here as $m' = 0$;
4. Both $\log K_s$ and $\log\alpha$ are multivariate normal (MVN) RSFs, ergodic over the region of interest, characterized by constant means, $F = E[\log K_s]$ and $A = E[\log\alpha]$, and by covariances $C_{ff}(\boldsymbol{\xi})$ and $C_{aa}(\boldsymbol{\xi})$, respectively, and cross-covariance, $C_{fa}(\boldsymbol{\xi})$ given by:

$$C_{pp}(\xi) = \sigma_{p*}^2\exp(-\xi') \qquad [3]$$

where f and a are the perturbations of $\log K_s$ and $\log\alpha$, respectively, $p = f$ or a, $p^* = f$, a, or fa; $\boldsymbol{\xi}' = (\mathbf{x}-\mathbf{x}'')/\mathbf{I}_p$, is the scaled separation vector, $\xi' = |\boldsymbol{\xi}'|$; σ_f^2 and σ_a^2, and $\mathbf{I}_f = (I_{f1},I_{f2},I_{f3})$ and $\mathbf{I}_a = (I_{a1},I_{a2},I_{a3})$ are the respective variances and correlation scales of $\log K_s$ and $\log\alpha$, σ_{fa}^2 is the cross-variance between perturbations of $\log K_s$ and $\log\alpha$, and $\mathbf{I}_{fa} = 2\mathbf{I}_f\mathbf{I}_a/(\mathbf{I}_f+\mathbf{I}_a)$;
5. The porosity, ϕ is constant and uniform throughout the formation.

Use of the aforementioned assumptions under ergodic conditions, by elimination of $\mathbf{q}'$ from eqn [1], expression of the various parameters and variables on the right-hand side (RHS) of eqn [2a] in terms of means and perturbations, i.e., $\log K_s = F+f$, $\log\alpha = A+a$, $\psi = H+h$, and use of the Taylor expansion with first-order terms retained, gives the first-order perturbation approximation of the steady flow equation as:

$$\frac{\partial^2 h}{\partial x_i^2} - \Gamma(2J_i - \delta_{1i})\frac{\partial h}{\partial x_i} - J_i(J_i - \delta_{1i})\Gamma a + J_i\left(\frac{\partial f}{\partial x_i} - \Gamma H\frac{\partial a}{\partial x_i}\right) = 0, \quad i = 1,2,3 \qquad [4]$$

where $\Gamma = \exp(A)$ is the geometric mean of α, $J_i = \partial(-H - x_1)/\partial x_i$ $(i=1,2,3)$ is the mean of the head-gradient vector and δ_{1i} $(i=1,2,3)$ is the Kronecker delta.

On the assumption that the flow is gravity-dominated, i.e., the mean pressure-head gradient is zero so that the mean head-gradient, J_i, is given by $J_i = \delta_{i1}$ $(i=1,2,3)$, the exact solutions of eqn [4] in terms of the (cross-)spectral relationships between the RSFs of f, a, and h, which are obtained by using Fourier–Stieljes integral representations, are:

$$\hat{C}_{hh}(\mathbf{k}) = \frac{k_1^2[\hat{C}_{ff}(\mathbf{k}) + \Gamma^2H^2\hat{C}_{aa}(\mathbf{k}) - 2\Gamma H\hat{C}_{fa}(\mathbf{k})]}{k^4 + \Gamma^2k_1^2} \qquad [5]$$

$$\hat{C}_{hf}(\mathbf{k}) = \frac{k_1(\Gamma k_1 - j'k^2)[\hat{C}_{ff}(\mathbf{k}) - \Gamma H\hat{C}_{fa}(\mathbf{k})]}{k^4 + \Gamma^2k_1^2} \qquad [6]$$

$$\hat{C}_{ha}(\mathbf{k}) = \frac{k_1(\Gamma k_1 - j'k^2)[\hat{C}_{fa}(k) - \Gamma H\hat{C}_{aa}(\mathbf{k})]}{k^4 + \Gamma^2k_1^2} \qquad [7]$$

where j' is the imaginary unit, $\mathbf{k} = (k_1,k_2,k_3)$ is the wave number vector, $k = |\mathbf{k}|$, and the inverse Fourier transform of the covariance of the formation properties, $C_{pp}(\boldsymbol{\xi})$ (eqn [3] with $p = \log K_s$ or $\log\alpha$), is given by:

$$\hat{C}_{pp}(\mathbf{k}) = \frac{\sigma_p^2 I_{p1}I_{p2}I_{p3}}{\pi^2(1 + I_{p1}^2k_1^2 + I_{p2}^2k_2^2 + I_{p3}^2k_3^2)^2} \qquad [8]$$

In a similar way, if the variables on the left-hand-side (LHS) of eqns [1] and [2b] are expressed in terms of means and perturbations, i.e., $\mathbf{q}' = \mathbf{Q}+\mathbf{q}$ and $\Theta = S+s$, respectively, and the Taylor expansion is used with first-order terms retained, the (cross-) spectral relationships between the RSFs of f, a, h, and s, and f, a, h, and q_i $(i=1,2,3)$, which are obtained by using Fourier–Stieljes integral representations, are:

$$\hat{C}_{ss}(\mathbf{k}) = B^2\{\hat{C}_{hh}(\mathbf{k}) + H^2\hat{C}_{aa}(\mathbf{k}) + H[\hat{C}_{ha}(\mathbf{k}) + \hat{C}_{ah}(\mathbf{k})]\} \qquad [9]$$

$$\hat{C}_{sh}(\mathbf{k}) = B[\hat{C}_{hh}(\mathbf{k}) + H\hat{C}_{ha}(\mathbf{k})] \quad [10]$$

$$\hat{C}_{sf}(\mathbf{k}) = B[\hat{C}_{hf}(\mathbf{k}) - H\hat{C}_{fa}(\mathbf{k})] \quad [11]$$

$$\hat{C}_{sa}(\mathbf{k}) = B[\hat{C}_{ha}(\mathbf{k}) - H\hat{C}_{aa}(\mathbf{k})] \quad [12]$$

$$\begin{aligned}\hat{C}_{q_{ij}}(\mathbf{k}) = K_g^2\{&\delta_{1i}\delta_{1j}[\hat{C}_{ff}(\mathbf{K}) + (\Gamma H)^2\hat{C}_{aa}(\mathbf{k}) + 2\Gamma H\hat{C}_{fa}(\mathbf{k})] \\ &+ \delta_{1i}[(\Gamma + j'k_i)(\hat{C}_{fh}(\mathbf{k}) + \Gamma H\hat{C}_{ah}(\mathbf{k}))] \\ &+ \delta_{1j}[\Gamma - j'k_j)(\hat{C}_{hf}(\mathbf{k}) + \Gamma H\hat{C}_{ha}(\mathbf{k})] \\ &+ (\delta_{1i}\delta_{1j}\Gamma^2 + k_i^2)\hat{C}_{hh}(\mathbf{k})\} \quad i,j = 1,2,3 \end{aligned} \quad [13]$$

where $B = (1/4)\ \Gamma^2 H\ \exp[-(1/2)\Gamma H]$, $K_g = \exp(Y)$ is the geometric mean conductivity, and $Y = F - \Gamma H$ is the mean of log$K$. Note that because $\psi = H + h$ is a function of water saturation, the (cross-)spectra $\hat{C}_{mn}$ $(\boldsymbol{k})$ (m, $n = f$, a, h, s; eqns [5–7] and [9–13]) depend on saturation. This dependence, however, is omitted for simplicity of notation. Furthermore, for $H \to 0$ (and $\Gamma \to 0$), the RHS of eqns [9–12] vanish, while eqns [5], [6], [7], and [13] reduce to their counterparts associated with steady flow in saturated formations.

The (cross-)covariances, $C_{mn}(\xi)$ associated with the (cross-)spectra $\hat{C}_{mn}(\mathbf{k})$ (m, $n = f, a, h, s$; eqns [5–13]), are then calculated by taking the Fourier transform of the respective $\hat{C}_{mn}(\mathbf{k})$, i.e.:

$$C_{mn}(\boldsymbol{\xi}) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \exp(j'\,\mathbf{k}\cdot\boldsymbol{\xi})\hat{C}_{mn}(\mathbf{k})d\mathbf{k} \quad [14]$$

Scaled forms of the pressure-head covariance, $C'_{hh}(\boldsymbol{\xi}) = C_{hh}(\boldsymbol{\xi})/I_{yv}^2\sigma_y^2$, independent of water saturation, Θ, are illustrated graphically in **Figure 1** as functions of the scaled separation distance in the direction parallel to the mean flow, $\xi'_1 = \xi_1/I_{yv}$ (**Figure 1a, c**), and in the direction perpendicular to the mean flow, $\xi'_2 = \xi_2/I_{yv}$ (**Figure 1b, d**), for selected values of the length-scale ratios, $\eta = \lambda/I_v$ and $\rho = I_h/I_v$, where $I_v = I_{f1} = I_{a1}$, and $I_h = I_{f2} = I_{f3} = I_{a2} = I_{a3}$. Here σ_y^2 and $I_{yv} = I_{y1}$ are the Θ-dependent, log$K$ variance and correlation length-scale in the vertical direction, respectively, given by:

$$\sigma_y^2 = \sigma_f^2 + \Gamma^2H^2\sigma_a^2 - 2\Gamma H\sigma_{fa}^2 - \Gamma^2\sigma_h^2 \quad [15a]$$

$$I_{yv} = \frac{\sigma_f^2 I_{f1} + \Gamma^2H^2\sigma_a^2 I_{a1} - 2\Gamma H\sigma_{fa}^2 I_{fa1} - \Gamma^2\int_0^\infty C_{hh}(\xi_1,0,0)d\xi_1}{\sigma_f^2 + \Gamma^2H^2\sigma_a^2 - 2\Gamma H\sigma_{fa}^2 - \Gamma^2\sigma_h^2} \quad [15b]$$

and $\sigma_h^2 = C_{hh}(0)$ is the variance of the pressure-head perturbations.

Note that the correlation length-scales, I_v and I_h, determine approximately the distances, parallel and perpendicular to the mean flow, respectively, at which property variations cease to be correlated. On the other hand, the macroscopic capillary length-scale,

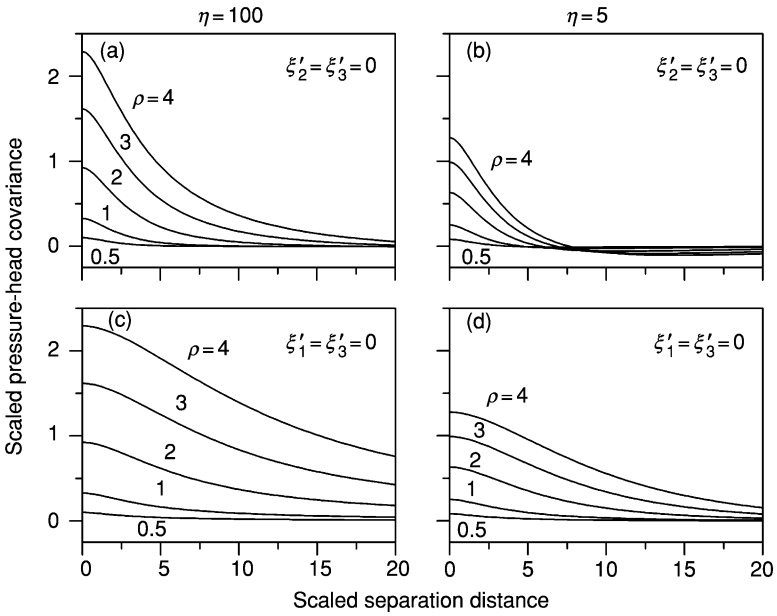


Figure 1 Scaled pressure-head covariance, C'_{hh}, independent of mean water saturation, S, as a function of the scaled separation distance parallel, ξ'_1 (a and b), and perpendicular, ξ'_2 (c and d) to the direction of the mean flow, for selected values of $\rho = I_h/I_v$ (the numbers labeling the curves), for various values of $\eta = \lambda/I_v$, and for $\nu = \sigma_a^2/\sigma_f^2 = 0.2$, $\sigma_f^2 = 0.5$, $\sigma_{fa}^2 = 0$, $I_{yv} = 0.2$ m, $Q/\phi S = 1$m day^{-1}, and $J = (1,0,0)$.

λ, determines the relative magnitude of the capillary forces in unsaturated flow; it may be regarded as a natural length-scale of the unsaturated soil. The results depicted in Figure 1 show that, for given variability in the soil properties, both the magnitude and the persistence of the pressure-head variability increase with increasing ρ and η. Furthermore, the persistence of the pressure-head variability in the direction perpendicular to the mean flow is larger than that in the direction of the mean flow, particularly when both ρ and η are relatively large. This can be explained as follows: when the mean flow is vertical and I_v is kept constant, an increasing ρ expresses an increase in the size of the typical flow barriers in a direction normal to the mean flow; on the other hand, an increasing η expresses an increase in the macroscopic capillary length-scale, λ, i.e., a transition from a coarse-textured soil material, associated with negligible capillary forces to a fine-textured soil material, associated with significant capillary forces. Consequently, as ρ increases, the streamlines are deflected less easily; similarly, as η increases, the lateral head perturbation gradients increase. In other words, increasing stratification and capillary forces are expected to enhance the lateral dissipatation of water and, concurrently, to enhance the variability in the pressure head.

In Figure 2, principal components of the scaled flux covariance tensor, independent of S, i.e., $q'_{ii}(\boldsymbol{\xi}') = Cq_{ii}(\boldsymbol{\xi}')/\sigma_y^2 Q^2$ $(i = 1,2,3)$, where $Q = |\mathbf{Q}|$, are depicted as functions of $\xi'_1 = \xi_1/I_{yv}$, for selected values of the length-scale ratios, $\eta = \lambda/I_v$ and $\rho = I_h/I_v$. Note that because of the assumption of axisymmetric anisotropy, and for $J_2 = J_3 = 0$, $q'_{22}(\boldsymbol{\xi}) = q'_{33}(\boldsymbol{\xi})$. Figure 2 shows that the longitudinal component of the scaled flux covariance tensor is much larger than its transverse components, and also that both the magnitude and the persistence of the transverse components of the flux covariance tensor may be greater in fine-textured soil material. However, in such soil, the magnitude of the longitudinal component of the flux covariance tensor increases and its persistence decreases. On the other hand, soil stratification may increase the persistence of the longitudinal and the transverse components of the flux covariance tensor, may decrease the magnitude of the longitudinal component, and may either increase (for $\rho < \pi/2$) or decrease (for $\rho > \pi/2$) the magnitude of the transverse components of the flux covariance tensor.

The behavior of the principal components of the flux covariance tensor depicted in Figure 2 is explained on the same basis as above, i.e., when the mean flow is vertical and I_v is kept constant, an increasing ρ expresses increasing size of the typical flow barriers in a direction normal to the mean flow, while an increasing η expresses increasing capillary forces. Consequently, $q'_{11}(0)$ approaches unity at the small ρ limit, $\rho \to 0$ (which implies heterogeneity in the horizontal directions only) and decreases with increasing ρ, vanishing at the large ρ limit, $\rho \to \infty$ (which implies a perfectly stratified formation), while $q'_{ii}(i = 2, 3)$ are nonmonotonic functions of ρ with a maximum at

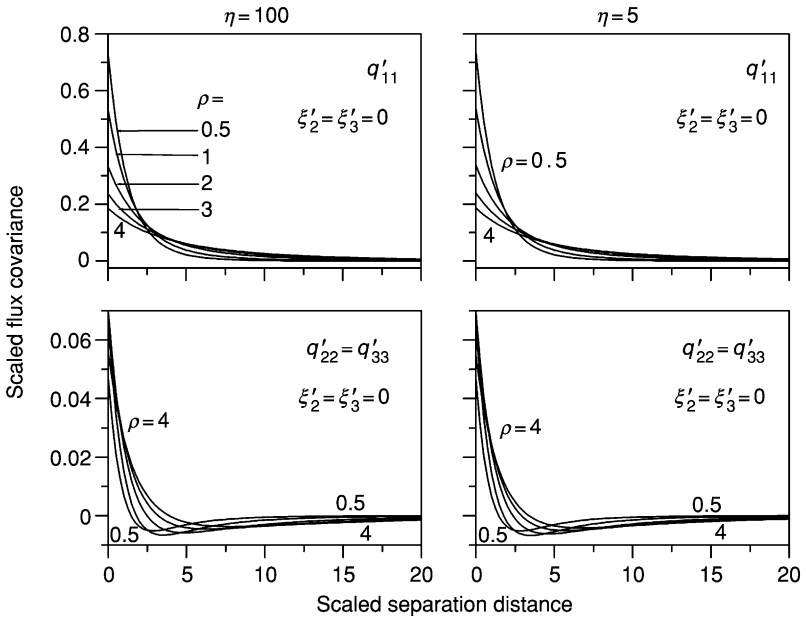


Figure 2 Principal components of the scaled flux covariance tensor, q'_{ii}, $i = (1,2,3)$, independent of mean water saturation, S, as a function of the scaled separation distance parallel to the direction of the mean flow, ξ'_1, for selected values of $\rho = I_h/I_v$ (the numbers labeling the curves), for various values of $\eta = \lambda/I_v$, and for $\nu = \sigma_a^2/\sigma_f^2 = 0.2$, $\sigma_f^2 = 0.5$, $\sigma_{fa}^2 = 0$, $I_{yv} = 0.2$m, $Q/\phi S = 1$m day^{-1}, and $J = (1,0,0)$.

$\rho = \pi/2$, that vanishes at the small ρ limit, $\rho \to 0$ and at the large ρ limit, $\rho \to \infty$. On the other hand, $q'_{ii}(0)$ $(i = 1,2,3)$ increases with increasing η.

When the mean flow is aligned with the x_1-axis only, the first term on the RHS of the Fourier transform of eqn [13] vanishes for $i,j = 2,3$, while both its second and third terms vanish at $\xi_1 = 0$ and $\xi_1 \to \infty$. Furthermore, the last term is antisymmetric in ξ_1. Consequently, integration of the principal components of the flux covariance tensor along the ξ_1-axis, when performed up to $\xi_1 \to \infty$, yields $\int_0^\infty Cq_{11}(\boldsymbol{\xi})d\xi_1 = Q^2\sigma_y^2 I_{yv}$ and $\int_0^\infty Cq_{ii}(\boldsymbol{\xi})d\xi_1 = 0$, $(i = 2$ to $3)$, irrespective of the value of ρ or η. Hence, the decrease in $q'_{11}(0)$ with increasing ρ and decreasing η must be compensated for by an increase in the separation distance over which $q_1(x_1)$ and $q_1(x_1 + \xi_1)$ are positively correlated. Consequently, the persistence of $q'_{11}(\xi_1)$ increases with increasing ρ and decreasing η. On the other hand, for $i = 2$ or $i = 3$, the increase in the separation distance over which $q_i(x_1)$ and $q_i(x_1 + \xi_1)$ are positively correlated with increasing ρ and η must be compensated by an increase in the separation distance over which $q_i(x_1)$ and $q_i(x_1 + \xi_1)$ are negatively correlated. Hence, the persistence of $q'_{ii}(\xi_1)$ $(i = 2,3)$ increases with increasing ρ and η.

Stochastic Analysis of the Transport

A first-order perturbation approach is used here in order to obtain the first two statistical moments of the PDF of the one-particle-trajectory for given statistics of the Eulerian velocity vector, $\mathbf{u}'(\mathbf{x})$, and, concurrently, for given statistics of the formation properties. In a general Lagrangian framework, the transport is described in terms of the motion of indivisible solute particles which are convected by the fluid. If pore-scale dispersion is neglected, the trajectory of a solute particle is related to $\mathbf{u}'(\mathbf{x})$ by the fundamental kinematic relationship:

$$\frac{d\mathbf{X}'}{dt} = \mathbf{u}'(\mathbf{X}') \quad \text{for} \quad t > 0; \mathbf{X}' = \mathbf{a} \quad \text{for} \quad t = 0 \qquad [16]$$

where t is time, and $\mathbf{X}' = \mathbf{X}'(t,\mathbf{a})$ is the trajectory of a particle which is at $\mathbf{X}' = \mathbf{a}$ when $t = 0$.

In order to relate the statistical moments of the PDF of the one-particle trajectory to those of the velocity PDF, the following assumptions are employed: (1) Lagrangian and Eulerian stationarity and homogeneity of the velocity field; (2) given statistics of the velocity field; (3) small fluctuations of the particle displacement about the mean trajectory; and (4) large Peclet numbers, i.e., local dispersion is omitted.

Under the aforementioned assumptions, for ergodic conditions and for a particle of a solute injected into the flow field at time $t = 0$ and location $\mathbf{x} = \mathbf{a}$, the solution of eqn [16] is:

$$\mathbf{X}'(t,\mathbf{a}) = \mathbf{a} + \mathbf{U}t + \int_0^t \mathbf{u}'(\mathbf{a} + \mathbf{U}\tau)d\tau \qquad [17]$$

where $\mathbf{U} = <\mathbf{u}'(\mathbf{x})>$ is the mean Eulerian velocity vector.

For fixed $\mathbf{a}$, the first two moments of $\mathbf{X}'$ (eqn [17]), the mean, $<\mathbf{X}'(t)>$ and, by a first-order approximation in the velocity variance, the covariance tensor, $X_{ij}(t) = <X_i(t)X_j(t)>$, $(i,j = 1,2,3)$, of the one-particle displacement at time t, where $\mathbf{X} = \mathbf{X}' - <\mathbf{X}'>$ is the fluctuation, are given by:

$$<\mathbf{X}'(t)> = \mathbf{U}t \qquad [18a]$$

$$X_{ij}(t) = 2\int_0^t (t - \tau)C_{u_{ij}}(\mathbf{U}\tau)d\tau \qquad [18b]$$

Here $C_{uij}(\boldsymbol{\xi}) = <u_i(\mathbf{x})u_j(\mathbf{x} + \boldsymbol{\xi})>$, $(i,j = 1,2,3)$, the velocity covariance tensor, where $u_i(\mathbf{x}) = u_i'(\mathbf{x}) - <u_i'(\mathbf{x})> = q_i(\mathbf{x})/\phi S - Q_i s(\mathbf{x})/\phi S^2$ is the velocity fluctuation, is given by:

$$C_{u_{ij}}(\boldsymbol{\xi}) = \frac{K_g^2 \delta_{1i}\delta_{1j}}{\phi^2 S^4} C_{ss}(\boldsymbol{\xi}) + \frac{C_{q_{ij}}(\boldsymbol{\xi})}{\phi^2 S^2} - \frac{K_g^2}{n^2\phi^3}\left[- \delta_{1i}\delta_{1j}[C_{sy}(\boldsymbol{\xi}) + C_{sy}(-\boldsymbol{\xi})] - \delta_{1j}\frac{\partial C_{sh}(\boldsymbol{\xi})}{\partial \xi_i} + \delta_{1i}\frac{\partial C_{sh}(-\boldsymbol{\xi})}{\partial \xi_j}\right] \qquad [19]$$

where $C_{sy}(\boldsymbol{\xi})$ is the cross-covariance between perturbations of $\log K$ and Θ given by:

$$C_{sy}(\boldsymbol{\xi}) = C_{sf}(\boldsymbol{\xi}) - \Gamma C_{sh}(\boldsymbol{\xi}) - \Gamma H C_{sa}(\boldsymbol{\xi}) \qquad [20]$$

Eqns [18a] and [18b] are of a general nature, consistent with the linearization of the flow equation. Furthermore, because the soil properties, $\log K_s$ and $\log\alpha$ are MVN, $\log K$, ψ, and Θ are also MVN. Consequently, the velocity is MVN, and, by eqn [17], the one-particle trajectory, $\mathbf{X}'(t)$, is also MVN at any time. This means that, generally, the PDF of $\mathbf{X}'(t)$ satisfies the Focker–Planck equation and the expected concentration satisfies the classical convection dispersion equation. Furthermore, under ergodic conditions, the one-particle trajectory statistical moments eqn [18], can be equated with those of the spatial moments of the resident concentration, $c(\mathbf{x};t)$ (defined as the mass of the solute per unit volume of aqueous solution) of a plume. For a passive solute which lies at $t = 0$, within a finite volume V_0, the ith coordinate of the centroid of the solute plume, $X_i^e(t)$, and the

moment of inertia of the plume with respect to the x_i-axis, $X^e_{ii}(t)$ $(i=1,2,3)$, are:

$$X^e_i(t) = M^1_i(t) = <X_i(t)> + \frac{C_s X'_i(t)}{S} \qquad [21a]$$

$$X^e_{ii}(t) = \left\{M^2_i(t) - [M^1_i(t)]^2\right\} = X_{ii}(t) - \left[\frac{C_{sX'_i}(t)}{S}\right]^2 \qquad [21b]$$

where $M^N_1(t)$, $N=1$, 2, are the first and the second normalized spatial moments of the distribution of $c(\mathbf{x};t)$ in the ith coordinate, $<X_i(t)>$ and $X_{ii}(t)$ are the principal components of eqns [18a] and [18b], respectively, and $C_{sXi}(t) = <X_i(t)s(0)>$ is the cross-covariance between perturbations of the ith component of $\mathbf{X}'$ and water saturation, Θ.

It is clear from eqn [21] that, because Θ is spatially variable, the velocity $\mathbf{u}'(\mathbf{x}) = \mathbf{q}'(\mathbf{x})/\phi\Theta(\mathbf{x})$ is not divergence-free, i.e., $\mathrm{div}[\mathbf{u}'(\mathbf{x})] \neq 0$. Consequently, $X^e_i(t)$ differs from $<X_i(t)>$ by the correction factor $C_{sXi}(t)/S$. This implies time dependence of the effective solute velocity, $v^e_i = \mathrm{d}(X^e_i)/\mathrm{d}t = U_i + C_{sui}(\mathbf{U}t)/S$, where $C_{sui}(\boldsymbol{\xi}) = <u_i(\mathbf{U}\tau)s(0)>$ $(i=1,2,3)$ is the cross-covariance between perturbations of the ith component of the Eulerian velocity vector, $\mathbf{u}'$ and water saturation Θ, given by:

$$C_{su_i}(\boldsymbol{\xi}) = \frac{K_g}{\phi S^2}\left[-\delta_{1i}C_{ss}(\boldsymbol{\xi}) + \delta_{1i}SC_{sy}(\boldsymbol{\xi}) - S\frac{\partial C_{sh}(\boldsymbol{\xi})}{\partial \xi_i}\right] \qquad [22]$$

On the other hand, if the second term on the RHS of eqn [21*b*] is regarded as a second-order term which can be neglected as a part of the first-order approximation, then $X^e_{ii}(t)$ and the effective macrodispersion, $D^e_{ii}(t) = (1/2)\mathrm{d}[X^e_{ii}(t)]/\mathrm{d}t$ are not affected by the fact that $\mathrm{div}[\mathbf{u}'(\mathbf{x})] \neq 0$, and are identical to their counterparts, $X_{ii}(t)$ and $D_{ii}(t)$, the principal components of eqns [18b] and [23] below, respectively.

Note that, from a physical point of view, the ratios $X_{ij}(t)/2t$ $(i,j=1,2,3)$ may be regarded as apparent dispersion coefficients, that, for $t\to\infty$, tend to the macrodispersion tensor, $D_{ij}(t)$, $(i,j=1,2,3)$ given by:

$$D_{ij}(t) = \frac{1}{2}\frac{\mathrm{d}X_{ij}(t)}{\mathrm{d}t} = \int_0^t C_{u_{ij}}(\mathbf{U}\tau)\mathrm{d}\tau \quad (i,j=1,2,3) \qquad [23]$$

In the case of the velocity covariance tensor eqn [19], the first term on the RHS of eqn [19] depends only on the variability of the water saturation, Θ, while the second term on the RHS of eqn [19] depends only on the variability of the water flux, $\mathbf{q}$. The third term on the RHS of eqn [19] represents the effect on the velocity of the interaction between the Θ and the $\mathbf{q}$ heterogeneities. Inasmuch as the term $\delta_{1i}\delta_{1j}[C_{sy}(\boldsymbol{\xi}) + C_{sy}(-\boldsymbol{\xi})]$ is negative and $|\delta_{1i}\delta_{1j}[C_{sy}(\boldsymbol{\xi}) + C_{sy}(-\boldsymbol{\xi})]| > -\delta_{1j}[\partial C_{sh}(\boldsymbol{\xi})/\partial\xi_i] + \delta_{1i}[\partial C_{sh}(-\boldsymbol{\xi})/\partial\xi_j]$, all three terms on the RHS of eqn [19] contribute to the variability in velocity, and, concurrently, to macrodispersion (eqn [23]). Note that when the mean flow is aligned with the x_1-axis only, i.e., $\mathbf{J} = (J_1,0,0)$, the first and third terms on the RHS of eqn [19] vanish for $i=j=2$ or 3. In other words, in this case the transverse components of the velocity covariance tensor (eqn [19]) and, concurrently, those of the macrodispersion tensor (eqn [23]), are independent of the spatial variability of Θ and are only influenced by its mean value, $S = <\Theta(\mathbf{x})>$. Note also that when the formation is saturated (i.e., when $H\to 0$ and $S\to 1$), and when Γ approaches its small limit, $\Gamma\to 0$, the first and third terms on the RHS of eqn [19] vanish; consequently, eqn [19] and, concurrently, eqn [23] reduce to their counterparts associated with steady flow in saturated, three-dimensional heterogeneous formations.

Scaled forms of the principal components of eqn [23], $D'_{ii} = \phi S D_{ii}/Q\sigma^2_y I_{yv}$, $(i=1$ to $3)$, are presented graphically in **Figure 3**, as functions of the scaled travel time, $t' = tQ/\phi S I_{yv}$, for selected values of mean water saturation, S, and the length-scales ratios, $\rho = I_h/I_v$ and $\eta = \lambda/I_v$. Note that because of the assumption of axisymmetric anisotropy, and for $J_2 = J_3 = 0$, $D'_{22}(t) = D'_{33}(t)$. **Figure 3** suggests that the longitudinal component of the scaled dispersion tensor is much larger than its transverse components, particularly in formations in which $\rho \leq 1$. Note that for $\mathbf{J} = (J_1,0,0)$, because of the first and the third terms on the RHS of eqn [19], D'_{11} is saturation-dependent, while $D'_{ii}(i=2,3)$ are not.

The behavior of the principal components of eqn [23] demonstrates the combined influence of the formation heterogeneity, the capillary forces and water saturation on solute spreading under unsaturated flow. This arises directly from their effects on the velocity covariance (eqn [19]). For given $\nu = \sigma^2_a/\sigma^2_f$, S, ρ, and η, and when $\mathbf{J} = (J_1,0,0)$, the time behavior of $D'_{ii}(i=1,2,3)$ describes a continuous transition from a non-Ficksian to a Ficksian regime, i.e., they grow monotonously and linearly with t' at small t' $(t' \ll 1)$. As t' increases, however, the behavior of D'_{ii} $(i=2,3)$ is completely different from that of D'_{11}. The latter is a monotonically increasing function of t', approaching a constant value at the large t' limit, $t'\to\infty$, while $D'_{ii}(i=2,3)$ are nonmonotonic functions of t', vanishing at the large t' limit.

Figure 3 suggests that in unsaturated flow, solute spreading is expected to increase with diminishing

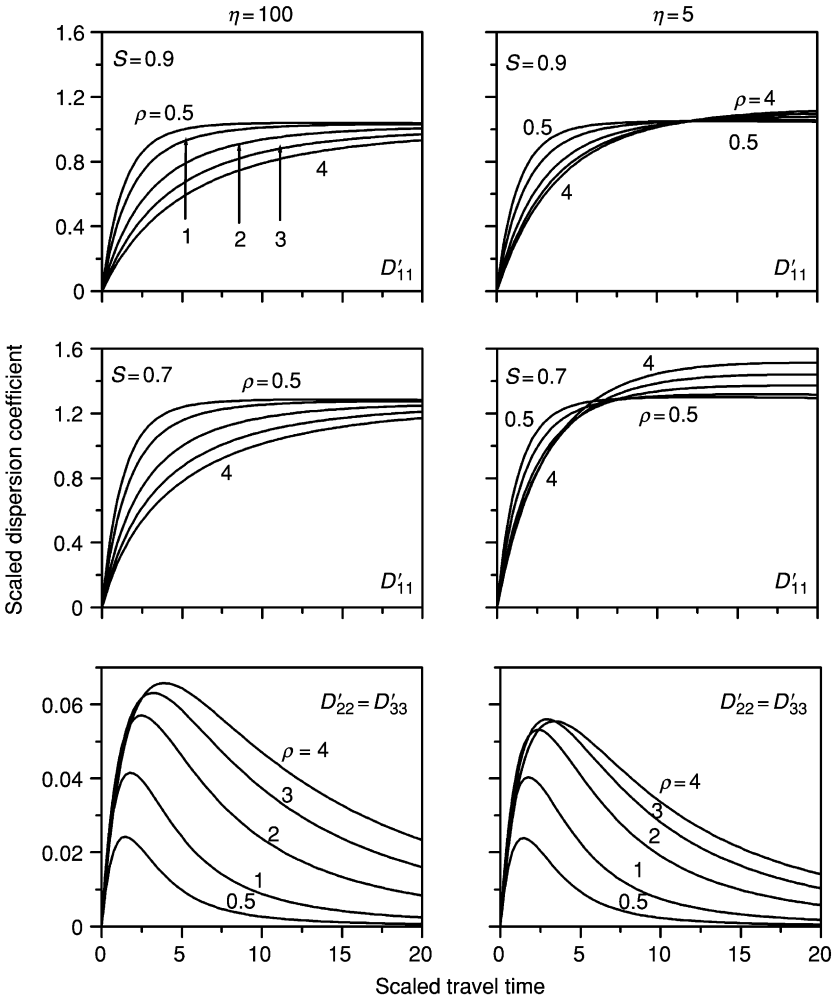


Figure 3 Principal components of the scaled macrodispersion coefficient tensor, D'_{i1}, $i=(1,2,3)$, as a function of the scaled travel time, t', for selected values of $\rho = I_h/I_v$ (the numbers labeling the curves), for various values of mean water saturation, S and $\eta = \lambda/I_v$, and for $\nu = \sigma_a^2/\sigma_f^2 = 0.2$, $\sigma_f^2 = 0.5$, $\sigma_{fa}^2 = 0$, $I_{yv} = 0.2$m, $Q/\phi S = 1$m day^{-1}, and $J=(1,0,0)$. Note that D'_{ii} ($i=(2,3)$) are independent of S.

water saturation. This is particularly so when the variances ratio, $\nu = \sigma_a^2/\sigma_f^2$, increases. Increasing stratification is expected to enhance solute spreading in the transverse directions in formations of both fine-textured and, especially, coarse-textured soil materials. On the other hand, increasing stratification is expected to reduce solute spreading in the longitudinal direction in formations of relatively fine-textured soil materials. In formations of relatively coarse-textured soil material, however, at a relatively small distance traveled, increasing stratification is expected to diminish solute spreading in the longitudinal direction while the converse is true when the travel distance exceeds a few log$K$ correlation length-scales. Furthermore, the travel distance required for the principal components of the macrodispersion tensor to approach their asymptotic values may be exceedingly large, particularly in relatively wet formations with significant stratification and with coarse-textured soil material that is associated with small capillary forces. Hence, in many practical situations of vadose-zone transport, the typical travel distance may be small compared with the travel distance required for the solute plume to reach its asymptotic, Ficksian behavior.

Summary

Before concluding, it is worthwhile emphasizing the limitations of the approach presented here. One limitation is due to the small-perturbation, first-order approximations of the velocity and the displacement covariance tensors, as a consequence of which the results are formally limited to formations with σ_y^2

smaller than unity. Other limitations stem from the fact that the calculations, which rely upon the assumption of ergodicity, are also restricted to conditions of constant mean-head gradient. The assumptions regarding the spatial structure of the formation (exponential covariance with axisymmetric anisotropy), the statistics of the relevant formation properties, and the flow-controlled attributes (statistically homogeneous), the local flow regime (steady-state flow), the local constitutive relationships for unsaturated flow (Gardner–Russo model), and the transport (Lagrangian stationarity, the neglect of both the fluctuations of the particle displacement about the mean trajectory and pore-scale dispersion), might also limit the applicability of the results of this approach. Nevertheless, the results of the first-order perturbation analysis are sufficiently reliable to indicate appropriate trends.

Most of the stringent assumptions of the first-order perturbation approach may be relaxed by using a numerical approach, which, in general, combines a statistical generation method that produces realizations of the heterogeneous formation properties in sufficient resolution, with an efficient numerical method to solve the partial differential equations that govern flow and transport in heterogeneous, variably saturated formations. However, the complexity and high degree of nonlinearity of such flow, compounded by the serious numerical problems which may be encountered when dealing with steep, spatial-head gradients, and the requirement that grid blocks must be much smaller than the length scale of the heterogeneity, all lead to a numerical problem of extreme difficulty which might demand a formidable computational capacity. This might substantially limit the capability of the numerical approach to solve flow and transport problems on a large, field scale.

List of Technical Nomenclature

$\boldsymbol{\alpha}$	Gardner–Russo parameter (m^{-1})
$\boldsymbol{\Gamma}$	Geometric mean of α (m^{-1})
$\boldsymbol{\delta}$	Kronecker delta (dimensionless)
$\boldsymbol{\eta} = \boldsymbol{\lambda}/I_{yv}$	Scale ratio (dimensionless)
$\boldsymbol{\Upsilon}$	Water saturation (dimensionless)
$\boldsymbol{\lambda}$	Macroscopic capillary length scale (m)
$\boldsymbol{\nu} = \boldsymbol{\sigma}_a^2/\boldsymbol{\sigma}_f^2$	Variance ratio (dimensionless)
$\boldsymbol{\xi}$	Separation distance vector (m)
$\boldsymbol{\xi}'$	Scaled separation distance vector (dimensionless)
$\boldsymbol{\rho} = I_h/I_v$	Length-scale ratio (dimensionless)
$\boldsymbol{\sigma}_p^2$	Variance of p' (units of p'^2)
$\boldsymbol{\tau}$	Dummy variable of integration (s)
$\boldsymbol{v}^e$	Effective solute velocity vector (ms^{-1})
$\boldsymbol{\phi}$	Porosity ($m^3\,m^{-3}$)
$\boldsymbol{\varphi}$	Hydraulic head (m)
$\boldsymbol{\psi}$	Pressure head (m)
$\mathbf{a}$	Coordinate vector of a point in V_0 (m)
a	A perturbation and mean of $\log\alpha$ (dimensionless)
$B = (1/4)\Gamma^2 H$	$\exp[-(1/2)\Gamma H]$ (m^{-1})
c	Resident concentration ($kg\,m^{-3}$)
$\mathbf{C}_{mn}$	(Cross-)covariance between perturbations m and n (units of $m \times n$)
$\hat{\mathbf{C}}_{mn}$	(Cross-)spectra between perturbations m and n (units of $m \times n$)
$\mathbf{C}_{q_{ij}}$	Flux covariance tensor ($m^2\,s^{-2}$)
$\mathbf{C}_{u_{ij}}$	Velocity covariance tensor ($m^2\,s^{-2}$)
$\mathbf{D}$	Macrodispersion tensor ($m^2\,s^{-2}$)
$\mathbf{D}'$	Scaled macrodispersion tensor (dimensionless)
D_{ii}^e	ith component of effective macrodispersion ($m^2\,s^{-1}$)
f, F	Perturbation and mean of $\log K_s$ (dimensionless)
h, H	Perturbation and mean of ψ (m)
$\mathbf{I}_p$	Correlation scale of p'(m)
I_v, I_h	Vertical and horizontal components of $\mathbf{I}_p$(m)
I_{yv}	Vertical correlation scale of $\log K$(m)
$\mathbf{J}$	Mean head-gradient vector (mm^{-1})
j'	The imaginary unit (dimensionless)
K	Unsaturated conductivity (ms^{-1})
K_g	Geometric mean of K (ms^{-1})
K_s	Saturated conductivity (ms^{-1})
$\mathbf{k}$	Wave number vector (m^{-1})
M_i^N	Nth normalized spatial moment in the ith coordinate (m^N)
$m = a, f, h, s$	(Units of a, f, h, or s)
m'	Parameter (eqn [2b]) (dimensionless)

$n = a, f, h, s$	(Units of a, f, h, or s)
$p' = f$ or a	(Dimensionless)
$p^* = f$, a, or fa	(Dimensionless)
$\mathbf{q}$, $\mathbf{Q}$	Perturbation and mean of $\mathbf{q}'$ (ms^{-1})
$\mathbf{q}'$	Water flux vector (ms^{-1})
q'_{ij}	Scaled flux covariance tensor (dimensionless)
s, S	Perturbation and mean of water saturation (dimensionless)
t	Time (s)
t'	Scaled value of t (dimensionless)
$\mathbf{u}$, $\mathbf{U}$	Perturbation and mean of $\mathbf{u}'$ (ms^{-1})
$\mathbf{u}'$	Eulerian velocity vector (ms^{-1})
V_0	Initial volume of the solute body (m^3)
$\mathbf{X}'$	Particle displacement vector (m)
X_{ij}	Particle displacement covariance tensor (m^2)
X_i^e	ith coordinate of the centroid of the solute plume (m)
X_{ii}^e	Moment of inertia of the plume with respect to the x_i-axis (m^2)
$\mathbf{x}$	Spatial coordinate vector (m)
$<\mathbf{X}>$ $\mathbf{X}$	Mean and perturbation of $\mathbf{X}'$ (m)
y, Y	Perturbation and mean of log$K$ (dimensionless)

Further Reading

Beckett PHT and Webster R (1971) Soil variability, a review. *Soils and Fertility* 34: 1–15.

Bras RL and Rodriguez-Iturbe I (1985) *Random Functions and Hydrology*. Reading, MA: Addison-Wesley Professional.

Dagan G (1989) *Flow and Transport in Porous Formations*. Berlin, Germany: Springer-Verlag.

Mantoglou A and Gelhar LW (1987) Stochastic modeling of large-scale transient unsaturated flow systems. *Water Resources Research* 23: 37–46.

Nielsen DR, Biggar JW, and Erh KH (1973) Spatial variability of field-measured soil-water properties. *Hilgardia* 42: 215–260.

Polmann DJ, Mclaughlin D, Luis S, Gelhar LW, and Ababou R (1991) Stochastic modeling of large-scale flow in heterogeneous unsaturated soils. *Water Resources Research* 27: 1447–1458.

Russo D (1998) Stochastic modeling of scale-dependent macrodispersion in the vadose zone. In: Sposito G (ed.) *Scale Dependence and Scale Invariance in Hydrology*, pp. 266–290. Stanford, UK: Cambridge University Press.

Russo D (1998) Stochastic analysis of flow and transport in unsaturated heterogeneous porous formation: effects of variability in water saturation. *Water Resources Research* 34: 569–581.

Russo D, Zaidel J, and Laufer A (1994) Stochastic analysis of solute transport in partially saturated heterogeneous soil. I. Numerical experiments. *Water Resources Research* 30: 769–779.

Russo D, Russo I, and Laufer A (1997) On the spatial variability of parameters of the unsaturated hydraulic conductivity. *Water Resources Research* 33: 947–956.

Russo D, Zaidel J, and Laufer A (1998) Numerical analysis of flow and transport in a three-dimensional partially saturated heterogeneous soil. *Water Resources Research* 34: 1451–1468.

Taylor GI (1921) Diffusion by continuous movements. *Proceedings of the London Mathematics Society* A20: 196–211.

Yeh T-CJ (1998) Scales issues of heterogeneity in vadose-zone hydrology. In: Sposito G (ed.) *Scale Dependence and Scale Invariance in Hydrology*, pp. 224–265. Stanford, UK: Cambridge University Press.

Yeh T-CJ, Gelhar LW, and Gutjahr AL (1985a) Stochastic analysis of unsaturated flow in heterogeneous soils. 1. Statistically isotropic media. *Water Resources Research* 21: 447–456.

Yeh T-CJ, Gelhar LW, and Gutjahr AL (1985b) Stochastic analysis of unsaturated flow in heterogeneous soils. 2. Statistically anisotropic media with variable α. *Water Resources Research* 21: 457–464.

STRESS–STRAIN AND SOIL STRENGTH

S K Upadhyaya, University of California–Davis, Davis, CA, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Concept of Stress

The concept of stress is the most fundamental concept of continuum mechanics and elasticity theory. When a body is subjected to external forces, it experiences deformation (strain) and stress. Consider a rectangular, parallelepiped-shaped volume element, ABCDEFGH, oriented such that its face BCGF is perpendicular to the X_1 axis; the face CDHG is perpendicular to the X_2 axis; EFGH is perpendicular to the X_3 axis; and where $X_1X_2X_3$ represents a right-handed Cartesian coordinate system such that the X_3 axis is vertical (Figure 1). Let ΔF_1, ΔF_2, and ΔF_3 respectively represent forces acting on these surfaces. Then the force per unit area acting on these surfaces (i.e., T_1, T_2, and T_3, respectively) as the size of the surface shrinks to zero can be represented by:

$$T_1 = \Delta F_1/(\Delta_2\Delta_3) \qquad [1a]$$

$$T_2 = \Delta F_2/(\Delta_1\Delta_3) \qquad [1b]$$

$$T_3 = \Delta F_3/(\Delta_1\Delta_2) \qquad [1c]$$

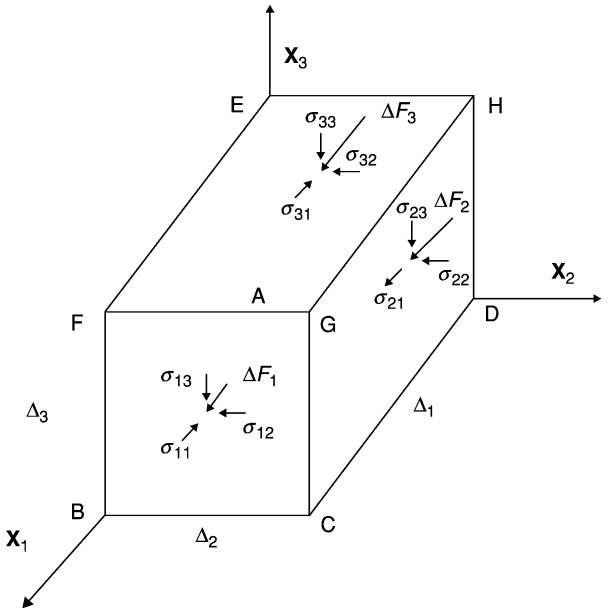


Figure 1 Components of a stress tensor.

in the limit Δ_1, Δ_2, and Δ_3 approach zero. In Eqns [1a]–[1c], Δ_1, Δ_2, Δ_3 are the linear dimensions of the volume element in X_1, X_2, and X_3 directions, respectively.

These three vectors can be decomposed along the coordinate axis as follows:

$$T_1 = \sigma_{11}\mathbf{i}_1 + \sigma_{12}\mathbf{i}_2 + \sigma_{13}\mathbf{i}_3 \qquad [2a]$$

$$T_2 = \sigma_{21}\mathbf{i}_1 + \sigma_{22}\mathbf{i}_2 + \sigma_{23}\mathbf{i}_3 \qquad [2b]$$

$$T_3 = \sigma_{31}\mathbf{i}_1 + \sigma_{32}\mathbf{i}_2 + \sigma_{33}\mathbf{i}_3 \qquad [2c]$$

where $\mathbf{i}_1$, $\mathbf{i}_2$, and $\mathbf{i}_3$ are unit vectors along the coordinate axes X_1, X_2, and X_3, respectively. The components of these stress vectors T_1, T_2, and T_3, form the elements of the stress tensor $[\boldsymbol{\sigma}]$, which is written as:

$$[\boldsymbol{\sigma}] = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} \\ \sigma_{21} & \sigma_{22} & \sigma_{23} \\ \sigma_{31} & \sigma_{32} & \sigma_{33} \end{bmatrix} \qquad [3]$$

The diagonal terms of this stress tensor are known as 'normal stresses' and the off-diagonal terms are known as 'shear stresses.' Figure 1 shows the various components of the stress tensor. As is usual in soil mechanics, compression is considered to be positive. Moreover, the subscripts of stress tensor component $\boldsymbol{\sigma}_{ij}$ are chosen so that the first subscript, 'i,' indicates the direction of the surface normal on which the stress component acts and the second subscript, 'j,' indicates the direction along which the stress component is directed. Thus σ_{13} indicates that this stress component acts on the surface whose normal is along the X_1 axis (i.e., BCGF) and σ_{13} is directed in the X_3 direction. In the absence of body couples, $\sigma_{ij} = \sigma_{ji}$, thus making the stress tensor symmetric. (Body couples, or distributed moments within the body, arise from the action of an electric field on polarized matter or the action of a magnetic field on magnetized particles. Such couples are not a concern in most mechanics problems and are ignored here.)

The first invariant of the stress tensor (the sum of the diagonal terms or the trace of the stress tensor, which does not change under a coordinate transformation) is related to the hydrostatic, spherical, or octahedral normal stress. It is given by:

$$J_1 = \mathrm{tr}(\sigma) = \sigma_{11} + \sigma_{22} + \sigma_{33} = 3\sigma_h = 3\sigma_{oct} \qquad [4]$$

where J_1 is the first invariant of the stress tensor, σ_h is the hydrostatic stress, and σ_{oct} is the mean octahedral normal stress. Mean pressure or bulk stress, p, is given by the mean of the diagonal elements, i.e.:

$$p = \sigma_h = \sigma_{oct} \quad [5]$$

The stress tensor can be decomposed into two parts in the following manner:

$$\begin{bmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} \\ \sigma_{21} & \sigma_{22} & \sigma_{23} \\ \sigma_{31} & \sigma_{32} & \sigma_{33} \end{bmatrix} = \begin{bmatrix} p & 0 & 0 \\ 0 & p & 0 \\ 0 & 0 & p \end{bmatrix} + \begin{bmatrix} \sigma_{11}-p & \sigma_{12} & \sigma_{13} \\ \sigma_{21} & \sigma_{22}-p & \sigma_{23} \\ \sigma_{31} & \sigma_{32} & \sigma_{33}-p \end{bmatrix} \quad [6]$$

The first tensor on the right-hand side is known as the 'hydrostatic' or 'spherical stress tensor,' and the second tensor is known as the 'deviatoric stress tensor' [S]. The second invariant of the stress tensor is related to the octahedral shear stress, τ_{oct}, and is given by:

$$J_{2D} = 1/6[(\sigma_{11}-\sigma_{22})^2 + (\sigma_{22}-\sigma_{33})^2 + (\sigma_{33}-\sigma_{11})^2] + \sigma_{12}^2 + \sigma_{23}^2 + \sigma_{13}^2 \quad [7]$$

where we have utilized the fact that the stress tensor is symmetric. The relation between octahedral shear stress, τ_{oct}, and J_{2D} is given by:

$$\tau_{oct} = [2/3\, J_{2D}]^{1/2} \quad [8]$$

This second invariant of the deviatoric stress tensor or the related quantity τ_{oct} plays an important role in describing the yield behavior in soil mechanics. In some cases the third invariant of the deviatoric stress tensor is also used to represent the yield behavior, such as in the extended Drucker–Prager yield criteria.

Principal Stress

For the symmetric version of the stress tensor $[\boldsymbol{\sigma}]$ given in Eqn [3], which is associated with the coordinate system shown in Figure 1, it is always possible to locate a set of mutually orthogonal planes along which all shear stress components will be zero, leaving only the normal components of the stress tensor. These normal components of stress tensor are known as principal stresses, the planes are called principal planes, and the normal vectors associated with these planes are known as principal axes or directions. If the principal axes are taken as the reference axes, then the stress tensor given in Eqn [3] becomes:

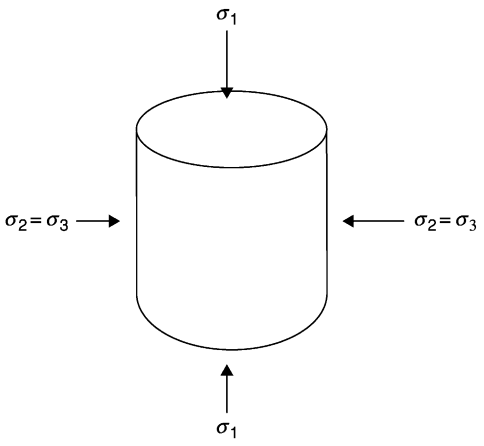


Figure 2 Cylindrical state of stress in soil mass.

$$[\boldsymbol{\sigma}] = \begin{bmatrix} \sigma_1 & 0 & 0 \\ 0 & \sigma_2 & 0 \\ 0 & 0 & \sigma_3 \end{bmatrix} \quad [9]$$

where σ_1, σ_2, and σ_3 are known as 'principal stresses.' The largest (in magnitude) of the principal stresses is known as the 'major principal stress,' the smallest one is known as the 'minor principal stress,' and the other one is known as the 'intermediate principal stress.' In terms of principal stresses, the second invariant of stress tensor and τ_{oct} become:

$$J_{2D} = 1/6[(\sigma_1-\sigma_2)^2 + (\sigma_2-\sigma_3)^2 + (\sigma_3-\sigma_1)^2] \quad [10]$$

For example, consider a cylindrical body of soil under a triaxial loading situation, as shown in Figure 2. For this case, σ_1, and $\sigma_2 = \sigma_3$ are the principal stresses. This example will be used throughout this discussion. The mean hydrostatic stress and the octahedral shear stress are given by:

$$p = (\sigma_1 + 2\sigma_3)/3 \quad [11]$$

$$\tau_{oct} = \left(\frac{2}{3}J_{2D}\right)^{1/2} = \frac{\sqrt{2}}{3}(\sigma_1 - \sigma_3) \quad [12]$$

In soil plasticity theory, often the deviator stress $q = (\sigma_1 - \sigma_3) = (3J_{2D})^{1/2}$ is used to describe shear behavior, and mean hydrostatic pressure, p, is used to describe volumetric compression.

Effective Stress

In soil mechanics usually effective stress rather than the total stress is used in constitutive relationships. For saturated soils, effective hydrostatic pressure, p_{eff}, is defined as the difference between the total

hydrostatic stress, p_t, minus the pore-water pressure, u_w (i.e., $p = p_{eff} = p_t - u_w$). For unsaturated soils the effective pressure is a function of soil suction or soil-moisture tension. The effective stress in unsaturated soils may be represented by:

$$p_{eff} = p_t - u_a + \chi(u_a - u_w) \qquad [13]$$

where u_a is pore air pressure and χ depends on water content. Often it is taken as a fraction of unit cross-sectional area of soil occupied by water. However, this is problematic from a continuum mechanics and thermodynamic point of view. There is a need to treat unsaturated soil as a four-phase medium consisting of air, soil particles, water, and 'contractile skin' which separates air from water. Instead of using a single effective stress as given in Eqn [13] to define the stress state of soil, it is possible to use two different stress-state variables: net stress $(p_t - u_a)$ and matric suction $(u_a - u_w)$. The additional stress-state variable, matrix suction, is related to soil-moisture content through the soil-water release characteristic, which is regarded as an important constitutive function.

A total stress approach has been used to determine critical state parameters for unsaturated, agricultural soils. This method is acceptable as long as it is recognized that the material parameters are treated as a function of soil-moisture content. Unsaturated agricultural soils can be modeled in terms of total stress as long as the effect of unsaturated conditions is duly recognized. Thus the effective hydrostatic pressure, p_{eff} (i.e., $p = p_{eff}$), is used for saturated soil, and total pressure, p_t (i.e., $p = p_t$), is used for unsaturated soil.

Concept of Strain

When a body is subjected to external forces, it experiences deformation. The ratios of these deformations to the respective original dimensions are known as 'strain.' Thus, if a slender rod of length L is subjected to a uniaxial compressive load, its length will decrease, leading to a compressive strain of $\epsilon_a = u/L$, where u is the decrease in length (Figure 3). This is known as the 'Cauchy strain.'

This concept can be easily extended to three dimensions. Just as in the case of stress, it is possible to represent the strain experienced by any body under the generalized external forces in terms of a strain tensor, $[\boldsymbol{\epsilon}]$ as follows:

$$[\boldsymbol{\epsilon}] = \begin{bmatrix} \epsilon_{11} & \epsilon_{12} & \epsilon_{13} \\ \epsilon_{21} & \epsilon_{22} & \epsilon_{23} \\ \epsilon_{31} & \epsilon_{32} & \epsilon_{33} \end{bmatrix} \qquad [14]$$

where each ϵ_{ij} is known as a normal strain if $i = j$ and known as a shear strain if $i \neq j$. If there are no body

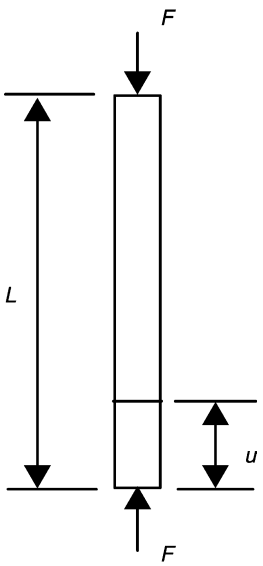


Figure 3 A slender rod under compressive axial strain.

moments, then $\epsilon_{ij} = \epsilon_{ji}$. This gives six strain components for a three-dimensional case. The subscripts 'i' and 'j' have the same implications as they did in the case of the stress tensor described above. If strain levels are small, these strain components can be expressed as:

$$\epsilon_{11} = \frac{\partial u_1}{\partial x_1}; \quad \epsilon_{22} = \frac{\partial u_2}{\partial x_2}; \quad \epsilon_{33} = \frac{\partial u_3}{\partial x_3}$$
$$\epsilon_{12} = \frac{1}{2}\left(\frac{\partial u_1}{\partial x_2} + \frac{\partial u_2}{\partial x_1}\right); \quad \epsilon_{13} = \frac{1}{2}\left(\frac{\partial u_1}{\partial x_3} + \frac{\partial u_3}{\partial x_1}\right);$$
$$\epsilon_{23} = \frac{1}{2}\left(\frac{\partial u_3}{\partial x_2} + \frac{\partial u_2}{\partial x_3}\right) \qquad [15]$$

where u_1, u_2, and u_3 and deformations of the body along three mutually orthogonal, rectangular coordinate system axes such as those in Figure 1. It is a common practice in engineering to use engineering shear strains, γ_{ij}, which are related to ϵ_{ij} as follows:

$$\gamma_{ij} = 2\epsilon_{ij} \qquad [16]$$

The first invariant of the strain tensor, $I_1 = \text{tr}\ (\epsilon) = \epsilon_1 + \epsilon_2 + \epsilon_3 = \epsilon_v$, is known as the 'volumetric strain.' Just as in the case of stress tensor, it is possible to decompose the strain tensor into two parts as follows:

$$\begin{bmatrix} \epsilon_{11} & \epsilon_{12} & \epsilon_{13} \\ \epsilon_{21} & \epsilon_{22} & \epsilon_{23} \\ \epsilon_{31} & \epsilon_{32} & \epsilon_{33} \end{bmatrix} = \begin{bmatrix} \epsilon_v/3 & 0 & 0 \\ 0 & \epsilon_v/3 & 0 \\ 0 & 0 & \epsilon_v/3 \end{bmatrix} + \begin{bmatrix} \epsilon_{11} - \epsilon_v/3 & \epsilon_{12} & \epsilon_{13} \\ \epsilon_{21} & \epsilon_{22} - \epsilon_v/3 & \epsilon_{23} \\ \epsilon_{31} & \epsilon_{32} & \epsilon_{33} - \epsilon_v/3 \end{bmatrix} \qquad [17]$$

The first tensor on the right-hand side is the volumetric strain tensor and the second one is the deviatoric strain tensor. The second invariant of the strain tensor, I_{2D}, is of particular interest and is given by:

$$I_{2D} = 1/6[(\epsilon_{11} - \epsilon_{22})^2 + (\epsilon_{22} - \epsilon_{33})^2 + (\epsilon_{33} - \epsilon_{11})^2] + \epsilon_{12}^2 + \epsilon_{23}^2 + \epsilon_{13}^2 \quad [18]$$

In engineering practice, a closely related term, 'octahedral shear strain,' is used and is given by:

$$\gamma_{oct} = \left[\frac{2}{3} I_{2D}\right]^{\frac{1}{2}} \quad [19]$$

Principal Strain

For the symmetric version of the strain tensor given in Eqn [14], which is represented in the $X_1X_2X_3$ coordinate system, it is always possible to find a set of three mutually perpendicular planes along which all shear-strain components will be zero. The directions normal to these planes are known as principal strain axes. In terms of the principal strains, the strain tensor reduces to:

$$[\boldsymbol{\epsilon}] = \begin{bmatrix} \epsilon_1 & 0 & 0 \\ 0 & \epsilon_2 & 0 \\ 0 & 0 & \epsilon_3 \end{bmatrix} \quad [20]$$

where ϵ_1, ϵ_2, and ϵ_3 are principal strains. In terms of these principal strains, the second invariant of the deviatoric strain reduces to:

$$I_{2D} = 1/6[(\epsilon_1 - \epsilon_2)^2 + (\epsilon_2 - \epsilon_3)^2 + (\epsilon_3 - \epsilon_1)^2] \quad [21]$$

For the example problem, shown in Figure 2, $\epsilon_2 = \epsilon_3$. Under these conditions, the volumetric strain becomes:

$$\epsilon_v = \epsilon_1 + 2\epsilon_3 \quad [22]$$

and the octahedral shear strain becomes:

$$\gamma_{oct} = \left[\frac{2}{3} I_{2D}\right]^{\frac{1}{2}} = \frac{\sqrt{2}}{3}(\epsilon_1 - \epsilon_3) \quad [23]$$

In soil plasticity, ϵ_v and ϵ_s (the latter of which is given by $2/3\ (\epsilon_1 - \epsilon_3) = \sqrt{2}\gamma_{oct}$) are often used to represent soil behavior.

Void Ratio and Volumetric Strain

Often volumetric strain is represented in terms of changes in void ratio, e, which is defined as the ratio of total pore space to the total volume of individual solid particles, i.e.:

$$e = \frac{V_0}{V_s} \quad [24a]$$

where V_0 is the total volume of voids and V_s is the total volume of solid particles. The volumetric strain ϵ_v is related to the changes in the void ratio by the following equation:

$$d\epsilon_v = -\frac{de}{1 + e_i} \quad [24b]$$

where e_i is the initial void ratio. Note that the negative sign in Eqn [24b] is owing to the sign convention that treats compressive strain as positive.

Constitutive Laws

The constitutive law has been defined as a general functional relationship between stress, strain, and their rates, i.e.:

$$f(\sigma, \dot{\sigma}, \epsilon, \dot{\epsilon}) = 0 \quad [25]$$

where $\dot{\sigma}$ and $\dot{\epsilon}$ represent stress and strain rates, respectively. For static equilibrium conditions, the most general linear relationship between stress and strain tensors takes the following form:

$$[\boldsymbol{\sigma}] = [\mathrm{C}][\boldsymbol{\epsilon}] \quad [26]$$

where [C] is a fourth-order tensor, known as 'material property tensor,' since it relates two second-order tensors. For homogeneous, elastic material, the material property tensor consists of 81 elements. However, symmetry of stress and strain tensors, and strain energy considerations reduce the number of independent constants required to 21. If this discussion is limited to isotropic elastic materials, only two independent material constants are required to relate stress and strain. The Young modulus, E, and Poisson ratio, ν, are usually selected as material properties to represent elastic isotropic behavior. With reference to Figure 4a, which shows a rod under uniaxial loading, the following are definitions for the Young modulus and Poisson ratio:

$$E = \frac{\sigma_1}{\epsilon_1} \quad [27a]$$

$$\nu = -\frac{\epsilon_3}{\epsilon_1} \quad [27b]$$

Now consider a component under a general state of stress. If we consider the principal stress and strain as shown in Figure 4b, then the strain component ϵ_1 is caused by the direct stress, σ_1, as well as two transverse stress σ_2 and σ_3 due to the Poisson effect. Similar arguments hold for strain components ϵ_2 and ϵ_3.

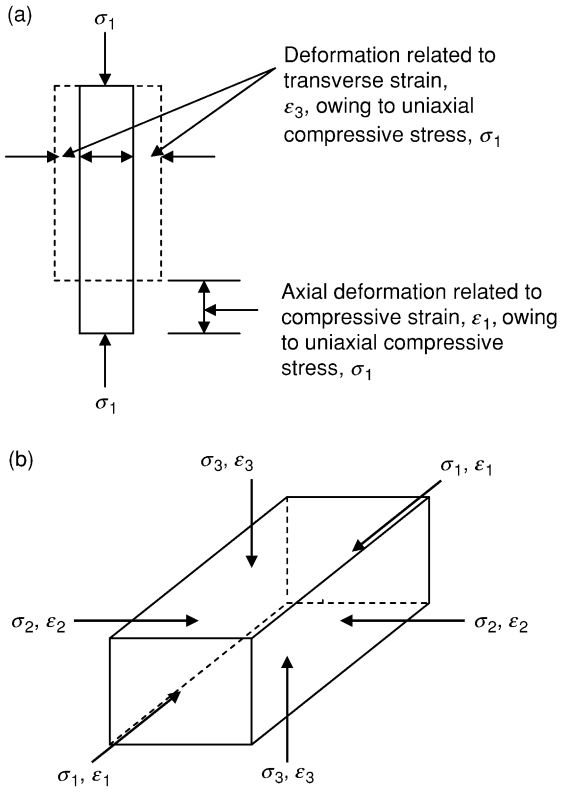


Figure 4 Behavior of (a) a rod under uniaxial loading and (b) a prismatic member under general loading.

Therefore:

$$\epsilon_1 = \frac{1}{E}\sigma_1 - \frac{\nu}{E}\sigma_2 - \frac{\nu}{E}\sigma_3 \qquad [28a]$$

$$\epsilon_2 = \frac{1}{E}\sigma_2 - \frac{\nu}{E}\sigma_3 - \frac{\nu}{E}\sigma_1 \qquad [28b]$$

$$\epsilon_3 = \frac{1}{E}\sigma_3 - \frac{\nu}{E}\sigma_1 - \frac{\nu}{E}\sigma_2 \qquad [28c]$$

Moreover, summing Eqns [28a]–[28c]:

$$\epsilon_1 + \epsilon_2 + \epsilon_3 = \epsilon_v = \frac{(1-2\nu)}{E}(\sigma_1 + \sigma_2 + \sigma_3) = \frac{3(1-2\nu)}{E}p \qquad [29]$$

where $\epsilon_v = \epsilon_1 + \epsilon_2 + \epsilon_3$ is the volumetric strain and $p = (\sigma_1 + \sigma_2 + \sigma_3)/3$ has been substituted (see Eqns [4] and [5]). Alternately Eqn [29] can be written as:

$$p = \frac{E}{3(1-2\nu)}\epsilon_v = K\epsilon_v \qquad [30]$$

where $K = E/3(1-2\nu)$ is known as the 'elastic bulk modulus.'

It is always true for an axisymmetric case such as cylindrical loading that $\sigma_2 = \sigma_3$. Subtracting Eqn [28c] from [28a]:

$$\epsilon_1 - \epsilon_3 = \frac{1+\nu}{E}(\sigma_1 - \sigma_3) \qquad [31a]$$

As mentioned earlier, $\epsilon_1 - \epsilon_3 = 2/3\ \epsilon_s$ and $\sigma_1 - \sigma_3 = q$. Therefore, Eqn [31a] becomes:

$$q = 3\left\{\frac{E}{2(1+\nu)}\right\}\epsilon_s = 3G \cdot \epsilon_s \qquad [31b]$$

where $G = E/2(1+\nu)$ is the shear modulus.

In soil mechanics, elastic bulk modulus, K, and shear modulus, G, are usually chosen as material parameters. If the symmetric properties of stress and strain tensors are utilized, Eqn [26] can be rewritten in terms of material parameters, K and G, as:

$$\begin{Bmatrix}\sigma_{11}\\ \sigma_{22}\\ \sigma_{33}\\ \sigma_{12}\\ \sigma_{23}\\ \sigma_{13}\end{Bmatrix} = \begin{bmatrix} K+\frac{4G}{3} & K-\frac{2G}{3} & K-\frac{2G}{3} & 0 & 0 & 0\\ K-\frac{2G}{3} & K+\frac{4G}{3} & K-\frac{2G}{3} & 0 & 0 & 0\\ K-\frac{2G}{3} & K-\frac{2G}{3} & K+\frac{4G}{3} & 0 & 0 & 0\\ 0 & 0 & 0 & 2G & 0 & 0\\ 0 & 0 & 0 & 0 & 2G & 0\\ 0 & 0 & 0 & 0 & 0 & 2G \end{bmatrix} \begin{Bmatrix}\epsilon_{11}\\ \epsilon_{22}\\ \epsilon_{33}\\ \epsilon_{12}\\ \epsilon_{23}\\ \epsilon_{13}\end{Bmatrix} \qquad [32a]$$

Sometimes it is useful to represent strain in terms of stress (i.e., inverse of Eqn [32a]). This inverse relationship is given by Eqn [32b]:

$$\begin{Bmatrix}\epsilon_{11}\\ \epsilon_{22}\\ \epsilon_{33}\\ \epsilon_{12}\\ \epsilon_{23}\\ \epsilon_{13}\end{Bmatrix} = \begin{bmatrix} \frac{1}{3G}+\frac{1}{9K} & \frac{1}{9K}-\frac{1}{6G} & \frac{1}{9K}-\frac{1}{6G} & 0 & 0 & 0\\ \frac{1}{9K}-\frac{1}{6G} & \frac{1}{3G}+\frac{1}{9K} & \frac{1}{9K}-\frac{1}{6G} & 0 & 0 & 0\\ \frac{1}{9K}-\frac{1}{6G} & \frac{1}{9K}-\frac{1}{6G} & \frac{1}{3G}+\frac{1}{9K} & 0 & 0 & 0\\ 0 & 0 & 0 & \frac{1}{2G} & 0 & 0\\ 0 & 0 & 0 & 0 & \frac{1}{2G} & 0\\ 0 & 0 & 0 & 0 & 0 & \frac{1}{2G} \end{bmatrix} \begin{Bmatrix}\sigma_{11}\\ \sigma_{22}\\ \sigma_{33}\\ \sigma_{12}\\ \sigma_{23}\\ \sigma_{13}\end{Bmatrix} \qquad [32b]$$

For the example problem of cylindrical state of stress, Eqns [30] and [31b] apply. These equations can be written in a matrix form as follows:

$$\begin{Bmatrix} p \\ q \end{Bmatrix} = \begin{bmatrix} K & 0 \\ 0 & 3G \end{bmatrix} \begin{Bmatrix} \epsilon_v \\ \epsilon_s \end{Bmatrix} \quad [33a]$$

Alternatively:

$$\begin{Bmatrix} \epsilon_v \\ \epsilon_s \end{Bmatrix} = \begin{bmatrix} \frac{1}{K} & 0 \\ 0 & \frac{1}{3G} \end{bmatrix} \begin{Bmatrix} p \\ q \end{Bmatrix} \quad [33b]$$

From Eqns [33a] and [33b], it is clear that volumetric and shear behavior are completely decoupled.

Material Behavior under Load

Response of a material to external loading can be linear-elastic, nonlinear-elastic, rigid-perfectly plastic, elastic-perfectly plastic, or elastoplastic. Figure 5 shows these responses.

Soil rarely exhibits a linear-elastic behavior (Figure 5a). A nonlinear-elastic constitutive relationship (Figure 5b) can be used to describe soil loading behavior along a particular loading or unloading path with reasonable accuracy. However, it is not a good indicator of soil behavior under wide varieties of loading and unloading conditions (i.e., general loading–unloading conditions). Rigid-perfectly plastic (Figure 5c) and elastic-perfectly plastic (Figure 5d) models are sometimes used in soil analysis, particularly in limit equilibrium analysis. Figure 5e represents a typical metal bar under uniaxial tension. Often, soil exhibits similar behavior under isotropic compression as well as triaxial compression. In the case of metals, when a specimen is loaded in a uniaxial direction, it usually exhibits a linear elastic behavior until the yield point (point A) is reached on its stress–strain curve. After yield it will further deform if loaded to a point B. If the specimen is now unloaded, it will usually follow an unloading path such as BCD. When the load is completely removed, it will not have recovered its original length but will have a permanent deformation or plastic deformation. If this specimen is reloaded, it may follow a path such as DEB, during which the material usually exhibits elastic behavior. At point B it will yield again and follow path BF. Unloading–reloading curves are generally close to each other, and the area under the curve BCDE represents hysterisis loss. Since point B corresponds to a higher stress level than point A, this type of material is often called a 'work-hardening' material. Note that beyond point B, along path BF, both elastic and plastic deformations will occur. If the specimen is unloaded and reloaded at point F it may trace an unloading–reloading curve FGHI similar to the curve BCDE. As mentioned earlier, soil exhibits similar behavior (overconsolidated soil may show a distinct hump) except that it seldom shows linear elastic behavior even when loaded from a stress-free, undeformed state. During unloading–reloading conditions, soil usually exhibits logarithmic (nonlinear) elastic behavior until the yield point is reached.

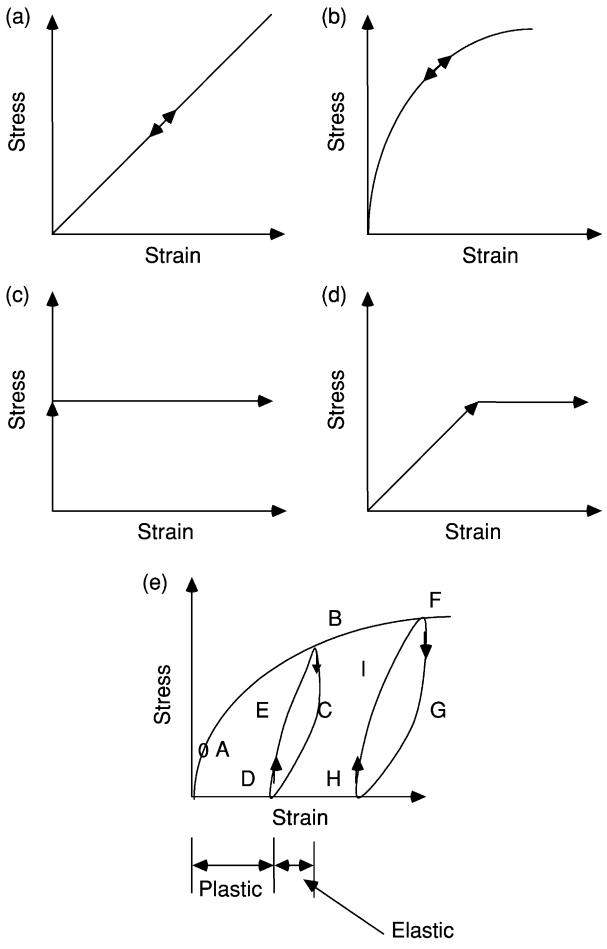


Figure 5 Idealized stress–strain curves for materials: (a) linear-elastic; (b) nonlinear elastic; (c) rigid-perfectly plastic; (d) elastic-perfectly plastic; (e) and elastoplastic behavior.

Nonlinear Elastic Behavior

In principle, it is possible to model the stress–strain behavior of soil using nonlinear elastic models for specific loading paths and conditions. Nonlinearity implies that the elastic parameters are no longer constants but depend on the states of stress and strain of the soil. For these conditions usually an incremental analysis is conducted using tangent moduli values: tangent bulk modulus, K_t, and tangent shear modulus, G_t. Figure 6 shows graphically the significance of tangent moduli. For the case of nonlinear elastic constitutive models, Eqns [32a] and [b] can be written in an incremental form to account for nonlinear behavior as follows:

$$\begin{Bmatrix} d\sigma_{11} \\ d\sigma_{22} \\ d\sigma_{33} \\ d\sigma_{12} \\ d\sigma_{23} \\ d\sigma_{13} \end{Bmatrix} = \begin{bmatrix} K_t + \frac{4G_t}{3} & K_t - \frac{2G_t}{3} & K_t - \frac{2G_t}{3} & 0 & 0 & 0 \\ K_t - \frac{2G_t}{3} & K_t + \frac{4G_t}{3} & K_t + \frac{2G_t}{3} & 0 & 0 & 0 \\ K_t - \frac{2G_t}{3} & K_t - \frac{2G_t}{3} & K_t + \frac{4G_t}{3} & 0 & 0 & 0 \\ 0 & 0 & 0 & 2G_t & 0 & 0 \\ 0 & 0 & 0 & 0 & 2G_t & 0 \\ 0 & 0 & 0 & 0 & 0 & 2G_t \end{bmatrix} \begin{Bmatrix} d\epsilon_{11} \\ d\epsilon_{22} \\ d\epsilon_{33} \\ d\epsilon_{12} \\ d\epsilon_{23} \\ d\epsilon_{13} \end{Bmatrix} \quad [34]$$

and the inverse relationship is given by:

$$\begin{Bmatrix} d\epsilon_{11} \\ d\epsilon_{22} \\ d\epsilon_{33} \\ d\epsilon_{12} \\ d\epsilon_{23} \\ d\epsilon_{13} \end{Bmatrix} = \begin{bmatrix} \frac{1}{3G_t} + \frac{1}{9K_t} & \frac{1}{9K_t} - \frac{1}{6G_t} & \frac{1}{9K_t} - \frac{1}{6G_t} & 0 & 0 & 0 \\ \frac{1}{9K_t} - \frac{1}{6G_t} & \frac{1}{3G_t} + \frac{1}{9K_t} & \frac{1}{9K_t} - \frac{1}{6G_t} & 0 & 0 & 0 \\ \frac{1}{9K_t} - \frac{1}{6G_t} & \frac{1}{9K_t} - \frac{1}{6G_t} & \frac{1}{3G_t} + \frac{1}{9K_t} & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{2G_t} & 0 & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{2G_t} & 0 \\ 0 & 0 & 0 & 0 & 0 & \frac{1}{2G_t} \end{bmatrix} \begin{Bmatrix} d\sigma_{11} \\ d\sigma_{22} \\ d\sigma_{33} \\ d\sigma_{12} \\ d\sigma_{23} \\ d\sigma_{13} \end{Bmatrix} \quad [35]$$

For the triaxial, cylindrical-loading example problem, Eqn [35] takes the following form (cf. Eqn [33b]):

$$\begin{Bmatrix} d\epsilon_v \\ d\epsilon_s \end{Bmatrix} = \begin{bmatrix} 1/K_t & 0 \\ 0 & 1/(3G_t) \end{bmatrix} \begin{Bmatrix} dp \\ dq \end{Bmatrix} \quad [36]$$

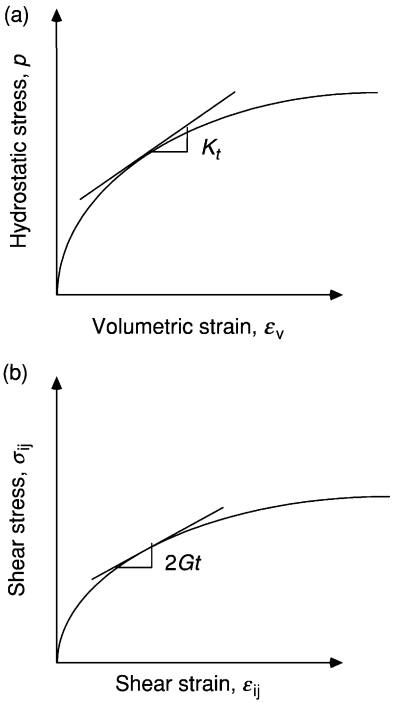


Figure 6 Stress–strain relationship for nonlinear elastic, isotropic material: (a) hydrostatic stress versus volumetric strain; (b) shear stress versus shear strain.

Variable-Moduli Models

Since both K_t and G_t are expected to be a function of stress and strain state of soil, these two tangent moduli can be represented as some explicit function of stress and strain, or stress and strain invariants. These are known as 'variable-moduli models' and many have been proposed for soil. One such model is:

$$K_t = K_0 + K_1 \cdot p \quad [37]$$

and

$$G_t = G_0 + G_1\sqrt{J_{2D}} \quad [38]$$

where K_0 and G_0 are the initial bulk and shear moduli, respectively, and K_1 and G_1 are stress-related material parameters. Here the tangent bulk modulus is assumed to depend on the mean hydrostatic stress and the tangent shear modulus is assumed to depend on the second invariant of the deviatoric stress tensor.

For the example problem using Eqns [36], [37], and [38], the following differential equations are obtained:

$$d\epsilon_v = \frac{dp}{(K_0 + K_1 p)} \quad [39]$$

$$d\epsilon_s = \frac{dq}{3(G_0 + G_1' q)} \quad [40]$$

where $G_1' = G_1/\sqrt{3}$, since $q = \sqrt{3J_{2D}}$.

Upon integration, Eqns [39] and [40] yield:

$$\epsilon_v = \frac{1}{K_1} \ln\left(\frac{K_0 + K_1 p}{K_0}\right) \quad [41]$$

$$\epsilon_s = \frac{1}{3G_1'} \ln\left(\frac{G_0 + G_1' q}{G_0}\right) \quad [42]$$

Under certain circumstances, this model allows the tangent shear modulus to become negative; but, in reality, material would have failed before this happens. From Eqn [42] it is clear that $\epsilon_s \to -\infty$, as $(G_0 + G_1' q) \to 0$. A modification of Eqn [39] is often used to describe the nonlinear elastic behavior of soil before it yields. If K_0 is zero, integration of Eqn [39] gives:

$$\epsilon_v = \frac{1}{K_1} \ln(p) + C$$

or:

$$e = e_i - \frac{(1 + e_i)}{K_1} \ln(p) = e_i - \kappa \ln(p) \quad [43]$$

where e_i is the void ratio at an initial pressure of 1 kPa, κ is the logarithmic bulk modulus which is related to K_1 and e_i (i.e., $\kappa = (1 + e_i)/K_1$). Eqn [43] is commonly used to describe the elastic behavior of soil. For normally consolidated soil (i.e., path A in Figure 7), isotropic compression is also described by a similar equation of the type:

$$e = \Gamma - \lambda \ln(p) \quad [44]$$

where Γ and λ are soil parameters. Eqns [43] and [44] are graphically represented in Figure 7.

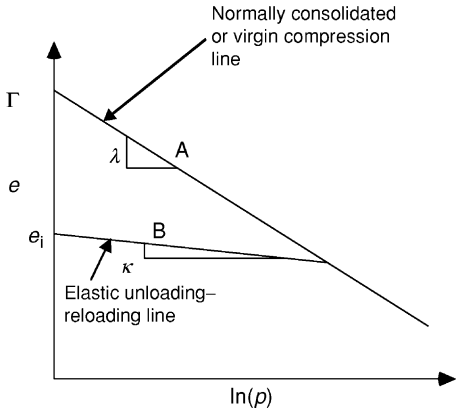


Figure 7 Isotropic consolidation for normally consolidated soil (curve A) and elastic unloading and reloading line (curve B).

The three soil parameters κ, λ, and Γ are used in describing soil behavior in the Cam-clay critical-state model (see the section Critical-State Soil Mechanics, below). Note that both κ and λ are associated with variable-moduli models that represent nonlinear behavior of soil in isotropic compression of normally consolidated soils and unloading–reloading characteristics of soils.

Duncan and Chang model

A well-known, hyperbolic-type stress–strain relationship has been proposed by Duncan and Chang. This model uses the tangent Young modulus and Poisson ratio. It assumes that the stress difference q is a hyperbolic function of axial strain. Figure 8 shows a typical stress–strain curve for both sandy and clayey soils.

This stress–strain curve can be represented by:

$$q = \sigma_1 - \sigma_3 = \frac{\epsilon}{a + b\epsilon} \quad [45]$$

where a is related to the initial tangent modulus ($E_i = 1/a$) and b is related to the asymptotic value of stress difference, $(\sigma_1 - \sigma_3)_{ult}$ ($(\sigma_1 - \sigma_3)_{ult} = 1/b$). It was found that compressive strength of soil at failure was always slightly smaller than this asymptotic stress value. The asymptotic value can be related to compressive failure stress by a factor, R_f, i.e.:

$$(\sigma_1 - \sigma_3)_f = R_f (\sigma_1 - \sigma_3)_{ult} \quad [46]$$

Combining Eqn [45] with [46]:

$$(\sigma_1 - \sigma_3) = \frac{\epsilon}{\left[\frac{1}{E_i} + \frac{\epsilon R_f}{(\sigma_1 - \sigma_3)_f}\right]} \quad [47]$$

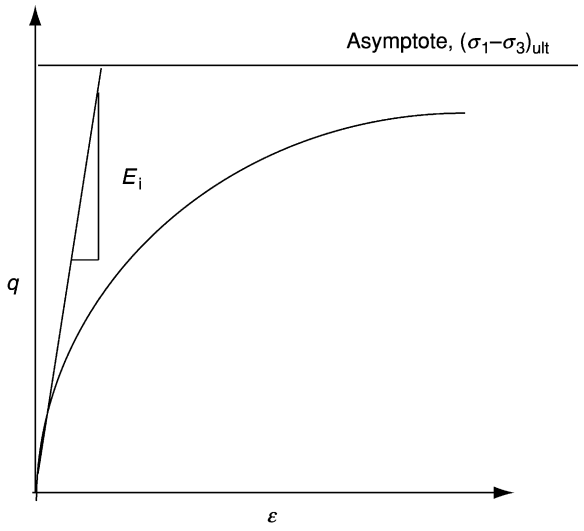


Figure 8 Variation of stress difference or deviator stress q as a function of axial strain ϵ.

The response of many soils depends on the confining pressure for a given stress path. Duncan and Chang used a relationship proposed by Janbu to represent initial tangent modulus as a function of confining stress as follows:

$$E_i = K' p_0 (\sigma_3 / p_0)^n \quad [48]$$

where K' and n are empirical constants, p_0 is the atmospheric pressure, and σ_3 is the minor principal stress. Moreover, the compressive pressure at failure in terms of the Mohr–Coulomb failure criterion is: funny initial moe

$$(\sigma_1 - \sigma_3)_f = \frac{2c\ \cos(\phi) + 2\sigma_3\ \sin(\phi)}{1 - \sin(\phi)} \quad [49]$$

where c is cohesion and ϕ is soil internal angle of friction. The tangent Young modulus can be obtained by differentiating Eqn [47] with respect to axial strain, i.e.:

$$E_t = \frac{\partial(\sigma_1 - \sigma_3)}{\partial \epsilon} \quad [50]$$

Eqns [47]–[50] show that:

$$E_t = (1 - R_f S)^2 E_i \quad [51]$$

where S is the fraction of mobilized stress strength and is given by:

$$S = \frac{(\sigma_1 - \sigma_3)}{(\sigma_1 - \sigma_3)_f} \quad [52]$$

or

$$E_t = \left[1 - \frac{R_f(1 - \sin(\phi))(\sigma_1 - \sigma_3)}{2c\ \cos(\phi) + 2\sigma_3\ \sin(\phi)}\right]^2 K' p_0 \left(\frac{\sigma_3}{p_0}\right)^n \quad [53]$$

This expression involves five parameters. It is relatively easy to implement such models into numerical analysis techniques such as the finite element technique (FEM), and a constant value of the Poisson ratio can be used. It should be noted that although these models can be readily incorporated into numerical analysis, the path dependency of the parameters used limits their widespread use. Since plastic strain dominates soil behavior following yield (point A in Figure 5), soil behavior can be better represented by elastoplastic models.

Elastoplastic Behavior of Soil

As mentioned previously, once the material yields (point A in Figure 5), there will be some irrecoverable (plastic) deformation. There are two important aspects of plastic behavior of materials: (1) a yield criterion, and (2) postyield behavior.

Yield Criteria

When the material arrives at a certain state of stress under the action of an external load, it may permanently lose the ability to regain its original dimensions for any further increase in that load. The relationship between various stress components at this limiting situation (transition from elastic to plastic region) is expressed in terms of a scalar function, f, known as the yield criterion, i.e.:

$$f = f(\sigma_{11}, \sigma_{22}, \sigma_{33}, \sigma_{12}, \sigma_{13}, \sigma_{23}) \quad [54]$$

For isotropic materials, this criterion can be expressed in terms of principal stresses or stress invariants as follows:

$$f = f(J_1, J_2, J_3) \quad [55]$$

Hydrostatic stress seldom plays a role in metal failure, therefore the yield criterion given in Eqn [54] is usually expressed only in terms of stress invariants of the deviatoric stress tensor as:

$$f = f(J_{2D}, J_{3D}) \quad [56]$$

where J_{3D} is the third invariant of the deviatoric stress tensor. J_{3D} is not dealt with here, since most of the widely used yield criteria are represented in terms of J_{2D} only. The well-known von Mises yield criterion depends on J_{2D} as follows:

$$J_{2D} = \frac{\sigma_y^2}{3} \quad [57]$$

where σ_y is the yield stress under uniaxial load. Another widely used yield criterion is the Tresca yield criterion, which is given by:

$$\sigma_1 - \sigma_3 = \sigma_y \quad [58]$$

Although, for the case of the cylindrical triaxial loading (i.e., the example problem), both von Mises and Tresca yield criteria reduce to the same equation, since $J_{2D} = (\sigma_1 - \sigma_3)^2/3$ (cf. Eqn [10]), in general these two yield criteria differ from each other.

Unlike for metals, the yield criterion for geologic materials depends on the first invariant of the stress tensor or the hydrostatic pressure. This is because these types of materials exhibit internal friction, which results in frictional forces that increase with normal load. The well-known Mohr–Coulomb failure criterion is a yield function that accounts for internal friction within the soil mass. This yield criterion is commonly expressed as:

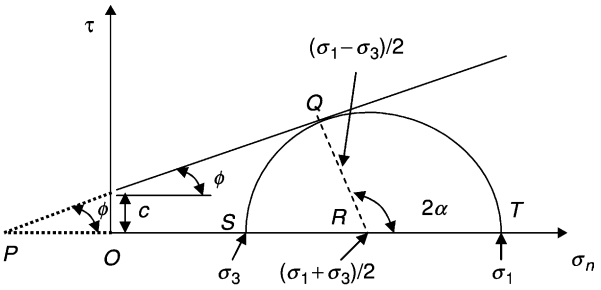


Figure 9 Mohr–Coulomb failure criterion for geologic materials. Point O is the origin of the coordinate axes, *R* is the center of the Mohr circle, *S* and *T* are intersection points of the Mohr circle with the normal stress axis, *Q* is the point of tangency between the Mohr circle and Mohr–Coulomb failure line, *P* is the point of intersection of the Mohr–Coulomb failure line with the normal stress axis.

$$\tau_{\max} = c + \sigma_{\mathrm{n}} \tan(\phi) \qquad [59]$$

where $\tau_{\max}$ is the maximum shear stress; σ_{n} is the normal stress; c is cohesion; and ϕ is the soil internal angle of friction. Figure 9 is a graphical representation of the Mohr–Coulomb failure criterion superimposed on a typical Mohr circle. Eqn [59] can also be written in terms of principal stress as:

$$\sigma_1 = \sigma_3 \tan^2\alpha + 2c \tan\alpha \qquad [60]$$

where α is $45^\circ + \phi/2$. It should be noted that, for frictionless materials (i.e., $\phi = 0$), the Mohr–Coulomb failure criterion reduces to the Tresca yield criterion. The Mohr–Coulomb yield criterion is difficult to implement in three-dimensional applications, because the yield surface is an irregular hexagonal pyramid with sharp corners (mathematical singularities). A generalization, known as the 'Drucker–Prager yield criterion,' that results in a yield surface in the form of a smooth right circular cone is commonly used in modeling granular materials. This yield criterion is expressed in terms of stress invariants as follows:

$$\sqrt{J_{2\mathrm{D}}} - \alpha J_1 - k = 0 \qquad [61]$$

where α and k are related to internal angle of friction and cohesion. Figure 10 shows a plot of the Drucker–Prager yield criterion in the $\sqrt{J_{2\mathrm{D}}}$ versus J_1 plane. For purely cohesive soils, α is zero and the Drucker–Prager yield criterion reduces to von Mises yield criterion.

Hardening Cap

Many geologic materials experience almost continuous plastic deformation from the very beginning of loading. Thus if a mass of soil previously loaded to point B in Figure 10 is unloaded along path BA and then loaded even hydrostatically (i.e., $\sigma_1 = \sigma_2 = \sigma_3 = p$) along the J_1 axis, it will deform elastically until it is loaded to point C, and any further loading leads to plastic deformation. The continuous curved surface, which is assumed to join the Drucker–Prager fixed-yield surface smoothly at point C, is called the 'hardening cap.' As the material is loaded to point E along the hydrostatic axis, the hardening cap is pushed out to location DE, thus extending the elastic region from OPBC to OPDE. In essence, the material has become harder under unloading and reloading. Thus, if the soil mass is loaded along an arbitrary path EFG after the hydrostatic loading to E, the soil will behave elastically up until it reaches a stress state defined by point F. Further loading will cause the soil to deform plastically until it reaches point G on the fixed failure surface defined by the Drucker–Prager yield condition. A state of stress above the line PG is not attainable according to this yield criterion. This idea of hardening caps has been incorporated in the development of an unified theory of soil mechanical behavior that links volume changes, stress state, and yielding into a single framework called the critical-state soil mechanics.

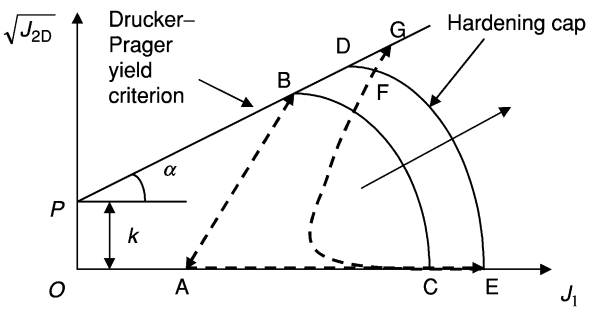


Figure 10 Drucker–Prager yield criterion with hardening caps.

Postyield Behavior

Upon yield, soil mass undergoes plastic deformation. The total strain can be decomposed into elastic and plastic strain, i.e.:

$$\epsilon_{\mathrm{ij}} = \epsilon^{\mathrm{e}}_{\mathrm{ij}} + \epsilon^{\mathrm{p}}_{\mathrm{ij}} \qquad [62]$$

where $\epsilon^{\mathrm{e}}_{\mathrm{ij}}$ is the recoverable elastic strain and $\epsilon^{\mathrm{p}}_{\mathrm{ij}}$ is the permanent plastic deformation. The elastic strain can be determined based on a constitutive relationship such as in Eqn [35]. The plastic strain is determined using the concept of plastic potential and flow rule.

Plastic potential and flow rule The plastic potential, g, is assumed to be a scalar function of the state of stress within the soil mass (i.e., $g = g(\sigma_{11}, \sigma_{22}, \sigma_{33}, \sigma_{12}, \sigma_{13}, \sigma_{23})$). The direction of the plastic strain is

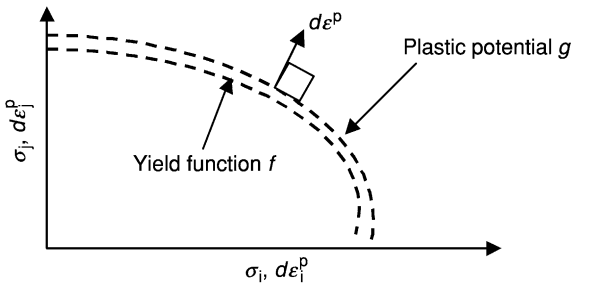


Figure 11 Plastic potential and yield function for a material governed by the associative flow rule.

defined by a flow rule. According to the flow rule, the incremental plastic strain vector is oriented in the direction of the normal to the yield potential (Figure 11). This condition is known as 'normality rule.' Using this definition, the plastic strain is given by:

$$\epsilon_{ij}^p = \chi \frac{\partial g}{\partial \sigma_{ij}} \qquad [63]$$

where χ is a positive scalar multiplier or loading index. For some materials, the plastic potential g is assumed to be the same as the yield function f. These materials are said to follow the associative flow rule; materials for which the plastic potential differs from the yield function are said to follow the nonassociative flow rule.

Hardening law An additional concept that is necessary to describe the plastic behavior completely is the hardening law. The hardening law describes the growth of yield function or hardening cap as the material undergoes plastic deformation (e.g., the rule for the growth of the hardening cap from BC to DE in Figure 10). A hardening parameter, h, is often included in the yield function to describe the changes in the yield function with plastic flow, i.e.:

$$f = f(\sigma_{11}, \sigma_{22}, \sigma_{33}, \sigma_{12}, \sigma_{13}, \sigma_{23}, h) \qquad [64a]$$

Critical-State Soil Mechanics

The critical-state concept is based on the observation that when a soil sample is subjected to increasing shear loading, it will yield and undergo plastic deformations (both volumetric and shear) and finally arrive at a critical volume, after which its volume remains unchanged, although its shape continues to change. The Cam-clay model for the elastoplastic behavior of wet clay includes this critical-state concept while accounting for the stress and volume changes during yielding. A slightly modified version of this original model has been used widely in numerical simulation studies. Here this 'modified Cam clay model' is considered for the particular case of a triaxial loading of a cylindrical soil sample to keep the mathematical treatment relatively simple. The hydrostatic stress ($p = (\sigma_1 + 2\sigma_3)/3$), shear stress ($q = (\sigma_1 - \sigma_3)$), and void ratio ($e$) are used as the state variables in developing the model.

Figure 12 graphically represents the soil behavior in the modified Cam-clay model. Figure 12a shows the variation of void ratio as a function of the hydrostatic pressure. Curve ABCD is the normal consolidation line (NCL) under hydrostatic loading. Curves EB and FC are elastic unloading–reloading curves (URL). The NCL and URL curves are mathematically represented by Eqns [43] and [44]. Figure 12b is the representation of NCL and URL in the void ratio versus $\ln(p)$ plane (cf. Figure 7). Figure 12c shows the critical-state line (CSL) and the elliptical yield surface, OPQ, that is hypothesized by the modified Cam-clay model in the q–p plane. The major axis of the ellipse coincides with the hydrostatic stress axis and is equal to $(p_0/2)$, which is the hardening parameter, h. The critical-state line is assumed to pass through the origin with slope M in the q–p plane and intersects the yield surface at point P. When the soil arrives at the critical-state line, it undergoes no more volume change, i.e., $d\epsilon_v$ is zero. This requirement along with the normality condition can be used to show that point P is directly above the center of the elliptical yield surface. Therefore the coordinates of this point are $(p_0/2, Mp_0/2)$, and $(Mp_0/2)$ is the minor axis of the ellipse:

$$\frac{\left(p - \frac{p_0}{2}\right)^2}{\left(\frac{p_0}{2}\right)^2} + \frac{q^2}{\left(\frac{Mp_0}{2}\right)^2} = 1 \qquad [64b]$$

Multiplying by $(Mp_0/2)^2$ and simplifying:

$$M^2[p(p - p_0)] + q^2 = 0 \qquad [64c]$$

Figure 12d is a three-dimensional (3-D) representation of the modified Cam-clay model which shows that the critical-state line is a space curve, the projection of which is a straight line on the q–p plane.

Modified Cam-Clay Constitutive Relationship

The following relationships are developed using the associative flow rule (i.e., plastic potential is the same as the yield function). The elastic portion of the total strain is given by Eqn [36], i.e.:

$$\begin{Bmatrix} d\epsilon_v^e \\ d\epsilon_s^e \end{Bmatrix} = \begin{bmatrix} 1/K_t & 0 \\ 0 & 1/(3G_t) \end{bmatrix} \begin{Bmatrix} dp \\ dq \end{Bmatrix} \qquad [65]$$

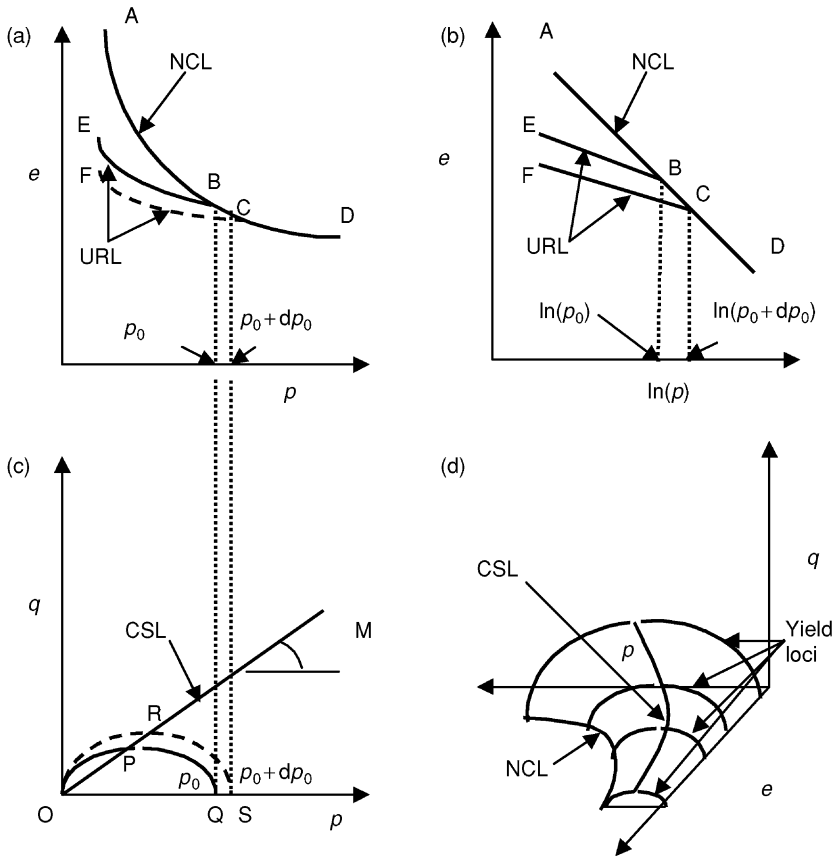


Figure 12 Modified Cam-clay model with elliptical yield surface: (a) deformation in the *e*-*p* plane; (b) deformation in the *e*-ln(*p*) plane; (c) behavior in the *q*-*p* plane; and (d) three-dimensional representation of the model. NCL, normal consolidation line; URL, unloading–reloading line; CSL, critical-state line.

where superscript 'e' represents elastic strain components. The plastic strain components are given by:

$$\mathrm{d}\epsilon_{\mathrm{v}}^{\mathrm{p}} = \chi \frac{\partial g}{\partial p}$$
$$\mathrm{d}\epsilon_{\mathrm{s}}^{\mathrm{p}} = \chi \frac{\partial g}{\partial q} \qquad [66]$$

where superscript 'p' stands for plastic strain components. Because of the assumption of associative flow, the plastic potential, g, is the same as the yield function, f. The elliptical yield function described in the previous section is given by (cf. Eqn [64c]):

$$f(p,q,p_0) = g(p,q,p_0) = M^2[p(p-p_0)] + q^2 = 0 \qquad [67]$$

Eqn [67] can be used to obtain an expression for loading index, χ, as follows. Taking the total differential of Eqn [67]:

$$\mathrm{d}f = \mathrm{d}g = \frac{\partial f}{\partial p}\mathrm{d}p + \frac{\partial f}{\partial q}\mathrm{d}q + \frac{\partial f}{\partial p_0}\mathrm{d}p_0 \qquad [68]$$

However, $\mathrm{d}p_0$, which is related to the size of the yield surface, i.e., changes in the hardening parameter, can be related to volumetric plastic strain using Eqns [43] and [44]. Referring to **Figure 12b**, from Eqns [43] and [44]:

$$e_0 = e_{\mathrm{i}} - \kappa \ln(p_0) \qquad [69a]$$

$$e_0 = \Gamma - \lambda \ln(p_0) \qquad [69b]$$

The elastic part of the volume change is related to that described in Eqn [69a], and the total volume change is related to that described in Eqn [69b]. Taking differentials of Eqns [69a] and [69b]:

$$\mathrm{d}e^{\mathrm{e}} = -\kappa \frac{\mathrm{d}p_0}{p_0} \qquad [70a]$$

$$\mathrm{d}e = -\lambda \frac{\mathrm{d}p_0}{p_0} \qquad [70b]$$

Therefore, the plastic part of the deformation is given by:

$$\mathrm{d}e^{\mathrm{p}} = \mathrm{d}e - \mathrm{d}e^{\mathrm{e}} = -(\lambda - \kappa)\frac{\mathrm{d}p_0}{p_0} \qquad [71]$$

From Eqns [24b] and [71]:

$$d\epsilon_v^p = -\frac{de^p}{1+e_i} = \left\{\frac{\lambda-\kappa}{1+e_i}\right\}\frac{dp_0}{p_0} \quad [72]$$

Eqn [72] can be solved for dp_0 as follows:

$$dp_0 = p_0\left\{\frac{1+e_i}{\lambda-\kappa}\right\}d\epsilon_v^p \quad [73]$$

Therefore, from Eqns [66], [68], and [73]:

$$\frac{\partial f}{\partial p}dp + \frac{\partial f}{\partial q}dq + \frac{\partial f}{\partial p_0}p_0\left\{\frac{1+e_i}{\lambda-\kappa}\right\}\chi\frac{\partial g}{\partial p} = 0 \quad [74]$$

Solving Eqn [74] for χ gives:

$$\chi = \frac{-\left(\frac{\partial f}{\partial p}dp + \frac{\partial f}{\partial q}dq\right)}{\frac{\partial f}{\partial p_0}p_0\left\{\frac{1+e_i}{\lambda-\kappa}\right\}\frac{\partial g}{\partial p}} \quad [75]$$

If $\eta = q/p$ is substituted as η, then for the Cam-clay model:

$$\frac{\partial f}{\partial p} = \frac{\partial g}{\partial p} = (M^2-\eta^2)p$$
$$\frac{\partial f}{\partial q} = \frac{\partial g}{\partial q} = 2\eta p$$
$$\frac{\partial f}{\partial p_0} = -M^2 p \quad [76]$$

where $p_0 = (M^2+\eta^2)p/M^2$ is substituted for from Eqn [67]. (From Eqn [67]: $f(p, q, p_0) = M^2\,[p\,(p - p_0)] + q^2 = 0$; therefore $M^2\,p^2 - M^2\,pp_0 + \eta^2\,p^2 = 0$, since $q = \eta p$; further simplification of this equation leads to $p^2\,(M^2+\eta^2) = M\,pp_0$, or $p_0 = (M^2+\eta^2)\,p/M^2$.) Substituting these partial derivatives in Eqn [75] and simplifying:

$$\chi = \left(\frac{\lambda-\kappa}{1+e_i}\right)\frac{(M^2-\eta^2)dp + 2\eta\,dq}{(M^2+\eta^2)(M^2-\eta^2)p^2} \quad [77]$$

Finally, from Eqn [66] and [76], upon simplification:

$$d\epsilon_v^p = \frac{(\lambda-\kappa)}{(1+e_i)(M^2+\eta^2)p}\left[(M^2-\eta^2)dp + 2\eta dq\right] \quad [78a]$$

$$d\epsilon_s^p = \frac{(\lambda-\kappa)}{(1+e_i)(M^2+\eta^2)p}\left[2\eta dp + \frac{4\eta^2}{(M^2-\eta^2)}\right] \quad [78b]$$

or in matrix form,

$$\left\{\begin{matrix} d\epsilon_v^p \\ d\epsilon_s^p \end{matrix}\right\} = \frac{(\lambda-\kappa)}{(1+e_i)(M^2+\eta^2)p}\begin{bmatrix} (M^2-\eta^2) & 2\eta \\ 2\eta & \frac{4\eta^2}{(M^2-\eta^2)} \end{bmatrix} \quad [79]$$

If the soil mass is within the elastic region, Eqn [65] can be used to compute elastic strain. However, if the soil mass has yielded and is undergoing elastic as well as plastic deformation, Eqn [79] can be used to obtain plastic strain while Eqn [65] provides the magnitude of the elastic strain. This model contains five soil parameters (κ, ν or G, λ, M, Γ). The first two parameters are related to the elastic behavior of soil mass whereas the last three parameters are related to the plastic behavior. Note that the parameter κ is related to soil bulk modulus K by the relationship, $\kappa = \frac{1+e_i}{\kappa}$, as shown previously.

Although the elliptic yield function represents the behavior of normally consolidated and lightly overconsolidated soils reasonably well, it overpredicts the strength of overconsolidated soils.

NSDL–AU Model

Based on numerous tests conducted using unsaturated agricultural soils at the National Soil Dynamics Laboratory (NSDL) and Auburn University (AU) in Alabama, USA, a model for plastic behavior of soils has been formulated, the NSDL–AU model, which is based on two key developments. The first is the development of a model to describe the hydrostatic compression of unsaturated agricultural soils. This equation for NCL differs from Eqn [43] and has the following form:

$$\bar{\epsilon}_v = (A + Bp)(1 - e^{-Cp}) \quad [80]$$

where $\bar{\epsilon}_v$ is the natural volumetric strain, $\ln(V/V_0)$, and A, B, and C are material parameters. Eqn [80] has been extended to represent cylindrical triaxial test results. The extended equation is given by:

$$\bar{\epsilon}_v = (A + Bp)(1 - e^{-Cp}) + \frac{\sqrt{2}}{3}D\frac{q}{p} \quad [81]$$

Note the use of $\tau_{oct} = \sqrt{2}/3\ q$ in this formulation. Moreover, this equation contains an additional material parameter, D. Eqn [81] represents the yield surface for unsaturated agricultural soils and is quite similar to the yield surface shown in **Figure 12d**. Furthermore, the model inherently contains a critical-state line. It should be noted that Eqn [81] represents plastic behavior of soil and does not account for any unloading–reloading action (i.e., elastic behavior). Most agricultural soils tend to be slightly overconsolidated and their deformation is primarily controlled by plastic deformation.

Composite Soil-Strength Parameters

Although the foregoing discussion treats mechanical behavior of soil from the veiwpoint of continuum mechanics, it has not been used widely by field researchers and practitioners due to its complexity and the difficulty in obtaining engineering soil parameters of undisturbed *in situ* soils. Devices such as cone penetrometers, and shear and sinkage devices have been developed to provide various measures of soil strength for use in tillage, traction, and soil-compaction studies. These devices provide soil-strength parameters that tend to depend on the geometry of the test device and type of loading applied. Moreover, these soil parameters often do not represent any single soil property, but are usually functions of several fundamental engineering properties of soil. Therefore soil parameters obtained using these devices are often called 'composite soil parameters'; e.g., penetrometers, shear-vane devices, and shear graphs. There are also various techniques used for determining soil stickiness, shatter resistance, and cutting resistance. The cone penetrometer is well-known among these devices because it is a simple device and very easy to use. Shear-vane devices and shear graphs have been used to obtain soil shear and sinkage parameters. They are useful in predicting tractive ability of wheels and tracks using a semiempirical approach.

Cone Penetrometer

Perhaps the most widely used device to measure soil strength in the field is the cone penetrometer. Although the cone penetrometer was developed to determine the mobility of off-road vehicles at the Waterways Experiment Station in Vicksburg, Mississippi, USA, it has been used to predict traction, draft requirements of tillage implements, and to quantify soil strength to indicate soil-compaction level and impedance to root growth. The most common form of this device consists of a polished steel cone, which is pushed against the soil and then the force of penetration is measured. The American Society of Agricultural Engineers (ASAE) has developed a standard (S313.3) that describes the geometry of a standard cone penetrometer, whereas a second standard (EP542) outlines the proper procedure for using this device.

Since the force needed to penetrate the soil is related to the geometry of the device, a cone with a base diameter of 20.27 mm and an apex angle of 30° is selected as the standard shape (Figure 13). For harder soil conditions, a smaller cone with a base diameter of 12.83 mm is used. A second key variable that influences the force of penetration is the penetration rate. ASAE Standard EP542 recommends a quasistatic rate of 1.83 m min^{-1}. Although it is difficult to control the insertion rate with handheld devices, hydraulically or electrically operated devices can be designed to operate at this standard speed. Force is usually measured using a load-sensing mechanism such as a load cell. Newer devices often include a depth-measuring sensor such as a potentiometer so that a soil penetration resistance profile can be obtained. The penetration resistance force is expressed as cone index, which is the ratio of the force to the base area of the cone. The soil cone index value obtained using a soil cone penetrometer is a composite value that depends on soil texture, bulk density, and moisture content. In terms of engineering properties of soil, it depends on cohesion, soil internal angle of friction, soil metal friction, and adhesion.

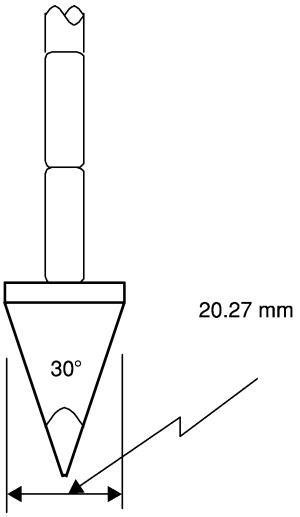


Figure 13 Standard cone penetrometer of the American Society of Agricultural Engineers.

One of the main concerns with the use of cone index to represent soil strength is its variability, especially in dry and cloddy conditions. ASAE standard EP542 recommends that at least 20 measurements must be taken near the field capacity of soil in a given location to obtain a representative measure of soil strength. With the advent of precision agriculture and the potential role of soil compaction in limiting water infiltration, drainage, and root growth, there is an increased interest in the cone penetrometer as a soil-strength mapping tool. Consequently, fully automated cone penetrometers with global positioning systems (GPS) to provide geographic position data are now commercially available.

Measurement of Soil Sinkage and Shear

Inadequacy of cone index values in representing soil characteristics relevant to tractive ability of wheels

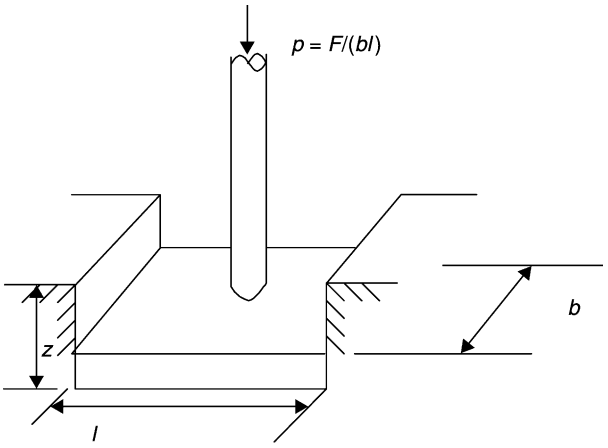


Figure 14 A rectangular sinkage plate.

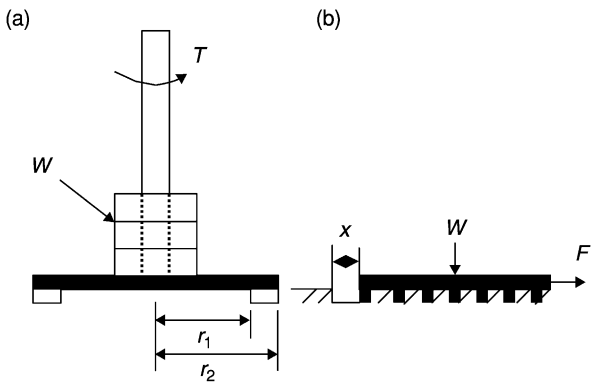


Figure 15 Soil-shear test devices: (a) shear ring; and (b) grouser plate.

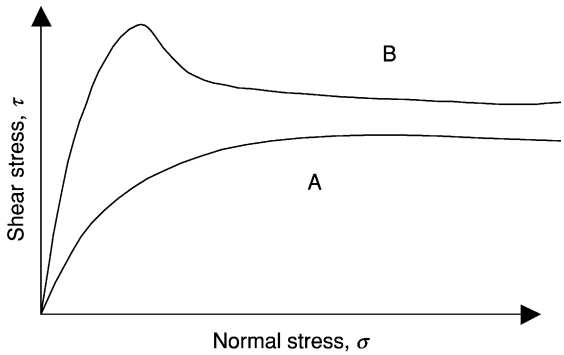


Figure 16 Shear characteristics of soil (A) normally consolidated or slightly overconsolidated and (B) overconsolidated.

and tracks has led to the development of sinkage and shear devices such as the bevameter.

Soil sinkage devices consist of either circular or rectangular plates that are pushed against soil, and their load deformation characteristics are recorded (Figure 14). The Bernstein equation is often used to relate applied load to soil deformation as follows:

$$p_s = kz^n \qquad [82]$$

where p_s is the applied compressive pressure, z is the soil sinkage, and k and n are constants. The sinkage parameter was found to depend on plate width by the following relationships:

$$k = \frac{k_c}{b} + k_\phi \qquad [83]$$

where k_c and k_ϕ are parameters related to soil cohesion and angle of internal friction, respectively, and b is the minimum plate dimension. Since the parameter n in Eqn [82] is usually not an integer, units of k are not straightforward. An alternate formulation that overcomes the problem of dimension of k is:

$$p = k_r \left(\frac{z}{b}\right)^n \qquad [84a]$$

and:

$$k_r = (k_1 + k_2\, b) \qquad [84b]$$

where k_1 and k_2 are once again parameters related to soil cohesion and angle of internal friction, respectively. Note that the unit of k_r in Eqn [84a] is the same as that of pressure. These plate sinkage relationships have been used to model rolling resistances of wheels and tracks.

Shear characteristics of soil such as cohesion and angle of internal friction have been used to model the tractive ability of wheels and tracks and soil cutting by tillage tools. Torsional shear devices (shear vane, shear cone, shear ring, shear graph) and rectangular grouser plates are often used to measure soil-shear characteristics. Figure 15 shows a circular shear head and a rectangular grouser plate. The maximum shear stress at a specific normal load, W, on a torsional device as shown in Figure 15a is given by:

$$\tau = \frac{3T}{2\pi(r_1^3 - r_2^3)} \qquad [85]$$

where T is the torque and r_1 and r_2 are the radii of the shear vane.

The original bevameter used a shear device similar to the one shown in Figure 15a. A recent development includes an instrumented soil test device that uses a grouser plate, sinkage plate, and a standard cone to obtain soil shear, sinkage, and cone index data. Figure 16 shows the shear characteristics of soil measured using a shear test device. The curve A in Figure 16 is a representation of a normally consolidated or slightly overconsolidated soil. The

curve B is a typical response of an overconsolidated soil. Most agricultural soils behave as shown in curve A, which can be represented by the empirical equation:

$$\tau = \tau_{max}(1 - e^{-x/k_s}) \quad [86]$$

where τ_{max} is the maximum shear stress, k_s is the soil-shear modulus, and x is the soil deformation in the direction of shear stress.

Moreover, the Mohr–Coulomb criterion (cf. Eqn [59]) can be used to represent τ_{max}. Note that $\tau_{max} = F/A$ and $\sigma_n = W/A$, where A is the plate area. If the shear test is repeated at two different values of the normal load, W, both c and ϕ can be determined.

Some Recent Developments

Increased interest in precision agriculture within the last decade has led to the investigation of causes of yield variability within a field. One factor believed to influence crop yield is soil compaction, since it has a direct impact on soil hydraulic conductivity. As mentioned previously, the cone penetrometer is the soil-strength measuring device that is being used increasingly to map soil compaction level. Since it is a highly variable point measurement, numerous cone index values are needed to obtain proper representation of a field. This limitation of the cone penetrometer has led to the development of alternative devices that can measure and map soil strength in a continuous manner. One such device consists of a texture–soil-compaction sensing system that consists of a simple tine that is instrumented with a load cell to measure soil-cutting force. It also incorporates a dielectric-based soil-moisture sensor, because soil-moisture content influences soil-cutting force significantly. The soil-cutting force, F, is a function of soil bulk density, ρ, texture, ξ, and moisture content, θ, when the device is operated at a constant speed and operating depth; i.e.:

$$F = f(\rho, \xi, \theta) \quad [87]$$

Based on the field measurements the soil-cutting force was determined as:

$$F = g(\rho, \xi) * e^{-c\theta} \quad [88]$$

where c is an empirical constant. The unknown function $g(\rho, \xi)$ is 'texture/soil compaction index' (TCI). Therefore, TCI is given by:

$$TCI = F/e^{-c\theta} \quad [89]$$

Note that this TCI value depends on both soil bulk density and texture. Since texture is a static property in the absence of cut-and-fill operation, TCI values can be used as an indicator of soil-compaction level. The TCI sensor has been interfaced to a differential global positioning system (DGPS) to obtain soil-strength maps of tomato fields and correlate them to tomato yield. The field test results indicate that, although the TCI sensor works reasonably well, it is not helpful in locating the compact layer within the soil mass. As of 2003, active research is currently in progress to develop a compaction profile sensor, which can measure the compaction level of soil with depth. Successful development of such a real-time soil-compaction profile sensor may contribute to the development of site-specific tillage (tilling only where there is a need to loosen soil) and limit tillage depth to the hardpan depth.

See also: **Compaction**; **Conservation Tillage**; **Cultivation and Tillage**; **Site-Specific Soil Management**; **Structure**; **Subsoiling**; **Swelling and Shrinking**

Further Reading

ASAE (2000) *Soil Cone Penetrometer.* ASAE Standard S313.3, pp. 832–833. St. Joseph, MI: ASAE.

ASAE (2000) *Procedures for Using and Reporting Data Obtained with the Cone Penetrometer.* ASAE Standard EP542, pp. 987–989. St. Joseph, MI: ASAE.

Bailey AC and Johnson CE (1989) A soil compaction model for cylindrical stress states. *Transactions of the American Society of Agricultural Engineers* 32(3): 822–825.

Bailey AC and Johnson CE (1994) *NSDL-AU Model, Soil Stress–Strain Behavior.* ASAE Paper No. 94-1074. St. Joseph, MI: ASAE.

Bailey AC, Johnson CE, and Schafer RL (1984) Hydrostatic compaction of agricultural soils. *Transactions of the American Society of Agricultural Engineers* 27(4): 925–955.

Das BJ (1983) *Advanced Soil Mechanics*. New York: McGraw-Hill.

Desai CS and Siriwardane HJ (1994) *Constitutive Laws for Engineering Materials with Emphasis on Geologic Materials*. Englewood Cliffs, NJ: Prentice Hall.

Duncan JM and Chang CY (1970) Non-linear analysis of stress and strain in soils. *Journal of Soil Mechanics and Foundations Division* 96: 1629–1653.

Hettiaratchi DRP (1987) A critical state soil mechanic model for agricultural soil. *Soil Use and Management* 3(3): 94–105.

Kirby JM (1989) Measurements of the yield surfaces and critical-state of some unsaturated agricultural soils. *Journal of Soil Science* 40: 167–182.

Roscoe KH and Burland JB (1968) On the generalized stress–strain behavior of "wet" clay. In: Heyman J and Leckie FA (eds) *Engineering Plasticity*, pp. 335–609. Cambridge, UK: Cambridge University Press.

Roscoe KH, Schofield A, and Worth CP (1958) On yielding of clay soils. *Geotechnique* 8: 22–53.
Upadhyaya SK, Chancellor WJ, Perumpral RL *et al.* (1994) *Advances in Soil Dynamics*, vol. 1. St Joseph, MI: ASAE.
Upadhyaya WJ, Chancellor JV, Perumpral RL *et al.* (2002) *Advances in Soil Dynamics*, vol. 2. St Joseph, MI: ASAE.
Wood DM (1990) *Soil Behavior and Critical State Soil Mechanics*. New York: Cambridge University Press.

STRUCTURE

V A Snyder and M A Vázquez, University of Puerto Rico Agricultural Experiment Station, San Juan, Puerto Rico 00926–1118

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

The term 'structure' of a granular medium refers to the spatial arrangement of solid particles and void spaces. In materials such as coarse sands and gravels, the particles are loosely bound and tend to arrange themselves in closely packed, minimum-energy configurations (Figure 1a). Most soils, however, tend to exhibit a hierarchical structure. That is, primary mineral particles, usually in association with organic materials, form small clusters or 'first-order aggregates.' These in turn form larger clusters or 'second-order aggregates,' and so on, as illustrated schematically in Figure 1b.

Aggregate hierarchy in soils is reflected not only in increasing aggregate size with each successive level, but also in the predominant mechanisms by which particles in aggregates are bonded together; i.e., aggregates at different hierarchical levels tend to bond together by different mechanisms. Hence, the term 'structure' in soil science generally carries a connotation of bonding mechanisms in addition to geometrical configuration of particles.

Without hierarchical structure, medium- and fine-textured soils such as loams and clays would be nearly impermeable to fluids and gases, and at 'typical' moisture contents would possess a mechanical strength prohibitive to growth of plant roots and soil organisms. Thus, structure plays a crucial role in the transport of water, gases, and solutes in the environment, and in transforming soil into a suitable growth medium for plants and other biological organisms. Physical appearance of structured vs unstructured 'puddled' soil is shown in Figure 2.

This article briefly describes different hierarchical levels of soil structure, and dominant processes and mechanisms through which structural bonding occurs. Also discussed are statistical models and geometric scaling concepts used to describe the hierarchical system as a whole. A final section summarizes the dynamics of structure in tilled soils. Useful books and reviews on soil structure are listed at the end.

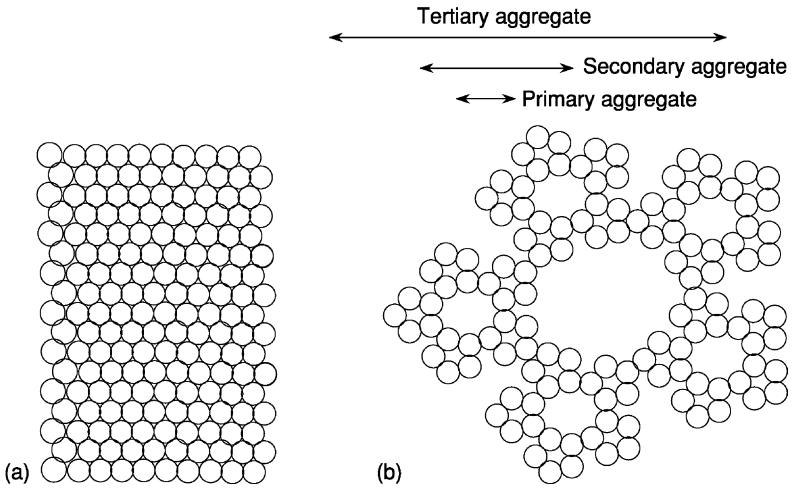


Figure 1 Idealized diagrams of (a) unstructured close-packed particles; and (b) system of hierarchically structured particles.

Hierarchical Levels of Soil Structure and Bonding Mechanisms

A fairly broad consensus exists for classifying soil aggregates into two main hierarchical categories, microaggregates (<250 μm diameter) and macroaggregates (>250 μm diameter). The microaggregates are typically subdivided into subclasses, <2 μm, 2–20 μm, and 20–250 μm. Salient properties of these aggregate categories are summarized below, and illustrated schematically in Figure 3. Comparison of aggregate size scales with other characteristic soil dimensions is presented in Figure 4.

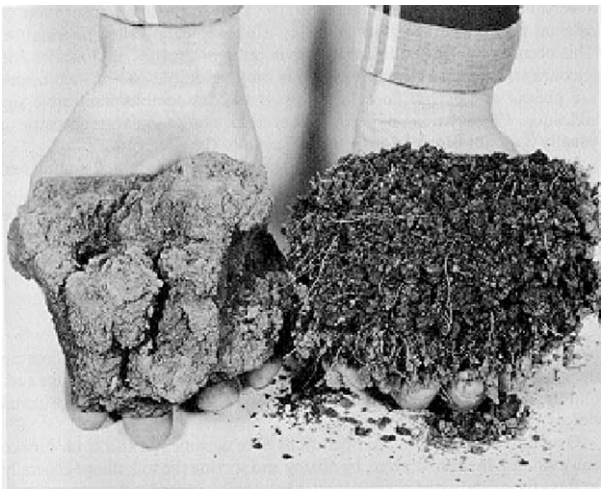


Figure 2 Photograph of structured soil and the same soil in a 'puddled' state where structure has been mostly destroyed. Courtesy of the USDA Natural Resources Conservation Service.

Microaggregates less than 2 μm in Diameter

These appear to be initially formed by flocculation of clay particles into domains or quasicrystals (Figures 3 and 5). Particularly in very small aggregates (<0.2 μm), most organic matter appears to be absorbed only onto external surfaces of the quasicrystals. Thus, bonding between particles in soil quasicrystals is governed by essentially the same van der Waals and electrical double-layer phenomena that produce quasicrystals in simple clay–water systems. Exclusion of organic matter from internal quasicrystal surfaces appears quite pronounced in montmorillonitic soils where clay particles exhibit a strongly oriented, mutually parallel structure (regions marked 'T' in Figure 5), and to a lesser extent in soils with illitic or kaolinitic mineralogy and correspondingly less-ordered domains. Very small microaggregates are highly resistant to mechanical disruption, typically requiring several minutes of ultrasonic dispersion, often with the aid of an oxidizing agent.

In a next hierarchical level, quasicrystals and other mineral particles coalesce around central bonding nuclei of highly processed organic materials of humic and polysaccharide nature. The clay domains and other mineral particles on the outside of these microaggregates protect the inner organic core against access by microorganisms. Figure 5 shows pockets of polysaccharide material (P) surrounded by clay tactoid coatings (T). The humic and polysaccharide materials typically consist of multiple strands and functional groups, which provide many

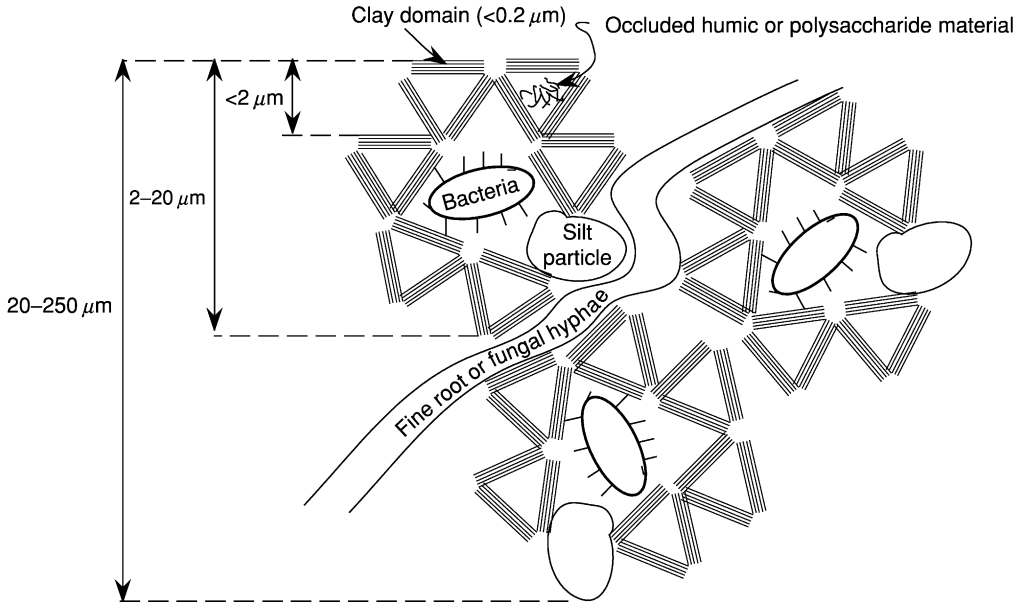


Figure 3 Schematic diagram of a hierarchical system of microaggregates illustrating special characteristics of each class.

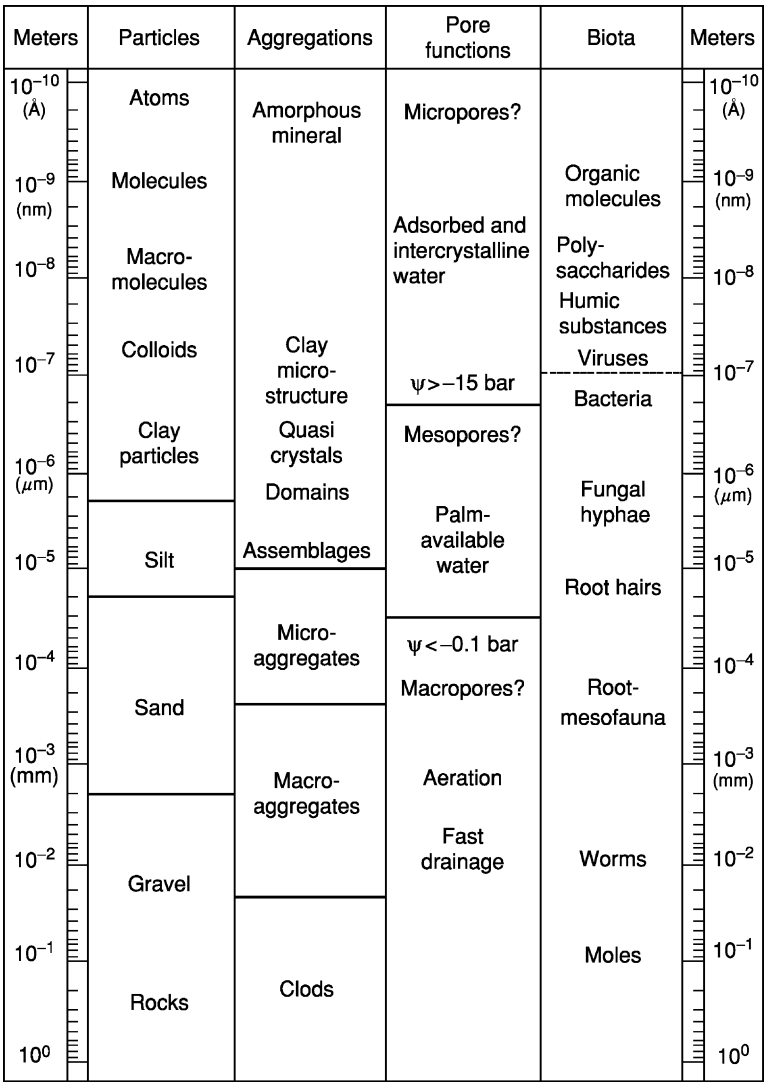


Figure 4 Comparative scales in soil structure. Reproduced with permission from Waters AG and Oades JM (1991) In: Wilson WS (ed.) *Advances in Soil Organic Research*, p. 164. Cambridge: Royal Society of Chemistry.

sites for bonding with the mineral surfaces. Since the predominant electrical charge of organic polymers at 'normal' soil pH values is usually negative, bonding with negatively charged inorganic colloids is largely achieved through 'bridging' by multivalent cations such as Ca^{2+}, Fe^{3+}, Al^{3+} and their hydrous oxides, which are able to complex with both mineral surfaces and organic functional groups (Figure 6 and Table 1). In highly weathered soils with abundant variable-charge minerals, ligand exchange between mineral surfaces and organic functional groups can produce particularly strong bonds. This is particularly notable in Oxisols, characterized by extremely strong micro-porous microaggregates resulting in a characteristic bimodal pore-size distribution (Figure 7).

In order for microaggregates to form effectively in soil, organic binding materials must be finely distributed throughout the soil, rather than deposited in isolated pockets. Particularly effective mechanisms appear to be *in situ* biosynthesis of organic materials by microorganisms associated with extensive networks of fine roots with high turnover rates, such as under grass vegetation.

Microaggregates tend to form slowly in soils, but once formed they also degrade slowly, even under unfavorable soil management systems. Organic substances in microaggregates have been observed to remain stable for hundreds and in some cases even thousands of years. Among the stabilized organic materials are enzymes that may contribute to processes such as N mineralization and herbicide degradation. The amount of carbon 'sequestered' in soil microaggregates constitutes a substantial fraction of the total amount organic carbon on Earth. This has

stirred much interest in the impact of soil structure management on atmospheric CO_2.

Microaggregates 2–20 μm in Diameter

These are formed from oriented clay domains, microaggregates $<2\,\mu$m in diameter, and/or coarse clay and fine silt particles, which coalesce around a central core of hyphal fragments and bacterial cells or colonies (Figures 3 and 5). Bonding is effected by microbial materials such as polysaccharide synthesized by the bacteria and hyphae. The outer layer of clay domains and microaggregates protects the bacteria from organisms such as nematodes and protozoa, which are too large to penetrate the outer layer. Thus, up to 40–60% of the microbial biomass in soil has been found associated with microaggregates 2–20 μm in diameter.

Like $<2\,\mu$m aggregates, 2–20 μm aggregates are very resistant to mechanical disruption, sometimes resisting up to 5 min of ultrasonic dispersion.

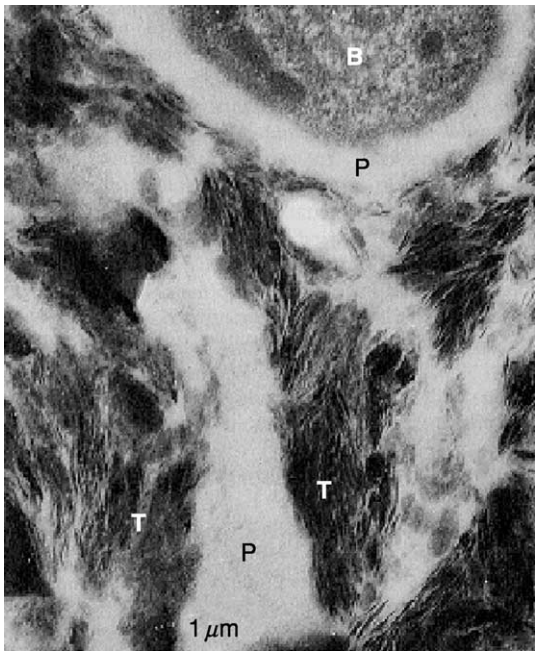


Figure 5 Micrograph of a $<2\,\mu$m microaggregate showing clay tactoids (T), occluded polysaccharide material (P), and bacterial capsule (B) surrounded by polysaccharide material. Reproduced with permission from Foster RC (1978) In: Emerson WW, Bond RD, and Dexter AR (eds) *Modification of Soil Structure*, p. 104. Chichester: Wiley.

Microaggregates 20–250 μm in Diameter

These are formed largely by particles or aggregates $<20\,\mu$m in diameter, bonded by polysaccharide material around central nuclei of fine roots and fungal hyphae, which may or may not be subsequently completely degraded by soil microorganisms. Bonding is generally strong enough that the aggregates are stable to slaking upon direct immersion of air-dry soil into water. Micrographs of two such aggregates are shown in Figure 8.

Macroaggregates (greater than 250 μm Diameter)

General properties of macroaggregates and their dynamic nature in soil management systems Due to their effect on size of the largest soil pores, these aggregates are very influential in determining macroscopic soil properties such as mechanical strength,

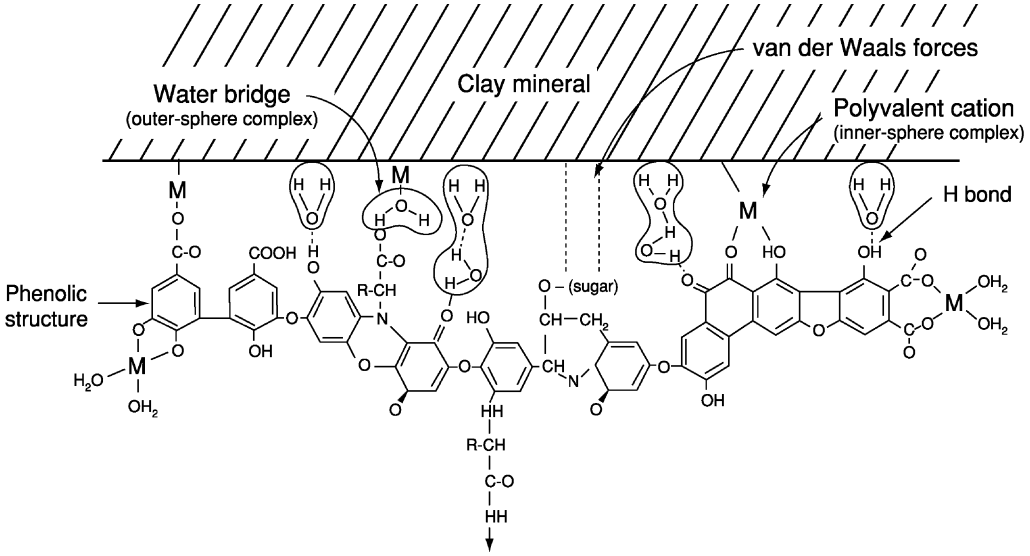


Figure 6 Illustration of bonding mechanisms between humic acid and mineral surfaces. Reproduced with permission from Haynes RJ and Beare MH (1996) In: Carter MR and Stewart BA (eds) *Structure and Organic Matter Storage in Agricultural Soils*, p. 217. Boca Raton, FL: CRC Press.

Table 1 Possible aggregate-bonding mechanisms

I. Clay domain–clay domain

A. Domain face–domain face

Cations bridge between negative faces. Mechanism similar to that for orientation of clay platelets into domains

$$\text{Face}^{-} \text{- - - } M^{n+} \text{ - - -}^{-} \text{ Face}$$

B. Domain edge—domain face

Positive edge site to negative face

$$\text{Edge Al—}OH_2^{+} \text{- - -}^{-} \text{ Face}$$

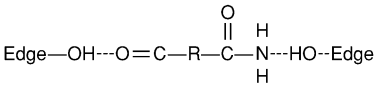
II. Clay domain—organic polymer—clay domain

A. Domain edge—organic polymer—(domain)

1. Anion exchange: Positive edge site to polymer carboxyl

$$\text{Edge Al—}OH_2^{+} \text{ - - - } {}^{-}\text{OOC—R—}$$

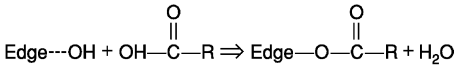
2. Hydrogen bonding between edge hydroxyl and polymer carbonyl or amide



3. Cation bridge between negative edge site and polymer carboxyl

$$\text{Edge—}O^{-}\text{—}M^{n+}\text{—}{}^{-}\text{OOC—R—COO}^{-}$$

4. van der Waals attraction between edge and polymer
5. Ligand exchange



B. Domain face—organic polymer—(domain)

$$\text{Face Si—O—HO—R—OH}$$

1. Hydrogen bonding between polymer hydroxyl and external or internal (expanding lattice minerals) face silicate oxygens

$$\text{External face}^{-} \text{- - -} M^{n+}\text{—OOC—R—COO}^{-}\text{—}$$

2. Cation bridge between domain external face and polymer carboxyl or other polarizable group
3. van der Waals attraction between face and polymer

III. Quartz—(silt, inorganic, and organic colloids)—quartz

A. Chemical bonds established between quartz surface gels of hydrated alumino silicates and active groups of other aggregate constituents

B. Quartz grains held in a matrix of silt and clay stabilized primarily by:

1. Oriented clay particles
2. Irreversibly dehydrated silicates, sesquioxides, or humic–sesquioxide complexes
3. Irreversibly dehydrated humic materials
4. Silt-size microaggregates stabilized by iron humates
5. Organic colloids and clay domains bonded by mechanisms cited under I and II

Adapted with permission from Hillel D (1998) *Environmental Soil Physics*. London: Academic Press.

hydraulic conductivity, and aeration status. They are also the aggregates most susceptible to breakdown under stresses due to tillage, compaction, raindrop impact, or wetting-induced slaking.

Macroaggregation seems primarily caused by reorientation and binding of clay particles and microaggregates by fine roots and hyphae, and further cementation by extracellular polysaccharides. Associations between extensive fine root systems and vesicular arbuscular mycorrhizal (VAM) fungi, which produce large amounts of polysaccharide, appear particularly effective in forming macroaggregates. Saprophytic fungi also stabilize aggregates, particularly when substrates such as combinations of straw and manure are added to the soil. Algal filaments, covered with slimy gels, have also been observed to be effective. It is often difficult to distinguish between the relative importance of binding by fine plant roots and fungal hyphae because, particularly in the case of VAM fungi, fungi and root growth are often positively

correlated. What does seem fairly clear is that plant species with large systems of fibrous roots, associated with VAM hyphae that secrete large amounts of polysaccharide gel, are effective stabilizers.

Bonding by polysaccharides apparently does not occur uniformly throughout macroaggregates, but rather primarily in the larger cracks or pores (15–50 μm in diameter), precisely where aggregates are most likely to rupture. Due to this efficient placement of bonds, even small increases or removals of polysaccharide can cause large increases or decreases in macroaggregate stability. Polysaccharides in large pores are more accessible to degradation by microorganisms than in microaggregates, so they are relatively labile and sensitive to changes in management.

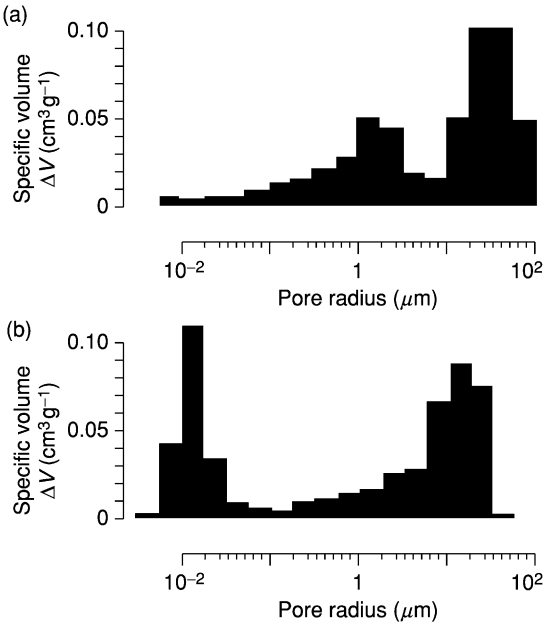


Figure 7 Pore size distribution for (a) silty soil and (b) an oxisol obtained by mercury porosimeter. Adapted from Bartoli F, Dutartre Ph, Gomendy V, *et al.* (1998) In: Baveye PJ, Parlange JV, and Stewart BA (eds) *Fractals in Soil Science*, p. 220. Boca Raton, FL: CRC Press, with permission.

Because labile polysaccharides constitute only a small fraction of the total amount of soil organic matter, it is not surprising that total soil organic matter content does not always correlate well with macroaggregate stability. This is particularly true shortly after sudden changes in management, where changes in total organic matter may be minimal but significant changes may have occurred in the network of fine roots, fungal hyphae, and associated polysaccharides, with associated rapid changes in aggregate stability. For example, Figure 9 shows that, after changing from continuous corn to an alfalfa cropping system, soil organic carbon accumulation lagged behind aggregate formation. In such situations, measurements of labile polysaccharide or related parameters are often better indicators of trends in macroaggregate stability than is total organic matter content (Table 2).

In cases where total soil organic matter content does correlate positively with macroaggregate stability, management has often been stable over long periods of time, e.g., continuous forest, pasture, or tillage. The high correlation probably reflects not only aggregate stabilization by organic matter, but also organic matter stabilization by aggregates. For a given type of soil management, correlation between organic matter and structural stability can be highly

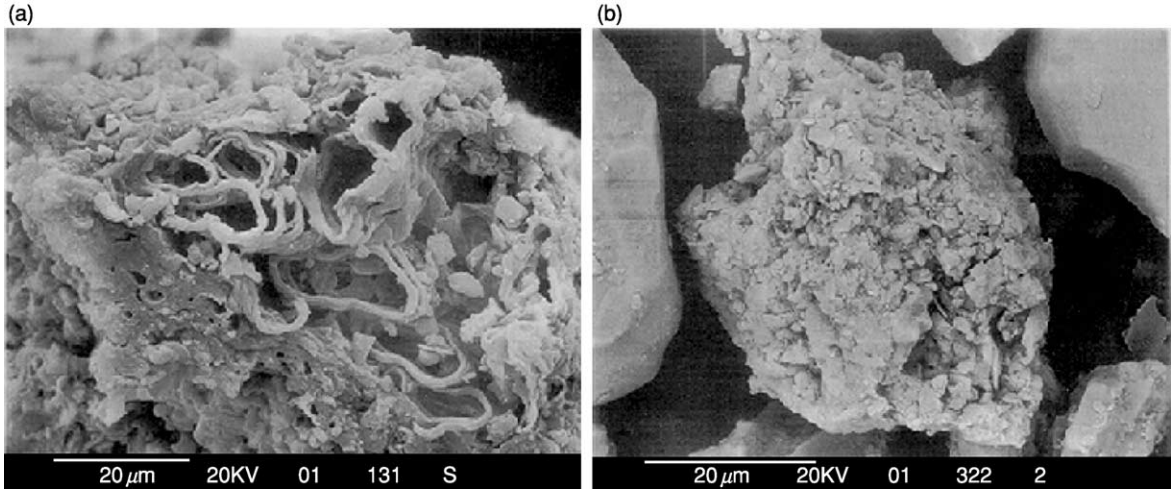


Figure 8 Micrographs of 20–250 μm aggregates showing (a) partly decayed vascular bundle surrounded by inorganics; (b) aggregates with elongated void running from top left to bottom right, with no remnants of plant anatomy evident. Reproduced with permission from Waters AG and Oades JM (1991) In: Wilson WS (ed.) *Advances in Soil Organic Research*, pp. 169–170. Cambridge: Royal Society of Chemistry.

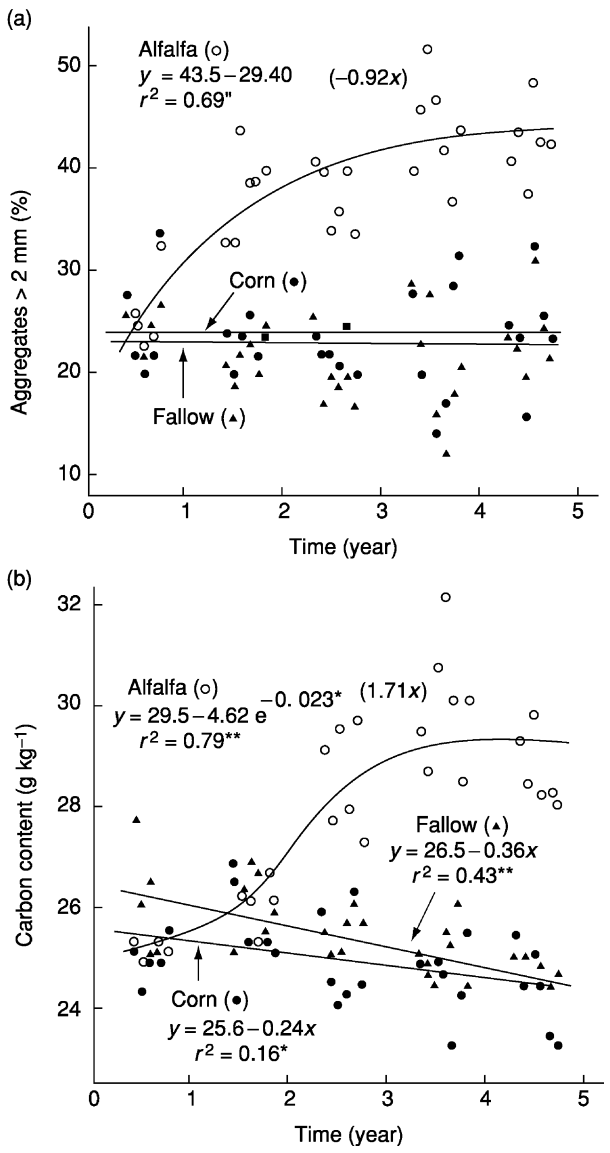


Figure 9 Data (open circles) showing how increases in soil organic carbon content (b) lag behind increase in aggregate stability (a) following changes in land management. Reproduced with permission from Angers DA and Carter MR (1996) In: Carter MR and Stewart BA (eds) *Structure and Organic Matter Storage in Agricultural Soils*, p. 199. Boca Raton, FL: CRC Press.

soil-dependent, with texture playing a major role. In many cases a certain threshold level of organic matter is necessary before macroaggregates begin stabilizing (Figure 10). The threshold value tends to increase with increasing soil clay content, suggesting that a critical amount of organic matter per unit mineral surface area must be exceeded in order for effective aggregation to occur.

Earthworm casts Earthworm casts are the indurated soil material surrounding channels left in the soil by earthworms. Casts are initially quite unstable, since they are essentially puddled soil emitted from the earthworm's gut. Stability comes later, as a result of microbial interaction with organic materials mixed with soil in the gut. Fungal hyphae also contribute to cast stability, but probably originate from fungi outside the casts since fungi are strict aerobes that cannot survive inside the earthworm gut. When disrupted by tillage or mild shaking in water, earthworm casts generally break down into macroaggregate-sized fragments, and thus are often classified as macroaggregates.

Casts provide stability to earthworm channels, which over time under zero-tillage conditions may occur in such numbers as to contribute significantly to soil hydraulic conductivity and aeration status. An important property of these channels is their continuity, which makes them particularly effective as preferential flow conduits. Tillage disrupts continuity of earthworm channels, rendering them largely ineffective for transport.

Role of Wetting and Drying on Soil Structure Development

A key element in the development of soil structure is the cyclic pattern of wetting and drying of field soils.

When wet soils dry, particles are drawn together by negative pressures or suctions that develop in the pore water. Microscopically, this promotes clay orientation and interparticle bonding. Macroscopically, the soil as a whole shrinks or decreases in volume, causing the development of crack networks and surfaces of weakness between neighboring soil elements. These to a large extent define the boundaries of soil aggregates, particularly macroaggregates, and form continuous channels that facilitate water infiltration and gas movement.

The extent of soil cracking depends primarily on the intensity of drying, and on the amount and type of soil clay minerals. Soils with large amounts of 2:1 clay minerals, such as vertisols, are notorious for networks of large shrinkage cracks. At the other extreme, soils with predominantly nonexpandable clay minerals, such as oxisols, exhibit very little shrinkage and cracking behavior. This to a large extent explains the characteristic 'massive' soil structure of oxisols, manifested by a lack of visually distinct macroaggregate boundaries.

If crack networks extend to the soil surface, as is usually the case in untilled soil, then water infiltration during rainfall events occurs preferentially down these cracks until swelling causes the cracks to seal up again. Dispersed clay in the infiltrating water may be deposited at crack boundaries as water infiltrates the

surrounding aggregates, forming clay skins or 'cutans' on the aggregate surfaces. Also, since the soil at crack surfaces is the first to wet during infiltration events, soil displacements during wetting may result in shearing or 'smearing' of these surfaces. Such shear zones, or 'slickensides,' are particularly notorious in vertisols.

Characterization of Soil Structure Based on Visual Assessment

Visual Inspection of Soil Aggregates

A routine component of soil surveys is the description of soil macroaggregates based on visual examination. The criteria used in this classification scheme are type or shape, class or size, and grade or distinctness. Different structure types are illustrated in Figure 11, and further details are given in Table 3.

A certain amount of subjectivity exists in deciding the precise category for a given soil structure, particularly with respect to shape and grade. Nevertheless, the classification system has been used successfully for grouping soils according to structure-related properties such as permeability and preferential flow of solutes. This is largely due to the fact that aggregate morphology is related to morphology of the interaggregate void spaces where most water and solute transport takes place. For example, blocky

Table 2 Sensitivity of various chemical indices in relation to changes in aggregate stability over time

Previous cropping history	*Aggregate stability (MWD, mm)*	*Organic C (%)*	*Acid-hydrolyzable carbohydrate (%C)*	*Hot water-extractable carbohydrate ($\mu g\ C^{-1}$)*	*Microbial biomass C ($\mu g\ C\ g^{-1}$)*
18-year pasture	2.7	3.2	0.35	208	1018
4-year pasture[a]	2.5	2.5	0.26	169	890
1-year pasture	2.0	2.4	0.25	152	801
1-year arable	1.3	2.4	0.23	140	738
4-year arable	1.2	2.4	0.23	134	712
10-year arable	1.0	2.0	0.19	127	610

[a]The 1-year and 4-year pasture and 1-year and 4-year arable soils come from a cropping rotation of 4-years arable followed by 4-years pasture. (Data from Haynes RJ, Swift RS, and Stephen RC (1991) Influence of mixed cropping rotation (pasture-arable) on organic matter content, water stable aggregation and clod porosity in a group of soils.) *Soil Tillage Res* 19: 77–87.
Reproduced with permission from Haynes RJ and Beare MH (1996). In: Carter MR and Stewart BA (eds) *Structure and Organic Matter Storage in Agricultural Soils*, p. 236. Boca Raton, FL: CRC Press.

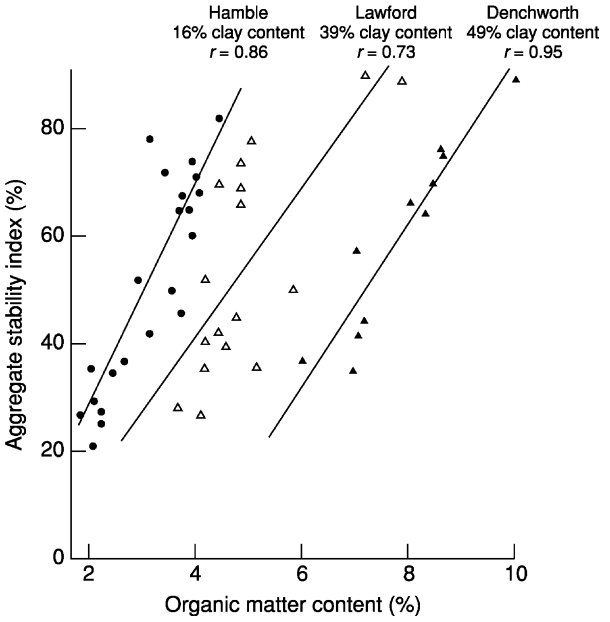


Figure 10 Relation between aggregate stability and organic matter content in soil with varying texture. Reproduced with permission from Haynes RJ and Beare MH (1996) In: Carter MR and Stewart BA (eds) *Structure and Organic Matter Storage in Agricultural Soils*, p. 232. Boca Raton, FL: CRC Press.

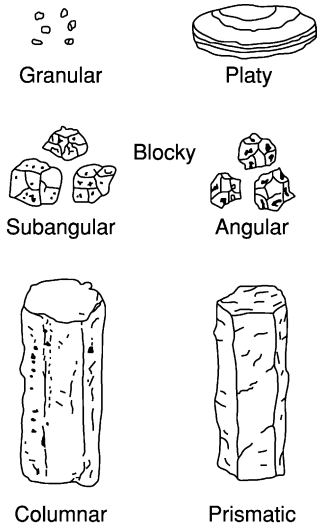


Figure 11 Classification of macroaggregates based on visual appearance. Reproduced with permission from Hillel D (1998) *Environmental Soil Physics*. London: Academic Press.

Table 3 Classification of soil structure according to US Department of Agriculture soil survey staff

A Type: shape and arrangement of peds

	Platelike	*Prismlike*		*Blocklike—polyhedral—spheroidal* Three approximately equal dimensions arranged around a point			
	Horizontal axes longer than vertical, arranged around a horizontal plane	Horizontal axes shorter than vertical. Arranged around vertical line. Vertices angular		*Blocklike—polyhedral* Plane or curved surfaces accommodated to faces of surrounding peds		*Spheroidal–polyhedral* Plane or curved surfaces not accommodated to faces of surrounding peds	
		Without rounded caps	With rounded caps	Faces flattened; vertices sharply angular	Mixed rounded, flattened faces; many rounded vertices	Relatively nonporous peds	Porous peds
B Class: size of peds	*Platy (mm)*	*Prismatic (mm)*	*Columnar (mm)*	*Blocky (mm)*	*Subangular blocky (mm)*	*Granular (mm)*	*Crumb (mm)*
1. Very fine or very thin	<1	<10	<10	<5	<5	<1	1
2. Fine or thin	1–2	10–20	10–20	5–10	5–10	1–2	1–2
3. Medium	2–5	20–50	20–50	10–20	10–20	2–5	2–5
4. Coarse or thick	5–10	50–100	50–100	20–50	20–50	5–10	
5. Very coarse (very thick)	>10	>100	>100	>50	>50	>10	
C Grade: durability of peds	0. Structureless		No aggregation or orderly arrangement				
	1. Weak		Poorly formed, nondurable, indistinct peds that break into a mixture of a few entire and many broken peds and much unaggregated material				
	2. Moderate		Well-formed, moderately durable peds, indistinct in undisturbed soil, that break into many entire and some broken peds but little unaggregated material				
	3. Strong		Well-formed, durable, distinct peds, weakly attached to each other, that break almost completely into entire peds				

Reproduced with permission from Hillel D (1998) *Environmental Soils Physics*. London: Academic Press.

and prismatic macroaggregates are often bounded by networks of large, continuous cracks, whereas interaggregate void spaces for granular or crumb structures are smaller and more finely distributed.

Image Analysis

Examination of photographs of soil crack patterns, or of thin sections, has long been used to obtain detailed information on the shape of aggregates and their spatial organization in relation to interaggregate void spaces. This has become increasingly feasible with the advent of powerful yet low-cost digital image acquisition and analysis technology. Three-dimensional image analysis is also possible using techniques such as computer-aided tomography.

Mathematical Models of Soil Structure

Aggregate Size Distributions

A common method for characterizing soil structure involves measuring the size distribution of soil fragments or 'aggregates' produced by a specified fragmentation method. The distributions are often fitted to two-parameter probability models such as in Table 4. In many cases one of the two parameters is nearly constant. For example, for 'fractal' distributions, the fractal dimension D may be relatively constant. Likewise for log-normal and Weibull distributions, the geometric standard deviation σ and exponent λ, respectively, often do not vary much. In such cases the distribution may be defined by specifying only the diameter X at some fixed probability P, or a probability-weighted diameter or 'mean weight diameter' defined over some fixed probability interval (P_1, P_2) as:

$$\mathrm{MWD} = \int_{P_1}^{P_2} X(P)\,dP \qquad [1]$$

Pore Size Distributions

Pore size distributions may be determined by microscopic examination of thin sections, or inferred indirectly from moisture release characteristics or mercury intrusion porosimetry. The latter two methods are based on the capillary relation between equivalent cylindrical pore radius (r) and gauge pressure head (h) at which liquid will just enter or recede from the pore:

$$r = 2\sigma\cos\Phi/\Delta\rho g h \qquad [2]$$

Here ρ is liquid density, σ is solid–liquid surface tension, g is the gravitational constant and Φ is the solid–liquid contact angle. Equation [2] allows inferring the total volume of voids in radius interval $r + dr$ from the measured volume of liquid drained or intruded in the pressure head range $h + dh$. The pore-size histograms in Figure 7 are an example of distributions inferred by this method.

For the common case of liquid water in non-hydrophobic soils, where $\rho \approx 1\ \mathrm{g\,cm^{-3}}$, $\sigma \approx 71\ \mathrm{dyn\,cm^{-1}}$, and $\Phi \approx 0$, eqn [2] reduces to:

$$r = 0.15/h \qquad [3]$$

where r and h are expressed in cm.

Scaling Models of Soil Structure

Soil structure often exhibits certain similitude or 'scaling' properties, which greatly simplify its mathematical representation. Two types of scaling have received wide attention in soil science, fractal scaling and Miller scaling (in honor of the brothers E.E. and R.D. Miller, who developed the original concepts). Fractal scaling concerns similarities between hierarchical levels in a given soil, whereas Miller scaling deals with structural similarities between different soils. In both cases, the underlying assumption is that the structural elements being compared are similar

Table 4 Common cumulative probability functions used to characterize aggregate size distribution in soils

Distribution	*Functional form*	*Definition of variables*
Fractal	$P(x) = \left(\frac{X}{\sigma}\right)^{3-D}$	$P(x)$ = cumulative mass fraction X = aggregate diameter σ, D = constants
Log-normal	$P(x) = \frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{\ln X} \exp\left\{-1/2\left[\ln\left(\frac{X}{\mu}\right)\right]^2\right\} d\ln X$	$P(x)$ = cumulative mass fraction X = aggregate diameter μ,σ = constants
Weibull	$P(x) = 1 - \exp\left\{-\left(\frac{X}{\sigma}\right)^{\lambda}\right\}$	$P(x)$ = cumulative mass fraction X = aggregate diameter σ, λ = constants

geometrically, differing only in some characteristic dimension or 'length.'

Fractal scaling relations between hierarchical levels in a given soil Fractals may be defined, somewhat simplistically, as hierarchical porous or irregular objects composed of successively nested similar elements. To illustrate, a simple fractal object is shown in Figure 12. The lowest hierarchical level (Figure 12a) plays the role of reference element or 'generator.' A second hierarchical level (Figure 12b) is established by forming a larger cluster from the primary clusters. These in turn form a third hierarchical level (Figure 12c), and so on.

The filamentous fractal object in Figure 12 is similar in many ways to particle clusters formed during diffusion-limited flocculation. Other more compact fractal objects, like the one shown earlier in Figure 1b, are perhaps more representative of soil aggregates.

A property of these and many other self-similar objects is that the number of primary particles N required to form a given hierarchical level is a non-integer power law function of the characteristic length L (aggregate diameter) at that hierarchical level, i.e.:

$$N \propto L^D = L^{E-1+f} \qquad [4]$$

Here E represents dimensionality of the problem ($E = 3$ for three-dimensional space, and 2 for two-dimensional objects in the plane) and f is a fraction $0 > f \leq 1$. The parameter $D = E - 1 + f$ is often known as the fractal dimension. The fraction f increases with the degree of 'space-filling' of the given fractal object. For example, f has a value of approximately 0.465 for the open filamentous object in Figure 12, and a value closer to unity (0.658) for the more compact 'aggregate-like' structure in Figure 1b. In the case of complete space filling (such as squares formed by placing smaller squares edge to edge, cubes formed by placing smaller cubes face to face, etc.), $f = 1$ and eqn [4] reduces to the familiar Euclidean relation $N \propto L^E$.

A consequence of scaling property eqn [4] is that the number of particles per unit 'characteristic' bulk volume, N/L^E, varies as L^{f-1}, which is an inverse relation whenever $f < 1$. Consistent with this result, aggregate bulk density often decreases in linear log-log fashion with increasing aggregate size (Figure 13).

Analysis of pore spaces in certain fractal structures has yielded theoretical power law relations between volumetric water content θ and matric suction h. At least qualitatively, this agrees with the well known empirical Brooks–Corey relation:

$$\theta(h) = \theta_{sat}(h/h_e)^b \qquad [5]$$

where θ_{sat} is saturated water content, h_e is suction at air-entry and b is a soil-dependent constant.

Fractal scaling models have proved particularly useful for describing the geometry of complicated crack networks in soils.

Miller scaling relations between different soils In Miller scaling, no *a priori* assumption is made regarding fractal hierarchy within a given soil (although fractal scaling is allowed under certain conditions). All that is required is similitude between

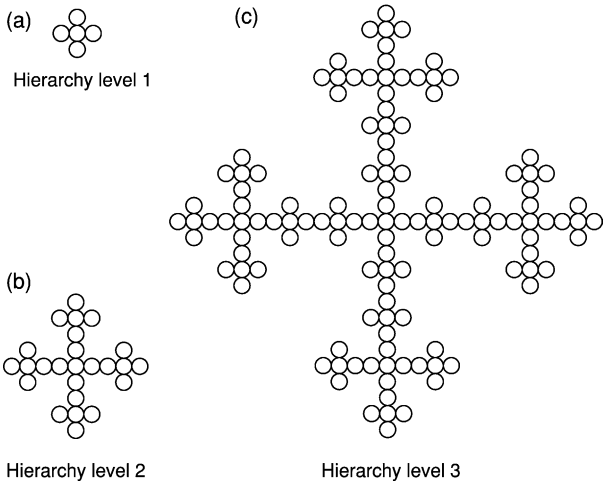


Figure 12 Self-similar filamentous objects of increasing hierarchical levels. Adapted with permission from Meakin P (1991) Fractal aggregates in geophysics. *Review on Geophysics* 29: 317–354.

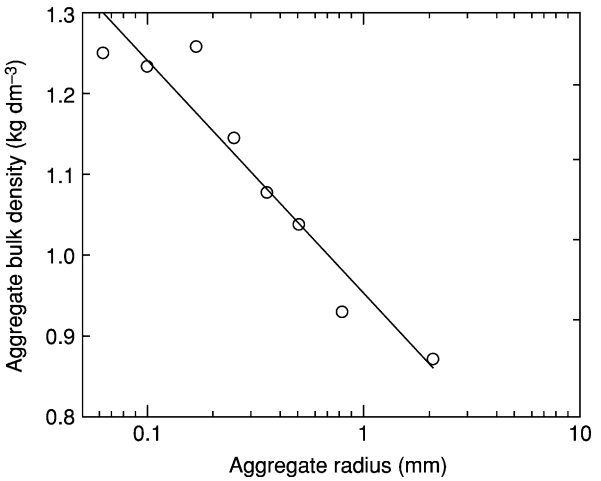


Figure 13 Bulk density of soil aggregates as a function of their radius. (Data from Chepil WS (1950) Methods of estimating apparent density of discrete soil grains and aggregates. *Soil Science* 70: 351–362, with permission.)

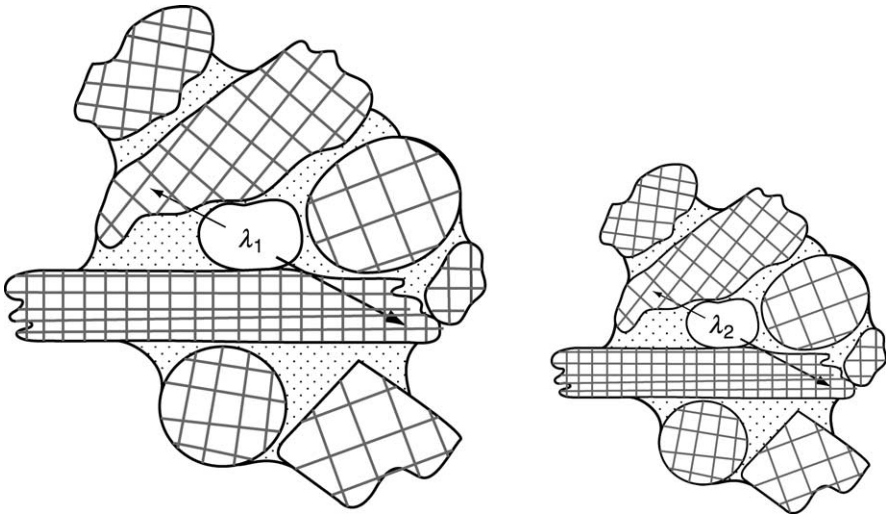


Figure 14 Illustration of two Miller similar soils with different characteristic lengths λ_1 and λ_2. After Miller EE and Miller RD (1956) Physical theory for capillary flow phenomena. *Journal of Applied Physics* 27: 324–332.

different soils, regardless of what the internal hierarchy may be.

In its strictest sense, Miller similitude assumes that the geometry of both elementary particles and their structural organization is identical across soils (Figure 14). Assuming dominance of capillary phenomena and creeping Newtonian fluid flow, this type of similarity imposes scale invariance on the dimensionless moisture release characteristic $\theta(h\lambda/\sigma)$ and the dimensionless hydraulic conductivity function $K(\theta)\eta/\lambda^2$, where K(θ) is hydraulic conductivity K at moisture content θ, η is fluid viscosity and σ is surface tension of the liquid–vapor interface. Miller similar soils always have identical porosity.

In practice, a weaker form of Miller similitude or 'generalized scaling' is often adopted, under recognition that it is usually the larger soil pores (i.e., those in upper hierarchical levels) that exhibit similitude and furthermore conduct most of the pore water. In generalized scaling, not only is the characteristic length λ of these large pores allowed to vary across soils, but also the soil volume fraction or 'effective' porosity P_e that they occupy. This is illustrated in Figure 15, where similarly shaped conducting or 'effective' structural units of different characteristic lengths λ_1 and λ_2 are embedded in a surrounding nonconducting soil matrix. For a given value of λ, the 'effective' porosity P_e is proportional to the number of elementary structural units per unit overall soil volume. Thus P_e may be considered a characteristic 'pore number' scaling factor, complementing λ which accounts for characteristic size. This contrasts with Miller similar soils, where λ is the only independent scale factor because P_e is always constant.

A consequence of generalized scaling behavior is invariance of the reduced moisture retention function $S_e(h\lambda/\sigma)$, where S_e is an 'effective' pore saturation defined as:

$$S_e = \theta_e/P_e = (\theta - \theta_o)/(\theta_{sat} - \theta_o) \qquad [6]$$

Here θ and θ_{sat} are the actual and saturated soil moisture contents, respectively, and θ_o is a 'residual' water content below which scaling relations no longer apply. θ_o may be considered as the water content when all the 'effective' structural units have drained and only the nonconducting matrix surrounding the 'effective' structural elements remains saturated. The parameter $\theta_e = \theta - \theta_o$ is known as the 'effective' water content and $\theta_{sat} - \theta_o$ is the effective porosity P_e.

Soils with invariant $S_e(h\lambda/\sigma)$ functions are frequently observed to have scale-invariant hydraulic conductivity functions $K(S_e)/\lambda_K^2$, where λ_K^2 is an empirical soil-dependent scale factor that may or may not correlate with the pore size scale factor λ. Only under strict Miller similarity does $\lambda = \lambda_K$.

Two fractal soils will also exhibit Miller scaling, if: (1) the shapes of the initial generating structures are the same in both soils, even though their characteristic lengths λ are different; and (2) both soils have the same number of hierarchical levels.

Structure of Tilled Agricultural Soils

Throughout history, modification of soil structure by tillage has played a central role in crop production. The tilled 'plow layer' plays a crucial role in determining plant growth and transport of gas, water, and

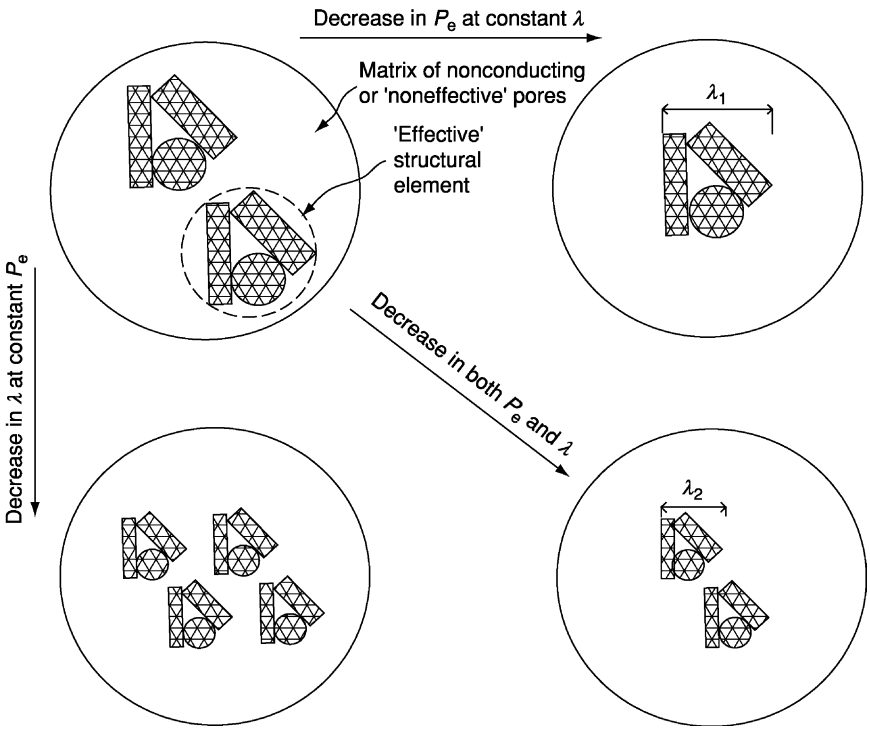


Figure 15 Illustration of four similar soils exhibiting generalized scaling.

chemicals in the environment. Thus, no review of soil structure would be complete without some discussion on structural dynamics in tilled soils.

Initial Soil Conditions Produced by Tillage: Soil Tilth

The structure or 'tilth' of freshly tilled soil is defined by the size and shape distribution, spatial arrangement, and internal structure of soil fragments produced by tillage. For a given soil and tillage implement, the tilth obtained depends primarily on soil moisture content at time of tillage.

Tillage under wet conditions, particularly in heavy-textured soils, generally results in large, plastically deformed fragments where internal structure may be seriously damaged or 'puddled.' Tillage of dry soils, on the other hand, generally results in minimal plastic deformation and structural damage, but undesirably large fragments may still be produced. The optimum soil moisture content for tillage is at some intermediate value, often close to the lower plastic limit, wherein maximum fragmentation occurs with little plastic deformation. Soils in this condition are said to exist in their most 'friable' state. Soils with predominantly large or intermediate pore sizes, associated with light texture or good structure or both, tend to be friable over fairly large ranges in moisture content. On the other hand, poorly structured fine-textured soils, particularly those dominated by high-activity clays, tend to present narrow moisture ranges for maximum friability.

Post-tillage Soil Structural Transformations

The loose structure produced by tillage tends to be highly unstable, so that under action of wetting and drying the soil resettles back toward a more stable structural state. The most important components of this resettlement process are crusting at the soil surface, and fracture and plastic deformation of aggregates deeper in the tilled layer.

Surface crusting of tilled soils The formation of thin (<2 mm) crusts at the soil surface, due to action of raindrops and sprinkler irrigation, is a common feature of cultivated soils throughout the world. Crusts are characterized by greater density, higher shear and tensile strength, finer pores and lower hydraulic conductivity than the underlying tilled soil. Important consequences are poor seedling emergence and increased runoff and erosion. A thin-section photograph taken from crusted soil is shown in Figure 16.

Crusting is thought to be a combined effect of two processes: (1) aggregate rupture caused by mechanical forces, such as raindrop impact and pore-air entrapment during rapid water infiltration; and (2) dispersion of clay particles. At low electrolyte concentrations, soils with Na^+-dominated exchange complexes tend to disperse spontaneously upon wetting. Ca^{2+}-dominated soils at low electrolyte concentrations may also disperse on wetting, but this

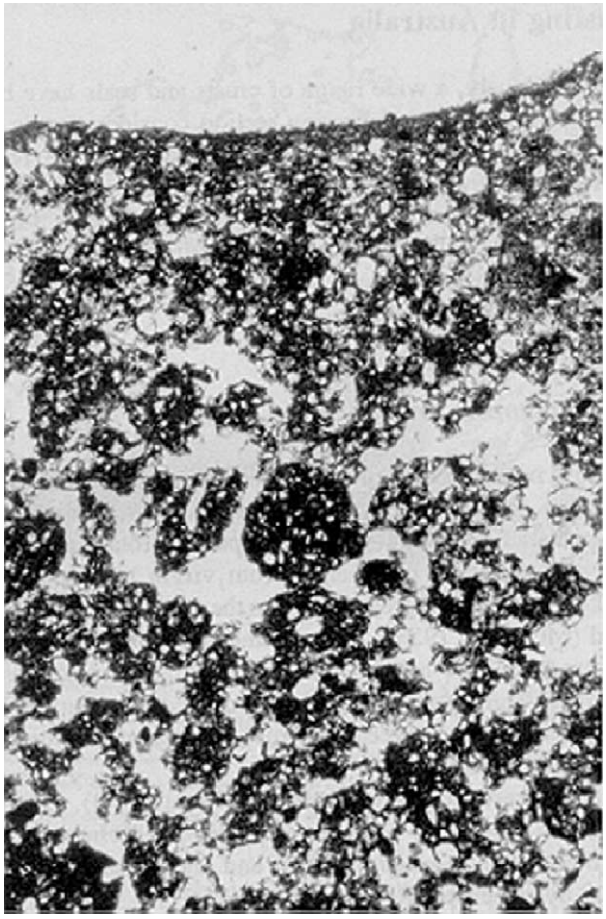


Figure 16 Photomicrograph of thin section from a crusted soil showing the denser layer near the soil surface. Frame length = 4 mm. Reproduced with permission from Chartres CJ (1992) In: Sumner ME and Stewart BA (eds) *Soil Crusting*, p. 344. Boca Raton, FL: CRC Press.

Figure 17 Fragmentation state of large clods after six wetting-drying cycles; (a) clayey vertisols; (b) clayey oxisols; and (c) clayey ultisols.

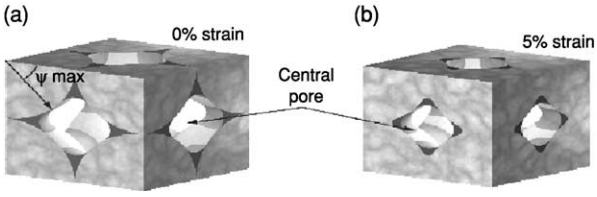


Figure 18 Schematic illustration of aggregate bed deformation resulting from plastic flattening of contact points: (a) initial undeformed aggregate bed; (b) aggregate bed after 5% volumetric strain. Courtesy of Dani Or, University of Connecticut.

frequently requires additional mechanical energy input, such as by recent tillage or raindrop impact. Dispersion in both Na^{+}- and Ca^{2+}-dominated soils is reduced at high electrolyte concentrations, but to a much greater degree in the Ca^{2+}-dominated case.

Soils with 10–30% clay content, with at least traces of smectitic clay and high amounts of silt, seem to be the most prone to crusting. Clay contents >30% tend to stabilize aggregates against disruption, whereas at <10% clay not enough fine material is present for dispersion to have significant effects.

Efforts to control crusting mainly involve ground cover to reduce raindrop impact, and chemical application to control dispersion and stabilize aggregates. Commonly used chemicals are moderately soluble Ca^{2+}-salts such as gypsum that maintain high Ca^{2+} concentrations in solution, together with polymers such as polyacrylamides (PAM) that interact with clay surfaces to promote flocculation and aggregation.

Fracture and plastic deformation of aggregates during wetting and drying cycles At depths greater than several millimeters in the plow layer, soil is not subject to direct raindrop impact, and wetting occurs more gradually and under greater tensions than in the surface layer. Thus, the mechanisms governing soil structural changes are different from the surface layer. The main mechanisms appear to be aggregate fragmentation due to differential soil swelling during water infiltration, and plastic deformation of wet aggregates due to concentrated shear stresses at interaggregate points of contact.

Wetting-induced aggregate fragmentation seems to increase with increasing soil shrink–swell potential. The most extreme and familiar manifestation is the 'self-mulching' of vertisols, wherein large soil clods completely break down to small fragments after only one or two wetting and drying cycles (Figure 17a). Other heavy-textured soils with lesser clay activity manifest the same phenomenon, but generally require more wetting and drying cycles to achieve comparable amounts of fragmentation (Figure 17b,c).

Plastic deformation of aggregates occurs when shear stresses at interaggregate contact points exceed the plastic yield strength (dependent on water content) of the soil material. The result is a gradual flattening and coalescence of interaggregate contact points (Figure 18), tending toward a state where aggregate boundaries disappear and the soil reaches its original

isotropic structure. The main causes of the concentrated stress at interaggregate contact points are believed to be overburden pressure and capillary forces due to water menisci between adjacent aggregates.

Long-Term Effects of Tillage

In the short term, tillage produces a loose soil structure conducive to plant growth and water infiltration. However, when intensively tilled soil is allowed to reconsolidate through the above mechanisms over an extended period, it often reaches a state of greater compactness than if it had never been tilled at all, typically requiring years in the undisturbed state to recover its original (pre-tillage) structure.

The Notion of Structural Quality and the Nonlimiting Water Range

An important question is: what constitutes a 'good' or 'poor' soil structure for plant growth? These attributes pertain to the general notion of structural quality. Such a notion is difficult to quantify, because plant growth in soils is strongly influenced by a number of structure-related physiological stress factors, such as soil aeration, mechanical impedance, and water availability, all of which vary in different ways with changing soil water content. For example, aeration stress is greatest at high water content, and decreases as the soil dries out, whereas stresses associated with water availability and mechanical impedance tend to increase with decreasing water content. Soil structure acts as a modifier of these relationships, by increasing or reducing the severity of each stress factor at a given soil water content.

Optimum soil physical conditions for plant growth generally reside within some intermediate range of water content where none of the stress factors is limiting. This range is known as the nonlimiting water range (NLWR). Plants growing in soils with a wide NLWR have a lesser probability of experiencing stress during their growth cycle than those growing in soils with a narrow NLWR. The NLWR has been proposed as a simple, measurable indicator of soil physical conditions or structural quality. In this framework, soils with a wide NLWR are said to have good structure, whereas those with a narrow NLWR are considered poorly structured.

Summary

The proper study of soil structure includes the definition of its essential nature, its role in the environment, methods of its characterization, and processes involved in its development and degradation. Especially important are the role and dynamics of soil structure in agricultural ecosystems, primarily because it is in these ecosystems that soil structure is most influenced by, and in turn influences, human management. Simple mathematical models of soil structure have proved useful in explaining hydrological and other macroscopic phenomena in soils.

See also: **Aggregation:** Microbial Aspects; Physical Aspects

Further Reading

Baveye P, Parlange JY, and Stewart BA (eds) (1998) *Fractals in Soil Science*. Boca Raton: CRC Press.

Carter MR and Stewart BA (eds) (1996) *Structure and Organic Matter Storage in Agricultural Soils*. Boca Raton: Lewis Publishers.

De Boodt MF, Hayes MHB, and Herbillon A (eds) (1990) *Soil Colloids and Their Associations in Aggregates*. New York: Plenum Press.

Kay BD and Angers DA (2001) Soil structure. In: Warrick AW (ed.) *Soil Physics Companion*. Boca Raton: CRC Press.

Pachepsky Y, Radcliffe DE, and Selim HM (eds) (2003) *Scaling Methods in Soil Physics*. Boca Raton: CRC Press.

Sumner ME and Stewart BA (eds) (1992) *Soil Crusting*. Boca Raton: Lewis Publishers.

SUBSOILING

R L Raper, USDA–ARS, Auburn, AL, USA

Published by Elsevier Ltd.

Introduction

'Subsoil' refers to the stratum of soil immediately below the surface soil or topsoil. Often this layer is overlooked, as most land management is focused on the topsoil, which can be altered drastically by tillage and other practices. However, the subsoil can have a large impact on a soil's potential productivity. If this layer of soil is extremely dense, roots may not penetrate, rooting volume will be decreased, nutrient uptake will be reduced, and plants may become susceptible to drought; also, water may not be able to infiltrate into the subsoil, thus limiting available water for plant growth and increasing surface runoff and potential soil erosion. Disrupting the subsoil to allow proper water infiltration and root growth may be necessary for optimum plant response.

'Subsoiling' refers to the process of soil tillage performed by a tool inserted into the soil to a depth of at least 350 mm. Tillage conducted by a narrow tool inserted to a more shallow depth is sometimes referred to as chisel plowing and is mostly used to loosen or level the soil surface and prepare a seedbed. Although tillage has been performed for several thousand years at a shallow depth, subsoiling is a relatively new operation, having only been performed since vehicles have excessively compacted the soil through their large mass and frequent traffic. Prior to the twentieth century, the ability to till deeper than just a few centimeters was not possible due to a lack of tractive force, nor was it necessary, because compaction due to repeated vehicle traffic had not yet been caused. In addition, naturally dense subsoils (e.g., hardpans, fragipans) require such treatment to allow proper root growth and drainage.

Shallow soil compaction caused by natural processes or field equipment can usually be alleviated by chisel plowing to shallow depths. However, if compaction penetrates deeper, more radical measures such as subsoiling may be necessary. This deeper compaction is often caused by repeated trafficking of the soil surface with large vehicle loads. In some locations, natural processes can lead to deep compaction, which can restrict root, water, and air movement in the soil.

Currently, subsoiling is practiced on a routine basis throughout the world. Many soils respond positively to subsoiling, resulting in yield improvements. Tillage tools used for subsoiling vary widely and result in differences in residue draft force requirements, remaining on the soil surface, and belowground soil disruption.

Measurement of Subsoiling

Determining when to subsoil requires some measurement of soil compaction. Cone index is the most accepted measure of soil compaction and has been used to determine when roots are restricted and can no longer expand into soil. This term is defined as the force required to insert a standard 30° cone into the soil. When values of cone index approach 1.5–2 MPa, root growth becomes limited and plants can start suffering the ill effects of soil compaction. After subsoiling, however, cone index values as low as 0.5 MPa are commonly found down to the depth of tillage (Figure 1).

Benefits of Subsoiling

The most obvious benefit of subsoiling is the disruption of deep, compacted subsoil layers. If soil compaction is excessive in these layers, roots cannot

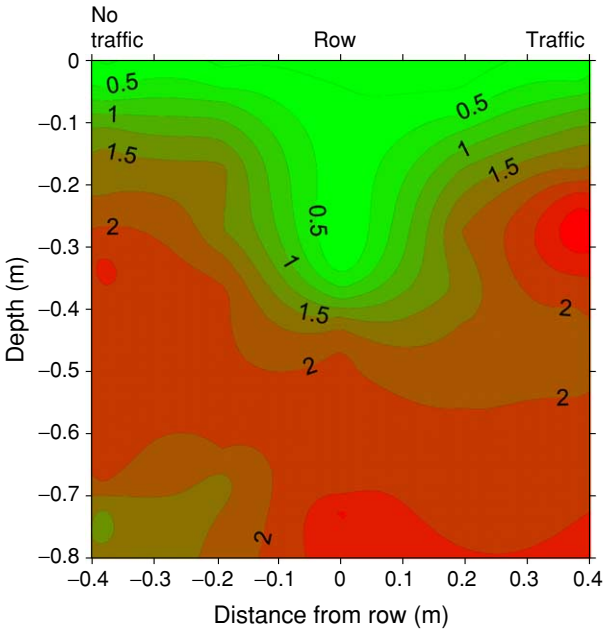


Figure 1 Cone index (mega pascals) isolines of soil showing location of nontrafficked row middle, row, and trafficked row middle. Minimal values of cone index are colored green and are shown near the center of the graph, representing root extension down to 0.3 m. Severely compacted zones are particularly evident under the trafficked row middle at the 0.3-m depth.

penetrate and are restricted to shallow depths. During times of drought, plants grown in a compacted soil are immediately susceptible, as their roots are confined to shallow zones that do not contain adequate soil moisture. Subsoiling excessively compacted soils lossens the soil for root growth, the depth of which is increased, so the plants are better able to withstand periods of drought.

Coupled with the increased root growth is the improved infiltration that usually accompanies subsoiling. Rainfall that previously exceeded infiltration capacity can be stored in the subsoil. The loosened soil provides pathways into the soil for rainfall to move quickly, instead of ponding on the soil surface and eventually evaporating or running off. Larger amounts of soil moisture may then be available to the plant during the growing season when moisture may otherwise be limited.

Increased numbers of macropores are also found after subsoiling. Even though some of these pores disappear as the soil reconsolidates, many stay open and provide increased storage of water and oxygen for plant roots. However, it is important that subsequent vehicle traffic be minimized to achieve long-lasting effects of subsoiling. Some research has reported that benefits of subsoiling are lost by the second pass of a vehicle tire. This could mean that subsoiling might not benefit a crop if traffic from a primary tillage operation and a planting operation were allowed to stray too close to the subsoiled channels. Maintaining the loosened soil profile and the increased storage capacity for water could be extremely valuable to plant roots during temporary summer droughts.

Ultimately, crop yields may improve from subsoiling, although the amount of improvement is difficult to estimate, because soil type, soil condition, plant species, and climate all have large effects. Many soils have shown benefits of being subsoiled; however, the amount of relative benefit may be offset by the expense of performing the operation. Some coarse-textured soils (sandy-to-loamy), which compact easily and require minimum tillage forces for subsoiling, show significant yield improvements when subsoiled. Some fine-textured soils are not economically subsoiled due to the lack of a yield improvement or because of the high draft forces necessary for subsoiling.

Subsoiler Design

Tillage tools used for subsoiling vary greatly in design and use. The individual vertical members that contact the soil and provide disruption are referred to as shanks. Their design varies greatly depending upon purpose, geographical location, soil type and depth of use.

Prior to 1950, most subsoiler shanks were straight, with a slight forward projection angle. However, research near the end of that decade recognized that other shapes, including curved and elliptical subsoiler shanks, can provide reduced draft forces in some soil types and soil conditions (Figure 2). Eventually, parabolic shanks became widely used and accepted as reducing draft forces. Some studies, however, have found that straight shanks mounted at an aggressive forward angle have reduced tillage draft when used in sandy soils. One negative effect of using curved shanks is that these shanks are designed to operate at a particular depth. When curved shanks are operated at depths either shallower or deeper than their intended depth, draft can increase, probably due to soil bodies that build up in front of the shanks, resulting in 'soil-on-soil' friction in contrast to the lesser 'metal-on-soil' friction (Figure 3).

Some subsoilers disrupt the soil in a symmetric manner, leaving equally disturbed soil on either side of the subsoiler as it moves in a forward direction. However, in the mid-1970s, the bentleg subsoiler was developed, which was designed to disturb the soil in a nonsymmetric manner. This shank is bent to one side by 45°, with the leading edge rotated forward by 25°.

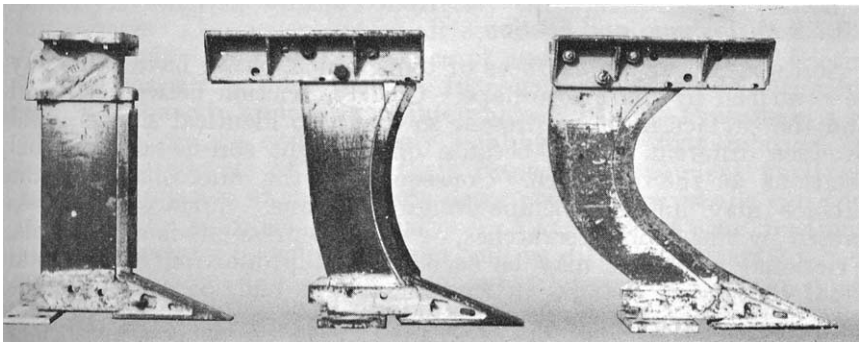


Figure 2 Straight, slightly curved, and deeply curved subsoilers, as tested in the 1950s. (Reproduced with permission from Nichols ML and Reaves CA (1958) Soil reaction to subsoiling equipment. *Agricultural Engineering* 39: 340–343.)

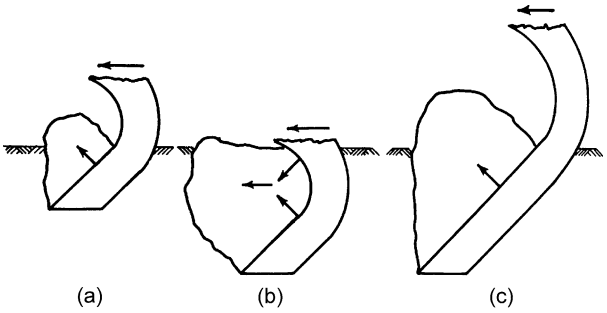


Figure 3 (a) Curved subsoiler operating at design depth; (b) curved subsoiler operating at deeper than design depth; (c) curved subsoiler scaled in size to operate at increased depth. (Reproduced from Gill WR and Vanden Berg GE (1966) *Soil Dynamics in Tillage and Traction*. Washington, DC: USDA.)

Figure 4 Side (left) and front (right) view of a bentleg subsoiler.

As the shank is traveling forward, it contacts the soil over a 216-mm width, which is substantially larger than any of the nonbentleg subsoilers. The main advantage of the bentleg subsoiler is its ability to pass through the soil leaving the surface relatively free of disturbance. For this reason, many producers have adopted this form of tillage as a method of alleviating soil compaction while maintaining large amounts of residue on the soil surface (Figure 4).

The bentleg subsoiler is commonly thought to require larger amounts of draft energy than traditional subsoilers. However, several studies have shown that comparable draft forces are generated for the bentleg subsoiler and traditional subsoilers if they are operated at similar depths.

Several other methods have been advocated for reducing draft on agricultural vehicles or increasing belowground disruption. These include vibrating the subsoiler, rotating the subsoiler shanks, placing a wing on or behind the subsoiler shank, using multiple shanks placed immediately behind each other and operating at different depths, or using a rolling coulter. Many of these experimental methods have shown promise and are undergoing further refinements, but none have been adopted by the agricultural machinery industry as of 2003.

Force Required for Subsoiling

Subsoiling requires a great amount of tillage energy. Based on experimental data obtained throughout the USA, the American Society of Agricultural Engineers' (ASAE) Standard D497.1 gives the following equation for calculation of draft force for subsoiling:

$$D = F_{\mathrm{i}}\left[A + C(S)^2\right]WT \qquad [1]$$

where D is implement draft (in newtons), F is dimensionless soil texture adjustment, i is fine (1), medium (2), or coarse-textured (3) soil, A and C are machine-specific factors, S is field speed (in kilometers per hour), W is number of rows or tools, and T is tillage depth (in centimeters).

Many factors influence draft forces of subsoiling and are represented in Eqn [1]. Soil texture certainly has a very large effect on draft force, with the factor F_{i} ranging from 0.45 for coarse-textured soils to 1.0 for fine-textured soils. (Coarse-textured soils are described as being sandy soils, medium-textured soils are described as being loamy soils, and fine-textured soils are described as being high in clay content.) Speed is also considered to be one of the most important factors. Because speed has a large effect on draft force, most subsoiling operations are conducted at relatively low speeds. The number of subsoiler shanks is also important, with the draft force increasing proportionally for each additional shank being pulled through the field.

One factor that is easily overlooked and that appears last in the equation is tillage depth. Often producers set their subsoiler depth at the deepest position that the tractor can pull. The subsoiler will only be moved upward toward the soil surface when excessive draft is sensed; after this area of excessive soil strength is passed, the original depth of subsoiling will be returned to. However, recent research reports that optimum yields are achieved when the depth of tillage is targeted to the depth of the root-impeding layer. Here, subsoiling energy is only expended to the problematic depth without wasting it by tilling too deeply. Also, excessive depths of subsoiling may increase compaction at these depths, as pressure from vehicle traffic could be propagated downward to depths below the tillage zone. As tillage is conducted more deeply every year, compaction moves downward through loosened soil horizons.

Eqn [1] only gives an approximate mean range (±50%) for subsoiler draft force, as many differences

can exist between subsoilers. One difference that is commonly found between subsoiler shanks is tip design. For Eqn [1], factors *A* and *C* have been designated as 226 and 1.8 for narrow points, but as 294 and 2.4 for 30-cm winged points. The effect of the point is to increase the draft force substantially, while increasing belowground disruption.

Management Practices

Subsoilers have been used mostly to totally disrupt entire fields or severely compacted parts of fields, especially headlands where turning and excessive trafficking has occurred. V-frame subsoilers have typically been designed with subsoiler shanks positioned closely over the width of operation (**Figure 5**). Their width is set so that the disrupted zone from a shank intersects the disrupted zone from nearby shanks. The compacted portion of the field or the entire field is loosened using this method of subsoiling with little consideration for future field activities. However, some producers subsoil at a slight angle to the previous year's rows so that excessively trafficked lanes of the field are mostly removed. Another reason for conducting the subsoiling operation at an angle is that it ensures that an entire row would not be missed if an error in subsoiler positioning occurred or if a shank were bent or broken and did not adequately disrupt the soil. Secondary tillage is normally required to prepare the soil surface for planting after a complete subsoiling disruption operation.

As many producers have begun to realize the benefits of maintaining an adequate surface-residue cover, they have become concerned that the subsoiling operation may bury excessive amounts of crop residue. To avoid this, one method that has been widely adopted is called 'strip-tillage' or 'in-row subsoiling.' This process involves subsoiling directly beneath the row. A single shank is pulled through the soil directly beneath each row to loosen the soil only in close proximity to the crop. Tillage energy is not wasted on loosening zones between rows that are not necessary for crop roots. Areas between rows are mostly left undisturbed with full residue coverage. One significant difference between strip-tillage and subsoiling as conducted by a V-frame subsoiler is the condition of the soil as it is left by the implement. Strip-tillage implements usually include a method of closing the subsoiling slot left by the shank to prepare a seedbed because planting may immediately follow the strip-tillage operation (**Figure 6**).

When to Subsoil

The frequency of subsoiling is dependent primarily upon the particular cropping system and the soil's needs. In some areas of the world, benefits of subsoiling have been found up to 10 years later. However, in the southeastern USA, most research indicates that this tillage operation could be performed every other year without limiting yields. To a large degree, it depends upon the management system that is put in place following the subsoiling operation. When

Figure 5 V-frame subsoiler used for complete disruption of soil compaction.

Figure 6 Strip-tillage subsoiler used for subsoiling beneath rows. Note closing wheels, which prepare the seedbed for planting.

controlled traffic and in-row subsoiling are both used, a soil may be able to withstand the compaction forces generated by traffic on the soil surface. The forces that tires place on the soil surface in the row middles will not propagate sideways into the row area, because the hardened soil in the row middles will be able to withstand the traffic without deformation. This process is likened to tires running on cement pads which are located between rows in the field. As long as the tires are kept on the cement pads, little compaction occurs under the rows. The loosened soil beneath the rows will be able to maintain their loose structure for longer periods of time and may not need to be subsoiled as often. Even when subsoiling is performed, the shank runs in soil that was previously loosened and not recompacted and the subsoiling forces are reduced. New automatic steering systems that enable precise control of the tractor as it passes through the field could enhance the adoption of controlled traffic as a management tool and could reduce the need for annual subsoiling (Figure 7).

Figure 7 Crops grown in a controlled-traffic system, where subsoiled zones beneath rows are kept separate from compacted zones between rows. Note deep rooting beneath rows.

Because of the large energy requirements necessary for subsoiling, some producers subsoil only a portion of their field during a year. The next year, they alternate and till other parts. This incremental approach is continued until the entire field has been subsoiled and the operation begins again. This type of rotation may provide enough overall loosening of the soil to enable good productivity. However, if deep tillage is not performed during a year that had a moisture-limiting condition, yield reductions could occur. To reduce overall risk from drought effects on their crops, many producers who must subsoil do it on an annual basis.

What time of year to subsoil is largely dependent upon the producer's schedule. Many producers subsoil in the autumn after their harvest is complete. Subsoiling at this time of the year can be efficient, because farmers have several months to prepare for planting the next season's crop. Waiting until spring to subsoil can delay planting if adverse weather conditions exist. Subsoiling can also remove compaction and rutting caused by the harvesting process, which can excessively traffic the field with very large loads (usually defined as greater than 10 Mg per axle) from

the harvesting or transport equipment. Subsoiling in the autumn can also be helpful due to the ability to have a loosened soil profile during winter months when rainfall is greater for many climatic regions. This loosened soil profile has increased surface roughness and increased infiltration, which enables more rainfall to penetrate and to be stored for future crop use.

Many producers who practice in-row subsoiling wait until spring to subsoil. They subsoil immediately before planting to give the crop the maximum benefit from the subsoiling operation. Combination planters and subsoilers have been produced for a one-pass operation. Producers who typically use this cropping system subsoil on an annual basis because their soils recompact very easily. Measurements made in the southeastern USA have shown that subsoiling conducted in the autumn does not provide adequate soil loosening necessary for optimum crop yields due to the soil's natural ability to reconsolidate over the winter months.

There is one disadvantage to subsoiling in the spring: occasionally when the subsoiling operation is closely followed by planting, a problem can develop relating to the proper emergence of seedlings. Intense rainfall events can cause the subsoiling channel to settle quickly and move the seed downward within the loosened zone. Replanting the crop may be necessary if the seeds have been excessively covered by soil.

Another consideration for deciding when to subsoil is the moisture content of the soil. Maximum disruption of the soil profile is usually provided when the soil is extremely dry, in contrast to subsoiling when the soil is almost saturated. However, subsoiling forces increase dramatically when soil is dry, and adequate tillage or tractive energy may not be available. A reasonable compromise seems to be to recommend subsoiling when soil moisture is near the permanent wilting point, but soil drying has not progressed to the point of the hygroscopic coefficient, when soil moisture is in vapor phase (i.e., soil moisture is bound tightly to soil solids and little is available for plant use). Some research indicates that little difference in soil disruption is measured between soil moisture of the hygroscopic coefficient point and that of permanent wilting point. However, the difference in tillage forces required for subsoiling between these two soil conditions can be significant.

Maintaining Surface-Residue Coverage

Consideration also has to be given to the amount of soil-surface disruption. Often producers are concerned with trying to reduce draft forces and they forget about leaving the soil in a smooth condition appropriate for planting. Also, efforts should be made to ensure that an adequate amount of crop residue is not buried by the subsoiling operation. This may be especially important to fragile residues when a crop such as soybean is followed by subsoiling. Research results also indicate that subsoiling when soil is near the hygroscopic coefficient causes maximum disturbance to the soil surface. Subsoiling near the permanent wilting point disrupts the soil surface less and probably results in less residue burial. Properly choosing a subsoiling shank that minimally disturbs crop residue and then operating it at the correct soil moisture and at a proper speed results in minimal surface disturbance and maximum subsurface disruption.

Even though most research on subsoiling has targeted force reduction, some early observations of the soil surface indicated that the overall effect of shank design had little effect on soil breakup. However, recent research indicates that bentleg subsoilers typically do a better job of maintaining surface residue than subsoilers, which are used for complete disruption. Consequently bentleg subsoilers have been readily adopted for use in many conservation tillage systems.

Subsoiling in Irrigated Fields

When irrigation is available, yield responses to subsoiling are less obvious. Increasing plant rooting depth by subsoiling may not be important, as the plants are likely to obtain all of their moisture through irrigation water. There are two advantages of using subsoiling with irrigation: the first is the ability of the soil to store additional moisture after subsoiling so less frequent irrigation is necessary; the second is that not all parts of a field respond positively to irrigation. Lower-lying areas of irrigated fields that are poorly drained show yield increases when subsoiled.

Subsoiling in Perennial Crops

Most subsoiling operations are conducted in row-cropping agriculture, although many other crops may experience potential benefits from this operation. In many forest locations a subsoiling operation is typically conducted before pine seedlings are planted. The subsoiling operation is conducted by large, single-row plows pulled either by bulldozers or log-skidders (Figure 8). Usually, this operation is conducted every 3 m and the trees are planted soon after the operation is completed. Several advantages are seen from subsoiling prior to forestry planting, for example: (1) pine tree seedlings gain from having

Figure 8 Subsoiling plow used for forest tillage to prepare soil before tree planting. Note backward sweep of subsoiler shank, which is used to allow the implement to pass over tree stumps.

compacted soil disrupted, enabling roots to penetrate to depths adequate for soil moisture; (2) the forest soil surface is smoother and more easily accessible to machinery; and (3) there is economic benefit from having the trees machine-planted rather than hand-planted.

Fields where other nonannual crops are grown such as pastures, vineyards, orchards, and fields used for sugar cane production that are prone to repeated heavy machinery traffic are frequently subsoiled prior to planting. After they are planted, the crops may be in place for many years and the opportunity for loosening is reduced. Subsoiling can be performed after the crop is established, but it is not advisable: plant growth and yields may be detrimentally affected due to excessive root pruning, potentially followed by periods of drought.

Considerations Before Subsoiling

Even though it may be possible to subsoil a field to remove compaction, care should be exercised before this potentially expensive operation is performed. Once soil is loosened by subsoiling, it will easily recompact if traffic is applied in the same area. Research indicates that two passes of a tractor in the subsoiled area will cause the soil to return to its previous state prior to subsoiling. If traffic is controlled, however, the benefits of subsoiling can be long-lasting.

Using a cover crop has been shown in some locations to replace the need for subsoiling. Winter cover crops are able to increase infiltration of winter rainfall and assist with water storage for use by the main cash crop the following summer. Evaporation of soil moisture is also hindered by the residue cover provided by the winter cover crop, even persisting several months later during summer months. The increased amount of soil moisture present under the cover crop reduces the overall soil strength and allows the plant roots to continue to grow downward. However, cover crops are not advisable under all climate and soil conditions, and subsoiling may still offer increased crop response in severely compacted soils.

The overall management of the production system should be examined to determine whether the soil compaction that is being alleviated by subsoiling is natural or whether it is traffic-induced. If it is natural, then subsoiling may have to be performed on an annual basis to give plants the maximum benefit of the operation. However, if a portion of the compaction is machine-induced, adoption of controlled traffic or a cover crop may enable the subsoiling operation to be performed less frequently.

Producers should be aware that some soils may not respond positively to subsoiling even though they may appear to be compacted. Yield improvements may not be realized in soils that are not severely

compacted or are not in need of subsoiling. Even some soils that are compacted may have adequate plant growth due to the presence of old root channels or significant earthworm activity and may not be improved significantly by subsoiling.

Summary

Subsoiling is a necessary tillage process for many fields used for crop production. Subsoiling disturbs the soil down to at least 350 mm and provides for increased rooting in soils compacted by either natural causes or by vehicle traffic. The potential success of subsoiling varies depending upon the design of the shank, the timing of the operation, the crop, the soil, and management decisions. Ongoing research is aimed at a better understanding of subsoiling so that producers can determine whether their soils might benefit from this tillage process.

List of Technical Nomenclature

draft force The horizontal force required to pull the implement through the soil

See also: **Compaction**; **Conservation Tillage**

Further Reading

ASAE (1998) *Agricultural Machinery Management Data D497.4*. ASAE Standards, pp. 360–367. St. Joseph, MI: American Society of Agricultural Engineers.

ASAE (1998) *Terminology and Definitions for Agricultural Tillage Implements SAE S414.1*. ASAE Standards, pp. 261–272. St. Joseph, MI: American Society of Agricultural Engineers.

ASAE (1999) *Procedures for Obtaining and Reporting Data with the Soil Cone Penetrometer EP542*. ASAE Standards, pp. 964–966. St. Joseph, MI: American Society of Agricultural Engineers.

ASAE (1999) *Soil Cone Penetrometer S313.2*. ASAE Standards, pp. 808–809. St. Joseph, MI: American Society of Agricultural Engineers.

Gill WR and Vanden Berg GE (1966) Design of tillage tools. In: Gill WR and Vanden Berg GE (eds) *Soil Dynamics in Tillage and Traction*, pp. 211–297. Washington, DC: USDA.

Larson WE, Eynard A, Hadas A, and Lipiec J (1994) Control and avoidance of soil compaction. In: Soane BD and van Ouwerkerk C (eds) *Soil Compaction in Crop Production*, pp. 597–625. Amsterdam, the Netherlands: Elsevier.

Nichols ML and Reaves CA (1958) Soil reaction to subsoiling equipment. *Agricultural Engineering* 39: 340–343.

SULFUR IN SOILS

Contents
Overview
Biological Transformations
Nutrition

Overview

M A Tabatabai, Iowa State University, Ames, IA, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

The average sulfur (S) content of the earth's crust is estimated to be between 0.06 and 0.10%. It is usually ranked as the 13th most abundant element in nature. Sulfur occurs in soils in organic and inorganic forms, with organic S accounting for more than 95% of the total S in most soils from the humid and semihumid regions. The proportion of organic and inorganic S in soils samples, however, varies widely according to soil type and depth of sampling. It is usually somewhat lower in subsurface than in surface soils.

Although it is well known that S in soils is present mainly in organic combinations, very little is known about the identities of these S compounds. The inorganic S fraction in soils may occur as sulfate (SO_4^{2-}) and compounds of lower oxidation state such as sulfide (S^{2-}), thiosulfate ($S_2O_3^{2-}$), tetrathionate ($S_4O_6^{2-}$), polysulfides (S_n^{2-}, where $n > 10$), sulfite (SO_3^{2-}), and elemental S (S^0). The last four are detected in soils treated with elemental S or certain pollutants. In well-drained, aerated soils, most of the inorganic S

normally occurs as sulfate, and the concentrations of reduced S compounds are generally 1%. There are several forms of sulfate in soils. These include easily soluble sulfate, adsorbed sulfate, insoluble sulfate, and sulfate coprecipitated (cocrystallized) with $CaCO_3$. Under anaerobic conditions, particularly in tidal swamps and poorly drained or waterlogged soils, the main form of inorganic S in soils is sulfide and, often, elemental S.

Carbon–Nitrogen–Phosphorus–Sulfur Relationships

Significant information is now available on the relationships between C, N, P, and S in soils around the world. Unlike P, which can be present in significant proportions as organic and inorganic combinations, inorganic N and S values are very small relative to organic forms of these elements in soils. Therefore, often the relationships between organic C, total N (instead of organic N), organic P, and total S (instead of organic S) are reported. Significant variation can occur in the C:N:P:S ratios of individual soils, but the mean ratios for groups of soils from different regions are similar. Agricultural soils, in general, have a mean C:N:P:S ratio of approximately 130:10:1.3:1.3. Soils under native grass have ratios of the order of 200:10:1:1. Peat and organic soils have intermediate ratios of approximately 160:10:1.2:1.2. These ratios are shown in Table 1 for six Brazilian surface soils and six Iowa surface soils.

Sources of Sulfur in soils

Minerals Sources

Many S-containing minerals occur in nature. The main S-bearing minerals in rocks and soils are present in two states: (1) as sulfate, such as in gypsum ($CaSO_4 \cdot 2H_2O$), anhydrite ($CaSO_4$), epsomite ($MgSO_4 \cdot 7H_2O$), and mirabilite ($Na_2SO_4 \cdot 10H_2O$); or (2) as sulfide, such as pyrite and marcasite (FeS_2), sphalerite (ZnS), chalcopyrite ($CuFeS_2$), cobaltite (CoAsS), pyrrhotite ($Fe_{11}S_{12}$), galena (PbS), arsenopyrite ($FeS_2 \cdot FeAs_2$), and pentlandite $(Fe,Ni)_9S_8$.

Fertilizers Sources

There are many fertilizer materials, both liquid and solid, which are used to supply S to growing crops. The S source selected for any particular situation is determined by the crop to be grown, the S level of the soil, the cost of the material, and the ease with which its use can be fitted into a particular fertilizer program.

Atmospheric Sources

Rainfall and the atmosphere constitute a third important source of S in soils. It is estimated that in the USA more than 25 million tons of SO_2 or 13 million tons of S are emitted annually into the atmosphere. Most of these amounts are derived from combustion of fossil fuels, but industrial processes such as ore smelting, petroleum-refining operations, and other such sources contribute approximately 20% of

Table 1 Carbon, nitrogen, phosphorus, and sulfur relationships in some Brazilian and Iowa surface soils

	Ratio				
Soil no.	*Organic C/total N*	*Total N/organic P*	*Total N/total S*	*Organic P/total S*	*Organic C/total N/organic P/total S (organic S)*
Brazilian soils					
1	21.0	11.8	3.4	0.3	210:10:0.9:3.0 (2.5)
2	21.3	9.7	7.0	0.9	213:10:1.0:1.4 (1.1)
3	12.8	13.3	8.3	0.7	128:10:0.8:1.2 (1.1)
4	13.9	3.3	6.1	2.0	139:10:3.1:1.6 (1.6)
5	23.6	18.7	12.0	0.7	236:10:0.5:0.8 (0.8)
6	23.8	10.2	6.3	0.7	238:10:1.0:1.6 (1.5)
Mean	19.4	11.2	7.2	0.9	194:10:1.2:1.6 (1.4)
Iowa soils					
7	9.4	6.2	9.1	1.6	94:10:1.6:1.1 (1.0)
8	10.8	7.0	9.2	1.3	108:10:1.4:1.1 (1.0)
9	10.4	7.0	6.9	1.0	104:10:1.4:1.4 (1.4)
10	10.3	8.0	7.4	0.9	103:10:1.2:1.3 (1.3)
11	12.4	8.2	8.0	1.0	124:10:1.2:1.3 (1.2)
12	13.1	7.6	6.9	0.9	131:10:1.3:1.5 (1.4)
Mean	11.1	7.3	7.9	1.1	110:10:1.4:1.3 (1.2)

Reproduced with permission from Neptune AML, Tabatabai MA, and Hanway JJ (1975) Sulfur fractions and carbon–nitrogen–phosphorus–sulfur relationships in some Brazilian and Iowa soils. *Soil Science Society of America Proceedings* 39: 51–55.

the total S emitted into the atmosphere. Another major source is volcanic activity around the world.

Chemical Nature of Organic Sulfur in Soils

Understanding the nature and properties of the organic S fractions in soils is important, because these compounds govern the release of plant-available S. Even though much of the organic S compounds in soils remains unidentified, three broad groups of S compounds are recognized. These groups have been classifed according to the nature of the reagents used or according to the groups of S compounds attacked by the reagents. Thus, three distinct groups of S-containing compounds have been identified (Figure 1):

1. Organic S that is not directly bonded to C and is reduced to H_2S by hydriodic acid (HI). This fraction is believed to be largely in the form of sulfate ester with C–O–S linkages. Examples of substances that contain these linkages include arylsulfate, alkylsulfates, phenolic sulfate, sulfated polysaccharides, choline sulfate, and sulfated lipids. Other organic sulfates could be present as sulfamates (C–N–S) and sulfated thioglycosides (N–O–S). On average, approximately 50% of the total organic S in humid and semihumid regions is present in this form, but it can range from 30 to 60% (Table 2). Values as high as 95% have been reported for Iowa subsoils. Unless otherwise indicated, the HI-reducible S includes the inorganic SO_4^{2-} fraction (Table 3);
2. Organic S that is directly bonded to C (C–S) and is reduced to inorganic sulfide by Raney Ni (50% each of Ni and Al powder) in an alkaline medium (NaOH). This fraction is believed to consist largely of S in the form of S-containing amino acids such as methionine and cysteine. Its concentration in soils ranges from 10 to 30% of the total organic S (Tables 2 and 3);

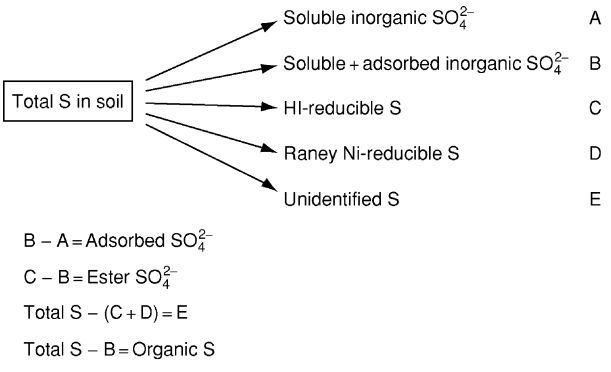


Figure 1 Organic and inorganic sulfur fractions in soils.

3. Organic S that is not reduced by either of the reagents employed in estimation of fractions 1 and 2. This unidentified fraction is inert to HI and Raney Ni. It is generally in the range of 30–40% of the total organic S (Tables 2 and 3). It is very stable, because it resists degradation by caustic chemical reagents; therefore, this fraction is of little importance as a potential source of S for plants.

Inorganic Sulfur in Soils

Inorganic S in soils may occur as sulfate and as compounds of lower oxidation states such as sulfide, polysulfides, sulfite, thiosulfate, and elemental S. In well-drained, well-aerated soils, most of the inorganic S normally occurs as sulfate, and concentrations of reduced S compounds are barely traceable, if present at all. The concentration of soluble SO_4^{2-} is estimated after extraction with 0.01% LiCl or 0.15% $CaCl_2$ solution, and the concentration of the soluble plus the insoluble fraction is estimated after extraction with a $Ca(H_2PO_4)_2$ solution containing 500 mg P l^{-1} (Table 2).

Fate of Inorganic Sulfate in Soils

The inorganic sulfate in soils may occur as water-soluble salts that can be leached from soils, adsorbed by soil colloids, or insoluble forms. Soil chemical properties such as pH, type of clay, and presence of cations are important factors in governing the leaching and adsorption of SO_4^{2-} in soils. The mechanism of SO_4^{2-} adsorption by soils involves coordination with hydrous oxides, exchange on edges of silicate clays, and molecular adsorption. Both the water-soluble and adsorbed SO_4^{2-} are available to plants. Under anaerobic (waterlogged) conditions, SO_4^{2-} is reduced to S^{2-} and precipitated as metal sulfide. This process is biological in nature.

Leaching losses

The movement of sulfate in soils determines the magnitude of losses of S in drainage water. Transport of sulfate within a soil profile is influenced by its concentration in soil solution, its reaction with the solid-phase components, and movement, velocity, and pattern of water movement within the soil. The relative magnitude of, and interactions among, these factors determines the physicochemical fate of the sulfate released from mineralization of organic S or added to soils as fertilizer, crop-residue decomposition, irrigation waters, and atmospheric deposition. From the information available on sulfate losses with percolating water, the following conclusions can be drawn:

Table 2 Total sulfur and percentage distribution of various forms of sulfur in some Brazilian and Iowa surface soils

		Percentage of total soil S in form specified						
		Inorganic sulfate S						
Soil no.	*Total S (mg kg^{-1} soil)*	*LiCl*	*Ca(H_2 PO_4)$_2$*	*Hl-reducible S*	*Ester sulfate S*	*Carbon-bonded S*	*Unidentified organic S*[a]	*Total organic S*
Brazilian soils								
1	59	5.1	15.3	35.6	20.3	5.1	59.3	84.7
2	43	9.3	23.3	51.2	27.9	4.7	44.1	76.7
3	72	6.9	12.5	52.8	40.3	11.1	36.1	87.5
4	214	3.7	5.1	70.1	65.0	5.6	24.3	94.9
5	209	1.4	4.8	53.6	48.8	12.4	34.0	95.2
6	398	2.0	6.3	43.0	36.7	5.0	52.0	93.7
Mean	166	4.7	11.2	51.1	39.8	7.3	41.7	88.8
Iowa soils								
7	55	7.3	7.5	50.9	43.4	18.2	30.9	92.5
8	174	5.8	4.9	42.9	47.0	9.2	37.9	94.1
9	331	5.4	5.5	57.7	52.2	12.1	30.2	94.5
10	338	2.4	2.5	55.3	52.8	8.9	35.8	97.5
11	438	1.8	1.8	61.6	59.8	7.3	31.1	98.2
12	580	3.5	3.5	48.6	45.1	12.1	39.3	96.5
Mean	319	4.4	4.3	52.8	50.1	11.3	34.2	95.6

[a]Unidentified organic S was calculated from total S − (Hl-reducible S + carbon-bonded S).

Reproduced from Neptune AML, Tabatabai MA, and Hanway J (1975) Sulfur fractions and carbon–nitrogen–phosphorus–sulfur relationships in some Brazilian and Iowa soils. *Soil Science Society of America Proceedings* 39: 51–55.

Table 3 Fractionation of sulfur in surface soils from different regions

	Percentage of total S in form specified					
	HI-reducible		*C-bonded*[a]		*Unidentified*	
Location[b]	*Range*	*Mean*	*Range*	*Mean*	*Range*	*Mean*
Alberta, Canada (15)	25–74	49	12–32	21	7–45	30
Australia (15)	32–63	47	22–54	30	3–31	23
Brazil (6)	36–70	51	5–12	7	24–59	42
Iowa, USA (34)	36–66	52	5–20	11	21–53	37
Quebec, Canada (3)	44–78	65	12–32	24	0–44	11

[a]Determined by reduction with Raney Ni.
[b]Numbers in parentheses indicate number of samples.
Reproduced from Tabatabai MA (1984) Importance of sulphur in crop production. *Biogeochemistry* 1: 45–62.

1. Losses of SO_4^{2-} are reduced by cropping and are less with rooting and perennial crops than with annual crops;
2. Leaching losses of SO_4^{2-} are greatest when monovalent ions such as K^+ predominate; next in order are the divalent ions such as Ca^{2+} and Mg^{2+}, and leaching losses are minimal when soils are acid and appreciable concentrations of Fe and Al hydrous oxides are present;
3. Under comparable soil and cropping conditions, the amount of SO_4^{2-} removed from the soil profile is generally directly related to the amount of leachate;
4. Sulfate adsorption would lead to more rapid and complete removal of Cl^- from acid soils;
5. Sulfate losses increase with liming or amendment with phosphate;
6. Sulfate losses are less when the S fertilizer is banded than when broadcast.

Complete separation of the influence of physical and chemical soil properties on transport of sulfate is not possible, because several factors affect the distribution of an ion such as sulfate in the soil pores. These include: (1) the electric field surrounding the individual soil particles, and (2) the magnitude of the repulsive (or attraction) forces which are dependent on the mineralogy and chemical composition of the solid phase, and the pH and salt content of the aqueous phase.

Losses of S by leaching vary widely: some drainage water contains more S than the rain supplies, even though little or none is added in fertilizer. The extra quantity may be deposited directly on plants and soils from the atmosphere or released from soil organic matter or minerals. Expressed in kilograms of sulfur per hectare, the annual losses from unfertilized fields by drainage water in the state of Illinois range from 1.5 to 65 kg S ha^{-1}, in Germany they average 33 kg S ha^{-1}, in Europe and North America they average 15 kg S ha^{-1}, in South America they average 4.5 kg S ha^{-1}, and in some areas of Australia they are less than 1 kg S ha^{-1}. It has been estimated that between 3 and 32 kg S ha^{-1} is lost by tile drainage in the state of Iowa. In general, the annual loss from soils by leaching varies from insignificant amounts to as much as 320 kg S ha^{-1} from soils treated with S fertilizers.

Sulfate Adsorption by Soils

Soils vary widely in their capacity to adsorb sulfate. Because sulfate adsorption occurs at low pH values (less than 6), its adsorption is negligible in most agricultural soils ($pH > 6$). Its adsorption in subsurface acid horizons plays an important part in contributing to S requirement of crops; conserving S from excessive leaching, and in determining S distribution in soil profiles. Sulfate adsorption is a reversible process and is influenced by a number of soil properties. These include:

1. *Clay content and type of clay mineral*. Sulfate adsorption usually increases with clay content of soils. Kaolin minerals retain more sulfate than montmorillonite clays;
2. *Hydrous oxide of Al and Fe*. Hydrous oxides of Al, and to a lesser extent of Fe, show marked tendencies to retain sulfate, especially the former in certain soils;
3. *Soil horizon or depth*. Most soils have some capacity to adsorb sulfate. The amounts of sulfate adsorption in surface horizons may be low but are often greater in lower soil horizons due to the presence of more clay and Fe and Al oxides;
4. *Soil pH*. Sulfate adsorption in soils is favored by strong acid conditions. It becomes almost negligible at $pH > 6$;
5. *Sulfate concentration and temperature*. The amount of sulfate adsorbed is concentration- and temperature-dependent. The amount of sulfate adsorbed is in kinetic equilibrium with sulfate concentration

in solution. Temperature has a relatively small effect on sulfate adsorption by soils;

6. *Effect of time*. Sulfate adsorption increases with the length of time it is in contact with soil;

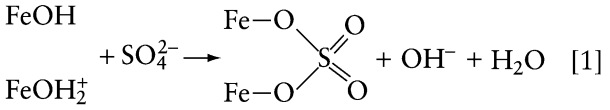
7. *Presence of other anions*. Sulfate is weakly held by soils. The retention decreases in the following order: hydroxyl > phosphate > sulfate = acetate > nitrate = chloride. Phosphate will displace or reduce the adsorption of sulfate;

8. *Effect of cations*. The amount of sulfate adsorbed is affected by the associated cation or by the exchangeable cation following the order: $H^+ > Sr^{2+} > Ba^{2+} > Ca^{2+} > Mg^{2+} > Rb^+ > K^+ > NH_4^+ > Na^+ > Li^+$. Both the cation and the sulfate from the salt are retained, but the capacity of adsorption of sulfate is different from that of the associated cation.

Mechanisms of Sulfate Adsorption by Soils

Several mechanisms have been proposed for adsorption of sulfate by soils.

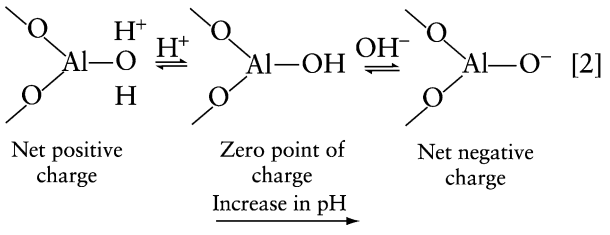
Coordination with hydrous oxides In acid soils sulfate adsorption essentially involves the chemistry of Fe and Al. The hydrous Fe and Al oxides tend to form coordination complexes due to the donor properties of oxygen. The adsorption involves the replacement of two surface OH groups (or OH_2^+) by one SO_4^{2-}. The two O atoms of the SO_4^{2-} are each bound to a different Fe^{3+}, resulting in a binuclear bridging surface complex:



[1]

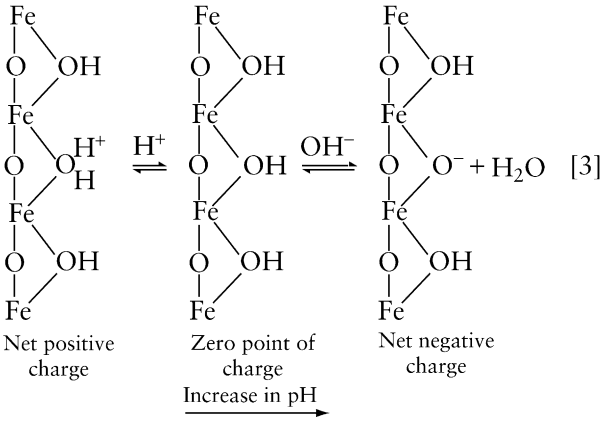
This reaction has been demonstrated to occur on the surfaces of goethite (α-FeOOH), and, because soils contain Fe oxides, it is assumed that a similar reaction takes place in soils under acid conditions.

The effect of pH on SO_4^{2-} adsorption by hydrous oxides in soil must consider the zero point of charge (the pH at which the change on the surface is zero). Deviation from this pH value involves protonation or deprotonation of Al oxides:



[2]

A similar protonation or deprotonation of Fe oxides is possible:



[3]

The positively charged Al and Fe hydrous oxides adsorb the SO_4^{2-} ions.

In addition to the mechanisms described above, homoionic Al-saturated clay coated with hydrated oxides R (Fe and Al) may adsorb sulfate ions as follows:

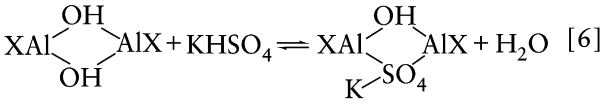
$$ {}^{y}K^+ + Al_x\text{—clay} + {}^{y}H_2O \longrightarrow Al_x(OH)_y^{Ky}\text{—clay} + {}^{y}H^+ \quad [4] $$

$$ SO_4^{2-} + R_x(OH)_y\text{—clay} \longrightarrow R_x[(OH)_{y-z}(SO_4)_z]\text{—clay} + {}^{z}OH^- \quad [5] $$

In this mechanism, it is assumed that K ion adsorption sites developed from the exchange and/or hydrolysis of Al on the clay surface. As a result of this hydrolysis, some Al ions are released into the solution. At the same time sulfate ions replace the OH ions from R(OH) coating on clay and substitutes for them. The replaced OH ions in turn react with H ions. According to this mechanism, whether the pH of the system increases or decreases depends on the relative rates of the two reactions: Al hydrolysis and OH ion exchange. It has been demonstrated that sulfate ion adsorption increases when sulfate-adsorbing soils are coated with Fe and Al oxides.

The adsorbed sulfate ions may be replaced by other anions of greater penetration (coordination) ability such as PO_4^{3-}. This has been demonstrated by displacing the adsorbed SO_4^{2-} from surface soils by application of phosphate fertilizers.

Exchange on edges of silicate clays This exchange presumably involves replacement of OH^- by SO_4^{2-} in terminal octahedral coordination with Al. The general effects and mechanism are similar to that discussed above for Al and Fe hydrous oxides. The following type of reaction can occur with both hydrated Al oxides and the Al layer kaolinite:



[6]

Molecular adsorption This mechanism is less understood than the others described above. It implies that SO_4^{2-} is adsorbed by some mechanism by which the associated cation is retained by soils. This mechanism has been referred to as 'salt' adsorption, 'molecular' adsorption, and imbibition.

Sulfur Transformations in Soils

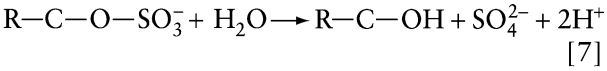
The transformations of S in soils are many and varied (i.e., oxidation, reduction, volatilization, decomposition and mineralization of plant and microbial residues), and often the changes are cyclic S changes from inorganic to organic forms (immobilization) and back again by living organisms.

Mineralization

The conversion of an element from organic form to an inorganic state as a result of microbial activity is termed 'mineralization.' As is the case with carbon and nitrogen, organic S in soils is mineralized to inorganic forms, mainly SO_4^{2-}, the form taken up by plant roots. The mechanisms involved in this transformation, however, are not clear. It appears that microorganisms are involved in this process, where they obtain their energy from the oxidation of carbonaceous materials in soils. During this process organic S is mineralized. Some of the mineralized S is used for synthesis of new microbial cell materials (immobilization), because the portion not required for synthesis is released as inorganic S. Mineralization and immobilization occur simultaneously in soils whenever organic material is undergoing microbial decomposition. The effect of temperature on the rate of N and S mineralization in 12 Iowa surface soils is shown in Table 4.

Sources of mineralizable S It is believed that ester sulfates in soils are the main sources of S mineralization in soils. However, carbon-bonded S (C–S) cannot be excluded, because this fraction contains the amino acids methionine, cystine, and cysteine, which can be converted to inorganic sulfate under aerobic conditions. The information available suggests that this form of S in soils can be a source for plant uptake.

Role of arylsulfatase in S mineralization Because a large proportion of the organic S in soils appears to be present as ester sulfate, it seems reasonable to expect that some organic S is mineralized by the action of arylsulfatase enzyme; indeed, this enzyme has been detected in soils. This enzyme catalyzes the hydrolysis of ester sulfate, releasing sulfate for plant uptake. The reaction is as follows:



[7]

Table 4 Comparison of rates of nitrogen and sulfur mineralization in Iowa surface soils; soil glass-bead columns were incubated at 20°C or 35°C and the mineral N and S produced were determined after leaching every 2 weeks with 0.01 mol l^{-1} KCl for a total of 26 weeks

	Rate of mineralization (kg ha^{-1} week^{-1})					
	20°C			*35°C*		
Soil	*N*	*S*	*N:S*	*N*	*S*	*N:S*
Lester	6.7	1.6	4.2	22.6	4.9	4.9
Ackmore	4.9	1.6	3.1	27.6	4.9	5.6
Fayette	5.6	1.8	3.1	23.5	5.2	4.5
Downs	8.1	2.7	3.0	35.9	7.0	5.1
Clarion	6.7	1.8	3.7	26.7	7.8	3.4
Muscatine	7.4	1.8	4.1	33.2	6.7	5.0
Nicollet	3.8	1.3	2.9	17.3	4.0	4.3
Tama	9.0	2.2	4.1	34.3	7.2	4.8
Webster	9.4	2.2	4.3	38.1	6.5	5.9
Canisteo	4.9	1.3	3.8	20.9	4.0	5.2
Harps	4.3	1.2	3.6	21.7	3.6	6.0
Okoboji	6.3	1.8	3.5	34.3	6.1	5.6
Mean	6.4	1.8	3.6	28.0	5.7	5.0

Reproduced from Tabatabai MA (1984) Importance of sulphur in crop production. *Biogeochemistry* 1: 45–62.

Pattern of sulfate release Because the opposing reactions of mineralization and immobilization can occur simultaneously, different patterns of SO_4^{2-} release have been observed, depending on the energy materials available for the microorganisms (Figure 2):

1. Immobilization of S during the initial stages of incubation followed by SO_4^{2-} release;
2. A steady linear release of SO_4^{2-} over the whole period of incubation;
3. A rate of release that decreases with incubation time.

The pattern of SO_4^{2-} release is not related to any specific soil properties, but is apparently caused by adjustment of the microbial populations to the incubation conditions and to the availability of the initial substrates.

Factors affecting sulfur mineralization Because S mineralization is microbiological in nature, any variable that affects microbial growth should affect S

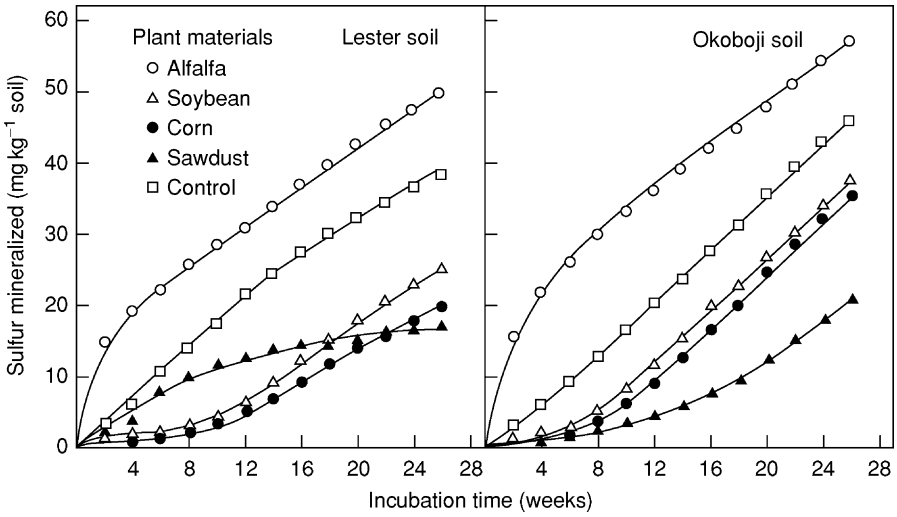


Figure 2 Effect of crop-residue treatment on sulfur mineralization in soils. (Reproduced from Tabatabai MA and Al-Khafaji AA (1980) Comparison of nitrogen and sulfur mineralization in soils. *Soil Science Society of America Journal* 44: 1000–1006.)

mineralization. Therefore, temperature, moisture, pH, and the availability of nutrients are most important.

Oxidation of Elemental S in Soils

Elemental S is one of the main sources of S added to soils. Before it can be untilized by crops, however, elemental S has to be oxidized to sulfate. Elemental S is oxidized in soils by chemical and biochemical processes, and a number of factors affect these processes. Microbial reactions dominate the processes. The microorganisms involved in oxidation of elemental S in soils belong to three groups: (1) chemolithotrophs (e.g., members of the genus *Thiobacillus*); (2) photoautotrophs (e.g., species of purple and green S bacteria); and (3) heterotrophs, which include a wide range of bacteria and fungi. Those listed in groups (1) and (2) are mainly responsible for oxidation of reduced S compounds in aerobic soils. Phototrophic bacteria are the predominant organisms responsible for oxidizing S^{2-} at the soil–water interface in flooded soils and in the rhizosphere of rice plants. The major reduced forms of inorganic S found in elemental S-treated soils are S^0, S^{2-}, and the oxyanions $S_2O_3^{2-}$, $S_4O_6^{2-}$, and SO_3^{2-}. These anion are oxidized, ultimately to SO_4^{2-} (Figure 3). The reactions involved in oxidation of elemental S in soils seem to be as follows:

$$S^0 \longrightarrow S_2O_3^{2-} \longrightarrow S_4O_6^{2-} \longrightarrow SO_3^{2-} \longrightarrow SO_4^{2-} \quad [8]$$

It is not clear whether these reactions are biochemical, occurring as a result of microbial processes in soils, or whether some of the intermediate products formed are the results of abiotic reactions. This is

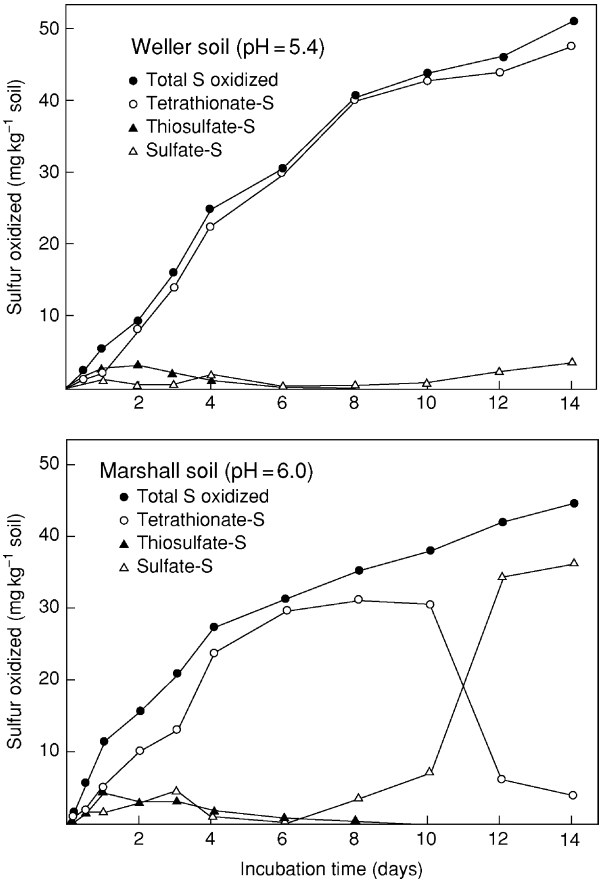


Figure 3 Amounts of thiosulfate-, tetrathionate-, and sulfate-S produced in two Iowa surface soils (Weller and Marshall soils) amended with elemental S (200 mg S kg^{-1} soil) and incubated at 30°C for various times. (Reproduced from Nor YM and Tabatabai MA (1977) Oxidation of elemental sulfur in soils. *Soil Science Society of America Journal* 41: 736–741.)

especially important in the case of the intermediates $S_2O_3^{2-}$ and $S_4O_6^{2-}$.

The enzyme rhodanese (thiosulfate-cyanide sulfotransferase; EC 2.8.1.1) catalyzes the conversion of the intermediate $S_2O_3^{2-}$ to SO_3^{2-} as follows:

$$S_2O_3^{2-} + CN^- \longrightarrow SO_3^{2-} + SCN^- \qquad [9]$$

Several factors affect of rate of oxidation of elemental S in soils. These include: (1) particle size: the finer the particles, the faster the reaction, which is because of the increase in surface area with decreasing particle size; (2) temperature: the higher the temperature, the greater the reaction rate (Figure 4); this is true between 10°C and 40°C; (3) time of contact with soil: the longer the reaction, the more the oxidation (Figure 5); and (4) effect of pH: the oxidation appears to be faster in alkaline than in acid soils.

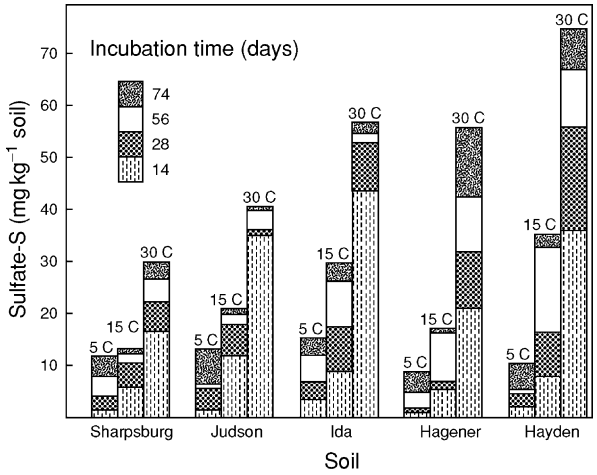


Figure 4 Effects of temperature and time of incubation on oxidation of elemental S (100 mg S kg^{-1} soil) in soils. (Reproduced from Nor YM and Tabatabai MA (1977) Oxidation of elemental sulfur in soils. *Soil Science Society of America Journal* 41: 736–741.)

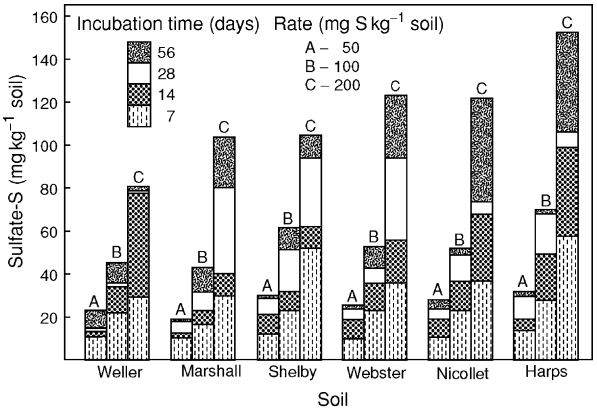


Figure 5 Effects of application rate and time of incubation on oxidation of elemental S in soils. The values for Weller and Marshall soils treated with 200 mg S kg^{-1} soil include the amounts of thiosulfate- and tetrathionate-S produced during 7 days and 14 days of incubation.

Reduction of Sulfate in Waterlogged Soils

The reduction of sulfate to H_2S is a process that occurs mainly by anaerobic bacteria; thus, it occurs only in anaerobic soils. This process is not important in aerobic agricultural soils, except perhaps in anaerobic microsites in soil aggregates. However, it is a major reaction in S cycling in waterlogged soils or periodic flooding, especially when in soils containing readily decomposable plant residues such as alfalfa (*Medicago sativa*). Bacterial reduction of sulfate involves either an assimilation or dissimilation process. In the former process, sulfate is reduced to the thiol (–SH) group of organic compounds for protein synthesis. In the latter process, the reduction leads to production of H_2S under very low redox potential (E_h) values. Under normal conditions, however, H_2S is not volatilized from soils, because it precipitates with Fe^{2+}, Mn^{2+}, Cu^{2+}, Cu^+, and/or Zn^{2+} present in soils. In the case of Fe^{2+}, it forms ferrous sulfide (FeS), and pyrite (FeS_2) is formed in severely reducing conditions by the reduction of sulfate to S^{2-} by the bacteria *Desulfovibrio desulfuricans*, which reacts with FeS to produce FeS_2.

Volatilization of S Compounds from Soils

Relative small amounts, if any, of S-containing gases, including H_2S, are released from aerobic agricultural soils, even when such soils are waterlogged. Substantial amounts of H_2S are released, however, from salt-marsh soils. Several gases are released from soils when treated with animal manures, sewage sludges, and protein-rich plant materials such as alfalfa, especially under waterlogged conditions. These include carbon disulfide (CS_2), which results from decomposition of the amino acids cystine and methionine; carbonyl sulfide (COS), released during decomposition of thiocyanate and isothiocyanate; and methyl mercaptan (CH_3SH), dimethyl sulfide (CH_3SCH_3), and dimethyl disulfide (CH_3SSCH_3), which result from decomposition of methionine-containing materials.

See also: **Acid Rain and Soil Acidification**; **Enzymes in Soils**; **Fertility**; **Greenhouse Gas Emissions**; **Minerals, Primary**; **Organic Matter:** Principles and Processes; **Organic Residues, Decomposition**; **Sorption:** Oxyanions; **Sulfur in Soils:** Biological Transformations; Nutrition

Further Reading

Freney JR (1967) Sulfur-containing organics. In: McLaren AD and Peterson GH (eds) *Soil Biochemistry*, vol. 1, pp. 229–259. New York: Marcel Dekker.

Germinda JJ and Gupta VVSR (1992) Biochemistry of sulfur cycling in soil. In: Stotzky G and Bollag J-M (eds) *Soil Biochemistry*, vol. 7, pp. 1–53. New York: Marcel Dekker.

McLachlan KD (ed.) (1975) *Sulphur in Australasian Agriculture*. Sydney, Australia: Sydney University Press.

Neptune AML, Tabatabai MA, and Hanway JJ (1975) Sulfur fractions and carbon–nitrogen–phosphorus–sulfur relationships in some Brazilian and Iowa soils. *Soil Science Society of America Proceedings* 39: 51–55.

Nor YM and Tabatabai MA (1977) Oxidation of elemental sulfur in soils. *Soil Science Society of America Journal* 41: 736–741.

Tabatabai MA (1984) Importance of sulphur in crop production. *Biogeochemistry* 1: 45–62.

Tabatabai MA (1985) Effect of acid rain on soils. *CRC Critical Reviews in Environmental Control* 15: 65–110.

Tabatabai MA (ed.) (1986) *Sulfur in Agriculture*. Agronomy monograph no. 27. Madison, WI: American Society of Agronomy.

Tabatabai MA (1987) Physicochemical fate of sulfate in soils. *Journal of the Air Pollution Control Association* 37: 34–38.

Tabatabai MA (1994) Soil enzymes. In: Weaver RW, Angle JS, and Bottomley PS (eds) *Methods of Soil Analysis*. part 2, pp. 775–783. *Microbiological and Biochemical Properties*. Book series no. 5. Madison, WI: Soil Science Society of America.

Tabatabai MA (1996) Sulfur. In: Sparks DL (ed.) *Methods of Soil Analysis*. part 3, *Chemical Methods*, pp. 921–960. Book series no. 5. Madison, WI: Soil Science Society of America.

Tabatabai MA and Al-Khafaji AA (1980) Comparison of nitrogen and sulfur mineralization in soils. *Soil Science Society of America Journal* 44: 1000–1006.

Tisdale SL, Nelson WL, and Beaton JD (1985) *Soil Fertility and Fertilizers*, 4th edn. New York: Macmillan.

Biological Transformations

S D Siciliano and J J Germida, University of Saskatchewan, Saskatoon, SK, Canada

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

In soil, sulfur exists in a variety of forms, some of which are beneficial to the environment, and other forms that are typically considered pollutants. The wide array of oxidation states and forms of sulfur results in a rich diversity of biological transformations involving this important element. These transformations are critical for many basic ecological processes such as plant growth, but also influence the movement of environmental contaminants.

The biological transformations of sulfur can be divided into four major classes: (1) oxidation; (2) reduction; (3) assimilation; and (4) mineralization (Figure 1). In addition, microbial metabolites can chemically fix or solubilize inorganic S compounds. The first two classes are related to the acquisition of energy for growth and the last two are typically related to nutrient acquisition. For energy acquisition, sulfur can act as an electron acceptor, an electron donor, and sometimes both at the same time. Sulfate (SO_4^{2-}) acts as an electron acceptor and is reduced to S^{-2} by a group of bacteria known as sulfate-reducing bacteria. These bacteria grow in the absence of oxygen. In contrast, S^{-2} can act as an electron donor and is oxidized by a group of bacteria known as sulfur oxidizers. These bacteria grow in the presence of oxygen. Interestingly, some bacteria can disproportionate sulfur, i.e., ferment sulfur. In this reaction, one of the two sulfur atoms, present in thiosulfate as S^{-1}, acts as an electron donor and the other sulfur atom, present as S^{+5}, acts as an electron acceptor.

The last two classes of sulfur transformation are directly linked to the nutritional status of the cell. Mineralization is a transformation in which organic sulfur contained in a growth substrate is released in a mineral form such as sulfate. In contrast, assimilation is a process is which sulfur present in the environment is incorporated into biomass. In the former case, the mineralized sulfur is available for other organisms to use, whereas in the latter the assimilated sulfur is no longer available to other organisms. Each of these transformations of sulfur has important side-effects on the environment. The mineralization of sulfur provides sulfate, which can potentially stimulate sulfate-reducing bacteria that play important roles in the arsenic and mercury biogeochemical cycles. The oxidation of reduced sulfur by sulfur oxidizers results in acid mine drainage, one of the primary detrimental impacts of mining.

The Global Sulfur Cycle

The global cycle of sulfur begins with sulfur volatilized largely as dimethylsulfide from marine algae, marshlands, mud flats, plants, and soil entering the atmosphere. This dimethylsulfide is converted photochemically to methanesulfonic acid, which then deposits on to the Earth and is rapidly converted by bacteria to carbon dioxide and sulfate. This sulfate

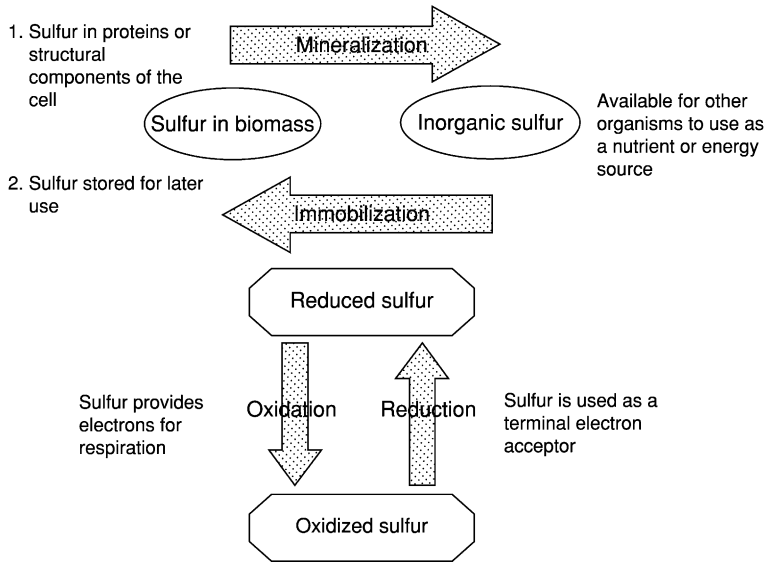


Figure 1 The four principal biological transformations of sulfur.

is assimilated by plants and immobilized into compounds like sulfoquinovosyl diacylglycerol, or alternatively the sulfur is released back to the atmosphere as dimethylsulfide. As plants die, bacteria metabolize the immobilized sulfur present in plant biomass and either immobilize this sulfur into their biomass or release it as sulfate for uptake by plants. In addition to this natural sulfur cycle, human activities have at least doubled the input of sulfur to terrestrial systems. Approximately 1.5×10^{11} kg S per year is deposited due to atmospheric pollution, largely in the form of sulfuric acid (sulfate), with small amounts of sulfite and biosulfite also being deposited. This deposited sulfate can either be used by bacteria as a terminal electron acceptor or assimilated by organisms requiring sulfur as a nutrient. Despite this recent disruption of the global sulfur cycle, the sulfur levels present in soil are largely due to pedogenic factors like climate, vegetation, and parent material. As a result, sulfur levels in soil range from 0.002 to 10%: most of this sulfur, 90%, is found in organic form.

The Biological Availability of Sulfur

Although 90% of sulfur in soil is present as organic sulfur, much of this organic S is not available for assimilation or dissimilation by microorganisms and plants. Microorganisms need to convert the organic sulfur in soil into sulfate before plants can take up the sulfate as a nutrient. This conversion of sulfate from organic sulfur is the end result of a complicated biogeochemical cycle. Sulfur enters the soil ecosystem and is converted and assimilated by microorganisms into their biomass. Predators then consume these bacteria and excess sulfur is released into the soil solution. The sulfur released into the soil solution is available for uptake by other organisms or plants. However, the predators do not mineralize all the S present in prey organisms and, as a result, a small fraction of S enters what is termed the 'resistant' pool of organic S present in soil (Figure 2). This resistant pool of organic S is found is large polymers that are in close association with clays. These large polymers resist microbial attack because it is not energetically favorable for microbes to attack these polymers. There are three distinct pools of sulfur in soil: (1) inorganic sulfate that is readily taken up by plants and microbes alike; (2) short-lived organic sulfur compounds like taurine and cysteine which are

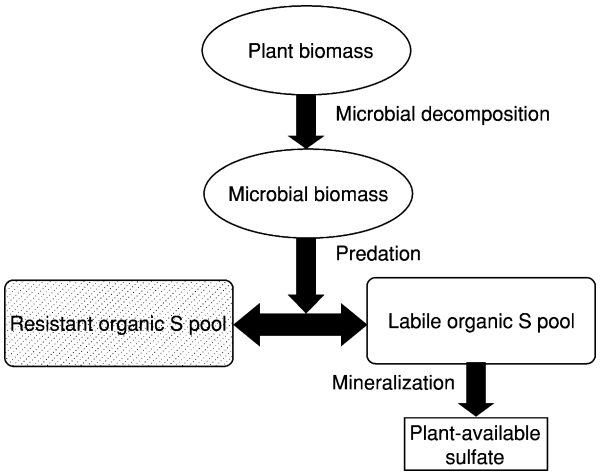


Figure 2 The different sulfur pools present in soil and their relationship to plant-available sulfur.

mineralized by microorganisms and either released as sulfate or incorporated into biomass; and (3) long-lived resistant organic sulfur polymers found in soil humus. Microbial transformations are the primary process by which sulfur is transferred between the pools of readily available, young organic sulfur and old organic sulfur. These transformations are discussed below.

Mineralization

Almost exclusively, microorganisms mediate sulfur mineralization. By definition, mineralization is the metabolic conversion of an organic form of an element to an inorganic form. Mineralization processes can be divided into cell-mediated or extracellular enzymatic processes. Cell-mediated mineralization occurs by oxidative decomposition under aerobic conditions or by desulfurization under anaerobic conditions. In cell-mediated mineralization, S present in carbon-containing compounds is mineralized as organisms consume the carbon to obtain energy. Thus, if there is more S in the substrate than the organism requires, the S will be mineralized and released to the environment. If there is not enough S, then any released S is consumed and immobilized by the organism. The break-even point for mineralization can be calculated based on the C:S ratio of the substrate, the C:S ratio of the organism, and the yield coefficient, i.e., how much of the consumed C is incorporated into biomass. As a general rule of thumb, if the C:S ratio of a substrate is 200 or less, then S is released. If the C:S ratio is greater than 400, then S is not going to be released because the microorganisms require the S for their own biomass. In this latter case, organisms rely on a second, enzymatic method of mineralization to obtain the sulfur essential for growth.

Enzymatic mineralization involves extracellular enzymes, such as arylsulfatases, released by the microorganism. These enzymes, once released from cells, will hydrolyze sulfate esters present in the soil. This process releases sulfate into the soil solution for use by the cells. The activity of these enzymes depends on a wide range of soil factors with reaction rates varying by a factor of 4 between different soils. These enzymes attack the vast amount of S present in the soil as organically bound S. Organic S in soil constitutes more than 90% of the total S present in soils. This organic S can be grouped into carbon-bonded S and organic sulfates. Organic sulfates, which comprise between 30 and 75% of sulfur in soil, typically include compounds such as sulfate esters (C-O-S), sulfamates (C-N-S), and sulfated thiogylcosides (N-O-S), and it is this form of sulfate that enzymatic mineralization processes attack.

Mineralization processes are biological transformations of sulfur that are intimately linked to cellular growth. Thus, factors that influence microbial activity, such as moisture, temperature, or plant growth, will influence mineralization. Increasing the growth of microorganisms by either adding a carbon source or plant growth typically results in reduced mineralization.

Assimilation

Assimilation is the process of converting inorganic sulfur into organic sulfur present in an organism. This assimilatory process is intimately linked to mineralization. In fact, these two processes, assimilation and mineralization, occur simultaneously. If the net effect is to release inorganic sulfate then this is termed mineralization; if instead the net effect is for the sulfur to be incorporated into the biomass, then this is termed assimilation. Thus, the first step of many assimilation processes is the conversion of sulfur into a form that can be incorporated into biomass. Taken alone, such steps might be considered mineralization processes but we term them assimilatory processes if: (1) the sum of the steps is to incorporate sulfur into biomass and (2) they are induced by sulfur limitation, suggesting that the organism is using the process to obtain sulfur. There are different biological assimilatory transformations that can occur depending on what form the available sulfur is in. For inorganic SO_4^{2-}, an enzyme called adenosine triphosphate (ATP) sulfurylase mediates the reaction between sulfate and ATP and sulfate to form adenosine-5′-phosphosulfate (APS) (Figure 3). In turn, this APS is transformed by APS kinase with another ATP molecule to form adenosine-3′-phosphate-5′-phosphosulfate (PAPS). This high-energy molecule reacts with PAPS reductase and two sulfhydryl groups to form sulfite, and sulfite is reduced by $NADPH_2$ to form sulfide which is incorporated into L-cysteine by *O*-acetylserine sulfhydrylase.

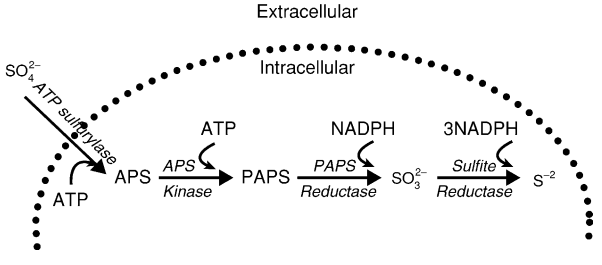


Figure 3 Assimilation of inorganic sulfur by bacteria. See text for a complete description of the reactions involved and definition of abbrevations. The sulfide produced is incorporated into cysteine by cysteine synthase.

This pathway will efficiently scavenge excess sulfate present in the soil solution that organisms require. So much so, that in the presence of an energy source, added SO_4^{2-} S is quickly incorporated into the organic fraction of the soil biomass and later found in the fulvic acid fractions. However, sulfate is not commonly found in soil solution and organisms have developed other pathways to assimilate sulfur for their metabolic needs.

Aromatic sulfonates, SO_3^{1-} groups bound to aromatic rings, can also serve as a sulfur source for assimilation. This assimilatory pathway is controlled by a genetic cluster called the 'sulfate-stimulation-induced stimulon' that encodes three distinct gene clusters, *asf*, *ssu*, and *ats*. Together these three genes control the assimilation of S by bacteria in soil. The *ats* gene cluster is responsible for the binding of aromatic sulfonates and transportation into the cell. The *ssu* genes encode for the cleavage for the C-S bond by a monooxygenase. The *asf* gene cluster provides the reducing equivalents from the oxidation of NADH necessary to assimilate aromatic sulfonates. The overall result of this pathway is the production of sulfite, which reacts with PAPS reductase to form sulfide, as described above. Alkane sulfonates are also a source of sulfur for assimilation. For most alkane sulfonates the pathway is very similar to that described above, with the exception that the transport mechanisms differ. There is a distinctly different pathway for taurine, involving an α-ketoglutarate-dependent dioxygenase which oxidizes both taurine and α-ketoglutarate to release sulfite. This pathway is regulated by *Tau*D and is specific for taurine, with little activity seen for other alkane sulfonates.

Sulfate esters are assimilated by the action of a group of enzymes termed 'arylsulfatases.' These enzymes are broadly grouped by their pH optima, with alkaline sulfatases having a pH optimum of 8.3–9.0 and acid sulfatases having a pH optimum of 6.5–7.1. In essence, these enzymes hydrate the S-O bond and release sulfate, which can then be assimilated by the pathway described above for sulfate. These assimilatory processes should be clearly differentiated from processes that use organo-sulfur compounds as a C and energy source. Organo-sulfur assimilation genes are regulated by the sulfate-stimulation-induced stimulon, which in turn is regulated by levels of cysteine, thiosulfate, and sulfite present in the cell. The presence of these compounds inside the cell will downregulate the sulfate-stimulated-induced genes and thereby repress expression of assimilatory enzymes. This is a logical regulatory control, since if there is excess cysteine and sulfate present in the cell, the cell likely does not need to assimilate sulfur.

Sulfur Oxidation

Many organisms, such as chemoautotrophs and photoautotrophs, use sulfur as a source of energy. This occurs by oxidizing S^0 to SO_3^{2-} and finally to SO_4^{2-} and in the process stripping six electrons from sulfur and using oxygen or nitrate as a terminal electron acceptor. A great variety of thiobacilli can perform these reactions. Some thiobacilli are chemoautrophs (only able to use S as an energy source), others are facultative chemoautotrophs (able to use C as an energy source if necessary) and some are mixotrophs (able to use C and S as an energy source at the same time). Most thiobacilli require the presence of oxygen but some, such as *Thiobacillus denitrificans*, can use nitrate as a terminal electron acceptor. Similarly, some thiobacilli can use Fe^0 in addition to sulfur as an electron donor.

The research on the importance of thiobacilli on sulfur oxidation comes largely from pure culture studies. The actual situation in soil involves a wide diversity of organisms. There are gliding sulfur oxidizers, including bacteria, the cells of which are arranged in trichomes, that show a gliding motion on the substrate. The most important members of this group in relation to S-oxidation in soils are species of *Beggiatoa*, bacteria that participate in sulfide oxidation in the root zone of rice. All strains of *Beggiatoa* deposit sulfur in the presence of H_2S. Phototrophic bacteria, such as *Chromatium* and *Chlorobium*, also play an important role in sulfide oxidation in rice paddy soil, but not in aerobic agricultural soils. A number of nonfilamentous, chemolithotrophic sulfur-oxidizing bacteria, such as *Sulfolobus*, *Thiospira*, or *Thiomicrospira*, have also been isolated from special habitats, but the importance of these bacteria in S oxidation in soils has yet to be determined. Some thiobacilli, such as *Thiobacillus denitrificans*, can oxidize sulfur or thiosulfate and use nitrate as an electron acceptor. Other aerobic bacteria, such as *Arthobacter*, *Bacillus*, and *Pseudomonas*, oxidize sulfur in soil during their normal metabolism of oxidizing organic compounds in soil and using oxygen as a terminal electron acceptor. These reactions have not been characterized but it is thought that this is occurring as a side reaction to the primary transformations being carried out by these organisms. In the soil ecosystem, it is widely assumed that thiobacilli are the dominant sulfur oxidizers but this view is based on the observation that if you add elemental sulfur to soil, the numbers of thiobacilli increase. However, no consistent correlation between S-oxidation rates and the prevalence of thiobacilli has been demonstrated, except in very broad terms. In general, it is assumed that initially heterotrophs oxidize elemental sulfur

until the pH is low enough that oxidation of sulfur chemoautotrophs is energetically favorable and then thiobacilli dominate. Thus, under energy-limiting conditions in an environment dominated by dissolved, reduced sulfur compounds with little organic matter, thiobacilli will dominate. In contrast, in a soil in which there are significant amounts of organic matter and less reduced sulfur, heterotrophs will dominate.

Disproportionation of Sulfur

In the latter part of the twentieth century, investigators discovered that not only can bacteria oxidize and reduce sulfur, but they can also disproportionate it. Disproportion reactions are best likened to a fermentation in which one portion of a molecule acts as an electron donor and the other, an electron acceptor. Bacteria are able to disproportionate thiosulfate by oxidizing the sulfur atom in thiosulfate that carries a +5 charge such that it now carries a charge of +6. This +6 sulfur is released as SO_4^{2-}. Simultaneously, bacteria oxidize the other sulfur atom in thiosulfate that carries a −1 charge such that it now carries a charge of −2. This −2 sulfur is released as hydrogen sulfide:

$$S_2O_3^{2-} + H_2O \rightarrow SO_4^{2-} + HS^- + H^+ \qquad [1]$$

This metabolic process has immense implications for the global sulfur cycle since typically in sediments sulfate predominates in the upper layers and sulfide in the lower layers. Hence, this metabolic reaction reveals an entirely new and, up until then, unsuspected ecological niche which is responsible for up to 60% of the thiosulfate transformations in sediment. Since then, other bacteria have been found that can disproportionate sulfonates such as taurine to sulfate and sulfide. The bacteria capable of this organic sulfur disproportation reaction form a new species called *Desulfonispora* since it appears that it is a distinct group of organisms that is capable of this reaction.

Reduction of Sulfur

The reduction of sulfur can occur for two reasons: assimilation and dissimilation. The assimilation pathway has been discussed above – bacteria convert sulfate to sulfide for inclusion in amino acids such as cysteine. Dissimilatory sulfate reductase refers to a process whereby sulfate is used as a terminal electron acceptor. Sulfate-reducing bacteria use such an enzyme, which allows them to oxidize organic substrates such as lactate, malate, and ethanol, and use sulfate as a terminal electron acceptor. Hence, these bacteria are primarily found in environments where there is little or no oxygen and live off the fermentation end products produced by other bacteria.

Sulfate-reducing bacteria catalyze the following process:

$$SO_4^{2-} + 5H_2 \rightarrow H_2S + 4H_2O \qquad [2]$$

Since hydrogen sulfide is extremely toxic, these organisms require some sort of metal to react with the hydrogen sulfide produced and precipitate as a nonsoluble metal sulfide. In soil, this metal is typically iron, with FeS being precipitated around areas of sulfate-reducing bacterial growth and activity. The reaction catalyzed by sulfate-reducing bacteria reaches an optimum at an E_h of −300mV at a pH of 7. These conditions are typically reached as one enters the highly anaerobic zone. Typically, terminal electron acceptors are depleted in the following order: oxygen, nitrate, nitrite, manganite, iron, sulfate, and finally, carbon dioxide. Thus, there are many alternative electron acceptors that will be used by a microbial community before sulfate is reduced. Despite this limitation, sulfate-reducing bacteria and activity are commonly found in soils, sediments, polluted water, oil-bearing strata, and shales.

Sulfate-reducing bacteria can implement an alternative survival strategy if sulfate levels in the water are too low to be used. The H^+ produced by organic matter oxidation is transferred directly to a methanogen which then consumes that H^+ to reduce CO_2 and produce methane. Because the methanogens keep the partial pressure of hydrogen so low, the sulfate-reducing bacteria can use this mechanism to obtain energy from organic matter degradation even when there is not enough sulfate around for them to transform to sulfide.

Sulfur Transformations and Environmental Quality

The biological transformations of sulfur can lead to significant environmental problems in the soil. The enzyme sulfate-reducing bacteria use to transfer methyl groups can also accidentally methylate mercury to form methylmercury. Sulfate-reducing bacteria have been identified as one of the primary causes of the increased mercury accumulation in many ecosystems because methylmercury is the form of mercury that most readily accumulates in food chains. Typically, this reaction occurs in anaerobic, lowland soils where the activity of sulfate-reducing bacteria is closely linked to methylmercury production. However, sulfate-reducing bacteria can also mitigate environmental pollution through the production of metal sulfides by the following reaction:

$$M^{2+} + S^{2-} \rightarrow MS \quad [3]$$

Sulfate-reducing bacteria like *Desulfovibrio* spp. can form sulfides of Sb, Co, Cd, Fe, Pb, Ni, and Zn during the reduction of sulfate. The extent of metal sulfide genesis depends on many factors, such as the amount of sulfate present and the relative toxicity of the metal ion. In nature, this toxicity is reduced when the metal ions are adsorbed on clays or complexed with organic matter.

The oxidation of metal sulfides in soil involves both chemical and microbial processes and, as a result, is a more complex process than is the oxidation of S^0 Chalcocite (Cu_2S), chalcopyrite ($CuFeS_2$), galena (PbS), pyrite (FeS_2), and nickel sulfide (NiS) are just a few examples of metal sulfides that are subject to microbial transformations. For example, the biological oxidation of pyrite (FeS_2) can follow one of two pathways. The first result follows a series of oxidation steps, described in eqns [4–7]. The second mechanism of pyrite oxidation involves thiosulfate as an intermediate but in the end requires the oxidation of elemental sulfur to sulfate by *Thiobacillus*. These biotic oxidations are responsible for the formation of acid mine drainage and acid soil formation in surface mine spoils. First, ferrous sulfate is formed as the result of an abiotic oxidation step:

$$2FeS_2 + 2H_2O + 7O_2 \rightarrow 2Fe_2SO_4 + 2H_2SO_4 \quad [4]$$

This reaction is then followed by the bacterial oxidation of ferrous sulfate, generally by *T. ferrooxidans*:

$$4FeSO_4 + O_2 + 2H_2SO_4 \rightarrow 2Fe(SO_4)_3 + 2H_2O \quad [5]$$

This reaction occurs chemically but can be accelerated 10^6–10^8 times by thiobacilli. This bacterial oxidation of ferric ion plays an important role in the bioleaching of metal sulfides in the environment because it cycles the iron between +2 and +3 and the iron is then free to oxidize sulfide minerals abiotically.

Subsequently, ferric sulfate is reduced and pyrite oxidized by a strictly chemical reaction:

$$Fe_2(SO_4)_3 + FeS_2 \rightarrow 3FeSO_4 + 2S^0 \quad [6]$$

$$2S + 6Fe_2(SO_4)_3 + 8H_2O \rightarrow 12FeSO_4 + 8H_2SO_4 \quad [7]$$

The elemental sulfur produced is finally oxidized by *T. thiooxidans*, and the acidity produced helps the whole process to continue.

$$2S^0 + 3O_2 + 2H_2O \rightarrow 2H_2SO_4 \quad [8]$$

Note the net production of 10 molecules of H_2SO_4 during the process.

Although several sulfur-oxidizing thiobacilli and heterotrophs can be isolated from acid sulfate soils in which pyrite is being oxidized, they appear not to play an important role in the process, with the exception of *T. ferrooxidans*. The biological oxidation of sulfides and other reduced S compounds can have severe consequences for the environment. For example, acid mine drainage contaminates several thousand kilometers of streams in the Appalachian coal-mining region of the USA.

Conclusion

There is a large amount of sulfur present in soil and most of it is found in a form susceptible to biological transformations. These transformations include mineralization, assimilation, oxidation, and reduction. Recently, anthropogenic activity has impacted the sulfur cycle by increasing the deposition of sulfate in the form of acid rain and by exposing more sulfur to oxidation during mining operations. The four basic transformations discussed in this article influence the fate of the movement of sulfur in terrestrial and aquatic ecosystems and the environmental impact of anthropogenic activities altering the sulfur cycle.

See also: **Organic Residues, Decomposition**; **Sulfur in Soils:** Overview; Nutrition

Further Reading

Benning C (1998) Biosynthesis and function of the sulfolipid sulfoquinovosyl diacylglycerol. *Annual Reviews in Plant Physiology and Molecular Biology* 49: 53–75.

Ehrlich HL (2002) *Geomicrobiology*. New York: Marcel Dekker.

Fenchel T, King GM, and Blackburn TH (1998) *Bacterial Biogeochemistry: The Ecophysiology of Mineral Cycling*. San Diego, CA: Academic Press.

Howarth RW, Stewart JWB, and Ivanov MV (eds) (1992) *Sulfur Cycling on the Continents: Wetlands, Terrestrial Ecosystems, and Associated Water Bodies. SCOPE 48*. Chichester, UK: John Wiley.

Huang PM, Bollag JM, and Senesi N (eds) (2001) *Interactions Between Soil Particles and Microorganisms: Impact on the Terrestrial Ecosystem*. Chichester, UK: John Wiley.

Paul EA and Clark FE (1989) *Soil Microbiology and Biochemistry*, 2nd edn. New York: Academia Press.

Sylvia DM, Furmann JJ, Hartel PG, and Zuberer DA (eds) (1998) *Principles and Applications of Soil Microbiology*. New Jersey: Prentice Hall.

Tabatabai MA (ed.) (1986) *Sulfur in Agriculture*. Madison, WI: American Society for Agronomy.

Nutrition

M A Tabatabai, Iowa State University, Ames, IA, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Sulfur(s) is considered either the fourth most important nutrient element, after N, P, and K, for plant growth, or the sixth, after C, H, O, N, and P, for protein composition. In addition to these six, there are 15 other elements that are essential for the growth of some plants. These are Ca, Mg, K, Fe, Mn, Mo, Ni, Cu, B, Zn, Cl, Na, Co, V, and Si. Not all these elements are essential for all plants, but all have been shown to be essential for some plants.

S has been recognized as an essential element for plant growth and development for more than 200 years. In 1859 Liebig was aware of the close relationship between N and S in many plants. The ancient Romans and Greeks demonstrated the beneficial use of gypsum ($CaSO_4 \cdot 2H_2O$) as a fertilizer. The history of the use of S fertilizers may be divided into three periods: (1) the first period is the reign of gypsum (1760–1845); during this period, gypsum was used widely and its beneficial effect was generally overestimated; (2) the second period is the reign of superphosphate (1845–1905), during which the need for S fertilization was ignored and the use of gypsum was discouraged by agronomists; (3) the third period is the Renaissance, or Modern period. The use of gypsum as a fertilizer in modern times can be traced back to the middle of the eighteenth century. A field trial with gypsum applied to a clover field was reported in Switzerland in 1768. Application of gypsum increased the yield of clover by more than two-fold.

Interest in fertility-related aspects of the S cycle is increasing, because S deficiencies in agronomic crops are observed with increasing frequency. Sulfur fertilization is now required for many crops around the world. Several factors are responsible for the increased need of S fertilization. These include: (1) increased use of high-analysis fertilizers that contain little or no S, (2) increased crop yields, (3) decreased use of S as a pesticide, (4) more intensive cropping, and (5) decreased gain of atmospheric S by soils and plants as a result of decreased combustion of coal and other S-containing fuels.

Sulfur Requirements of Crops

The importance of S in crop production is obvious, because plants require S for synthesis of essential amino acids and proteins, certain vitamins and coenzymes, glucoside oils, structurally and physiologically important disulfide linkages and sulfhydryl groups, and activation of certain enzymes.

Sulfur is absorbed by plant roots almost exclusively as the sulfate ion, SO_4^{2-}. Typical concentrations of S in plants range from 0.1 to 0.4%. As with N, much of the SO_4^{2-} taken up by plants is reduced in plants, and the S is found in –S–S– and –SH forms. Sulfate-S in large concentrations may occur in plant tissues and cell sap. Normally, S is present in equal or lesser concentration than phosphorus in such plants as wheat, corn, beans, and potatoes, but in greater concentrations in alfalfa, cabbage, and turnips. Generally, agronomic crops require about the same amount of S as they do for P.

Although the S content of plants varies depending on the supply available, some crops have greater S requirement than others. An average yield of forage crops removes 17–50 kg S ha^{-1}, and the cereal grains generally remove more than 30 kg ha^{-1}. Other crops such as cabbage, turnip, and alfalfa have a particularly high requirement for S. Such crops commonly need from 45 to 70 kg S ha^{-1}. One of the crops that require a very high amount of S is sugarcane. A yield of 224 tons ha^{-1} removes approximately 100 kg S (Table 1).

Another factor that affects the S requirement of plants is the available N. S and N are closely associated in protein synthesis, thus S requirements vary with the supply of N to crops. Therefore, when S becomes limiting for plant growth, addition of N does not increase the yield or protein concentration of plants.

The N:S ratios of many crops at their maximum yields have been assessed. Alfalfa requires 1 part of S for every 11–12 parts of N to ensure maximum production, while wheat, corn, beans, and sugarbeet leaf blades require 1 part of S for every 12–17 parts of N. The N:S ratios of grains such as oats and barley are 13:1 and 9:1, respectively.

Crop plants obtain their S requirements from a number of sources. These include: (1) soils, crop residues, and manures; (2) irrigation waters; (3) rainfall and the atmosphere; and (4) fertilizers and soil amendments. The order of importance of each of these sources varies with the type of crop, location, and management practices. The effects on S requirements are particularly marked in nonindustrialized areas where soil supplies are already low and additions from precipitation are being further reduced by shifts in energy sources (burning coal for energy production).

Functions of Sulfur in Plants

Sulfur has numerous functions in plant growth and metabolism. Among those are the following:

Table 1 Sulfur content of crops

Crop	Yield (tons ha^{-1})	Total S content (kg ha^{-1})
Grain and oil crops		
Barley	5.4	22
Corn	11.2	34
Grain sorghum	9.0	43
Oats	3.6	22
Rice	7.8	13
Wheat	5.4	22
Peanuts	4.5	24
Soybeans	4.0	28
Hay crops		
Alfalfa	17.9	45
Clover grass	13.4	34
Bermuda grass		
Common	9.0	17
Coastal	22.4	50
Brome grass	11.2	22
Orchard grass	13.4	39
Pangola grass	26.4	52
Timothy grass	9.0	18
Cotton and tobacco		
Cotton (lint + seed)	4.3	34
Tobacco		
Barley	4.5	21
Flue cured	3.4	50
Fruit, sugar, and vegetable crops		
Beets		
Sugar	67	50
Table	56	46
Cabbage	78	72
Irish potatoes	56	27
Onions	67	41
Oranges	52	31
Pineapple	40	16
Sugarcane	224	96

Reproduced from Terman GL (1978) *Atmospheric Sulfur – The Agronomic Aspects.* Technical Bulletin no. 23. Washington, DC: The Sulfur Institute.

1. It is required for the synthesis of S-containing amino acids, cystine, and cysteine, and methionine, which are major components of proteins;

2. It is required for the synthesis of important metabolites, including coenzyme A, biotin, thiamine, or vitamin B_1, glutathione, sulfolipids;

3. It is a vital component of ferredoxin, a type of nonheme Fe–S protein present in chloroplasts. This is involved in photosynthesis, nitrite reduction, sulfate reduction, the assimilation of N_2 by roots of nodule bacteria and free-living N_2-fixing soil bacteria, and in mitochondrial electron transport;

4. It is an essential component (as cysteinyl, a–SH group) of the active sites of many enzymes;

5. Although it is not a component of chlorophyll, it is required for its synthesis by plants;

6. It occurs in volatile compounds responsible for the characteristic taste and smell of plants in the mustard and onion families.

Sulfur in Soils

Soil as a living system is a dynamic ecosystem supporting the life of microorganisms, plants, and animals. As is the case with carbon and nitrogen, S is continuously being cycled between organic and inorganic forms, resulting in synthesis and degradation of a variety of S compounds. The reactions and the nature of the S compounds formed depend on the environmental conditions of the soil. These include aeration, water content, pH, presence and absence of metal ions, tillage systems applied, crop residues incorporated, liming, type of crops, and crop rotations used.

The relative proportion of inorganic and organic forms of S in soils varies with the soil type, depth of sampling, cropping system, season, and soil conditions (field-moist or air-dried). No direct chemical method is available for determination of total organic S in soils, and this is normally determined from the difference between total S and inorganic S determined separately. The inorganic S fraction is small, ranging from 2 to 6% of the total S in soils from the humid and semihumid regions. This is normally present as inorganic SO_4^{2-} in well-aerated soils. In calcareous soils, much of this fraction is occluded by precipitate of calcium carbonate, leading to a smaller fraction of the total S in organic combinations. Similarly, because of lack of percolating water, inorganic SO_4^{2-} accumulates in surface layers of arid soils in the form of gypsum, leading to a small fraction of organic S in such soils. Organic S concentration decreases with depth in the soil profile, and this decrease is associated with a decrease in organic C, except in soils where organic matter accumulates in subsurface or B horizons such as podzolic soils.

The chemical nature of soil organic matter in soils is important, because this is the reservoir that supplies SO_4^{2-} to plants. Even though complete chemical forms of S soil organic matter are not completely characterized, three broad groups of S compounds are recognized. These are:

1. Hydriodic acid-reducible S. This fraction consists of organic S that is reduced to H_2S upon boiling with hydriodic acid (HI) under N_2 gas. This form of S is bonded to a carbon atom through an oxygen atom (i.e., ester sulfate, C–O–S). Examples of such forms of S are phenolic sulfate, sulfated polysaccharides, choline sulfate, and sulfated lipids. On average, approximately 50% of the organic S in soils is in this form; it can range from 30 to 60%, and values as high as 95% have been reported for some Iowa subsurface

soils. It is mineralized to SO_4^{2-} through the enzyme arylsulfatase, as follows:

$$R\text{–}O\text{–}SO_3^- + H_2O \xrightarrow{\text{Arylsulfatase}} R\text{–}OH + HSO_4^- \quad [1]$$

2. Carbon-bonded S. This fraction consists of S atoms directly bonded with carbon atoms (i.e., C–S). This form of S is hydrolyzed by Raney Ni (50% each of Al and Ni powder) under alkaline conditions (NaOH). When treated with HCl, this mixture releases the S atoms in the form of H_2S. Sulfur-containing amino acids, cystine and methionine, are the main components of this fraction, which accounts for approximately 20–30% of the soil organic S. Other S-containing compounds such as sulfoxide, sulfones, sulfenic, sulfinic, and sulfonic acids, and S-containing heterocyclic compounds are also constituents of this fraction;

3. Inert or residual S. This fraction is the organic S that is not reduced by either of the reagents described above. This fraction represents approximately 30–40% of the total organic S in soils. Because this fraction is resistant to drastic chemical treatment, it is of little importance as a source of S for plants.

Sources of Mineralizable Sulfur in soils

It is believed that both HI-reducible and carbon-bonded S fractions of soil organic S are sources of SO_4^{2-} for plant uptake.

Organic Nitrogen and Sulfur Relationship in Soils

The presence of large concentrations of organic S in surface soils is indicated by the close relationship between organic N and S. In many of the soils studied in North America and other parts of the world, the mean ratio of organic N:S is approximately 8:1. But, this ratio may range from 5:1 to 13:1 (Table 2). The ratio appears to be the same for virgin soils and their cultivated counterparts (Table 3), suggesting that the organic forms of these elements mineralize in about the same ratio as they occur in soil organic matter. However, results of some studies indicate that organic S in soils is depleted faster than that of organic N.

As is true of N, when land is first cultivated, the S content of the soil declines rapidly and an equilibrium level is reached which is influenced by climate, cultural practices, and soil type. At the equilibrium level, soil organic matter essentially ceases to act as a source of S for plant growth. Before reaching this equilibrium, however, the rate of S mineralization is so slow that it cannot cope with the plant's need for this element. This results in the appearance of S-deficiency symptoms on the plants. Because S is a mobile element in plants, the S-deficiency symptoms appear on the older leaves. Recent studies indicate that lack of S to meet plant's requirements may not only reduce yields, but it can reduce quality (reduction of the amino acids cysteine and methionine in protein) in grain legumes by changing gene expression of storage proteins in developing seeds.

Table 2 Mean ratios of nitrogen to sulfur in surface horizons of soils

Location	*Description*[a]	*N:S ratio*
Virgin soils		
Alberta, Canada	Brown Chernozems (4)	7.1
	Black Brown Chernozems (4)	7.7
	Gray Wooded (7)	12.5
	Gleysols (6)	5.0
USA (several states)	Not specified (10)	8.7
Cultivated soils		
NSW, Australia	Pasture soils (5)	8.5
	Clover pasture, podzolic (44)	7.1
Eastern Australia	Acid soils (128)	8.3
	Alkaline soils (27)	6.6
Sao Paulo and Parana, Brazil	Agricultural, varied (6)	7.7
Canterbury, New Zealand	Grassland, unfertilized	7.7
North Scotland	Agricultural, noncalcareous (40)	7.1
USA		
Iowa	Agricultural, varied (37)	6.5
Iowa	Agricultural, varied (6)	7.7
Minnesota	Brown Chernozems (6)	6.4
	Black prairie soils (9)	6.1
	Podzols (24)	8.5
Mississippi	Podzols (4)	13.1
Oregon	Agricultural, varied (16)	9.9
Several states	Agricultural, varied (10)	8.0

[a]Numbers in parentheses are number of soils examined.
Reproduced from Tabatabai MA (1984) Importance of sulphur in crop production. *Biogeochemistry* 1: 45–62.

Sulfur Availability Indexes

It is generally accepted that S is taken up by plants in the SO_4^{2-} form. Studies involving nutrient culture experiments have shown that low-molecular-weight organic-S compounds are utilized by plant. But, the availability of soil organic S depends primarily on its mineralization, which, in turn, depends on climatic factors such as temperature and moisture, and on the chemical nature of organic S present in soils.

Assessment of the plant-available S in soils is complicated by the fact that, in addition to soils, several other sources contribute to the S needs of plants. These sources include S in rainfall and irrigation

Table 3 Organic nitrogen and sulfur ratios in virgin soils (V) and their cultivated (C) counterparts

Location		*N:S ratio*
Big Springs, TX	V	12.0
	C	11.1
Colby, KS	V	6.5
	C	6.8
Mays, KS	V	8.0
	C	6.8
Moccasin, MT	V	7.6
	C	6.0
Dalbart, TX	V	9.6
	C	11.7
Madan, ND	V	9.0
	C	9.6
North Platte, NE	V	9.9
	C	9.0
Lawton, OK	V	10.3
	C	10.4
Archer, WY	V	8.5
	C	8.0
Havre, MT	V	9.1
	C	8.2
Mean		8.9

Reproduced from Stewart BA and Whitefield CJ (1965) Effects of crop residue, soil temperature, and sulfur on the growth of winter wheat. *Soil Science Society of America Proceedings* 29: 752–755.

water, S in the atmosphere, and S in fertilizers and pesticides. Of the sources of S, the contribution of S in rainfall and direct absorption by soils and plants from the atmosphere are the most difficult to evaluate. In addition, both plants and soils absorb SO_2, and most likely other S gases, directly from the atmosphere. For example, studies in Wisconsin in the early 1970s, by using radioactive ^{35}S, have shown that under optimal yield 14% of the S in alfalfa (*Medicago sativa* L.) is derived from atmospheric sources.

There is no general agreement on the best methods for using an index of plant availability. The procedures used fall into one of the following two groups: (1) plant analyses and (2) soil analyses. The use of plant analysis for assessing S nutrition of plants is based on the notion that an essential element should be present in the plant as a concentration just sufficient for unrestricted plant growth.

Numerous methods and procedures have been proposed for evaluation of the plant-available S in soils. These methods include extraction with water, extraction with various salts or acid solutions, S mineralization during incubation, microbial growth, and plant growth and composition. The concentration of S removed by the various extractants normally falls in one of the following groups: (1) readily soluble sulfate, (2) readily soluble and some of the adsorbed sulfate, and (3) readily soluble and some of the adsorbed sulfate and some of the organic S.

Incubation procedures for assessing plant-available S paralleling those used in the estimation of available N have met with little success, because precise determination of the small amount of inorganic SO_4^{2-} released during incubation is difficult. In addition, the amount of S mineralized during incubation is affected by the presence or absence of plants.

Sulfur Requirement of Plants

The requirement of S or any other nutrient by crops is defined as "the minimum content of that nutrient associated with the maximum yield" or "the minimum rate of the uptake of the nutrient associated with maximum growth rate." The first definition refers to that the total amount of the element in the crops (normally expressed in kilograms per hectare) or the concentration of the element in the plant or plant part. The second definition is related to the minimum concentration of the element taken up from the soil or nutrient solution that is associated with the maximum growth. Information on the S content of crops is useful in estimating the S-fertilizer requirement.

Both the uptake and requirements for S differ greatly among plant species, among cultivars within species, and with the stage of development of the crop. Assessment of the S requirement of crops is more complicated than that of any other nutrient. That is because: (1) there are several S sources for plants, (2) each source has a different efficiency for its utilization by crops, (3) S has limited reuse within the crop, and (4) relatively large concentrations of S can be accumulated within the crop. The supply of S during the growing period can be changed by the atmospheric conditions and precipitation, impurities in fertilizers, addition of pesticides, and mineralization of organic S in soil organic matter. In spite of the difficulties associated with assessment of S status on crops, several approaches have successfully been used for this purpose. These include: (1) chemical soil tests, (2) a number of plant-tissue chemical analyses, and (3) crop-deficiency symptoms. The first of these approaches is the most commonly used.

Sulfur Metabolism in Plants

S metabolism in plants is very complex. The significant steps are as follows: (1) sulfate is 'activated,' that is, enzymatically converted to adenosine-5′-phosphosulfate (APS) and 3′-phosphoadenosine-5′-phosphatsulfate (PAPS). In plants, APS functions as substrate for SO_4^{2-} reduction and as a precursor of PAPS, while PAPS acts as the SO_4^{2-} donor in the formation sulfate esters (organic sulfates); (2) the

activated sulfate is reduced; (3) reduced S is incorporated into cysteine; (4) the cysteine-S is transferred into methionine and other essential compounds; (5) methionine is transferred into S-adenosylmethionine, which is a methyl ($-CH_3$) donor and a precursor of important non-S-containing compounds; and (6) cysteine and methionine are incorporated into proteins. Sulfate transport in plants is inhibited by structurally related compounds or ions such as SO_3^{2-}, $S_2O_3^{2-}$, $S_2O_5^{2-}$, SeO_4^{2-}, CrO_4^{2-}, MoO_4^{2-}, and WO_4^{2-}. Studies have shown that SeO_4^{2-} is a competitive inhibitor of SO_4^{2-} transport in plants.

Sulfur-Containing Materials Added to Soils

A number of materials added to soils contain S. These range from animal manures to sewage sludge, crop residues, and irrigation water. The S concentration in waste materials and irrigation water varies with the sources and quality of the material. Therefore, knowledge of the composition of such materials is needed before its application or use on soils.

Fertilizers

Many sources of S-containing fertilizer are available, most of which are old and established sources of S. Their behavior in soils is determined by the chemical nature of the S atom and the reactions involved when added to soils. The following are examples of dry fertilizer containing sulfur:

1. Ammonium sulfate. This compound is widely used as a source of N, but $(NH_4)_2SO_4$ contains 21% N in the ammoniacal form and approximately 24% of S in the SO_4^{2-} form;
2. Single superphosphate. This compound has been manufactured by the fertilizer industry since 1850. It contains approximately 8% of P. It is produced by reacting sulfuric acid (H_2SO_4) with phosphate rock. The reaction is:

$$3[Ca_3(PO_4)_2] \cdot CaF_2 + 7H_2SO_4 = 3Ca(H_2PO_4)_2 + 7CaSO_4 + 2HF \quad [2]$$

3. Ammonium phosphate sulfate. This is essentially a mixture of monoammonium phosphate and ammonium sulfate. The most common of these products contains 16% N, 8% P, and 13% S;
4. Gypsum. Large quantities of gypsum ($CaSO_4 \cdot 2H_2O$) are produced as a by-product of manufacturing concentrated phosphate fertilizer. It is a low-analysis fertilizer (13–14% S in the by-product forms and 18.6% S in the pure form);
5. Elemental S. When added to soils, elemental S is converted to SO_4^{2-}. The reactions involved are affected by particle size (the finer the particles, the faster the reaction rate), temperature, moisture and aeration, soil pH, soil type and properties, fertilizer interaction, characteristics of the elemental S, rate and placement, and time of application. The reactions involved are:

$$S^0 \longrightarrow S_2O_3^{2-} \longrightarrow S_2O_6^{2-} \longrightarrow SO_3^{2-} \longrightarrow SO_4^{2-} \quad [3]$$

6. Sulfur-bentonite. This product consists of 90% of elemental S and 10% of bentonite. When this material is added to soils, the bentonite imbibes moisture, causing the granules to disintegrate, releasing very fine S particles, which are rapidly converted to SO_4^{2-};
7. Elemental sulfur suspension. This material is made of finely ground elemental S (40–60% of S^0) and a small amount of attapulgite clay. It can be applied directly to soils or mixed with other fertilizer materials;
8. Phosphate-elemental sulfur materials. This material is a mixture of elemental S and mono- or diammonium phosphate. It is a source of S and P for plants;
9. Other dry S-fertilizer mixtures are available; these include urea-ammonium sulfate, ammonium nitrate–ammonium sulfate, and potassium sulfate mixtures;
10. Urea–S. This material contains 20% of S and 36% of N. It has excellent physical properties and easy to handle.

Fluid fertilizers containing sulfur include:

1. Ammonium thiosulfate. This compound is a clear solution containing 26% of S and 12% of N. It is the most commonly used S fertilizer. This material can be used for direct application to soils and in irrigation waters;
2. Ammonium-polysulfide. Polysulfides are used as fertilizers, as soil conditioners, and for treatment of irrigation waters to improve percolation into soils. Ammonium polysulfides are used for all these purposes. When added to soil, the polysulfide is changed to colloidal S and sulfide. Calcium polysulfide is marketed as soil conditioner and for treatment of irrigation water. The reaction involved is:

$$CaS_6 + H_2O + CO_2 = H_2S + CaCO_3 + 4S^0 \quad [4]$$

3. Ammonium bisulfite. This is a clear liquid, low-analysis fertilizer, containing 17% of S and 8.5% of N. It has a strong odor of SO_2 and is marketed to a very limited extent;

4. Urea-sulfuric acid mixtures. Two grades of this material are available on the market. One grade contains 10% of N and 18% of S, and the other contains 28% of N and 9% of S. Because this material is highly acidic (pH between 0.5 and 1.0), it is essential that all application equipment is made from stainless steel and extreme precautions must be taken in handling the material.

In addition to the S-containing materials described above, aluminum sulfate is added to soil for soil acidification. Acid soil is required for plants such as azaleas and camellias. When added to soils, the reaction is:

$$Al_2(SO_4)_3 + 6H_2O = 2Al(OH)_3 + 3H_2SO_4 \quad [5]$$

When using any of the above materials, the user should adhere to the recommendations provided by the manufacturer in terms of safety, rate of application, method and time of application, and storage conditions.

Sulfur in the Atmosphere

The atmosphere is an important compartment of the S cycle for agricultural crops and forest ecosystems. Sulfur gases and aerosols are found in the atmosphere of urban, rural, and remote areas. The sulfur species in the atmosphere range from oxidized forms such as sulfur dioxide (SO_2) and sulfate (SO_4^{2-}) to reduced forms such as hydrogen sulfide (H_2S), carbonyl sulfide (COS), dimethyl sulfide ($(CH_3)_2S$, and carbon disulfide (CS_2). These are derived from natural and anthropogenic sources. The lifetimes of the S compounds in the atmosphere range from a few hours to many years and can be found in various chemical forms from the troposphere to the stratosphere. The relationship between atmospheric S and the terrestrial ecosystem is very complex. To understand the interactions, the following information is needed: (1) the rate and chemical forms of S entering the atmosphere, (2) the chemical transformations occurring in the atmosphere, (3) the transport from the emission sources to the receptor (plant), (4) the amount and rate of transfer to soil and plants, and (5) the impact on crop production. The natural sources of S in the atmosphere include: (1) volcanic activity, (2) input from ocean spray, (3) bacterial decomposition of plant material in soils, and (4) animal manure and sewage sludge. Man-made sources include: (1) combustion of S-containing fossil duels (coal, oil, and gas), and (2) reduced S gases released from many industrial processes (e.g., pulp mills).

Atmospheric S gases are absorbed by plants, surface water, and surface soils or deposited in precipitation in the SO_4^{2-} form which, in turn, is absorbed by plants, retained in soils, or leached through groundwater.

Sulfur in Precipitation

Sulfur in the atmosphere occurs in gaseous, solid, and liquid form, with transformation among the different forms accomplished by chemical, biological, and photometric processes. As is the case with N, S is highly transitory among these forms in air and water. The reactions involved in these transformations determine the concentration of S in precipitation, which, in turn, affects its presence in surface and groundwater, with ultimate impact on soils and plants.

The data available on S content of rainfall indicate wide variation by geographic locations. Expressed in kilograms of sulfur per hectare, the annual addition of S in precipitation in North American ranges from 0.5 to 14. In many areas, most of this (approximately 60–80%) additional S is deposited during the crops' growing season. Laboratory and greenhouse experiments have shown that many soils of the USA do not contain sufficient plant-available S to meet the crops' requirements, yet no S-deficiency symptoms have been reported. This is because, in many areas, uptake of atmospheric SO_2 and of that in precipitation can compensate for the soil S deficiency. Indeed, it has been demonstrated that sunflowers can use SO_2 and H_2S as their only source of S without their normal growth being affected.

See also: **Acid Rain and Soil Acidification**; **Enzymes in Soils**; **Fertility**; **Leaching Processes**; **Minerals, Primary**; **Organic Matter:** Principles and Processes; **Organic Residues, Decomposition**; **Soil–Plant–Atmosphere Continuum**; **Sulfur in Soils:** Overview; Biological Transformations

Further Reading

Freney JR (1967) Sulfur-containing organics. In: McLaren AD and Peterson GH (eds) *Soil Biochemistry*, vol. 1, pp. 229–259. New York: Marcel Dekker.

Germinda JJ and Gupta VVSR (1992) Biochemistry of sulfur cycling in soil. In: Stotzky G and Bollag J-M (eds) *Soil Biochemistry*, vol. 7, pp. 1–53. New York: Marcel Dekker.

McLachlan KD (ed.) (1975) *Sulphur in Australasian Agriculture*. Sydney, Australia: Sydney University Press.

Nor YM and Tabatabai MA (1977) Oxidation of elemental sulfur in soils. *Soil Science Society of America Journal* 41: 736–741.

Stewart BA and Whitefield CJ (1965) Effects of crop residue, soil temperature, and sulfur on the growth of winter wheat. *Soil Science Society of America Proceedings* 29: 752–755.

Tabatabai MA (1984) Importance of sulphur in crop production. *Biogeochemistry* 1: 45–62.

Tabatabai MA (1985) Effect of acid rain on soils. *CRC Critical Reviews in Environmental Control* 15: 65–110.

Tabatabai MA (ed.) (1986) *Sulfur in Agriculture*. Agronomy monograph no. 27. Madison, WI: American Society of Agronomy.

Tabatabai MA (1987) Physicochemical fate of sulfate in soils. *Journal of the Air Pollution Control Association* 37: 34–38.

Tabatabai MA (1994) Soil enzymes. In: Weaver RW, Angle JS, and Bottomley PS (eds) *Methods of Soil Analysis*, part 2, pp. 775–783. *Microbiological and Biochemical Properties*. Soil Science Society of America Book Series No. 5. Madison, WI: Soil Science Society of America.

Tabatabai MA (1996) Sulfur. In: Sparks DL (ed.) *Methods of Soil Analysis*, part 3, pp. 921–960. *Chemical Methods*. Soil Science Society of America Book Series No. 5, Madison, WI: Soil Science Society of America.

Tabatabai MA and Al-Khafaji AA (1980) Comparison of nitrogen and sulfur mineralization in soils. *Soil Science Society of America Journal* 44: 1000–1006.

Terman GL (1978) *Atmospheric Sulfur – The Agronomic Aspects*. Technical Bulletin no. 23. Washington, DC: The Sulphur Institute.

Tisdale SL, Nelson WL, and Beaton JD (1985) *Soil Fertility and Fertilizers*, 4th edn. New York: Macmillan.

SURFACE COMPLEXATION MODELING

S Goldberg, USDA–ARS, Riverside, CA, USA

Published by Elsevier Ltd.

Introduction

Adsorption is the process through which ions are removed from solution and accumulate at a solid surface. The ion accumulation takes place at the interface between the surface and the solution forming a two-dimensional structure. If adsorption continues and leads to a three-dimensional structure, the process is called precipitation. The general loss of ions from solution to a surface is called sorption.

Adsorption can occur either specifically or nonspecifically. Specific adsorption occurs when ions have a high affinity for the surface and it results in the formation of inner-sphere surface complexes. Inner-sphere surface complexes are complexes that contain no water molecules between the adsorbing ion and the surface functional group. Examples of surface functional groups are reactive hydroxyl groups on oxide or clay minerals and carboxyl or phenol groups on organic matter. Such surface functional groups are a source of solid surface charge since they undergo dissociation and/or protonation reactions as a result of changes in solution pH. Specific anion adsorption occurs via ligand exchange where the adsorbing ion replaces a reactive surface hydroxyl from the surface functional group. Nonspecific adsorption is dominated by electrostatic attraction and results in outer-sphere complex formation or in adsorption in the diffuse ion swarm. Adsorption in the diffuse ion swarm is the weakest type of adsorption since the ion does not attach to a specific surface functional group but remains free in the aqueous solution, neutralizing surface charge only by its proximity to the charged solid surface. Outer-sphere surface complexes are also formed through nonspecific adsorption and contain at least one water molecule between the adsorbing ion and the surface functional group.

A model is a simplified representation of reality that considers only those characteristics of the system that are pertinent to the problem at hand. A chemical model provides a description of a chemical system consistent with its chemical properties while simultaneously being as simple and as chemically correct as possible. The ideal chemical model is realistic, effective, comprehensive, and predictive. A realistic model conforms to accepted theories of chemical behavior, an effective model closely describes experimental observations, a comprehensive model applies to a wide range of experimental conditions without modification, and a predictive model can be applied to various different chemical conditions.

Description of Models

Surface complexation models are chemical models that give a molecular description of adsorption phenomena using an equilibrium approach. Analogous to complex formation in solution, surface complexation models define surface species, chemical reactions, equilibrium constants, mass balances, and charge balances and their molecular features can be given thermodynamic significance. One of the major

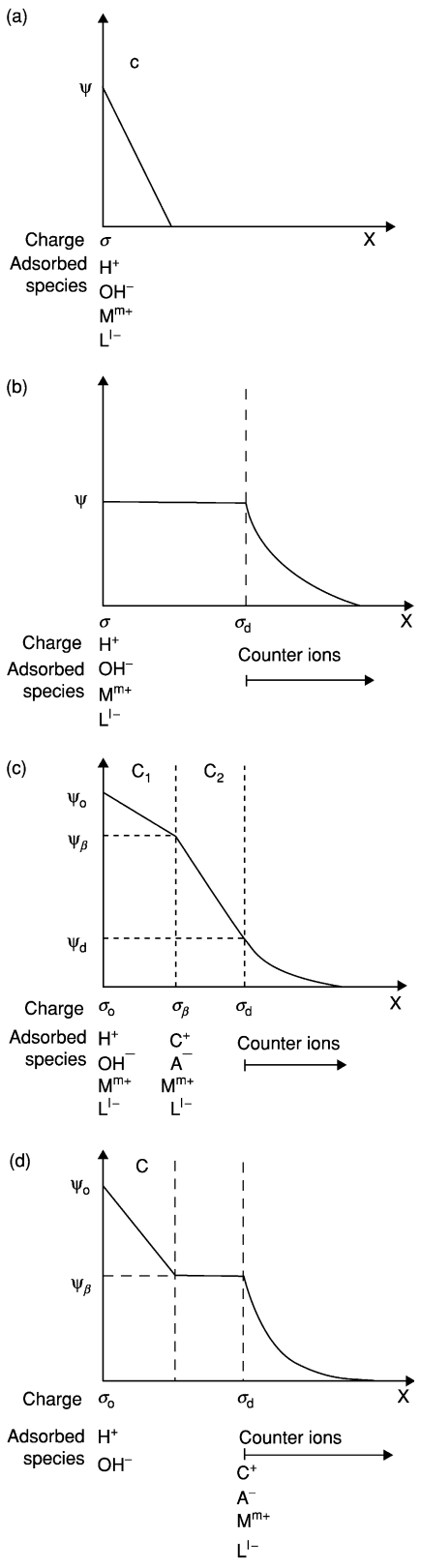


Figure 1 Placement of ions, potentials, charges, and capacitances for the (a) constant capacitance model, (after Westall (1986); (b) diffuse layer model, (after Dzombak DA and Morel FMM (1990) *Surface Complexation Modeling. Hydrous Ferric Oxide*. New York: John Wiley); (c) triple layer model, after Westall (1980); (d) one-pK model, (after Westall (1986)). (a) and (d) from Davis JA and Hayes K (eds) *Geochemical Processes at Mineral Surfaces*, ACS symposium Series 323: 54–78, Copyright (1986) American Chemical Society; (b) reprinted with permission of John Wiley from Dzombak DA and Morel FMM (1990) *Surface Complexation Modeling. Hydrous Ferric Oxide*; (c) reprinted with permission from Kavanaugh MC and Leckie JO (eds) *Particulates in Water Characterization, Fate, Effects and Removal*. ACS Advances in Chemistry Series 189: 33–44, Copyright (1980) American Chemical Society.

advancements of surface complexation models is consideration of the charge on both the adsorbing ion and the solid adsorbent surface. Surface complexation models constitute a family of models having many common characteristics and adjustable parameters. The models differ in their structural representation of the solid–solution interface, i.e., the location and surface configuration of the adsorbed ions.

Surface Configuration of the Solid–Solution Interface

Members of the surface complexation model family include the two-pK models: constant capacitance, diffuse-layer, triple-layer. Two-pK models are based on a reactive surface functional group, SOH, that undergoes both protonation and dissociation:

$$\mathrm{SOH} + \mathrm{H^+} \leftrightarrow \mathrm{SOH_2^+} \qquad [1]$$

$$\mathrm{SOH} \leftrightarrow \mathrm{SO^-} + \mathrm{H^+} \qquad [2]$$

hence the term two-pK model.

Comparable models can be written based on the one-pK concept. So far, the one-pK model has been developed based on the Stern model. In the one-pK model, surface functional groups carry either one or two protons, SOH and $\mathrm{SOH_2}$, respectively. Surface charging can be represented with one reaction:

$$\mathrm{SOH^{1/2-}} + \mathrm{H^+} \leftrightarrow \mathrm{SOH_2^{1/2+}} \qquad [3]$$

The location and surface configuration of the adsorbed ions for the various surface complexation models are presented in Figure 1. In the constant capacitance model and the diffuse layer model all surface complexes are inner-sphere and are located in a single surface plane (Figure 1a, b). The diffuse-layer model includes a diffuse layer that commences at the d-plane and extends into solution. In the triple-layer model ions forming inner-sphere surface complexes adsorb in the surface o-plane and ions forming outer-sphere surface complexes adsorb in a β-plane located between the o-plane and the d-plane (Figure 1c). In the representation of the one-pK model based on the Stern model indicated in Figure 1d, protons and hydroxyls form inner-sphere surface complexes

located in the o-plane; all other ions form outer-sphere surface complexes and are located in the d-plane.

Surface Complexation Reactions

In the two-pK models chemical reactions for inner-sphere surface complexation are eqns [1], [2], and:

$$\mathrm{SOH} + \mathrm{M^{m+}} \leftrightarrow \mathrm{SOM^{(m-1)}} + \mathrm{H^+} \quad [4]$$

$$2\mathrm{SOH} + \mathrm{M^{m+}} \leftrightarrow \mathrm{(SO)_2M^{(m-2)}} + 2\mathrm{H^+} \quad [5]$$

$$\mathrm{SOH} + \mathrm{L^{l-}} + \mathrm{H^+} \leftrightarrow \mathrm{SL^{(l-1)-}} + \mathrm{H_2O} \quad [6]$$

$$2\mathrm{SOH} + \mathrm{L^{l-}} + 2\mathrm{H^+} \leftrightarrow \mathrm{S_2L^{(l-2)-}} + 2\mathrm{H_2O} \quad [7]$$

$$\mathrm{SOH} + \mathrm{L^{l-}} + 2\mathrm{H^+} \leftrightarrow \mathrm{SHL^{(l-2)-}} + \mathrm{H_2O} \quad [8]$$

where M is a metal ion of charge m^+ and L is a ligand of charge l^-. Surface complexes can be monodentate or bidentate. In monodentate complexation one bond is formed between the adsorbing ion and the surface functional group. Bidentate complexes contain two bonds between the adsorbing ion and two surface functional groups. Equations [4] to [7] are used in the constant capacitance model. In the diffuse-layer model, reaction [8] is used in place of reaction [7] since bidentate complexes for adsorbed anions have not been considered.

The following chemical reactions for outer-sphere surface complexation are considered in the triple-layer model in addition to eqns [1], [2], and [4] to [7]:

$$\mathrm{SOH} + \mathrm{M^{m+}} \leftrightarrow \mathrm{SO^-} - \mathrm{M^{m+}} + \mathrm{H^+} \quad [9]$$

$$\mathrm{SOH} + \mathrm{M^+} + \mathrm{H_2O} \leftrightarrow \mathrm{SO^-} - \mathrm{MOH^{(m-1)}} + 2\mathrm{H^+} \quad [10]$$

$$\mathrm{SOH} + \mathrm{H^+} + \mathrm{L^{l-}} \leftrightarrow \mathrm{SOH_2^+} - \mathrm{L^{l-}} \quad [11]$$

$$\mathrm{SOH} + 2\mathrm{H^+} + \mathrm{L^{l-}} \leftrightarrow \mathrm{SOH_2^+} - \mathrm{LH^{(l-1)-}} \quad [12]$$

$$\mathrm{SOH} + \mathrm{C^+} \leftrightarrow \mathrm{SO^-} - \mathrm{C^+} + \mathrm{H^+} \quad [13]$$

$$\mathrm{SOH} + \mathrm{H^+} + \mathrm{A^-} \leftrightarrow \mathrm{SOH_2^+} - \mathrm{A^-} \quad [14]$$

where C^+ is the cation and A^- is the anion of the background electrolyte.

In the one-pK models, chemical reactions for surface complexation are eqn [3] and:

$$\mathrm{SOH^{1/2-}} + \mathrm{M^+} \leftrightarrow \mathrm{SOH^{1/2-}} - \mathrm{M^{m+}} \quad [15]$$

$$\mathrm{SOH^{1/2-}} + \mathrm{M^{m+}} + \mathrm{H_2O} \leftrightarrow \mathrm{SOH^{1/2-}} - \mathrm{MOH^{(m-1)}} + \mathrm{H^+} \quad [16]$$

$$\mathrm{SOH_2^{1/2+}} + \mathrm{L^{l-}} \leftrightarrow \mathrm{SOH_2^{1/2+}} - \mathrm{L^{l-}} \quad [17]$$

$$\mathrm{SOH^{1/2-}} + \mathrm{C^+} \leftrightarrow \mathrm{SOH^{1/2-}} - \mathrm{C^+} \quad [18]$$

$$\mathrm{SOH_2^{1/2+}} + \mathrm{A^-} \leftrightarrow \mathrm{SOH_2^{1/2+}} - \mathrm{A^-} \quad [19]$$

Equilibrium Constants for Surface Complexation

The equilibrium constants describing inner-sphere surface complexation in the two-pK models are:

$$K_+ = \frac{[\mathrm{SOH_2^+}]}{[\mathrm{SOH}][\mathrm{H^+}]}\exp[\mathrm{F}\Psi/RT] \quad [20]$$

$$K_- = \frac{[\mathrm{SO^-}][\mathrm{H^+}]}{[\mathrm{SOH}]}\exp[-\mathrm{F}\Psi/RT] \quad [21]$$

$$K_\mathrm{M}^1 = \frac{[\mathrm{SOM^{(m-1)}}][\mathrm{H^+}]}{[\mathrm{SOH}][\mathrm{M^{m+}}]}\exp[(\mathrm{m}-1)\mathrm{F}\Psi/RT] \quad [22]$$

$$K_\mathrm{M}^2 = \frac{[\mathrm{(SO)_2M^{(m-2)}}][\mathrm{H^+}]^2}{[\mathrm{SOH}]^2[\mathrm{M^{m+}}]}\exp[(\mathrm{m}-2)\mathrm{F}\Psi/RT] \quad [23]$$

$$K_\mathrm{L}^1 = \frac{[\mathrm{SL^{(l-1)-}}]}{[\mathrm{SOH}][\mathrm{L^{l-}}][\mathrm{H^+}]}\exp[-(\mathrm{l}-1)\mathrm{F}\Psi/RT] \quad [24]$$

$$K_\mathrm{L}^2 = \frac{[\mathrm{S_2L^{(l-2)-}}]}{[\mathrm{SOH}]^2[\mathrm{L^{l-}}][\mathrm{H^+}]^2}\exp[-(\mathrm{l}-2)\mathrm{F}\Psi/RT] \quad [25]$$

$$K_\mathrm{L}^3 = \frac{[\mathrm{SHL^{(l-2)-}}]}{[\mathrm{SOH}][\mathrm{L^{l-}}][\mathrm{H^+}]^2}\exp[-(\mathrm{l}-2)\mathrm{F}\Psi/RT] \quad [26]$$

where F is the Faraday constant, Ψ is the surface potential, R is the molar gas constant, T is the absolute temperature, and square brackets represent concentrations. The exponential terms are correction factors accounting for the effect of surface charge on surface complexation. Equations [20] to [25] are considered in the constant capacitance model where $\Psi = \Psi_o$. In the diffuse layer model, eqn [25] is replaced with eqn [26] and $\Psi = \Psi_d$.

Equilibrium constants for outer-sphere surface complexation in the triple layer model are:

$$K_{\mathrm{M}}^{3} = \frac{[\mathrm{SO^{-} - M^{m+}}][\mathrm{H^{+}}]}{[\mathrm{SOH}][\mathrm{M^{m+}}]} \exp[\mathrm{F}(m\Psi_{\beta} - \Psi_{\mathrm{o}})/RT] \quad [27]$$

$$K_{\mathrm{M}}^{4} = \frac{[\mathrm{SO^{-} - MOH^{(m-1)}}][\mathrm{H^{+}}]^{2}}{[\mathrm{SOH}][\mathrm{M^{m+}}]} \exp[\mathrm{F}((m-1)\Psi_{\beta} - \Psi_{\mathrm{o}})/RT] \quad [28]$$

$$K_{\mathrm{L}}^{4} = \frac{[\mathrm{SOH_2^{+} - L^{l-}}]}{[\mathrm{SOH}][\mathrm{H^{+}}][\mathrm{L^{l-}}]} \exp[\mathrm{F}(\Psi_{\mathrm{o}} - l\Psi_{\beta})/RT] \quad [29]$$

$$K_{\mathrm{L}}^{5} = \frac{[\mathrm{SOH_2^{+} - LH^{(l-1)-}}]}{[\mathrm{SOH}][\mathrm{H^{+}}]^{2}[\mathrm{L^{l-}}]} \exp[\mathrm{F}(\Psi_{\mathrm{o}} - (l-1)\Psi_{\beta})/RT] \quad [30]$$

$$K_{\mathrm{C}} = \frac{[\mathrm{SO^{-} - C^{+}}][\mathrm{H^{+}}]}{[\mathrm{SOH}][\mathrm{C^{+}}]} \exp[\mathrm{F}(\Psi_{\beta} - \Psi_{\mathrm{o}})/RT] \quad [31]$$

$$K_{\mathrm{A}} = \frac{[\mathrm{SOH_2^{+} - A^{-}}]}{[\mathrm{SOH}][\mathrm{H^{+}}][\mathrm{A^{-}}]} \exp[\mathrm{F}(\Psi_{\mathrm{o}} - \Psi_{\beta})/RT] \quad [32]$$

In the one-pK model the equilibrium constants for surface complexation are:

$$K_{\mathrm{H}} = \frac{[\mathrm{SOH_2^{1/2+}}]}{[\mathrm{SOH^{1/2-}}][\mathrm{H^{+}}]} \exp[\mathrm{F}\Psi_{\mathrm{o}}/RT] \quad [33]$$

$$K_{\mathrm{M}}^{1} = \frac{[\mathrm{SOH^{1/2-} - M^{m+}}]}{[\mathrm{SOH^{1/2-}}][\mathrm{M^{m+}}]} \exp[\mathrm{mF}\Psi_{\mathrm{d}}/RT] \quad [34]$$

$$K_{\mathrm{M}}^{2} = \frac{[\mathrm{SOH^{1/2-} - MOH^{(m-1)}}][\mathrm{H^{+}}]}{[\mathrm{SOH^{1/2-}}][\mathrm{M^{m+}}]} \exp[(m-1)\mathrm{F}\Psi_{\mathrm{d}}/RT] \quad [35]$$

$$K_{\mathrm{L}} = \frac{[\mathrm{SOH_2^{1/2+} - L^{l-}}]}{[\mathrm{SOH_2^{1/2+}}][\mathrm{L^{l-}}]} \exp[-l\mathrm{F}\Psi_{\mathrm{d}}/RT] \quad [36]$$

$$K_{\mathrm{C}} = \frac{[\mathrm{SOH^{1/2-} - C^{+}}]}{[\mathrm{SOH^{1/2-}}][\mathrm{C^{+}}]} \exp[\mathrm{F}\Psi_{\mathrm{d}}/RT] \quad [37]$$

$$K_{\mathrm{A}} = \frac{[\mathrm{SOH_2^{1/2+} - A^{-}}]}{[\mathrm{SOH_2^{1/2+}}][\mathrm{A^{-}}]} \exp[-\mathrm{F}\Psi_{\mathrm{d}}/RT] \quad [38]$$

Mass and Charge Balances

The mass balance expression for the surface functional group, SOH, in the two-pK models is:

$$\begin{aligned} S_{\mathrm{T}} = {} & [\mathrm{SOH}] + [\mathrm{SOH_2^{+}}] + [\mathrm{SO^{-}}] + [\mathrm{SOM^{(m-1)}}] \\ & + 2[(\mathrm{SO})_2\mathrm{M^{(m-2)}}] + [\mathrm{SL^{(l-1)-}}] + 2[\mathrm{S_2L^{(l-2)-}}] \\ & + [\mathrm{SHL^{(l-2)-}}] + [\mathrm{SO^{-} - M^{m+}}] + [\mathrm{SO^{-}} \\ & - \mathrm{MOH^{(m-1)}}] + [\mathrm{SOH_2^{+} - L^{l-}}] + [\mathrm{SOH_2^{+}} \\ & - \mathrm{LH^{(l-1)-}}] + [\mathrm{SO^{-} - C^{+}}] + [\mathrm{SOH_2^{+} - A^{-}}] \end{aligned} \quad [39]$$

The mass balance for the surface functional groups, SOH and SOH_2, in the one-pK model is:

$$\begin{aligned} S_{\mathrm{T}} = {} & [\mathrm{SOH^{1/2-}}] + [\mathrm{SOH_2^{1/2+}}] + [\mathrm{SOH^{1/2-} - M^{m+}}] \\ & + [\mathrm{SOH^{1/2-} - MOH^{(m-1)}}] + [\mathrm{SOH_2^{1/2+} - L^{l-}}] \\ & + [\mathrm{SOH^{1/2-} - C^{+}}] + [\mathrm{SOH_2^{1/2+} - A^{-}}] \end{aligned} \quad [40]$$

The mass balance represents a summation of all surface species considered in the particular surface complexation model.

The charge balance expressions for the two-pK models are:

$$\begin{aligned} \sigma_{\mathrm{o}} = {} & [\mathrm{SOH_2^{+}}] + [\mathrm{SOH_2^{+} - L^{l-}}] + [\mathrm{SOH_2^{+} - LH^{(l-1)-}}] \\ & + (m-1)[\mathrm{SOM^{(m-1)}}] + (m-2)[(\mathrm{SO})_2\mathrm{M^{(m-2)}}] \\ & + [\mathrm{SOH_2^{+} - A^{-}}] - [\mathrm{SO^{-}}] - [\mathrm{SO^{-} - M^{m+}}] \\ & - [\mathrm{SO^{-} - MOH^{(m-1)}}] - (l-1)[\mathrm{SL^{(l-1)-}}] \\ & - (l-2)[\mathrm{S_2L^{(l-2)-}}] - (l-2)[\mathrm{SHL^{(l-2)-}}] \\ & - [\mathrm{SO^{-} - C^{+}}] \end{aligned} \quad [41]$$

$$\begin{aligned} \sigma_{\beta} = {} & m[\mathrm{SO^{-} - M^{m+}}] + (m-1)[\mathrm{SO^{-} - MOH^{(m-1)}}] \\ & + [\mathrm{SO^{-} - C^{+}}] - l[\mathrm{SOH_2^{+} - L^{l-}}] - (l-1) \\ & [\mathrm{SOH_2^{+} - LH^{(l-1)-}}] - [\mathrm{SOH_2^{+} - A^{-}}] \end{aligned} \quad [42]$$

$$\sigma_{\mathrm{o}} + \sigma_{\beta} + \sigma_{\mathrm{d}} = 0 \quad [43a]$$

$$\sigma_{\mathrm{o}} + \sigma_{\mathrm{d}} = 0 \quad [43b]$$

where σ is the surface charge. The charge balances, eqns [41] and [42], represent the summation of all charge contributions in a particular plane of adsorption. All of the models consider charge balance in the surface plane, eqn [41]. Charge balance in the β-plane, eqn [42], is restricted to the triple-layer model. Charge balance eqn [43a] is considered in the triple-layer model while eqn [43b] is used in the diffuse-layer model. The charge balance expressions for the one-pK model are eqn [43b] and:

$$\sigma_o = \frac{1}{2}([SOH_2^{1/2+}] - [SOH^{1/2-}]) \quad [44]$$

Charge-Potential Relationships

All surface complexation models contain relations between surface charges and surface potentials. In the constant capacitance model the charge-potential relationship is:

$$\sigma = C\Psi \quad [45]$$

where C is the capacitance. The charge potential relationship for the diffuse layer model is:

$$\sigma_d = -(8\epsilon_o DRTI)^{1/2} \sinh(F\Psi_d/2RT) \quad [46]$$

where ϵ_o is the permittivity of vacuum, D is the dielectric constant of water, and I is the solution ionic strength. In the triple-layer model the charge potential relationships are eqn [46] and:

$$\sigma_o = C_1(\Psi_o - \Psi_\beta) \quad [47]$$

$$\sigma_d = C_2(\Psi_d - \Psi_\beta) \quad [48]$$

The charge-potential relationships for the one-pK model are eqn [46] and:

$$\sigma_o = C(\Psi_o - \Psi_d) \quad [49]$$

Obtaining Values of Adjustable Parameters

Surface Site Density

The total number of reactive surface functional groups, S_T is an important adjustable parameter in the surface complexation models and is related to the surface site density:

$$S_T = \frac{Sa10^{18}}{N_A} N_S \quad [50]$$

where S is the surface area, a is the particle concentration, and N_A is Avogadro's number. Experimental methods for determining surface site density include: tritium exchange, potentiometric titration, fluoride adsorption, and maximum adsorption. Values of this parameter can also be calculated from crystal dimensions or optimized to fit experimental adsorption data. Various determinations of surface site density vary by an order of magnitude; the lowest values are obtained from crystallographic calculations while tritium exchange yields the highest values. Uncertainty in the value of the surface site density is a major limitation in the use of surface complexation models since the ability of the models to describe adsorption is sensitively dependent on this value. To standardize surface complexation modeling, a fixed value of 2.31 sites nm^{-2} has been used for many natural materials. Applications of the diffuse-layer model to metal adsorption have split the total number of reactive surface functional groups into a 'strong,' S_s, and a 'weak,' S_w, set of adsorption sites. This approach greatly increases the number of adjustable parameters since each set of sites, S_i, has its own protonation, dissociation, and metal surface complexation constants.

Capacitances

Some values of capacitance (C in the constant capacitance and one-pK model and C_1 in the triple layer model) can be obtained graphically from slopes of protonation-dissociation constants versus surface charge. Alternatively, both capacitances, C_1 and C_2, in the triple-layer model can be obtained using an electrokinetic extrapolation technique. Capacitance values obtained experimentally usually exhibit great variability; therefore, capacitances have generally been optimized to fit the titration data.

Surface Complexation Constants

Values of the protonation and dissociation constants in the constant capacitance model and the triple-layer model can be obtained from the same graphs used to obtain values of capacitance. These constants can also be obtained by optimizing titration data using a computer program. Values of the surface complexation constants for ion adsorption are obtained using computer optimization. An advantage of computer optimization, in addition to ease of use, is that it yields bias-free parameters with standard deviations and quality-of-fit criteria. Individual optimized equilibrium constant values can be weighted to obtain overall best estimates of the parameter:

$$\overline{\log K} = \sum \frac{(1/\sigma_{\log K})_i}{\sum(1/\sigma_{\log K})_i} [\log K]_i \quad [51]$$

For the diffuse layer model a set of best estimates of $\log K$ are available for a variety of adsorbing cations and anions. The advantage of this data set is that the surface complexation constants are all self-consistent; i.e., all ion surface complexation constants were optimized using the same values of protonation-dissociation constants and surface site density. This is an important point since parameter values in the surface complexation models are interdependent. Additionally, since each surface complexation model contains a different set of assumptions for the solid–solution interface, surface complexation

constants from one model must not be used in any other model.

Applications to Ion Adsorption on Natural Samples

All surface complexation models were originally developed to describe charging behavior and ion adsorption of ions on oxide minerals. Various curves are commonly used to describe adsorption behavior: adsorption isotherms, adsorption edges, and adsorption envelopes. Adsorption isotherms describe ion adsorption as a function of equilibrium ion concentration, usually at fixed solution pH. Adsorption edges and adsorption envelopes both describe ion adsorption as a function of solution pH at a fixed total ion concentration. Adsorption edge is the term generally applied to cation adsorption, while the term adsorption envelope is used to describe anion adsorption.

The most commonly studied oxide surfaces with surface complexation models have been the iron oxides goethite and ferrihydrite. Subsequently, the models were extended to include adsorption on clay minerals, organic materials, and soil samples. In extending the models to natural samples certain approximations and modifications are necessary. In the application to natural systems, such as clay minerals or soils, the assumption is made that ion adsorption occurs through interaction with the hydroxyl groups at the edges of the clay particles. The effect of permanent negatively charged sites at the clay basal planes on this adsorption process is ignored. This simplification may not be appropriate, especially for anions whose edge adsorption may be affected by this negative charge.

The surface complexation models contain the assumption that ion adsorption takes place on one or at most two sets of reactive surface sites. This is clearly an oversimplification since even simple oxide minerals contain several sets of reactive hydroxyl groups. However, this simplification is necessary to maintain the number of adjustable parameters at a reasonable level. Natural materials such as soils are complex, multisite mixtures having a variety of reactive surface functional groups. Thus surface complexation constants determined for soils represent average composite values for all these sets of reactive surface functional groups.

Constant Capacitance Model

The constant capacitance model has been used to describe adsorption on silicon, aluminum, iron, and titanium oxides, kaolinite, montmorillonite, and illite clay minerals, plant cell walls, and soils. Adsorbing ions that have been investigated include the cation and metal ions: calcium, cesium, lead, copper, cadmium, zinc, nickel, cobalt, aluminum, iron, manganese, silver, mercury, lanthanum, europium, ytterbium, and the anions: phosphate, sulfate, arsenate, arsenite, selenite, selenate, borate, molybdate, silicate, fluoride, phthalate, salicylate, benzoate, citrate.

Examples of the fit of the constant capacitance model to trace metal adsorption edges are provided in Figure 2 for iron, lead, copper, and cadmium adsorption on silica. As for many trace metal cations, the amount of adsorption increases rapidly from 0 to 100% over a narrow pH range. The model is well able to describe these changes in adsorption for the four different metal ions. Figure 3 indicates the ability

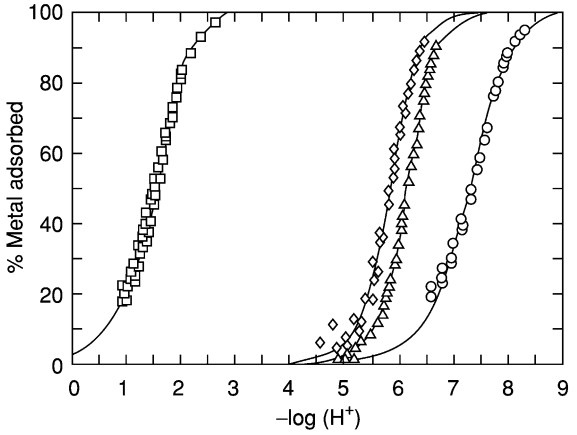


Figure 2 Fit of the constant capacitance model to metal adsorption on silica. Model results are represented by solid lines. □ Fe; ◇ Pb; △ Cu; ○ Cd. Reprinted with permission from Schindler PW, Fürst B, Dick R, and Wolf PU (1976) Ligand properties of surface silanol groups, I. Surface complex formation with Fe^{3+}, Cu^{2+}, Cd^{2+}, and Pb^{2+}. *Journal of Colloid Interface Science* 55: 469–475.

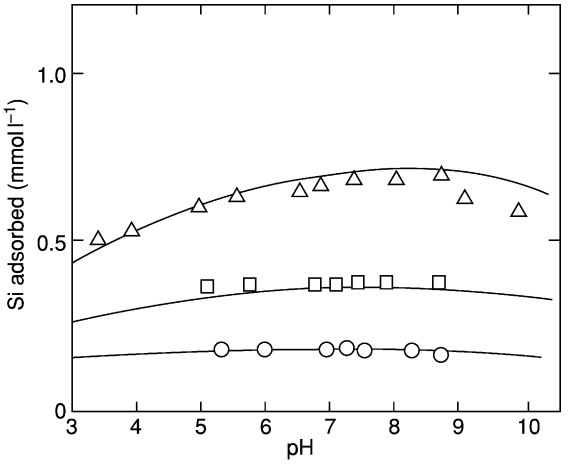


Figure 3 Fit of the constant capacitance model to silicate adsorption on goethite. Model results are represented by solid lines. ○ 2×10^{-4} M; □ 4×10^{-4} M; △ 8×10^{-4} M. Reprinted from Sigg LM and Stumm W (1981) The interaction of anions and weak acids with the hydrous goethite (α-FeOOH) surface. In: *Colloids and Surfaces*, vol. 2, pp. 101–117. Amsterdam: Elsevier Science.

of the constant capacitance model to fit adsorption envelopes for the anion, silicate, on to the iron oxide, goethite at various initial silicate concentrations. Silicate adsorption is nearly constant over most of the pH range. The model is able to describe the adsorption, including the pH dependence observed at the highest initial silicate concentration.

Applications of the constant capacitance model to soils have been restricted to anions: phosphate, selenite, borate, and molybdate. For the application of the model to selenite adsorption by soils, two sets of reactive surface functional groups were postulated; monodentate surface complexes were formed on one set of sites and bidentate surface complexes on the other. The model was initially applied to one Californian soil. As can be seen in Figure 4a, the fit of the model to the data is good. Subsequently, the model parameters obtained in fitting this soil were used to predict adsorption on additional Californian soils. Figure 4b shows that this prediction was qualitatively successful and indicates some predictive capability of the model for soils of somewhat similar chemical and physical characteristics.

An alternative approach has been developed for describing borate adsorption on soils. From the fitted surface complexation constants for a set of soils, a

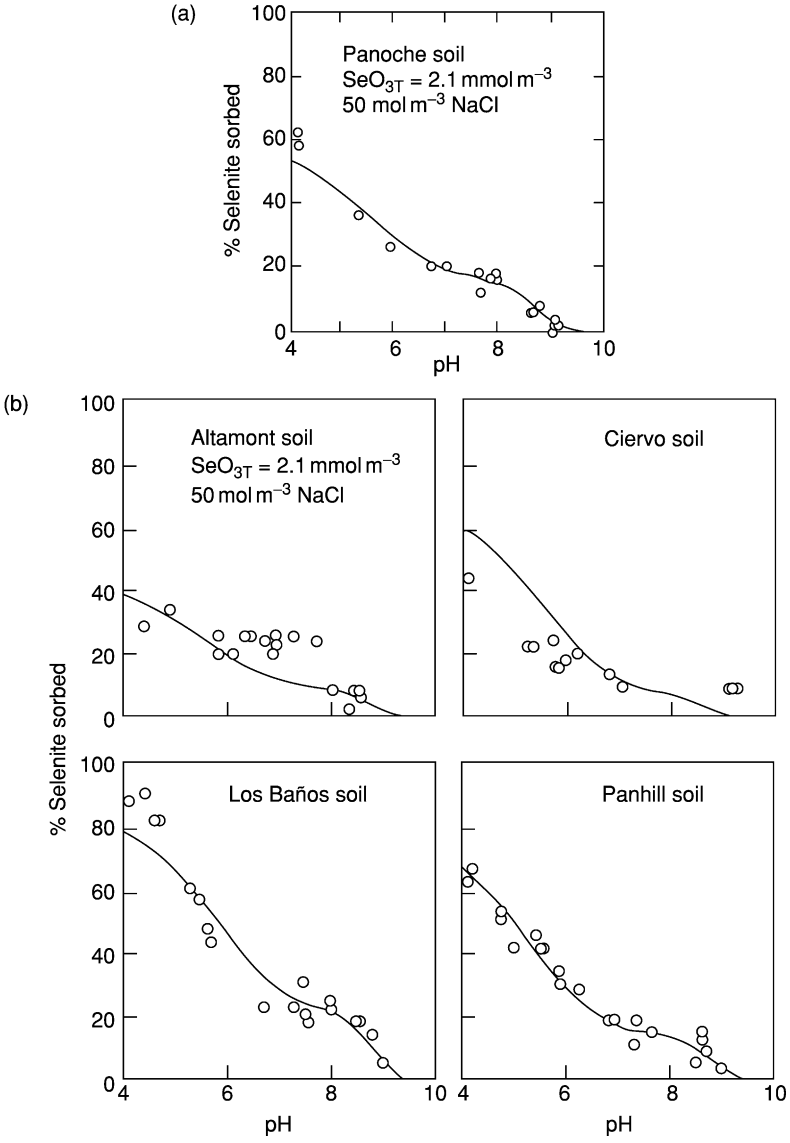


Figure 4 (a) Fit of the constant capacitance model to selenite adsorption on a California soil. Model fit is represented by a solid line. (b) Constant capacitance model predictions of selenite adsorption by California soils. Model predictions are represented by solid lines. Reprinted with permission from Sposito G, de Wit JCM, and Neal RH (1988) Selenite adsorption on alluvial soils: III. Chemical modeling. *Soil Science Society of America Journal* 52: 947–950.

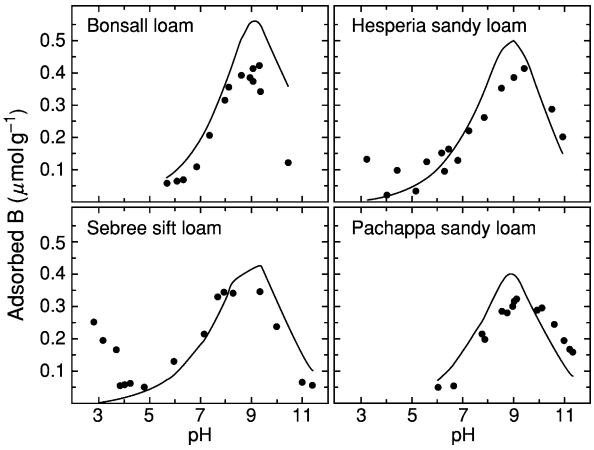


Figure 5 Constant capacitance model predictions of boron adsorption by soils of various soil chemical properties and diverse soil orders. Model predictions are represented by solid lines. Modified from Goldberg S, Lesch SM, and Suarez DL (2000) Predicting boron adsorption by soils using soil chemical parameters in the constant capacitance model. *Soil Science of Society of America Journal* 64: 1356–1363.

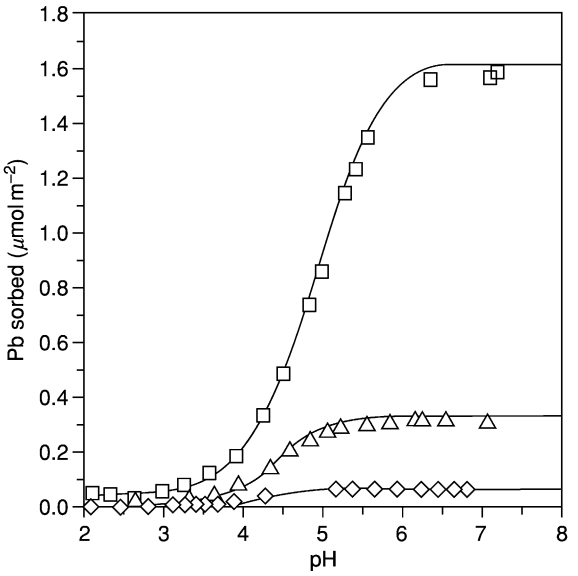


Figure 6 Fit of the diffuse layer model to lead adsorption on hematite. Model fits are represented by solid lines. □ 100 μmol l^{-1}; △ 20 μmol l^{-1}; ◇ 4 μmol l^{-1}. Reproduced with permission from Christl DI and Kretzschmar R (1999) Competitive sorption of copper and lead at the oxide–water interface: implications for surface site density. *Geochimica et Cosmochimica Acta* 63: 2929–2938.

general regression model was obtained that predicts the surface complexation constants for new soils from easily measured chemical parameters: surface area, organic carbon content, inorganic carbon content, and aluminum oxide content. These surface complexation constants were then used in the constant capacitance model to predict borate adsorption on the new soils. This approach constitutes a completely independent model evaluation that was able to predict borate adsorption on a diverse set of soils having a wide range of chemical properties, as indicated in Figure 5.

Diffuse-Layer Model

The diffuse-layer model has been used to describe adsorption on iron, aluminum, manganese, titanium, and silicon oxides, kaolinite, montmorillonite and biotite minerals, natural organic matter, bacterial cell walls, and sediments. Adsorbing ions that have been investigated include the cation and metal ions: calcium, strontium, barium, copper, nickel, zinc, cadmium, lead, cobalt, aluminum, chromium, silver, mercury, uranium and the anions: phosphate, sulfate, selenite, selenate, arsenate, arsenite, borate, chromate, fluoride, vanadate, thiosulfate, oxalate, phthalate, salicylate, benzoate, and fulvate.

The ability of the diffuse-layer model to describe metal adsorption edges is indicated in Figure 6 for lead adsorption on the iron oxide, hematite. The model was able to describe the data very well at three significantly different initial lead concentrations over a wide range of solution pH values. Figure 7 demonstrates the ability of the diffuse-layer model to fit

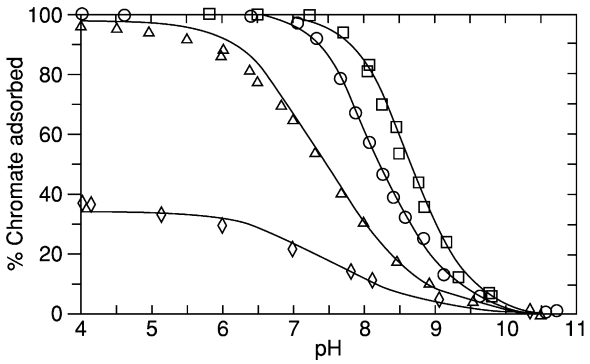


Figure 7 Fit of the diffuse layer model to chromate adsorption on goethite. Model fits are represented by solid lines. □ 0.01 mmol l^{-1}; ○ 0.05 mmol l^{-1}; △ 0.2 mmol l^{-1}; ◇ 0.8 mmol l^{-1}. Reprinted with permission from Mesuere K and Fish W (1992) Chromate and oxalate adsorption on goethite: I. Calibration of surface complexation models. *Environmental Science and Technology* 26: 2357–2364. Copyright American Chemical Society.

adsorption envelopes for the chromate anion on goethite. In order to fit the two intermediate chromate concentrations it was necessary to add a third chromate surface complex, [$SCrO_4^{3-}$]. With this addition, the model is well able to describe chromate adsorption at all initial chromate concentrations over a wide range of solution pH.

Applications of the diffuse-layer model to soil systems have not been carried out to date. Adsorption

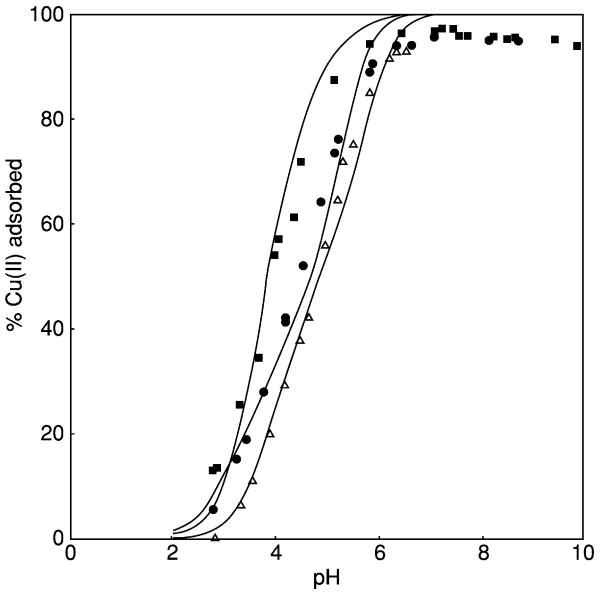


Figure 8 Fit of the diffuse layer model to copper adsorption on lignocellulose extracted from wheat bran. Model fits are represented by solid lines. △ $[Cu(II)]_{ini} = 2.10^{-4}$ mol l^{-1}, I = 0.1; ● $[Cu(II)]_{ini} = 2.10^{-4}$ mol l^{-1}, I = 0.01; ■ $[Cu(II)]_{ini} = 2.10^{-5}$ mol l^{-1}, I = 0.1. Reproduced with permission from Ravat C, Dumonceau J, and Monteil-Rivera F (2000) Acid/base and Cu(II) binding properties of natural organic matter extracted from wheat bran: modeling by the surface complexation model. *Water Research* 34: 1327–1329.

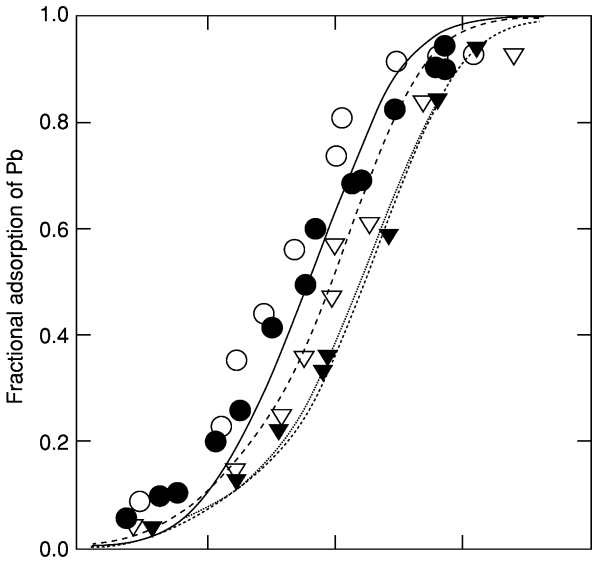


Figure 9 Fit of the triple-layer model to lead adsorption on goethite. Model fits are represented by solid lines. ○ 9.65 μmol l^{-1}; ● 24.1 μmol l^{-1}; 48.3 μmol l^{-1}; ▼ 72.4 μmol l^{-1}. Reproduced with permission from Kooner ZS, Cox CD, and Smoot JL (1995) Prediction of adsorption of divalent heavy metals at the goethite/water interface by surface complexation modeling. *Environmental Toxicology and Chemistry* 14: 2077–2083. Copyright SETAC, Pensacola, Florida, USA.

of various metal ions has been investigated on a lignocellulose organic substrate extracted from wheat bran. In these applications, two sets of reactive surface functional groups are considered representing carboxylic and phenolic sites. Figure 8 presents the ability of the diffuse-layer model to describe copper adsorption edges on this organic material as a function of initial copper concentration $[Cu_{ini}]$, solution pH, and ionic strength, I. In this application, the model is fitted to the data for $[Cu_{ini}] = 2 \times 10^{-4}$ M, $I = 0.1$. The model parameters resulting from this optimization were then used to predict the remaining model results depicted in Figure 8. The diffuse layer model is well able to describe copper adsorption on this natural material.

Triple-Layer Model

The triple-layer model has been used to describe adsorption on iron, aluminum, manganese, and silicon oxides, kaolinite and smectite clay minerals, and soils. Adsorbing ions that have been investigated include the cation and metal ions: sodium, potassium, calcium, magnesium, lead, zinc, cadmium, copper, cobalt, silver, mercury, uranium, plutonium, thorium, neptunium, and the anions: chloride, nitrate, perchlorate, sulfate, selenite, selenate, arsenate, arsenite, molybdate, borate, chromate, silicate, fluoride, carbonate, oxalate, phthalate, salicylate, lactate, acetate, formate, and humate.

The ability of the triple-layer model to describe metal ion adsorption edges is depicted in Figure 9 for lead adsorption on goethite. In this application an inner-sphere surface complex is assumed. The triple-layer model was well able to describe lead adsorption at various initial concentrations as a function of solution pH. Figure 10 shows the fit of the triple-layer model to anion adsorption envelopes for selenate adsorption on the iron oxide, ferrihydrite. With the assumption of two outer-sphere surface complexes, the model describes the adsorption data quantitatively for four vastly differing initial selenium concentrations as a function of solution pH.

The triple-layer model was able to fit calcium, magnesium, and sulfate adsorption on a Brazilian oxisol and sulfate adsorption on an acidic forest soil. Molybdate adsorption on two Californian soils as a function of solution pH and ionic strength could be described using either an inner-sphere or an outer-sphere adsorption mechanism. The quality of the fit was slightly better with the inner-sphere adsorption mechanism. In this application it was assumed that aluminol groups on the clay edges are the reactive surface functional groups in the soils and that

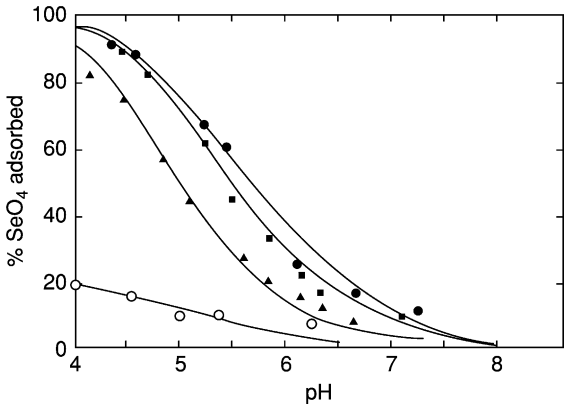


Figure 10 Fit of the triple-layer model to selenate adsorption on amorphous iron oxide. Model fits are represented by solid lines. Total Seo_4 added ● 2×10^{-7} mol l^{-1}; ■ 2×10^{-5} mol l^{-1}; ▲ 2×10^{-4} mol l^{-1}; ○ 2×10^{-3} mol l^{-1}. Reproduced with permission from Davis JA and Leckie JO (1980) Surface ionization and complexation at the oxide/water interface. III. Adsorption of anions. *Journal of Colloid Interface Science* 74: 32–43.

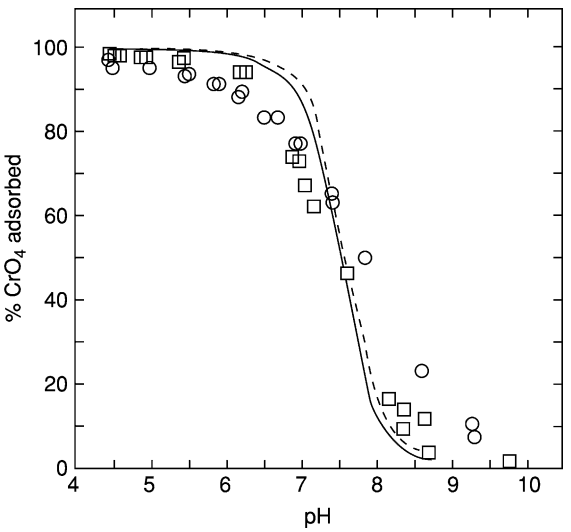


Figure 11 Fit of the triple-layer model to chromate adsorption on two soils. Model fits are represented by solid lines. □ Holston/Cloudland series soil; ○ Cecil/Pacolet series soil. Reprinted with permission from Zachara JM, Ainsworth CC, Cowan CE, and Resch CT (1989) Adsorption of chromate by subsurface soil horizons. *Soil Science of America Journal* 53: 418–428.

surface complexation constants determined for reactive surface hydroxyls of aluminum oxide can be used to describe the surface complexation reactions undergone by these aluminol groups.

In a similar approach, surface complexation constants previously determined for the reactions undergone by aluminum-substituted goethite were used to describe chromate adsorption by two soils. It was assumed that only the iron sites on the aluminum-substituted goethite are involved in chromate

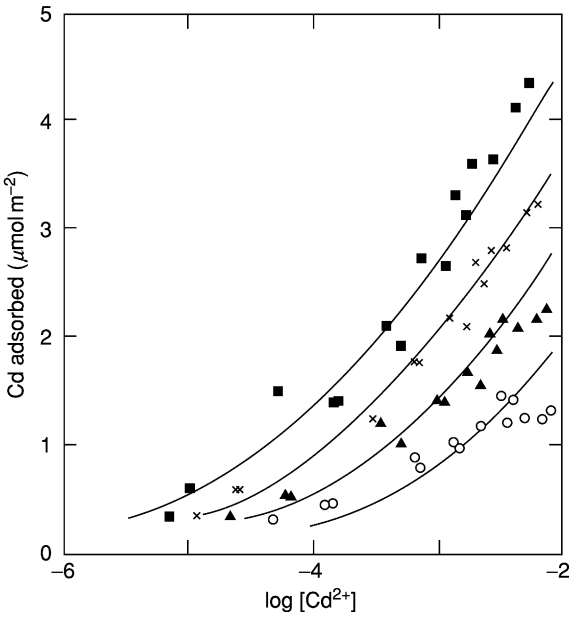


Figure 12 Fit of the one-pK model to cadmium adsorption on amorphous iron oxide. Model fits are represented by solid lines. pH ○ 7.0; ▲ 7.4; × 7.7; ■ 8.0. Reproduced with permission from van Riemsdijk WH, de Wit JCM, Koopal LK, and Bolt GH (1987) Metal adsorption on heterogeneous surfaces: adsorption models. *Journal of Colloid Interface Science* 116: 511–522.

adsorption, forming monodentate outer-sphere surface complexes. The ability of the model to describe chromate adsorption on the two soils as a function of solution pH is indicated in Figure 11. The fit of the model on these heterogeneous materials is qualitatively correct.

One-pK Model

The one-pK model has been used to describe adsorption on iron and aluminum oxides. The vast majority of studies to date have used goethite as the adsorbent material. Adsorbing ions that have been investigated include the cations: potassium, calcium, cadmium, copper, and anions: phosphate, arsenate, selenite, sulfate, chromate, lactate, oxalate, malonate, phthalate, citrate, and fulvate.

The ability of the one-pK model to fit cation adsorption isotherms is indicated in Figure 12 for cadmium adsorption on hematite. The model was well able to describe cadmium adsorption at various solution pH values and initial cadmium concentrations. In this application, consideration of background electrolyte surface complexes was neglected. Figure 13 indicates the ability of the one-pK model to fit anion adsorption isotherms for phosphate adsorption on goethite. The model was well able to describe phosphate adsorption over a very wide range of solution pH values

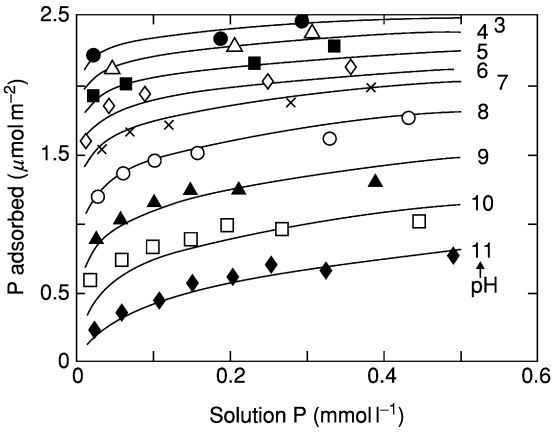


Figure 13 Fit of the one-pK model to phosphate adsorption on goethite. Model fits are represented by solid lines. Reproduced with permission from van Riemsdijk WH and van der Zee SEATM (1991) *Comparisons of Models for Adsorption, Solid Solution and Surface Precipitation*. Dordrecht, the Netherlands: Kluwer (based on experimental data of Bowden *et al.* (1980)) In: Bolt GH *et al.* (eds) *Interactions at the Soil Colloid–Soil Solution Interface*, pp. 241–256. Kluwer.

and initial phosphate concentrations. In this application, a surface complexation constant for potassium adsorption was also optimized.

List of Technical Nomenclature

ϵ_o	Permittivity of vacuum [$C^2\,J^{-1}\,m^{-1}$]
σ	Surface charge density [$C\,m^{-2}$]
σ_β	Surface charge density in the β-plane [$C\,m^{-2}$]
σ_d	Surface charge density in the d-plane [$C\,m^{-2}$]
σ_o	Surface charge density in the o-plane [$C\,m^{-2}$]
$(\sigma_{logK})_I$	Standard deviation for logK of the Ith data set
Ψ	Surface potential [V]
Ψ_β	Surface potential in the β-plane [V]
Ψ_d	Surface potential in the d-plane [V]
Ψ_o	Surface potential in the o-plane [V]
A^-	Anion of the background electrolyte
a	Particle concentration [$g\,l^{-1}$]
C	Capacitance [$F\,m^{-2}$]
C_1	Capacitance between the o- and the β-plane [$F\,m^{-2}$]
C_2	Capacitance between the β- and the d-plane [$F\,m^{-2}$]
C^+	Cation of the background electrolyte
D	Dielectric constant of water
F	Faraday constant [$C\,mol^{-1}$]
I	Ionic strength
K_I	Equilibrium constant
L	Ligand
l^{l-}	Charge on the ligand
M	Metal ion
m^{m+}	Charge on the metal
N_A	Avogadro's number
N_S	Maximum surface site density [$sites\,nm^{-2}$]
R	Molar gas constant [$J\,mol^{-1}\,K^{-1}$]
S	Surface area [$m^2\,g^{-1}$]
SOH	Reactive surface hydroxyls bound to a metal ion S
S_s	High-affinity sites
S_T	Total number of reactive surface functional groups
S_w	Low-affinity sites
T	Absolute temperature [K]

See also: **Clay Minerals**

Further Reading

Bolt GH and van Riemsdijk WH (1982) Ion adsorption on inorganic variable charge constituents. In: Bolt GH (ed.) *Soil Chemistry, Part B. Physicochemical Methods*, pp. 459–503. Amsterdam: Elsevier.

Davis JA and Kent DB (1990) Surface complexation modeling in aqueous geochemistry. *Reviews in Mineralogy* 23: 177–260.

Dzombak DA and Morel FMM (1990) *Surface Complexation Modeling. Hydrous Ferric Oxide*. New York: John Wiley.

Goldberg S (1992) Use of surface complexation models in soil chemical systems. *Advances in Agronomy* 47: 233–329.

Goldberg S (1993) Constant capacitance model: chemical surface complexation model for describing adsorption of toxic trace elements on soil minerals. *American Chemical Society Symposium Series* 518: 278–307.

Goldberg S (1995) Adsorption models incorporated into chemical equilibrium models. In: Loeppert R, Schwab AP, and Goldberg S (eds) *Chemical Equilibrium and Reaction Models*. Soil Science Society of America Special Publication 39, pp. 37–60.

Goldberg S (1998) Ion adsorption at the soil particle-solution interface: modeling and mechanisms. In:

Huang PM, Senesi N, and Buffle J (eds) *IUPAC Series on Analytical and Physical Chemistry of Environmental Systems*, vol. 4. *Structure and Surface Reactions of Soil Particles*, pp. 377–412. Chichester: John Wiley.

James RO and Parks GA (1982) Characterization of aqueous colloids by their electrical double-layer and intrinsic surface chemical properties. *Surface and Colloid Science* 12: 1119–216.

Sposito G (1983) Foundations of surface complexation models of the oxide–aqueous solution interface. *Journal of Colloid and Interface Science* 91: 329–340.

Sposito G (1984) *The Surface Chemistry of Soils*. Oxford: Oxford University Press.

SUSTAINABLE SOIL AND LAND MANAGEMENT

J L Berc, USDA Natural Resources Conservation Service, Washington, DC, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

The concept of sustainability has evolved to meet changing needs, changing realities, technological development, and new understanding of ecosystem functions. At any time in history it is likely that people have hoped to manage their land to sustain its productivity throughout their lifetime and for their offspring. In a new land with abundant resources, it is difficult to imagine depleting resources. Early national policy in the USA encouraged people to settle in the west, to claim the territory, and relieve population pressures in the east. If land was worn out, people could move to the western frontier. By the end of the nineteenth century, however, most of the land was settled, the frontier was 'closed,' and people could no longer count on more available land to replace worn-out land. Many families lacked sufficient knowledge, land, and water to sustain their livelihood on the land. Land was often pushed to produce crops that the soil and climate could not support. By the 1930s the nation was faced with much ruined and abandoned land in the east and a dust bowl and more abandoned farms in the west. From this ecological and human disaster came the public recognition of the need to help farmers and ranchers better balance long-term conservation needs and short-term economic needs so that agricultural production could be sustainable.

The Birth of Soil Conservation in US Agriculture

The vision of the early soil conservationists in the USA still inspires us today:

> Nature treats the earth kindly. Man treats her harshly. He overplows the cropland, overgrazes the pastureland, and overcuts the timberland. He destroys millions of acres completely. He pours fertility year after year into the cities, which in turn pour what they do not use down the sewers into the rivers and the ocean. The flood problem insofar as it is man-made is chiefly the result of overplowing, overgrazing, and overcutting of timber. This terribly destructive process is excusable in a young civilization. It is not excusable in the United States in the year 1938.
>
> *Henry A. Wallace, Secretary of Agriculture*
>
> Among the causes of this misuse of the soil and consequent waste of human resources, lack of knowledge on the part of individuals undoubtedly plays a large part. But even if all farmers thoroughly understood the consequences to themselves of the type of land use they are practicing and had perfect knowledge of soil-conservation techniques, a large number would still be unable to put the knowledge fully into practice. Social and economic limitations would prevent them from doing certain things they knew ought to be done.
>
> History is largely a record of human struggle to wrest the land from nature, because man relies for sustenance on the products of the soil. Yet too frequently man's conquest of the land has been disastrous: over extensive areas, his culture of the earth has resulted in extreme impoverishment or complete destruction of the soil resource on which he is dependent.... Conservation of the soil, in a national sense, requires the adoption of sound land-use principles and practices by agriculture as a whole. The attainment of this objective involves the widespread use of physical measures of land defense and the adjustment of certain economic and social forces tending to encourage exploitation of the soil.
>
> *Hugh H. Bennett, Chief, Soil Conservation Service, and W.C. Lowdermilk, 1938*
>
> Thou shalt inherit the Holy Earth as a faithful steward, conserving its resources and productivity from generation to generation. Thou shalt safeguard thy fields from soil erosion, thy living waters from drying up, thy forests from desolation, and protect thy hills from overgrazing by thy herds, that thy descendants may have abundance forever. If any shall fail in this stewardship of the land thy fruitful fields shall become sterile stony ground and wasting

> gullies, and thy descendants shall decrease and live in poverty or perish from off the face of the earth.
>
> *Walter Clay Lowdermilk, 1939*

W.C. Lowdermilk wrote this 'eleventh commandment' after reporting on the 'graveyard of empires' throughout the 'Old World,' where city after city was ruined not by climate change, but by the failure of their populations to conserve the soil and water resources from which they grew. His report, *Conquest of the Land Through 7000 Years*, was undertaken to help the USA in its struggle with the Dust Bowl and the gullied South. He warned that, if Americans failed to conserve their soil and water resources, the future of the nation would be at risk. He described erosion control and water management practices used at the time to increase water infiltration and reduce storm flow damage; and he called for improvement and adaptation of practices to address increasingly intensive land use.

The concept of sustainability in the 1930s focused on maintaining farm productivity to supply adequate food and fiber for the growing nation. Agricultural production was dependent on the inherent soil and water qualities of the land. Great progress was made during this era, with the introduction of contour cultivation, cover crops, and use of manure and crop rotations to restore soil nutrients and organic matter. The need for zoning, education, financial relief, and improved practices was recognized at the time to address holistically the roots of problems causing soil degradation and loss.

Emerging Agricultural Technology Shifts National Conservation Priorities

The synthetic fertilizer industry was in its early stages of development at the time of the beginning of soil conservation policy in the USA. Over the next generation, commercial nutrients began replacing crop rotations to maintain productivity. Water management at the watershed level was introduced, and soil drainage, irrigation, and land-leveling became major conservation efforts. By the late 1950s, more than 60 million hectares of farmland had been drained. Along with mechanization and new improved crop breeds, these technologies contributed to the doubling of crop and animal production between 1950 and the early 1990s. In this new technology frontier, the link between productivity and land quality was muted, and the impetus to conserve inherent soil productivity declined.

Erosion control was still encouraged to maintain the storage capacity of dams being constructed in watersheds throughout the country to prevent flooding and maintain adequate water supplies. The concept of 'tolerable soil loss' was developed in the 1950s from the perspective of the cost to the farmers of nutrients that would be lost from the eroded soil. The *1957 Yearbook of Agriculture: Soil* recommended that in the Corn Belt, depending on soil depth, mean annual soil losses should be less than 4.5–11 t ha^{-1} to avoid damaging the land or causing excessive silting on lower fields or in streams and reservoirs.

Nevertheless, erosion remained a major concern. The 1957 yearbook noted that "in the Great Plains alone, about 14 million acres not suited for permanent cultivation were cultivated in 1955. Much of this land offers low returns and is subject to severe erosion even in average years." The acreage damaged from erosion was equal to that damaged in the dust storms of the 1930s. Despite the long-term economic benefits to the farmer of conservation practices and because increasing numbers of farms were on rented land, long-term conservation investment was often not made.

The Conservation Reserve Program of the Soil Bank was established in 1956 to set aside cropland into soil-conserving crops such as grass and legumes for up to 10 years. By 1960 approximately 11.6 million hectares were enrolled. Objections from rural communities about disruptions to their economies caused by the enrollment of whole farms into this program caused it to be ended. The Agricultural Conservation Program, under which farmers received cost-share assistance for conservation, was made a permanent program in 1957. Conservation practices that were promoted included rotation cropping, terraces, shifts from row crops to pasture, improved grazing management, crop-residue management, and grassed waterways, as well as irrigation and drainage.

The importance of land-use planning and the use of the soil survey to develop land suitable for its intended purpose were introduced in the early 1960s. Concerns related to salt buildup in irrigated crops also emerged at that time.

Recognition of New Environmental Issues

The consolidation and specialization of farms, separation of crop from animal production, and increasing use of agrichemicals to manage production continued into the 1970s. No-till agriculture was introduced in conservation systems and helped to achieve conservation goals without relying on crop rotations or contour cultivation. Agricultural production systems today have further evolved, for example, incorporating genetically modified crops that tolerate herbicides used in reduced tillage. At the same time,

environmental awareness has grown. Soil and water quality are of concern not only for on-farm productivity, but also for human health on the farm, and in both rural and urban communities. The health of the ecosystem itself has been raised as a concern. However, public investment in conservation has not kept pace with these emerging conservation concerns.

The *1970 Yearbook of Agriculture, Contours of Change* noted that total conservation investment had not increased since 1962. A conservation-needs inventory was conducted, and it was found that almost two-thirds of cropland, pastureland, and rangeland needed conservation treatment; water pollution was severe.

In the *1983 Yearbook of Agriculture, Using Our Natural Resources*, erosion was cited as a national problem, despite 50 years of government programs to combat it. Concerns about maintaining farmland in agricultural production were discussed along with economic and environmental issues regarding irrigation and agrichemical use. Fish and wildlife habitat degradation resulting from reduced water levels from irrigation expansion, wetland conversion to cropland, planting fence row to fence row, and removing windbreaks was also raised. The advances of stubble-mulch and no-till management systems (developed in the 1950s) were discussed. Further, the Land Evaluation and Site Assessment system to aid land-use decision-making was introduced to support the 1981 National Farmland Protection Act, designed to slow urbanization of agricultural land.

The *1987 Yearbook of Agriculture, Our American Land* noted that, while technical assistance for conservation had remained stable since 1969, financial assistance declined by 77% between 1969 and 1985. New 'carrot and stick' provisions of the 1985 Farm Bill protected highly erodible land and wetlands through conservation compliance, and a new Conservation Reserve Program (CRP) removed approximately 15 million hectares of highly erodible cropland from production for 10 years.

Defining Sustainable Development

In 1987 The World Commission on Environment and Development (The Brundtland Commission) defined sustainable development as actions taken "... to meet the needs of the present without compromising the ability of future generations to meet their own needs." Agenda 21, The United Nations Programme of Action, developed at the 1992 Rio Earth Summit, declared:

> the destruction and degradation of agricultural and environmental resources is a major issue. Techniques for increasing production and conserving soil and water resources are already available but are not widely or systematically applied. A systematic approach is needed for identifying land uses and production systems that are sustainable in each land and climate zone, including the economic, social and institutional mechanisms necessary for their implementation.

The Agenda set the objective:

> To prepare and implement comprehensive policies and programmes leading to the reclamation of degraded lands and the conservation of areas at risk, as well as improve the general planning, management and utilization of land resources and preserve soil fertility for sustainable agricultural development.

In 1994, the US President's Council on Sustainable Development, Sustainable Agriculture Task Force, set forth the goal of management of agricultural activities to protect soil, air, and water quality, and biodiversity to increase agriculture's long-term productivity and profitability and enhance human health and well-being. The task force recommended that assistance to farmers be linked to voluntary implementation of farm and ranch plans for natural resource conservation in agricultural production. They also recommended implementation of land-use policies to preserve the prime land base for US agriculture.

The 1990 Farm Bill defined sustainable agriculture as:

> an integrated system of plant and animal production practices having a site-specific application that will, over the long term, satisfy human food and fiber needs; enhance environmental quality and the natural resources base upon which the agricultural economy depends; make the most efficient use of nonrenewable resources and on-farm/ranch resources; and integrate, where appropriate, natural biological cycles and controls; sustain the economic viability of farm/ranch operations; and enhance the quality of life for farmers/ranchers and society as a whole.

Policy Aspects

During the early 1990s, the US Department of Agriculture (USDA) also developed policy on sustainable development:

> The USDA is committed to working toward the economic, environmental, and social sustainability of diverse food, fiber, agriculture, forest, and range systems. USDA will balance goals of improved production and profitability, stewardship of the natural resource base and ecological systems, and enhancement of the vitality of rural communities. USDA will integrate these goals into its policies and programs, particularly through collaboration, partnerships and outreach.

In the 1991 yearbook, *Agriculture and the Environment*, Secretary of Agriculture Edward Madigan noted that national expenditure for pollution control increased fourfold from 1972 to 1990. Erosion had been greatly reduced by the CRP, and adjustments in the commodity programs in the 1990 Farm Bill were designed to encourage farmers to plant alternative crops and use crop rotations to reduce soil erosion and the use of agrichemicals. Advances in conservation tillage were reported, and associated environmental concerns such as increased application of agrichemicals were addressed.

A holistic, ecosystem-based concept of sustainable soil and land management was developed in technical and policy arenas throughout the decade. The Soil and Water Conservation Society published a collection of essays, *Soil Management for Sustainability*, in 1993. In it, the editors discussed applying agroecologic principles to farming to contribute positively to the quality of air, water, and soil resources and to meet the economic and social needs of the food and fiber producer. They offered the concept of farming according to the specific productivity and vulnerability of the soil in its immediate and larger landscape position.

In 1993, the National Research Council's Committee on Long-Range Soil and Water Conservation published *Soil and Water Quality, An Agenda for Agriculture*. Threats to soil resources were identified and criteria to guide soil management were recommended. The fate and transport of agricultural chemicals were analyzed to identify changes in farming systems required to improve water quality. Policy and program options to improve long-term conservation of soil and water quality were recommended. The committee identified these objectives for soil and water resource management: conserve and enhance soil quality as a fundamental first step to environmental improvement; increase nutrient, pesticide, and irrigation use efficiencies in farming systems; increase the resistance of farming systems to erosion and runoff; and make greater use of field and landscape buffer zones to capture otherwise uncontrolled off-site flow of agrichemicals.

The USDA released *Food and Agricultural Policy: Taking Stock for the New Century* in 2001. In the section on conservation and the environment, the report pointed out that, although soil erosion had declined by 40% over the previous 15 years, farmland was still losing 1.72 billion metric tons of soil every year. It stated that the array of conservation issues had grown with changes in the structure of agriculture, in farm practices, and in public concerns. Soil was recognized as an emerging challenge to manage as a national strategic asset. More than one-third (52.25 million hectares) of US cropland was reported to be in need of improved soil quality. It recognized that the biological, chemical, and physical processes that drive agricultural productivity cannot be bypassed with inputs. Soil degradation, such as compaction, crusting, salination, and loss of organic matter, needed to be addressed to maintain multiple soil benefits, including reduced runoff and erosion, increased carbon sequestration, and improved productivity and sustainability. The need to increase support for conservation on land in agricultural production was recognized, and innovative, market-based environmental benefits and/or stewardship policy tools were suggested.

The 2002 Farm Bill increased conservation on land in agricultural production and introduced a new program to pay farmers directly for the environmental benefits they produce with their conservation systems. The Farm Bill also directs a large portion of conservation resources to address environmental issues caused by livestock operations.

Sustainable and Organic Agricultural Systems

Sustainable agricultural systems strive to restore and enhance inherent soil qualities and productivity through improved conservation and management practices, while reducing use of commercial inputs. The USDA has supported a Sustainable Agriculture Research and Education Program since 1988. A number of colleges and universities throughout the USA have developed sustainable agriculture programs during this time frame.

Interest in organic agriculture has also grown dramatically. The production of organic food and fiber is largely dependent on building fertile, productive soils without the use of synthetic inputs. The public interest in agricultural production and products free of fertilizers and pesticides, expressed through the marketplace, both for human health and ecosystem health, has caused organic and sustainable production systems to flourish. US retail sales of organic agricultural products have increased by 20–25% each year since 1990 to US $7.8 billion in 2000, with global sales topping US $17.5 billion. US land in organic production grew from approximately one-quarter of a million hectares in 1992 to close to one million hectares in 2001. In 2002 the USDA National Organic Label Standard went into effect. This is expected to expand markets for organic products further.

Inventories and Assessment

Soil Loss

Soil losses recognized in the 1930s included both physical and chemical soil properties: loss of structure, loss of nutrients by crop removal and leaching, and loss of the soil itself through erosion by wind or water. Good conservation practices included primarily maintaining soil organic matter by crop rotations, cover crops, contour cultivation, and use of manure.

The 1937 appraisal of the 712.3 million hectares of agricultural land in the USA reported 114 million (16%) ruined or severely damaged by erosion, and another 314 million (44%) moderately damaged. Of the 168 million hectares of cropland then in production (1935 Census of Agriculture), 81 million were ruined, severely damaged, or had lost at least half of the topsoil, and erosion had begun on another 40.5 million hectares. Due to soil depletion from overgrazing of the 295 million hectares of rangeland in 1935 in the western USA, 239 million hectares (80%) were eroding more or less seriously, further reducing productive capacity.

Wind erosion was reported each year from 1954 to 1957 as damaging more than 6.1 million hectares each year, and destroying another 0.5–2 million hectares of crops. The annual cost of water-borne sediment damage in the mid-1980s was estimated between US $4 and 5 billion. The 1982 USDA National Resources Inventory (NRI) reported that 57.2% of nonfederal rural land needed conservation for erosion, drainage, irrigation management, or vegetative cover.

Changes in USDA conservation programs initiated in the 1985 Farm Bill created new progress in erosion control through the late 1980s into the 1990s. Conservation goals, however, were tied to earlier concepts of rates of tolerable soil loss (T), which emphasized farm-level economics more than ecosystem-wide sustainability. Conservation programs were targeted to the most highly erodible land, and ignored land that was eroding at close to T-values. Despite progress in controlling 'excess erosion,' the 1992 NRI reported that 82 755 000 ha of cropland and 23 661 000 ha of pastureland needed conservation treatment. The 1997 NRI reported that of the total 152 684 000 ha of cropland, 26 204 000 ha were eroding at more than

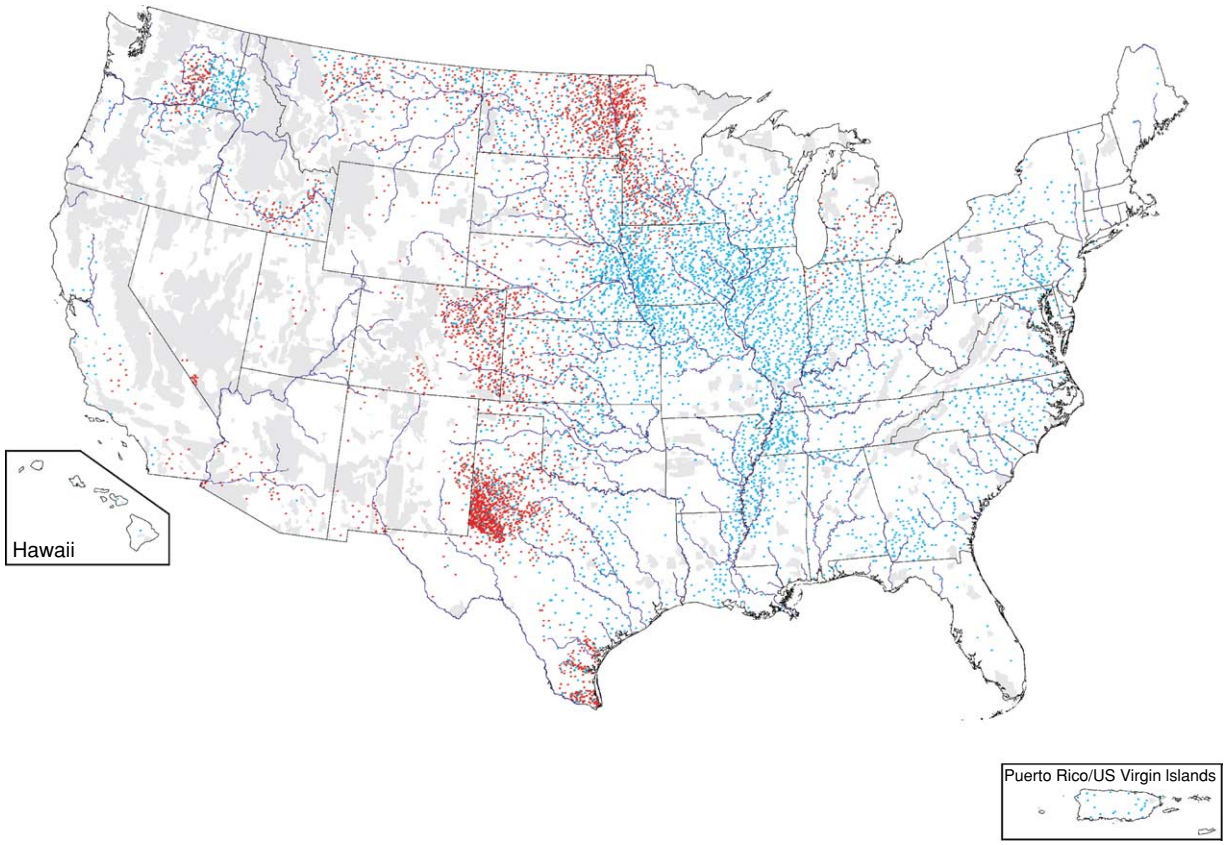


Figure 1 This dot map shows tons of erosion due to wind and water on cropland and Conservation Reserve Program land in the US. Each blue dot represents 970 million metric tons of mean annual erosion due to water. Each red dot represents 760 million metric tons of mean annual erosion due to wind. The combined erosion for the USA is 1.72 billion metric tons per year.

T owing to water erosion, and 19 372 000 ha of cropland were eroding at more than *T* owing to wind erosion (Figure 1).

Water Quality

Water pollution by nutrients or other agrichemicals from agricultural soils was not nationally recognized as a concern until the 1970s. From 1974 to 1981, river and stream water quality monitoring data showed that nitrate concentration increases were widespread nationally. From 1980 to 1989, the data showed that concentrations tended to decrease as often as they increased. These fluctuations are attributable to several farm economy crises during the period that severely curtailed application of purchased inputs. However, the cropland application of commercial nitrogen fertilizers continues to contribute to excess nitrogen in ground and surface waters. One example of the impact is the hypoxia, or 'dead' zone in the Gulf of Mexico, considered symptomatic of excessive nitrogen runoff (Figure 2).

From 1982 to 1992, commercial phosphorus consumption dropped 22% nationally, and river and stream monitoring data from 1982 to 1989 showed widespread declines in total phosphorus concentrations. However, increasing concentration of animal agriculture production systems resulted in increased localized problems associated with animal waste. As livestock production shifted to fewer, larger operations, more manure nutrients were spread on smaller areas. The number of counties without enough land to spread safely the manure generated within their boundaries increased significantly.

The 2002 Report, *State of the Nation's Ecosystems, Measuring the Lands, Water, and Living Resources of the United States*, evaluated chemical, physical, and biological conditions on US farmland. Approximately 20% of groundwater wells and 10% of stream sites on farmlands sampled by the US Geological Survey National Water Quality Assessment Program had nitrate concentrations that exceeded the federal drinking water standard (10 ppm or $10\,\mu g\,ml^{-1}$). About three-quarters of sampled farmland stream sites had phosphorus levels above the maximum concentration (0.1 ppm or $0.1\,\mu g\,ml^{-1}$) recommended to prevent algal growth in streams. More than 80% of the farmland streams had at least one pesticide above aquatic life guidelines and about 4% had one or more compounds that exceeded human health standards or guidelines. Approximately 60% of groundwater wells sampled in farmland areas had at least one detectable pesticide. It was noted that drinking water or aquatic standards or guidelines do not exist for approximately half of the 76 pesticides analyzed. In addition, mixtures of pesticides and intermittent high-concentration events are

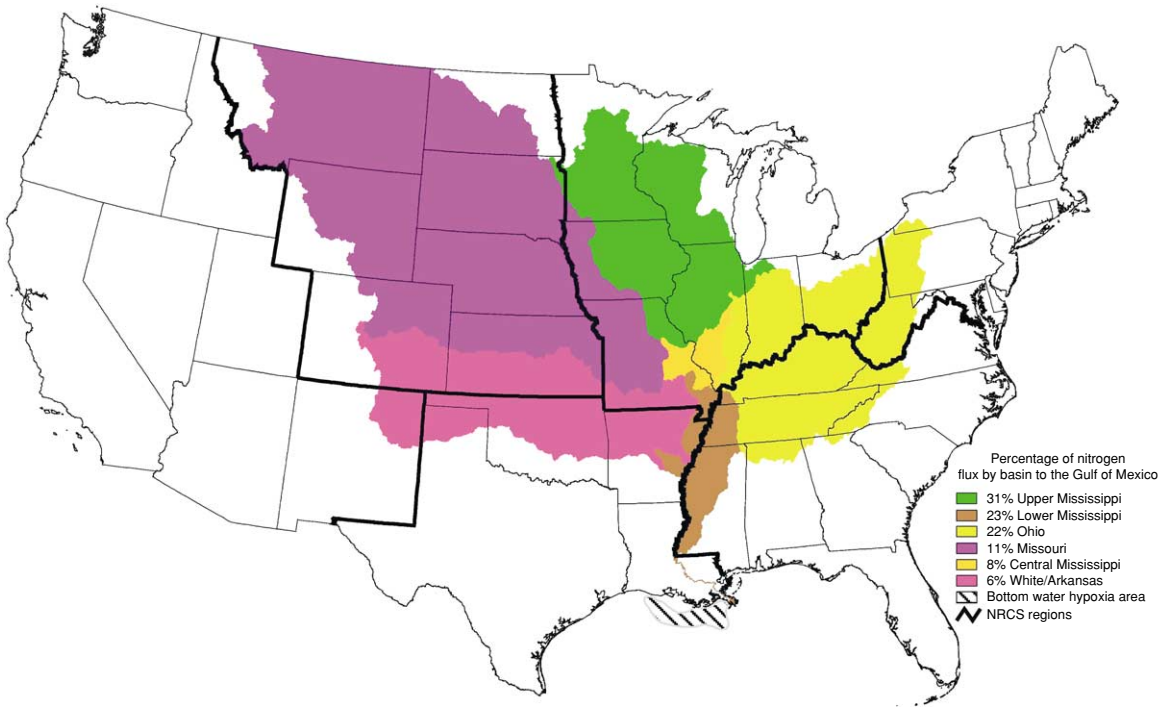


Figure 2 Areas where nitrogen leaves the soil and joins the water system, creating the 'bottom water hypoxia area' in the Gulf of Mexico. NRCS, Natural Resources Conservation Service.

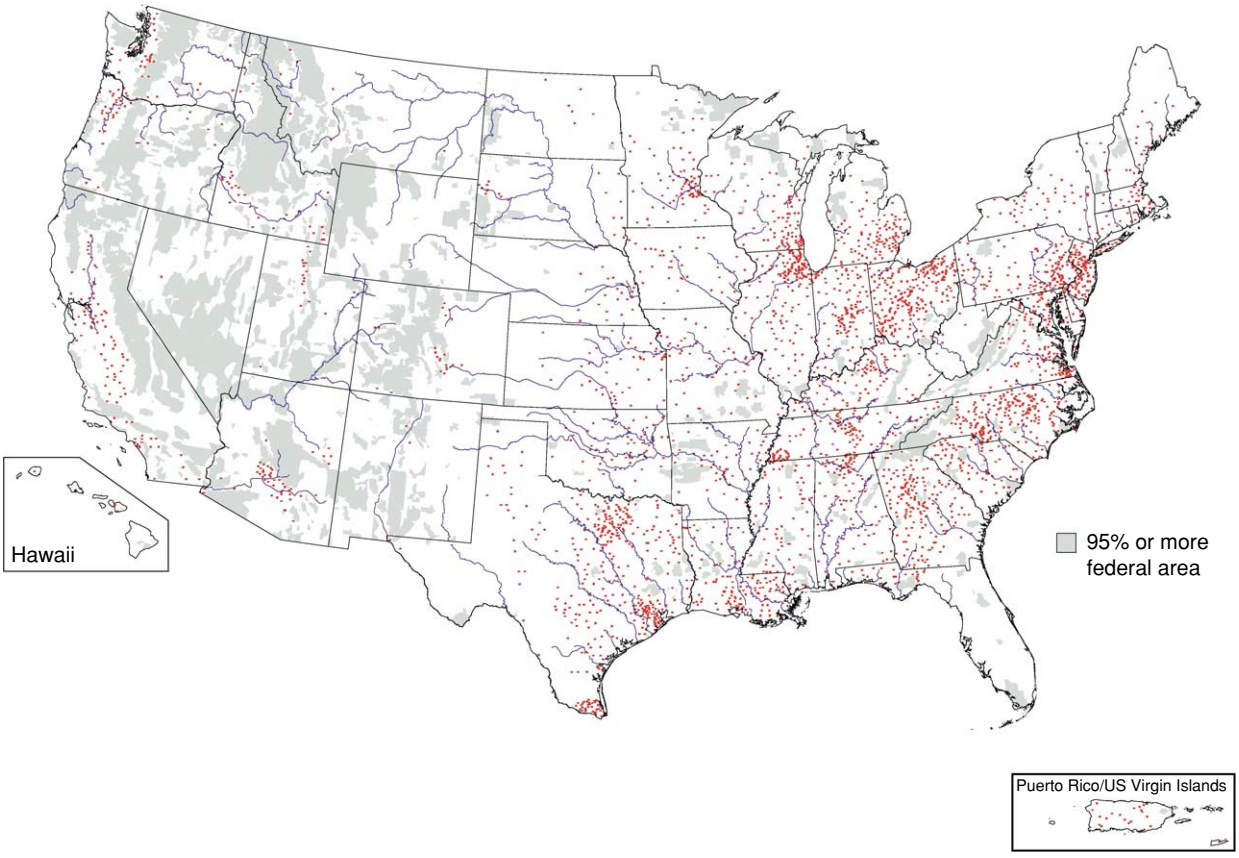


Figure 3 Of the 170 500 000 ha of cropland in 1982, 2 975 500 ha of prime farmland were developed for nonagricultural use between 1982 and 1997. Each red dot represents 810 hectares of newly developed land.

not covered or well understood. The report further identified the need for additional data and indicators to characterize soil organic matter, soil biology, and soil salinity.

Loss of Farmland

The loss of prime farmland to nonagricultural development emerged in the 1980s as a key concern to sustainable soil and land management. In 15 years, 8% of the 93 500 000 ha of prime farmland farmed in 1982 were lost, primarily due to nonagricultural development. This loss caused marginal land with less-favorable characteristics to be brought into production to meet agricultural demand. Production on marginal land requires more inputs and can create more negative offsite environmental impacts. A Federal Farmland Protection Program and state and county programs to preserve agricultural land were implemented and preserved about 0.5 million hectares of farmland by 2002, but the accelerating loss of agricultural land has not been significantly slowed (Figure 3).

Contributing to concerns about the supply of agricultural land is the recognition of the ecologic value of wetlands to sustain water quality, fish, and wildlife. Once considered a conservation practice, the drainage of wetlands is no longer supported or condoned. There is an effort to restore wetlands on agricultural lands, to achieve no net loss of wetlands in the USA. This puts further pressure on remaining croplands to provide the agricultural production demanded by increasing population throughout the world.

At the same time, the rising demand for and profitability of locally produced food, demonstrated by the growth of farmers' markets, 'pick-your-own' farms, and Community Supported Agriculture (contracts between farmers and consumers to supply farm production throughout the growing season for an annual fee), have helped to preserve some farmland, primarily located near urban consumers. These small, direct-market farms often use sustainable or organic production systems.

See also: **Civilization, Role of Soils**; **Degradation**; **Desertification**; **Land-Use Classification**; **Pesticides**; **Salination Processes**; **Salinity:** Management

Further Reading

Heinz Center (2002) *The State of the Nation's Ecosystems. Measuring the Lands, Water, and Living Resources of the United States*. Cambridge, UK: Cambridge University Press.

Kellogg RL, Lander CH, Moffitt DC, and Gollehon PEN (2000) *Manure Nutrients Relative to the Capacity of Cropland and Pastureland to Assimilate Nutrients: Spatial and Temporal Trends for the United States*. Washington, DC: US Department of Agriculture.

Lal R and Pierce FJ (eds) (1991) *Soil Management for Sustainability*. Ankeny, IA: Soil and Water Conservation Society.

Lowdermilk WC (1953) *Conquest of the Land Through 7000 Years*. Agriculture Bulletin No. 99, US Department of Agriculture, Natural Resources Conservation Service. Washington, DC: US Government Printing Office.

Sanders JH, Southgate DD, Lee JG (1995) *The Economics of Soil Degradation: Technological Change and Policy Alternatives*. SMSS Technical Monograph No. 22. US Department of Agriculture, Natural Resources Conservation Service, World Soil Resources. Washington DC: US Government Printing Office.

Sorensen AA, Greene RP, and Russ K (1997) *Farming on the Edge*. Washington, DC: American Farmland Trust.

UN Department of Public Information (1994) *Agenda 21: Earth Summit. The United Nations Programme of Action from Rio, 1992*. New York: UN Publications.

US Department of Agriculture (2001) *Food and Agricultural Policy: Taking Stock for the New Century*. Washington, DC: US Government Printing Office.

US Department of Agriculture, Soils and Men (1938) *1938 Yearbook of Agriculture*. Washington, DC: US Government Printing Office.

US Department of Agriculture (1957) *1957 Yearbook of Agriculture: Soil*. Washington, DC: US Government Printing Office.

US Department of Agriculture (1983) *Using Our Natural Resources: 1983 Yearbook of Agriculture*. Washington, DC: US Government Printing Office.

US Department of Agriculture (1998) *Ten Years of SARE: A Decade of Programs, Partnerships and Progress in Sustainable Agriculture Research and Education*. Washington, DC: US Government Printing Office.

US National Research Council Committee on Long-Range Soil and Conservation (1993) *Soil and Water Quality: An Agenda for Agriculture*. Washington, DC: National Academy Press.

US President's Council on Sustainable Development (1996) *Sustainable America: A New Consensus for Prosperity, Opportunity, and a Healthy Environment*. Sustainable Agriculture Task Force Report. Washington, DC: US Government Printing Office.

USDA–NRCS (1996) *America's Private Land: A Geography of Hope*. Washington, DC: US Government Printing Office.

USDA–NRCS (1997) *Water Quality and Agriculture*. Working Paper No. 16. Washington, DC: US Government Printing Office.

USDA–NRCS/Iowa State University Statistical Laboratory (2000) *Summary Report 1997 National Resources Inventory*. Washington, DC: US Government Printing Office.

World Commission on Environment and Development (The Brundtland Commission) (1987) *Our Common Future*. Oxford, UK: Oxford University Press.

SWELLING AND SHRINKING

D Smiles, CSIRO Land and Water, Canberra, Australia
P A C Raats, Wageningen University, Wageningen, The Netherlands

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

After a century of study, hydrologic theory of non-swelling soils is applied routinely to irrigation and drainage, and to other problems of land and water management where quantitative insights and predictions relating to soil water and solute movement are required. By contrast, water movement in soils that change volume with water content is not well understood, and management of swelling soils remains problematic.

Difficulties arise because volume change complicates measurement of material balance, and both the liquid and the solid phases must be considered; volume change also introduces ambiguity, for example, in water profile measurements based on neutron probes when compared with, say, time-domain reflectometry, where the probes may 'float' with the solid phase. Then, swelling or shrinking accompanying soil water content change results in vertical displacement of the wet soil, which involves gravitational work and contributes an overburden component to the total potential of the soil water. In addition, the total flux of the water has a component advected with the moving solid as well as the water potential-driven, darcian flux relative to the solid. Finally, many swelling soils crack and the network of cracks provides

alternative, preferential pathways for rapid flow of water which prejudice application of theory simply based on darcian flow.

Swelling Soils

The Vertisol and the Histosol orders in Soil Taxonomy represent soils that change volume with change in water content. The Vertisols are the more significant, with great areas in India (approx. $7.9 \times 10^5\,km^2$), Australia (approx. $7 \times 10^5\,km^2$), and Sudan (approx. $5 \times 10^5\,km^2$), while China, Ethiopia, and the USA each have approximately $1.5 \times 10^5\,km^2$. They tend to be chemically fertile, but physically they are generally very intractable; very sticky when wet and strong when dry. Characteristically, they have high contents of smectite clay mineral. In Australia in the Murray Darling Basin alone, there are approximately $2.5 \times 10^5\,km^2$ of these soils. They represent approximately 10% of the basin and almost 50% of its irrigated soils, and they support irrigated and dryland crops which are currently worth more than A\$$1.5 \times 10^9$ annually.

The Histosols of concern include at least $10^6\,km^2$ of low-lying clay soils of marine or estuarine origin. These soils, which are mainly of the Hemist suborder, are characterized by permanently high water tables and the possible presence of sulfuric or sulfidic materials. They occupy low-lying positions of the landscape and many have been laid down over the past 10 000 years in the littoral zone of most continents and islands. Their natural water volume fraction may be higher than 0.8 and their bulk density may be as low as $0.2\,t\,m^{-3}$. They occupy areas of high and increasing agricultural and urban importance. Their development often involves consolidation and the sulfidic soils among them generate acid drainage.

The hydrology of these soils is not well understood, and neither is the reaction and 'chromatographic' chemistry associated with transfer of nutrients, or the residence and degradation of agricultural chemicals. The physicochemical consequences of oxidation of sulfides in the Histosols are also unclear.

In addition to soils, there are also particulate industrial and mining slurries that present local environmental problems. The physics of swelling soils has much to offer with respect to dewatering and consolidating these materials to rehabilitate the landscape; it also has much to offer in engineering soils to support built structures.

Volume change accompanying water movement is central to these activities and, in the case of the Histosols, oxidation of organic matter may also contribute significantly to irreversible volume change accompanying development. Present approaches to hydrology of swelling soils, however, are generally based in nonswelling soil theory; they rarely account for volume change and its consequences, and their use may result in significant errors in estimations of local water and solute flux, water balance, and aquifer recharge. The consequences of oxidation of part of the soil matrix, for example, on material balance estimates are rarely considered.

History

Schübler measured soil volume change in the early nineteenth century and Tempany provided a contemporary perspective early in the twentieth century. Tempany introduced the concept of the shrinkage curve and formally related linear and 'cubical' contraction of soil aggregates in relation to water loss. At the same time, anchored rods were used to measure profile swelling and water content change.

One-dimensional flow of water in soils that change volume with water content was first analyzed in a modern context by Terzaghi, who sought to describe soil consolidation following the imposition of loads associated with built structures. He perceived that the soil water initially carries an imposed load in a saturated soil, and that the soil gradually compresses as this load is transferred to the soil matrix when the water redistributes. Terzaghi thus associated consolidation with water flow and his approach anticipated that of Richards for rigid unsaturated soils, in being based on material continuity and the Darcy law. The outcome was a linear diffusion equation that Terzaghi solved using Fourier series. A decade later and independently, Nicholson and Childs applied a diffusion approach to water flow in clay soils in an agricultural context.

Terzaghi's presentation was obscure: papers still appear in civil engineering journals speculating on what he meant; and civil engineers did not appreciate his formulation and use of solid-based, space-like material coordinates. As a result, the theory was recast in physical space specifically identified with small strain behavior and Terzaghi's classic textbook discarded the material insights of the 1923 work. In 1960, however, Terzaghi's solid-based, space-like coordinates were explicitly reintroduced by McNabb to deal with soil consolidation. Raats and Klute and Smiles and Rosenthal independently developed this approach and the latter also experimentally tested it. In addition, Philip analyzed the one-dimensional problem using conventional spatial coordinates. He dealt well with mechanics of swelling, but his formulation was complicated and he subsequently abandoned it in favor of the simpler mathematics associated with the material approach.

Focus on water flow relative to the solid, described by the Darcy law, lay at the core of these analyses. It also emerged that, when the Darcy law was combined with material continuity using solid-based space-like coordinates, a flow equation analogous to the Richards equation resulted so, at least for one-dimensional problems, solutions to the Richards equation became transferable to flow in swelling systems.

Contemporaneously, chemical engineering schools in Japan and the USA working on filtration of very wet particulate suspensions and civil engineers working on soil consolidation identified similar approaches. Focus on flow of water relative to the solid, rather than solid movement relative to the water, also provides a unifying principle that extends the analysis to sedimentation in very wet particulate systems.

Elements of Theory

Water flow in unsaturated nonswelling soils is described by combining the Darcy law with an equation of continuity that accounts for the water in the system during steady and unsteady flow; the Richards equation results. These equations are formulated at the macroscopic scale and they apply to materials where the water potential and the hydraulic conductivity are well-defined functions of the water content.

Two distinct but complementary philosophies emerge to deal with soil profile volume change and the nature of flow in swelling soils. The first approach is mesoscopic and it focuses on behavior of individual soil-structural elements. It envisages water movement through macropores to or from these elements which then swell or shrink three-dimensionally to produce one-dimensional (vertical), macroscopic profile swelling or shrinkage. The approach is concerned with darcian or nondarcian flow in the fractures and three-dimensional volume change of aggregates, which ultimately results in vertical soil profile movement.

The second approach is macroscopic. It recognizes three-dimensional volume change of soil-structural units, but argues that the detail of this behavior is irrelevant and that we need concern ourselves only with the net one-dimensional, vertical volume change and flow that is observed macroscopically. It asserts that if the area of cross section is large enough and if there is no lateral net transfer of material from the control area, then behavior of structural entities beneath it may be described by a representative average vertical displacement. This approach is analogous to the Darcy scale approach to soil water movement in nonswelling soil, which also relates to volume averages and is agnostic about flow detail in individual pores and aggregates. It results in a one-dimensional analysis analogous to that of Richards.

Swelling and shrinking of individual structural elements need to be considered explicitly if the characteristic time of these mesoscopic processes approximates that of the macroscopic processes. In formulating a basic theory, it is presumed that, in general, they do not concur, and focus is on the macroscopic approach. That approach has been extensively tested with saturated materials and, to a lesser extent, with unsaturated soils. It exemplifies the principal issues of equilibrium and flow in a swelling material.

The practical challenges in applying the macroscopic approach are to:

1. Define a scale of discourse that permits meaningful measurements of macroscopic properties of the soil, such as water content and bulk density;
2. Meet Darcy-like requirements that the water potential and the hydraulic conductivity of the soil are well-defined functions of the water content.

To use the Richards equation, the potential and conductivity characteristics are again necessary but, in addition, a 'shrinkage curve' is needed to define the way the volume of soil per unit area of cross section, i.e., the elevation, changes with the soil water content.

Analysis is based on the material balance equations for the solid as well as the water, because both may be in motion relative to an observer, and on the Darcy law, which describes flow of water relative to the particles in response to a space gradient of total head. Both these elements are critical, but material balance must be retained even when the applicability of the Darcy equation is uncertain.

Material Balance in Swelling Systems

One-dimensional flow of water in a swelling system requires material balance equations for both the aqueous and solid phases:

$$\frac{\partial \theta_w}{\partial t} = -\frac{\partial F_w}{\partial z} \qquad [1]$$

$$\frac{\partial \theta_s}{\partial t} = -\frac{\partial F_s}{\partial z} \qquad [2]$$

where z is a distance coordinate, t is the time, F_w and F_s are the volume flux densities of the water and of the solid (cubic meters per square meter per second) relative to the observer, and θ_w and θ_s are the volume fractions of the water and solid. The volume fractions are defined per unit area of horizontal cross section of the soil and the reference volume includes the cracks.

Recognizing that the flux of water, seen by the observer, occurs relative to and with the moving soil solid:

$$F_w = u + \theta_w F_s/\theta_s = u + \vartheta F_s \quad [3]$$

where u is the volume flux of water relative to the particles, and the second term on the right describes transfer of water advected with the moving solid. The ratio of volume fractions, ϑ, is called the moisture ratio:

$$\vartheta = \theta_w/\theta_s = \theta_g \rho_s \quad [4]$$

with θ_g the water mass fraction (kilograms per kilogram) and ρ_s the specific gravity of the soil solid. In saturated soils, ϑ is equal to the void ratio, e.

Eqns [1], [2], and [3] may be manipulated to derive a material balance Eqn [5] for the water based on a solid-based, space-like material coordinate, $m(z, t)$, viz:

$$\frac{\partial \vartheta}{\partial t} = -\frac{\partial u}{\partial m} \quad [5]$$

For one-dimensional flow, the coordinate, $m(z, t)$, is formally defined by:

$$dm(z,t) = \frac{\partial m}{\partial z}dz + \frac{\partial m}{\partial t}dt = \theta_s dz - F_s dt \quad [6]$$

This definition of $m(z, t)$ satisfies the material balance Eqn [2] for the solid and reflects the distribution of the solid in space. Furthermore, $m(z, t)$ is independent of soil displacement that might accompany water content change and can be determined by integrating Eqn [6]. This integration is simplified if it is based on a surface, e.g., $z = 0$, where $F_s = 0$. In the absence of sedimentation or erosion, the soil surface provides such a datum. Then:

$$m = \int_0^z \theta_s dz - \int_0^t F_{s_{z=0}} dt = \int_0^z \theta_s dz = \int_0^z (\rho/\rho_s) dz \quad [7]$$

where m is the cumulative volume of solid, per unit area of cross section, measured from $z = 0$ or the cumulative oven-dried mass per unit area measured away from the soil surface and divided by ρ_s. Thus m is, for example, the cumulative amount of oven-dried solid per unit area measured from the top of a continuous field core sample taken for water content measurement. The final integral in Eqn [7] shows how m is related to the soil specific gravity profile, $\rho(z)$.

Flux Laws in Swelling Systems

The Darcy law describes flow of water relative to the solid in response to a space gradient of the total potential, Φ, of the water in the soil:

$$u = F_w - \vartheta F_s = -k(\theta_w)\frac{\partial \Phi}{\partial z} \quad [8]$$

in which the hydraulic conductivity, $k(\theta_w)$, is a function of the water content but also of the structure. Here it is identified as a function just of θ_w. In nonswelling soils the total potential, Φ, consists of a gravitational component and a capillary component, the latter arising as a result of the interaction of the water with the soil surfaces and their geometry. If Φ and its components are defined as work per unit weight of water, they have dimensions, L, and units meters of H_2O. The hydraulic conductivity, $k(\theta_w)$, then has dimensions LT^{-1} and units meters per second.

Defined this way, Φ is identical to the engineer's hydraulic head. Specifically:

$$\Phi = z + p_w \quad [9]$$

with the gravitational head, z, the elevation of the point of concern in the soil above a convenient datum, and p_w the pressure in the water relative to the pressure of the ambient gas phase measured with a manometer. Eqn [9] applies to both swelling and nonswelling materials. In unsaturated, nonswelling soils, the pressure head, p_w, is negative and reflects only the capillary potential.

Water Potential in a Swelling Soil

The components of Φ in swelling systems have been considered from a thermodynamic perspective and mechanistically. The latter approach argues that Φ represents the work involved in transferring unit weight of water from a reference flat surface of water at atmospheric pressure to a swelling material at some height, z, relative to that reference. Thus Φ is the sum of:

1. The gravitational potential, z, representing elevation of the element of water from the datum to the height z;
2. The unloaded matric or capillary potential, ψ, representing interaction between the water and the soil solid surfaces and their geometry;
3. The overburden potential, Ω, representing vertical displacement of the wet soil accompanying unit change in water content at z.

If the soil is saturated, unit change in water content produces unit change in elevation of the soil and, if the volume of the system is parameterized by the void ratio, e, this implies that $de/d\vartheta = 1$. If the soil is unsaturated, $de/d\vartheta$ tends to be less than 1, unit change in water content does not produce unit height change, and the work involved in elevating the wet profile is

less than it would be in the saturated case. The total potential, Φ, is then defined by the equation:

$$\begin{aligned}\Phi &= z + p_w = z + \psi + \Omega \\ &= z + \psi + \alpha\left(\int_z^T \gamma dz + P\right)\end{aligned} \quad [10]$$

The unloaded capillary potential ψ is the potential a tensiometer would measure at the existing moisture ratio in the absence of overburden. In Eqn [10], γ is the wet specific gravity of the soil, P is any static load on the soil surface, and α reflects the degree to which the soil is elevated by unit change in water content. In a saturated soil, $\alpha = 1$; when the soil is unsaturated $1 \geq \alpha$; and in a nonswelling soil, $\alpha = 0$. In civil engineering contexts, P may be very important, but we neglect it here without any loss of generality. Notice that, in Eqn [10], p_w includes the overburden component and may be positive.

The overburden, Ω, is related to the civil engineers' effective stress according to the equation

$$\Omega = \alpha\sigma = \sigma' + p_w \quad [11]$$

in which σ is the total normal stress and σ' is the effective or interparticle stress. Comparison of Eqns [10] and [11] shows that $\sigma' = -\psi$, and hence is related to ϑ through $\psi(\vartheta)$.

The nature of α has generated some controversy. At one time civil engineering practice equated it with the slope of the shrinkage curve. A more correct but more complicated definition followed some argument in which it was accepted that α must be a value of $de/d\vartheta$ averaged from zero load to that actually experienced at the point in question. The mathematical consequences of this definition are cumbersome, however, and it seems to be agreed that the civil engineering approximation remains the most practicable way forward.

It has been argued that, because the manometric pressure, p_w, is measurable, while ψ, an 'unloaded' capillary potential, is experimentally inaccessible, the water content should be defined as a function of p_w rather than ψ. The world remains indifferent to the issue. Figure 1 compares water content profiles and components of the total potential of the soil water for idealized nonswelling and swelling soils.

Potential Gradient

It suffices that the driving force in the Darcy equation is the space gradient of hydraulic head (or total potential), which is the sum of the manometric pressure (which may be negative relative to atmospheric pressure) plus a gravitational component relative to a convenient height datum. In this approximation, and if $\vartheta(\psi)$ is single-valued, $\partial\Phi/\partial z$ from Eqn [10] becomes:

$$\begin{aligned}\frac{\partial\Phi}{\partial z} &= \frac{\partial}{\partial z}\left(\psi + z + \alpha\int_z^T \gamma dz\right) \\ &= \theta_s\left(\frac{\partial\psi}{\partial m}\right) + (1 - \alpha\gamma) \\ &\quad + \theta_s\left(\frac{d\alpha}{d\vartheta}\right)\left(\int_m^0 \gamma/\theta_s dm\right)\left(\frac{\partial\vartheta}{\partial m}\right)\end{aligned} \quad [12]$$

In Eqn [12], the second equality on the right-hand-side is derived by differentiating the first equality by parts, and using the definition of m, i.e., $dm/dz = \theta_s$ to eliminate z.

Darcy Law and Hydraulic Conductivity

Substitution of the last of Eqns [12] in the Darcy law, in m-space, yields:

$$\begin{aligned}u = -k(\theta_w)\theta_s\Bigg(&\left(\frac{\partial\psi}{\partial m}\right) + \left(\frac{1 - \alpha\gamma}{\theta_s}\right) \\ &+ \left(\frac{d\alpha}{d\vartheta}\right)\left(\int_m^0 \gamma/\theta_s dm\right)\left(\frac{\partial\vartheta}{\partial m}\right)\Bigg)\end{aligned} \quad [13]$$

In Eqn [13], $k(\theta_w)\theta_s = k_m$ is identified as a 'material' hydraulic conductivity. Philip derived this equation 30 years ago, while Sposito derived the same equation from thermodynamic considerations.

The hydraulic conductivity, k_m, has conventionally been treated as a material characteristic parameterized by ϑ. However, it must be a function of both the structure, parameterized by e as well as the water content, ϑ, in a swelling material. No difficulty arises in saturated swelling systems where $e = \vartheta$, but, when $e > \vartheta$, the hydraulic conductivity must be represented by a surface, $k(e, \vartheta)$. This issue has escaped analysis hitherto and the magnitude of the effect needs to be established.

Flow Equation

Substitution of Eqn [13] in Eqn [5] yields an equation analogous to the Richards equation for a nonswelling soil. For saturated clays subject to large, constant imposed loads, gravity can be neglected and $\alpha = 1$. This latter condition characterizes many important practical chemical engineering situations. Numerical methods have been used to solve the full equation together with a nonlinear shrinkage curve to illustrate the solution for flow in unsaturated marine clays.

The principal physical differences, which distinguish equilibrium and flow in saturated swelling systems from those in nonswelling soils, arise from the overburden effect in these systems. Overburden

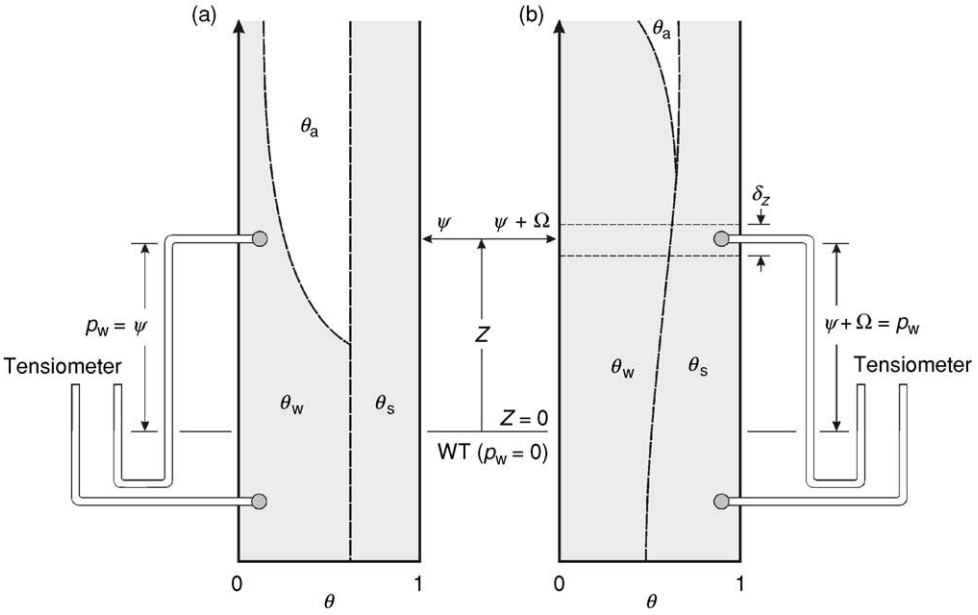


Figure 1 Comparison of the vertical, static equilibrium water content profiles in nonswelling soils (a) and swelling soils (b) in contact with a free water surface at atmospheric pressure. The *z*-coordinate is positive upwards, with its origin at the free water surface, and θ_w, θ_s, θ_a are the volume fractions of the water, solid, and air; they sum to unity. In the swelling soil, θ_s increases with depth toward some maximum density determined by the weight of the wet soil it supports and the water table height. A corresponding decrease is observed in θ_w. Desaturation in a swelling soil above the water table represents cracks and then desaturation of aggregates. The components of total potential, Φ, of the water in the two systems differ. Thus the arrows show the steps in transferring unit amount of water from the free water surface, WT, to the system at elevation *z* above it. Defined per unit weight of water, *z* is then the gravitational potential relative to the free water surface at WT; and p_w the pressure a manometer measures when inserted at height *z*. The manometric pressure is equal to ψ, the 'capillary' potential in the nonswelling soil, but is the sum of an unloaded 'capillary' or moisture potential and the overburden potential, Ω, which arises because of the vertical displacement of the wet profile, δz, that must accompany insertion of the water at height *z*. In the saturated region, unit displacement accompanies insertion of unit amount of water. In the unsaturated region of the swelling soil, the vertical displacement is less than unity. At static equilibrium, the manometer/tensiometer water level must have the same elevation as the water table. Reproduced from Smiles DE (2000) Hydrology of swelling soils: a review. *Australian Journal of Soil Research* 38: 501–521 (corrigendum 40: 1467).

reverses (and diminishes) the effect of gravity in systems where $\rho_s > 1$:

1. Infiltration into these systems resembles capillary rise in nonswelling soils;
2. Steady upward flow of water proceeds more readily in swelling than nonswelling soils;
3. Static equilibrium water-content profiles in the saturated region are drier at the bottom than at the top, in contrast to profiles in nonswelling soils;
4. The period of time for which a 'gravity-free' analysis applies to transient flow may be greatly extended;
5. The shape of the $\psi(\vartheta)$ characteristic together with a relatively weakly varying $k_m(\vartheta)$ often results in a material moisture diffusivity that decreases with increasing water content, so desorption in muds and slurries resembles absorption in nonswelling soil.

Overview

Water flow accompanied by volume change remains one of the most challenging areas of porous material physics. Some areas of application have been well tested, at least in one-dimensional flow, and theory has been usefully applied to a number of unit processes in chemical engineering and to consolidation in civil engineering. Examples of unit processes include filtration under constant pressure and constant rate conditions and effects of filter membranes that significantly resist flow. They also include centrifugation and centrifugal filtration of wet particulate slurries. Some of these analyses are based on well-known solutions of the Richards equation, but some required and tested solutions that had not previously been available in soil science. Constant-rate filtration and filtration through an impeding membrane exemplify the latter class of problem. Most of this work has involved saturated systems.

In soil science, however, one-dimensional theory remains incomplete and its test is uncertain. Curiously, it represents a bridging problem between well-founded and tested analysis of flow in unsaturated nonswelling systems and that in saturated swelling ones. In fact, Eqns [1]–[13] apply equally well to swelling and nonswelling materials, although application of the concepts and terminology of swelling soil theory to nonswelling soils is unusual. Thus, both 'limiting' systems are described well by the Richards equation and both demand that $\psi(\vartheta)$ and $k_{\mathrm{m}}(\vartheta)$ be well defined, although the swelling system requires a shrinkage curve also. The consequences of volume change in unsaturated soils, however, remain the challenge. It is useful to identify some important issues of concern.

Scale of Discourse

Because of the size of the structural units in swelling soils, the scale of discourse that permits application of an approach analogous to the approach of Richards for nonswelling soils, where properties represent measurable spatial averages, is much larger than is generally the case in nonswelling soils. As a result, the scale of measurement of water content and water potential requires careful consideration and it may be that many current methods fail to provide appropriate volume averages at the scales involved. The existence of cracks need not complicate measurement, however, provided the cracks are included in the representative volumes and areas of cross section. Water content defined per unit amount of solid, consistent with a solid-based, space-like coordinate, is a preferred basis for water balance in these soils. This convention ensures that volume change does not prejudice estimates of water content change.

Material Balance and Coordinates

In addition to requirements that material properties be consistent with the scale of the space-like coordinate if integral quantities such as the profile water content are to be interpreted without ambiguity, differences over time will be incorrect if the change in volume of a profile is ignored. Volume change accompanying water content change is irrelevant and material balance is exact, however, if the water content is expressed per unit amount of soil solid (mass or volume) and 'space' is defined as the cumulative amount to soil solid per unit area, measured, for example, from the soil surface. This approach extends to organic soils where the organic matter is being oxidized and to acid sulfate soils where drainage consolidates the profile as well as removing organic matter. The 'irreducible' mineral content of the soil provides a secure reference frame in these circumstances. It also provides an appropriate basis for intrinsic expression of the concentrations of water, organic matter, or sulfide.

Observations of consistent systematic behavior of heavy clay soils over almost 1000 days confirm that α of the soil profile is relatively constant at approximately one-third and is thus consistent with three-dimensional normal volume change of the soil aggregates. The effect is not greatly depth-dependent and permits quite accurate estimate of the soil water balance from surface displacement measures.

Water Flow

Water flow in swelling systems generally occurs in response to a space gradient of the total head. At the same time, preferential flow in macropores of cracked soils may dominate flow in the early stages of heavy rain or flood irrigation when the soil is very dry. Both phenomena and their relative importance have been considered in the mesoscopic analysis of flow, although a general approach has yet to be formulated. In addition, laboratory measurement must recognize and mimic the constraints experienced by swelling field soils, and procedures common in civil engineering may be obligatory in laboratory studies of flow and volume change in these soils. Early theory tended to identify these constraints carefully. Some recent experiments have been less careful and experimental artifacts inconsistent with theory have prejudiced analyses.

Multidimensional Flow

Application and test of theory have been confined largely to one-dimensional flow and volume change, although a long-available, three-dimensional analysis has also been used in some approximate analyses. Cracked systems have also been dealt with by using a single parameter to characterize the transversely isotropic deformation of the solid phase and combining this with one-dimensional, vertical flow of the aqueous phase. Experiments that model the constraints on, and test theory of, multidimensional flow remain elusive, however.

Measurement

Problems of measurement and its interpretation are endemic. For example, there has been concern about tensiometry and its interpretation in swelling systems with anisotropic constraints which have not been resolved. There is great need for laboratory and field measurement, with specific focus on noninvasive measurement of the time course of the spatial distributions of solid, liquid, and gaseous phases in the soil.

In the laboratory, civil engineers in particular systematically use equipment such as triaxial consolidation cells that are designed to characterize volume change with well-defined constraints in engineering soils. Their terminology and theory may not always be familiar to soil scientists but the technology offers much to soil physics. At the same time, many soil physical principles are unfamiliar to engineers, but each subdiscipline can learn from the other to the ultimate benefit of both.

Water Retention and Hydraulic Conductivity Functions in Unsaturated Swelling Soils

Both the water retention and hydraulic conductivity functions in swelling systems still present significant challenges for resolution. In relation to water retention, four parameters, e, ϑ, p_w, and P, of which three are independent, characterize an unsaturated, swelling porous material. The relations $e(P, \vartheta)$ and $p_w(P, \vartheta)$ have been preferred, mainly because of the wealth of experimental data that exists in relation to $e(P, \vartheta)$; this approach results in the notion of an unloaded moisture potential, as shown in Eqn [10]. The nature of the hydraulic conductivity function remains to be resolved. For nonswelling soils, hydraulic conductivity seems to be a nonhysteretic function of the volumetric water content, θ_w, with the fixed volume fraction of the solid, θ_s, as a parameter. Analogy suggests that the hydraulic conductivity, k, for an unsaturated swelling system must be similarly defined in a $k-e-\vartheta$ diagram. Few, if any, data are available to illustrate the nature of this surface. At the same time, there is an extensive data set for saturated montmorillonite in equilibrium with a range of solution salt concentrations. These data confirm that $k_m(\vartheta)$ is not unique in pure clays but reflects colloid structures associated with the cation suite and concentrations with which the clay is in equilibrium. At the same time, these data reveal strong unifying benefits across systems of different solution salt concentrations, when k is related to ψ. These observations await systematic study and explanation.

Summary

Water flow in swelling soils is described within a macroscopic, one-dimensional theory because field volume change, in the large, is vertical. The theory accounts for material balance of both the solid and liquid phases of the system, both of which tend to be in motion. It also recognizes that the total flux of water has a darcian component relative to the solid as well as a component advected with the moving solid. The theory identifies an overburden component of the water potential in addition to those due to gravity and capillarity. Material properties required by the approach require unambiguous measurement at the scale of application. Several issues of theory and measurement remain unresolved and skills currently available in civil and chemical engineering would benefit the science greatly.

List of Technical Nomenclature

α	Load factor
γ	Specific gravity of wet soil
θ_s	Volume fraction of solid
θ_w	Volume fraction of liquid
ϑ	Moisture ratio
ρ	Soil specific gravity
ρ_s	Soil solid specific gravity
σ	Total stress (L)
σ'	Effective stress (L)
Φ	Total water potential (L)
ψ	Capillary potential (L)
Ω	Overburden potential (L)
e	Void ratio
F_s	Volume flux density of solid (LT^{-1})
F_w	Volume flux density of water (LT^{-1})
k	Hydraulic conductivity (LT^{-1})
k_m	Material hydraulic conductivity (LT^{-1})
m	Solid-based space-like coordinate (L)
p_w	Manometric pressure (L)
t	Time (T)
u	Volume flux of water relative to particles (LT^{-1})
z	Vertical distance and gravitational potential (L)

See also: **Clay Minerals**; **Stress–Strain and Soil Strength**; **Structure**

Further Reading

Ahmad N and Mermut M (eds) (1996) *Vertisols and Technologies for their Management. Developments in Soil Science*. Amsterdam, the Netherlands: Elsevier.

Bouma J and Raats PAC (eds) (1984) *Proceedings of ISSS Symposium on Water and Solute Movement in Heavy Clay Soils*. International Institute for Land Reclamation

and Improvement Publication No. 37. Wageningen, The Netherlands: International Institute for Land Reclamation and Improvement.

McGarity JW, Hoult EH, and So HB (eds) (1984) *The Properties and Utilisation of Cracking Clay Soils. Reviews in Rural Science 5*. Armidale, Australia: University of New England Press.

McNabb A (1960) A mathematical treatment of one-dimensional consolidation. *Quarterly of Applied Mathematics* XVIII: 337–347.

Philip JR (1968) Kinetics of sorption and volume change in clay-colloid pastes. *Australian Journal of Soil Research* 6: 249–267.

Philip JR (1969) Hydrostatics and hydrodynamics in swelling soils. *Water Resources Research* 5: 1070–1077.

Philip JR (1970) Hydrostatics and hydrodynamics in swelling soils: reply to Youngs and Towner. *Water Resources Research* 6: 1248–1251.

Philip JR (1995) Phenomenological approach to flow and volume change in soils and other media. *Applied Mechanics Review* 48: 650–658.

Raats PAC (1998) Kinematics of subsidence of soils with a non-conservative solid phase. *Proceedings of a Symposium on New Concepts and Theories in Soil Physics and Their Importance for Studying Changes Induced by Human Activity*. The 16th World Congress of Soil Science, Aug. 20–26 1998, Montpellier, France. CD-ROM.

Raats PAC (2002) Flow of water in rigid and non-rigid, saturated and unsaturated soils. In: Capriz G, Ghionna VN, and Giovine P (eds) *Modeling and Mechanics of Granular and Porous Materials*, pp. 181–211. *Modeling and Simulation in Science, Technology and Engineering*. Boston, MA: Birkhäuser.

Raats PAC and Klute A (1968) Transport in soils. *Soil Science Society of America Proceedings* 32: 161–166 and 452–456.

Smiles DE (1986) Principles of constant-pressure filtration. In: Cheremisinoff NP (ed.) *Encyclopedia of Fluid Mechanics*, pp. 791–824. Houston, TX: Gulf Publishing Company.

Smiles DE (2000) Hydrology of swelling soils: a review. *Australian Journal of Soil Research* 38: 501–521 (corrigendum: 40: 1467).

Sposito G (1975) Steady vertical flows in swelling soils. *Water Resources Research* 11: 461–464.

Tempany HA (1917) The shrinkage of soils. *Journal of Agricultural Science* 8: 312–330.

TEMPERATE REGION SOILS

E A Nater, University of Minnesota, St. Paul, MN, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

A multiplicity of definitions for temperate regions exist, with a central concept that these regions have continental climates with significant seasonal differences in temperature. The US Department of Agriculture–Natural Resources Conservation Service (USDA–NRCS) defines temperate regions, and hence temperate region soils, as having mesic, thermic, or hyperthermic temperature regimes; iso regimes (those experiencing less than a 6°C annual difference in soil temperature at 50 cm depth) are normally excluded. Although aridic and xeric moisture regimes are often considered to occur in temperate regions, mainly subhumid (ustic) and humid (udic and perudic) temperate region soils are discussed here.

Temperate regions, and hence temperate region soils, fall mainly in the mid- to lower-latitude regions of the globe. The length and temperature of the growing season are sufficient for growth of annual grain and oilseed crops. In humid and subhumid temperate regions, precipitation is sufficient for growth of crops without irrigation, though there may be a dry season toward the end of the growing season in the drier parts of this region, thus necessitating early-maturing crops such as winter wheat. The majority of lands cultivated by intensive, mechanized agricultural practices occur in temperate regions.

Temperate regions are bounded by boreal regions, where soil temperatures are lower and the growing season is shorter, and by tropical regions, where temperatures are warmer and frosts rarely, if ever, occur. Humid and subhumid temperate regions are bounded by arid regions, where moisture is insufficient to support the growth of most crops without irrigation, and Mediterranean regions, where the majority of precipitation occurs during the cool season and the growing season is generally very dry. Overall, temperate regions constitute almost 40% of the Earth's ice-free terrestrial surface (Table 1).

State Factors

Climate

Climate is the initial bounding parameter for any discussion of temperate regions. Temperate regions encompass gradients in both moisture and temperature, but also in atmospheric moisture deficit, which has profound effects on evapotranspiration and consequently on soil moisture status and soil properties. In addition, the relative timing between annual temperature cycles, wet and dry cycles, and atmospheric moisture deficit are also important determinants for the plant communities that thrive in these landscapes. Climate strongly influences the development of soil properties, both directly through its effect on soil moisture status and indirectly through its effect on organisms.

Other factors being equal, soils in wetter parts of the region tend to have deeper sola and experience stronger leaching, lower base saturation, losses of

Table 1 Areal distribution of soil orders worldwide and in temperate regions

	Area		*Proportion*	
	Earth's surface (10^6 km^2)	*Temperate regions (10^6 km^2)*	*Earth's surface (%)*	*Temperate regions (%)*
Gelisols	11.26	–	8.61	–
Histosols	1.53	0.121	1.17	0.23
Spodosols	3.35	0.592	2.56	1.14
Andisols	0.91	0.202	0.7	0.39
Oxisols	9.81	0.197	7.5	0.38
Vertisols	3.16	1.642	2.42	3.17
Aridisols	15.73	11.560	12.02	22.29
Ultisols	11.05	3.405	8.45	6.57
Mollisols	9.01	4.491	6.89	8.66
Alfisols	12.62	4.868	9.65	9.39
Inceptisols	12.83	5.383	9.81	10.38
Entisols	21.14	14.355	16.16	27.68
Shifting sands	5.32	4.68	4.07	9.02
Rock	12.08	0.363	10.0	0.70
Totals		51.865	39.65	100.00

soluble salts, gypsum, and carbonates, and higher rates of mineral weathering. Soils in drier parts of the region, particularly those with high seasonal atmospheric moisture deficits, tend to have shallower sola, less leaching, and consequently higher base saturation, and are more likely to have accumulations of secondary carbonates and gypsum. Salts more soluble than gypsum tend to accumulate only in hydrologic discharge areas in these regions, though their accumulation is more common in aridic moisture regimes. Drier soils also tend to have lower rates of mineral weathering.

Likewise, soils in warmer parts of the region tend to be more weathered, have lower base saturation, and typically have less organic matter accumulation in the A horizon due to increased rates of organic matter decomposition. Increases in soil temperature enhance the rate of many chemical and biochemical pedogenic processes, particularly soil mineral weathering. Typically, a 10°C increase in temperature will produce a two- or three-fold increase in the rate of many biochemical reactions.

Soils in cooler parts of the region tend to have less mineral weathering, higher base saturation, and higher organic matter contents in the A horizon. These trends are reinforced by the occurrence of large areas of younger parent materials in the cooler parts of the region due to direct or indirect effects of continental glaciation, and also by the fact that the soil surface may be frozen for a significant portion of the year in colder climates, thus reducing the duration of action of many biologically and hydrologically driven pedogenic processes.

The balance between precipitation (or, more correctly, infiltration) and evapotranspiration controls the balance between leaching and chemical precipitation, and thus accumulation of carbonates, gypsum, and more soluble salts. Where precipitation exceeds evapotranspiration, leaching of relatively soluble materials occurs as water in excess of field capacity can move out of the system in response to gravity. Over time, leaching will remove all soluble salts, carbonates, and gypsum from soils, even those formed on carbonate bedrock.

Where precipitation is less than evapotranspiration, soil moisture status is less likely to exceed field capacity, and less water is available to be transported out of the solum. Consequently, carbonates, gypsum, and soluble salts will be moved downward in the profile with the wetting front, but may not be leached entirely out of the solum. Over time, horizons of carbonate, gypsum, or even soluble salts may form under horizons that have lost most, if not all, of the carbonates, gypsum, or soluble salts by leaching. Because of their high solubility, and hence mobility, salts more soluble than gypsum are easily removed from soils.

This simple analysis should be taken only as a rough guide; episodic precipitation events, snow melt, noncoincidence of precipitation and high evapotranspirative demand, and even bypass flow may all produce significant leaching even when annual total precipitation is much less than evapotranspiration.

Parent Materials

Because of their large areal extent, temperate region soils encompass a broad array of parent materials, ranging from residuum to glacially derived sediments to loess, alluvium, and colluvium. Although this broad array exists, a few generalities can be observed.

In the colder parts of temperate regions, particularly in the northern hemisphere, a significant proportion of parent materials were derived either directly or indirectly from processes associated with late Quaternary glaciation. Many of these soil landscapes formed on glacial drift, glaciofluvial, or glaciolacustrine sediments. Extraglacial areas near glacial margins were strongly affected by periglacial processes and may have experienced extensive wind erosion or fluvial reworking. Still other landscapes, often far removed from the glacial margins, were blanketed by thin-to-thick loess deposition derived mainly from riverine sediment sources associated with the wastage of continental ice sheets. Landscapes formed on recent glacial sediments have deranged (disconnected) drainage networks with numerous closed depressions and poor internal drainage.

In warmer parts of the region, soil landscapes not affected by recent glaciation or loess deposition form mainly on residuum, alluvium, or colluvium. Properties of these parent materials are closely related to local bedrock sources and thus may vary widely from one region to another. Soil landscapes in warmer temperate regions may show considerably more development and weathering due to the more stable nature of the surfaces, the longer duration of their exposure to pedogenic processes that alter them, and warmer soil temperatures. These landscapes generally have evolved highly dissected, well-connected, dendritic surficial drainage networks with few closed depressions.

The majority of andic parent materials are associated with continental margins, and particularly the Pacific rim. Restriction of this discussion to humid and subhumid temperate regions eliminates xeric, Mediterranean climates strongly affected by oceanic thermal masses, and thus effectively excludes nearly all of the soils dominated by andic parent material. The remaining parent materials, then, were derived from crystalline rocks.

Organisms

The majority of landscapes in humid and subhumid temperate regions are dominated by one of two ecosystem types: prairie or steppe, and deciduous or mixed deciduous–coniferous forest.

Low-relief landscapes, particularly in the drier parts of the region, are dominated by prairies or steppes, which range from shortgrass prairies, with a significant annual grass component in drier temperate regions, to tallgrass prairies, dominated by perennial bunch grasses in wetter temperate regions. Prairies can withstand and more readily regenerate from frequent fires than most forests can. Fires started in low-relief landscapes are often windswept and can cover very large areas, whereas higher-relief landscapes tend to slow or stop the spread of fires.

Deciduous or mixed deciduous–coniferous forests are also common in humid and subhumid temperate regions, particularly in areas of higher relief. They transition to coniferous or broadleaf evergreen forests in the warmer parts of the region, and to coniferous forests near the boreal region border.

Other ecosystem types such as spruce bog or fen, and savanna, are interspersed within temperate regions, but the prairie and forest general types predominate. Bounding this region are chaparral or desert scrub ecosystems in more arid regions; savanna, chaparral, or annual grasslands in Mediterranean regions; coniferous forests in boreal regions; and broadleaf evergreen forests or savannas in the tropics. These ecosystem types strongly affect soil properties.

The formation of thick, dark, organic-rich A horizons (mollic epipedons) is common under grassland- or savanna-type vegetation. Grasses have a deep, fibrous root distribution and a rapid root turnover rate (approximately one-third of grass roots die and are replaced annually) which, in combination, produce significant inputs of organic matter at depth, requiring no translocation to move the organic matter deep into the solum. In addition, the efficient base-cycling present in grassland ecosystems leads to high base saturation, a requirement for mollic horizon classification. Soils with mollic epipedons typically do not have identifiable E horizons owing to the lack of strong leaching in subsurface horizons and to the depth of the A horizon which masks them if present. Notable exceptions occur in Albolls, where distinct albic E horizons are present, often bisecting the mollic epipedon into a surficial and a subsurface portion.

Deciduous and mixed deciduous–coniferous forests typically have roots that are distributed much closer to the soil surface and have a much lower root-turnover rate. The majority of organic inputs in forests are from leaf litterfall, which occurs at the soil surface. Incorporation of leaf litter into the soil requires the activity of worms or other invertebrates. Where leaf materials are not incorporated, the soil typically develops a thin-to-thick layer of partially decomposed organic materials at the soil surface, termed the forest floor or the litter layer. This O horizon can be several centimeters thick and usually rests on an underlying E horizon formed by eluviation of clays and chelation, and subsequent leaching of iron and other coatings from the surface of silt and clay grains in the matrix. It may also rest on a thin A horizon. If earthworms or other invertebrates incorporate the forest floor into the underlying mineral soil horizons, then a relatively thin, light-colored A horizon (ochric epipedon) is typically observed. Introduction of nonnative earthworm species into forested landscapes in the northern USA has all but eliminated the thick O horizons once commonly observed in many forests.

Likewise, B horizon development is also strongly related to the organism factor. B horizon development is often minimal in prairie soils. Many have no B horizon at all (A/C morphology) or else have cambic B horizons that display a moderate degree of development. Older, more well-developed prairie soils may have argillic or natric horizons, but they are not nearly as common under prairie as they are under forests. Several factors tend to inhibit clay dispersal and movement in prairie soils, including the presence of organic coatings in the A horizon that tend to bind clays together; higher exchangeable Ca contents and higher pH generally present in grasslands, which promote flocculation; and the lack of stemflow common to deciduous forests, which tends to concentrate a high proportion of the precipitation falling on a tree directly under its bole, thus producing a localized effect similar to greatly increased precipitation.

Soils formed under deciduous or mixed deciduous–coniferous forest typically display prominent B horizon development. Argillic horizons are commonly observed under forest, with kandic horizons present in the more highly weathered soils or in those that formed on acidic parent materials; spodic horizons occur in coarse-textured soils dominated by coniferous or other vegetation that contributes large amounts of soluble fulvic acids to the soil surface.

Soils under coniferous forests may have spodic B horizons, resulting from podzolization processes. Coniferous leaf litter is more acidic than that of most deciduous forests and releases large quantities of fulvic acids, which are soluble in acidic environments. These organic compounds can chelate Fe, Mn, and Al from metal (hydr)oxide coatings on the soil

grains. In relatively coarse-textured soils, the fulvic acid–metal chelates are transported with the soil solution as long as they remain soluble, producing a distinct E horizon where metal (hydr)oxides have been removed. The chelated metals are deposited and accumulate lower in the profile due to changes in pH or other factors, producing distinctive red or black horizons of accumulation of Fe, Mn, and Al chelates.

Topography

Soil properties and soil-landscape relations are strongly affected by topography, in large part because topography exerts significant control on soil moisture relations and on erosional and depositional processes. There is a clear interaction between rates of infiltration and precipitation that strongly influences these relationships. Where the infiltration rate is lower than the rate of precipitation, runoff will occur. Runoff will be minimal on relatively flat summit positions due to their low slope, thus maximizing time for infiltration to occur. Runoff increases on shoulder and backslope positions, where slopes are higher, thus decreasing infiltration. These landscape positions are also the most susceptible to erosion by water due to the higher runoff and steeper slopes. Water and materials transported from the upper landscape positions accumulate on footslope and toeslope landscape positions. These are often the wettest landscape positions and are sites for deposition of eroded sediments. In more arid regions, they may also be sites for accumulation of soluble salts leached from higher landscape positions.

Soils in summit positions often show relatively more development than soils in other landscape positions due to increased infiltration and limited erosion in the summit position. These soils typically have deeper sola and are more leached than other soils in other landscape positions.

Infiltration into soils on shoulder and backslope landscape positions is proportionally lower than on summits due to increased runoff. Soils in these landscape positions are generally less developed than others in the same landscape, owing in part to their lower effective moisture status and to erosion. Erosion in these landscape positions is high because of the steeper slopes and the volume of runoff that occurs on them. These soils usually have limited B horizon development, thinner A horizons, and shallower sola than soils at the summit.

Soils in footslope and toeslope positions accumulate water and materials that were transported from above. In drier climates, these soils may show the greatest degree of development because they receive the most water. In wetter climates, footslope and especially toeslope landscape positions may be seasonally or permanently saturated. They may also be sites for accumulation of eroded sediments or soluble salts. Often the sola are very thick owing to addition of eroded materials.

If the rate of infiltration exceeds that of precipitation, runoff will not occur. If soil moisture exceeds field capacity, it will move in response to gravity. Depending on the presence or absence of subsurface strata or horizons that restrict the flow of water, water will either move downward to groundwater or may move laterally through the soil or underlying sediments by throughflow or lateral flow. Far more detailed hydrologic analyses are available now using digital terrain modeling and flownet analyses.

Landscape drainage patterns can significantly affect soil landscapes and the catenary relationships between soils. They have less effect on soils in upper landscape positions, more in lower landscape positions.

Relatively young landscapes typically have poorly dissected, deranged drainage patterns, where surficial connectivity is limited, drainage is mostly internal, and numerous closed depressions exist that collect water during the wet season and form seasonal-to-permanent depressional wetlands with hydric soils and wetland vegetation. This is particularly true in glacial landscapes, where the hummocky terrain commonly associated with glacial drift produces numerous closed depressions. The hydrology of landscapes with deranged drainage patterns is often quite complex. Soil properties are strongly affected not only by landscape position, but also by position with respect to recharge or discharge hydrology. Soils in footslope and toeslope positions are generally saturated for significant periods during the year and experience alternating reducing and oxidizing environments. The soils commonly have high organic matter contents and display a variety of redoximorphic features. Wetland rims may be water discharge sites and can sometimes contain significant salt, gypsum, or carbonate accumulations.

As landscapes evolve and mature, their surficial drainage patterns become more highly dissected, intersecting and draining closed depressions. They develop dendritic drainage patterns that have higher-order streams, increasingly more surficial connectivity, and fewer closed depressions. Lakes, ponds, and depressional wetlands are less common. The landscapes have more surficial and less internal drainage.

Significant differences between deranged and dendritic landscapes occur in the footslope and toeslope positions. In deranged landscapes, toeslopes are typically seasonal or permanent wetlands, with high concentrations of poorly decomposed organic materials, at least seasonal anaerobic conditions, and accumulation of erosional materials. Footslope

morphologies may be quite different depending on the hydrologic flow associated with them.

Recharge hydrology (soil water moves vertically to the groundwater) in wetlands is associated with leaching of soluble materials and removal of colloids from horizons or soils. Argillic horizon formation is a common result of recharge hydrology. Discharge hydrology (water moves from the groundwater table to or near the soil surface) is associated with chemical precipitation and is best observed in semiarid regions where wetland rims may display high salt concentrations resulting from capillary rise from saline groundwaters. Seasonal flow reversals are also common.

In dendritic landscapes, footslope and toeslope landscape positions often terminate in or near seasonal-to-perennial streams or alluvial floodplains that seasonally accumulate sediments. These landscape positions may also experience seasonal or permanent saturation, but it is less common than in deranged drainage systems.

Soils subjected to seasonal-to-permanent flooding or saturation can develop strongly anaerobic conditions because the rate of oxygen diffusion in the aqueous phase in soils is much lower than it is in the gas phase. If oxygen demand in the soil, generated mainly by root and microbial respiration, exceeds the rate of oxygen diffusion into the soil, reducing conditions can develop. Reduction requires both saturation of the soil and sufficient microbial activity to deplete oxygen from the soil so that alternate electron acceptors such as Fe^{3+} become favored. Saturation for insufficient duration or when microbial respiration is very slow (low soil temperatures or lack of suitable microbial substrate) will not produce sufficient oxygen demand in soils to consume free oxygen and thus will not produce reducing conditions or redoximorphic features.

Reducing conditions can lead to the dissolution of Fe and Mn (hydr)oxides, transport of reduced Fe^{2+} and Mn^{2+} with soil solution, and reprecipitation of Fe and Mn (hydr)oxides in other locations in the soil where pH and redox conditions favor precipition. Both Fe^{3+} and Mn^{4+} (hydr)oxides are very insoluble; the reduced forms (Fe^{2+}, Mn^{2+}), however, are very soluble in the aqueous phase and thus can be transported with the soil water and either removed entirely from a location in the soil, thus producing redox depletions, or moved to a different zone or region in the soil, where they are reoxidized to form redox concentrations. Because even small amounts of Fe and Mn coatings can mask the color of the underlying soil matrix, their removal provides a clear visual indicator of redoximorphic processes.

Saturated soils typically have higher organic matter contents than soils in more aerobic environments due to lower rates of organic-matter decomposition under anaerobic conditions. Histic epipedons may develop over mineral soil surfaces under these conditions. If the duration of saturation extends throughout most of the year, Histosols (organic soils) may develop.

Age

Soil age refers to the time since exposure of a surface and, as such, is a measure of the length of time a soil or soil landscape has been affected by pedogenic processes. Soil surfaces in whole regions may be exposed by processes such as continental glaciation or loess deposition, or over much smaller areas as a result of smaller-scale processes such as hillslope erosion or deposition.

Initially, soil properties strongly reflect the properties of the parent materials on which they form. Over time, they are altered by pedogenic processes and, given sufficient time, may little resemble their initial condition. Formation of horizons takes periods ranging from a few hundred years (formation of A horizons) to millennia (argillic, spodic, calcic horizons), to hundreds of thousands or even millions of years (oxic horizons).

Mineral fractions in young soils are derived from the parent materials and are largely unaltered. As soils age, mineral weathering occurs and secondary minerals begin to form. Over time the mineral fraction may be significantly altered from its initial state and be more closely related to weathering products controlled by thermodynamic stability fields. Over sufficiently long periods of time, soils with relatively different parent materials may end up with relatively similar mineralogies and compositions. Kaolin group minerals, which form by intense weathering of primary silicate minerals, often dominate the clay mineral fraction in highly weathered soils. In high-intensity, highly leached, base-poor environments, less thermodynamically stable clays such as illites and smectites are removed by chemical weathering and disappear entirely from older landscapes unless protected by microenvironments that are more suitable for their stability.

Effects of Humans

Humans also affect soils and their properties, particularly through agricultural and construction activities, but also through more subtle means. Temperate region soils have undergone extensive change due to the high density of intensive, mechanized agriculture. Agricultural tillage has caused accelerated erosion on many landscapes by leaving soil surfaces poorly protected against raindrop impact. Tillage and agricultural drainage of wetlands or seasonally wet soils have increased aeration of the solum and enhanced the rate of organic matter oxidation, releasing vast

quantities of CO_2 to the atmosphere. Many agricultural soils in temperate regions have lost more than 50% of their precultivation organic matter contents, with resultant losses in tilth and native fertility. Soil acidification is another widespread effect, resulting both from atmospheric deposition of nitric and sulfuric acid related to fossil fuel burning, but also from the application of nitrogen fertilizers.

Classification

Table 1 provides a generalized view of the areal extent of soils by US taxonomic order. The definition of 'temperate region' used in Table 1 is broadly defined to include both aridic and xeric moisture environments.

The dominant soil orders in temperate regions, as broadly defined, are Entisols and Aridisols, which comprise 27.7% and 22.3% of the total area, respectively. Other common orders include Inceptisols (10.4%), Alfisols (9.4%), Mollisols (8.7%), and Ultisols (6.6%). Nonsoil surfaces (shifting sands and rock) make up almost 10% of the total area of temperate regions. The remaining orders in toto constitute approximately 6% of the region.

Similar data portraying the areal extent of soil orders only in humid and subhumid temperate regions are not readily available. Because of the uncertainties in distribution of many of these orders between humid and subhumid and aridic and xeric moisture regimes, it is difficult to provide accurate relative distributions of these orders. However, some trends in soil distribution in these regions can be estimated from knowledge of soil distributions in aridic and xeric moisture regimes. Excluding soils from aridic moisture regimes would eliminate all Aridisols and nearly all of the shifting sands, along with large areas of Entisols. All of these soils and/or terrestrial surfaces are common in aridic moisture regimes. Mollisols, Vertisols, and Oxisols are also present in aridic moisture regimes, but constitute a smaller proportion of the total areas. Alfisols and Inceptisols are, by definition, excluded. It is more difficult to make generalizations about the areal extent of soil orders in xeric moisture regimes as compared to humid and subhumid temperate regimes, as all of the orders common to one region are also present in the other, though in varying proportions. The only clear distinction is that the majority of Andisols are located along coastal margins, particularly around the Pacific Rim; excluding xeric moisture regimes from the discussion would effectively exclude nearly all Andisols also.

For the most part, humid and subhumid temperate regions are dominated by Inceptisols, Entisols, Alfisols, Mollisols, and Ultisols. Other orders are present but generally constitute a smaller relative portion of humid and subhumid temperate regions.

Some general trends for the distribution of soil orders in humid and subhumid temperate regions follow. Alfisols and Ultisols are mainly associated with forested landscapes and Mollisols with prairies. Ultisols occur in warmer, wetter, and more stable landscapes of the region where soils have weathered more and base saturations are consequently lower. Alfisols are more common under forests in cooler, drier climates and younger landscapes where weathering, leaching, and removal of bases is not as extensive. Spodosols occur near the boundaries with boreal and tropical regions, where they are much more common.

Entisols and Inceptisols are interspersed throughout the region on erosional or depositional landscape positions, particularly shoulder and backslopes, where soil development is limited or new parent materials have been exposed or deposited. They are also common in wetter portions of the landscape where soil development processes are slow. Histosols occur in closed depressions and other very poorly drained landscape positions, particularly in glaciated regions, where closed depressions abound. Vertisols develop in clayey soils with high shrink–swell clays and seasonally wet and dry periods. They are common in alluvial and lacustrine settings, or form on decomposing shales, and are more common in the drier portions of the region.

List of Technical Nomenclature

Deranged drainage	Drainage patterns characterized by a lack of surficial connectivity and the presence of numerous closed depressions. Typically associated with relatively young landscapes
(Hydr)oxides	A term used to describe, in aggregate, oxides, hydroxides, and oxyhydroxides of metals, particularly iron and manganese, in soils. Used particularly when one is referring to materials which may fall in any of those three classes

All taxonomic terms and soil diagnostic horizon terms are from USDA–NRCS *Soil Taxonomy*.

See also: **Cold-Region Soils**; **Mediterranean Soils**; **Tropical Soils:** Arid and Semiarid; Humid Tropical

Further Reading

Birkeland PW (1999) *Soils and Geomorphology*, 3rd edn. London, UK: Oxford University Press.
Buol SW, Hole FD, McCracken RJ, and Southard RJ (1997) *Soil Genesis and Classification*, 4th edn. Ames, IA: Iowa State University Press.
Daniels RB and Hammer D (1992) *Soil Geomorphology*. New York: John Wiley.
Jenny H (1980) *The Soil Resource*. New York: Springer-Verlag.
Ruhe RV (1975) *Geomorphology*. Boston, MA: Houghton Mifflin.
Soil Science Society of America (1997) *Glossary of Soil Science Terms*. Madison, WI: Soil Science Society of America.
Soil Survey Staff (1999) *Soil Taxonomy*, 2nd edn. A basic system of soil classification for making and interpreting soil surveys. USDA Soil Conservation Service, Agricultural Handbook No. 436. Washington, DC: US Government Printing Office.
Wright HE Jr. (ed.) (1983) *Late-Quaternary Environments of the United States*, vol. 2. The Holocene. Minneapolis, MN: University of Minnesota Press.

Temperature Regime *See* **Thermal Properties and Processes**

TENSIOMETRY

T K Tokunaga, E.O. Lawrence Berkeley National Laboratory, Berkeley, CA, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

The practical application of equilibrium and non-equilibrium concepts to the hydrostatics and hydrodynamics of soil water in laboratory and field environments depends on the ability to measure relevant potentials. In unsaturated soils, the matric and gravitational potentials are usually regarded as the hydraulically important energy components. A tensiometer is a device used to measure the matric potential, ψ_m, the component of the soil water chemical potential that is directly dependent on water content. When the local soil gas-phase pressure differs from that of the reference pressure (usually taken as the local atmospheric pressure), the tensiometer reading also includes this difference as an additional air pressure potential. Thus, the tensiometer most generally equilibrates to the sum of the matric and air pressure (pneumatic) potentials, also referred to as the tensiometer potential. In addition, tensiometers can be sensitive to shrinking and swelling of soils. For simplicity, this discussion will be restricted to conditions where the soil gas phase is in equilibrium with the atmosphere, and soil volume changes are negligible, such that the tensiometer equilibrates simply with the soil matric potential.

The basic components of a tensiometer consist of a porous tip or cup, an internal cavity, and a device to measure pressure within the cavity (Figure 1). The water-saturated porous boundary (the tensiometer tip or cup) is placed in contact with soil. This permits exchange of soil water between the region of interest and the tensiometer cavity until equilibrium is reached. Thus, the measurement of matric (or capillary) potential is generally achieved through determining the pressure of a fluid phase within the tensiometer, after equilibrium with the surrounding soil has been established. If the water table rises up to or above the tip, the tensiometer provides a measurement of the local pressure potential rather than the matric potential. Because most tensiometers rely on measuring the pressure of liquid water inside the tensiometer body, the lowest attainable equilibrium reading is imposed by the vapor pressure of water. Although this suggests that the lower operational limit ranges from -100 to -95 kPa for temperatures ranging from 5 to 35°C, a practical lower limit of approximately -85 kPa is commonly encountered because of exsolution of gases (air). The lower operating limit should be taken into account in selecting the porous tensiometer tip. Hydrophilic ceramics are most commonly used, but water-wettable porous glass, stainless steel, plastic, cloth, and even paper are used for tensiometer tips in various applications. The air-entry value of the porous tip is the gauge pressure at which the air phase displaces water-filled pores. Thus, when the matric potential decreases

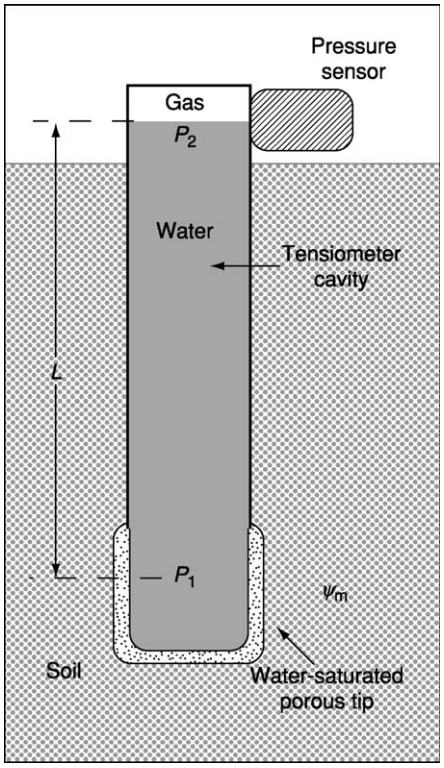


Figure 1 Basic components of a tensiometer. When significant, the hydrostatic pressure difference acting over the vertical distance L between the measurement point and the tensiometer tip must be accounted for. P_1, gauge pressure at elevation of interest.

(becomes more negative) such that its magnitude is larger than the air-entry value, air leaks in through the tip, causing the tensiometer to fail. Ceramics with air-entry values of approximately −100 kPa are typically used for tensiometer tips in field applications. More permeable ceramics with lower-magnitude air-entry suctions (e.g., −50 to −5 kPa) are used in some laboratory applications.

The original tensiometers, developed in the early twentieth century, consisted of a water-filled porous interface connected to a manometer. Since introduction of these early instruments, a wide variety of tensiometer designs have been developed. In field applications, tensiometers are used for irrigation scheduling, characterization of flow in soil profiles and deeper unsaturated sediments, and monitoring of contaminated soils. In the laboratory, tensiometers are used in soil physics, agricultural, and vadose-zone hydrologic research.

Equilibrium

By measuring the pressure of a fluid phase within a tensiometer, a measurement of the soil matric potential is obtained. The reliability of this measurement depends on establishing local equilibrium between the tensiometer and the soil, and on the condition that the matric potential is within the instrument's operational range, criteria that are discussed later. Equilibrium between the soil water and tensiometer can be viewed mechanically as a balance between capillary and adsorptive forces retaining water in the soil, and the hydrostatic pressure (suction) of water within the tensiometer. More generally, net exchange of soil water between the tensiometer (system) and soil (reservoir) occurs until differences in soil water chemical potential vanish. At this state, the gauge pressure in the tensiometer at the elevation of interest (denoted P_1 in Figure 1) is equal to ψ_m. When the pressure is recorded at an elevation different from that of the tensiometer tip, a hydrostatic correction is applied. In the example shown in Figure 1, $\psi_m = P_2 + \rho_w g L$, where ρ_w is the density of water, and g is the acceleration due to gravity. The time necessary to attain such equilibrium is generally determined by hydraulic properties of the soil, the soil–tensiometer contact, and response characteristics of the tensiometer.

Response Time

The time required to attain soil–tensiometer equilibrium depends on two basic factors: the rate of soil water transfer, and the change in tensiometer pressure per unit mass of soil water exchanged. These two factors can be parameterized in terms of the tensiometer–soil conductance and tensiometer gauge sensitivity. The overall conductance is determined by the (saturation-dependent) soil hydraulic conductivity, hydraulic contact between the soil and tensiometer tip, and the conductance of the tensiometer tip. Here, only the strictly tensiometer-dependent factors of tip conductance and gauge sensitivity will be addressed. In order to obtain a tensiometer reading rapidly that accurately reflects the local soil matric potential, it is desirable to maximize both the tip conductance and the gauge sensitivity. As shown below, the product of these two parameters is equal to the inverse of the instrument's response time.

The tensiometer conductance, K, is a measure of the volumetric rate that water can flow at through a tensiometer tip per unit hydraulic potential difference. When the potential difference (outside minus inside of the tensiometer) is expressed in terms of hydraulic head, K has dimensions of (L^2T^{-1}). The preferred maximization of K in order to shorten tensiometer response times can be achieved through increasing the permeability and bulk area of the porous tip, and decreasing the thickness of the tip. Each of these strategies is subject to constraints. Increasing the permeability of the tip material is achieved by

increasing its characteristic pore-size (since permeability scales to the square of pore diameter), and narrowing its pore-size distribution. However, the extent to which the characteristic pore-size can be increased is severely limited by the fact that the air-entry pressure is inversely proportional to pore-size. Thus more permeable tensiometer tips will not permit measurements of lower (more negative) matric potentials. The bulk surface area of the ceramic can be increased in order to provide more area through which soil water is exchanged. This option is limited when spatially resolved measurements of potential distributions are needed, and when constraints are associated with installation (labor and/or disturbance of soil profiles). The final option available for increasing the tensiometer-tip conductance involves decreasing the thickness of the porous ceramic or membrane. Compromised mechanical integrity imposes the primary limit on this option. In addition, thinner membranes may require more structural support to minimize flexing in response to pressure changes. This latter effect compromises the tensiometer's sensitivity.

The change in a tensiometer's pressure reading per unit volume (V_w) exchange of soil water defines its gauge sensitivity, S. When the pressure reading is expressed in head ($h(L)$) units, S has dimensions of (L^{-2}). Different types of pressure indicator have vastly different gauge sensitivities. Assuming the tensiometer cavity is rigid and completely water-filled, a pressure transducer can provide $S > 10^9\,m^{-2}$. In contrast, a simple water manometer (practical only for low-magnitude ψ_m measurements) has $S = dh/dV_w = A^{-1}$, where A is the cross-sectional area of the manometer tube. A 2-mm inner-diameter water manometer thus has a sensitivity of $3 \times 10^5\,m^{-2}$. Improving the sensitivity of manometer-based tensiometry through decreasing A is limited due to also increasing the uncertainty of the contact angle-dependent capillary rise (water) or capillary depression (mercury). Although the aforementioned values span a very wide range, most mechanical (Bourdon tube vacuum gauge) and electronic pressure transducers used in tensiometry have S in the range of 10^7–$10^9\,m^{-2}$. It should be noted that, because gases are highly compressible, the presence of air within tensiometers decreases S. The S associated with a gas cavity of volume V is equal to $P(\rho_w g V)^{-1}$, where P is the absolute pressure of the gas. Thus, a 1-cm^3 volume of air inside a tensiometer will allow S to be no greater than $10^7\,m^{-2}$, even when a much more sensitive pressure transducer is used to obtain that pressure reading.

The tensiometer's response time, τ, indicates how rapidly the instrument equilibrates with its environment when the rate of soil water exchange is controlled by the tip conductance (rather than the soil or soil–tensiometer contact). Under such conditions of tensiometer-limited response, the time parameter is related to K and S through $\tau = (KS)^{-1}$. The rate at which the tensiometer reading R relaxes towards its equilibrium value R_{eq} is then proportional the instantaneously magnitude of disequilibrium:

$$\frac{dR}{dt} = -\frac{(R - R_{eq})}{\tau} \quad [1]$$

In most tensiometers, τ is practically constant over the operational range, such that R_{eq} is approached according to:

$$\frac{R - R_{eq}}{R_0 - R_{eq}} = e^{-t/\tau} \quad [2]$$

where R_0 is the initial tensiometer reading. Thus τ is equal to the time require to diminish the difference between an arbitrary initial reading and the equilibrium reading to e^{-1} (approximately 37%) of the original discrepancy. By $t = 5\tau$, the tensiometer reading is within 1% of R_{eq} relative to the original discrepancy. Although tensiometers relying on pressure transducers or vacuum gauges have τ-values in the range of fractions of seconds to seconds, much slower responses are obtained in most applications due to low, unsaturated hydraulic conductivity of the soil and poor tensiometer–soil hydraulic contact.

Range of Applications

Given the wide variety of environments in which it is desirable to have information on matric potentials, it is worth addressing limitations in ranges of tensiometer applications as well as approaches that have been developed to diminish or circumvent limitations. Here, lower ranges in matric potential, tensiometry in deep unsaturated zones, and temperature-related problems are briefly discussed.

Although water-filled tensiometers have a lower practical limit of about -85 kPa, much lower matric potentials can be measured with osmotic tensiometers. In an osmotic tensiometer, a polymer solution within the tensiometer body equilibrates with soil water through a semipermeable membrane (ideally impervious to the polymer). The polymer solution equilibrates at a positive gauge pressure (in order to compensate for water potential lowering resulting from interactions with the polymer), thereby avoiding negative gauge pressure limitations of conventional tensiometers. Problems arising from high sensitivity to thermal fluctuations and slow leakage of the polymer have severely limited use of osmotic tensiometers, but recent advances may lead to wider interest in this alternative.

Interest in flow through deep unsaturated soils and sediments has grown over the past two decades, and this has necessitated the development of deep tensiometry methods for field applications. Use of a conventional field tensiometer with a water-filled column to provide the hydraulic connection between the buried tensiometer tip and the pressure gauge at the soil surface (Figure 1) permits a decreasing range of matric potential measurement with deeper installations because of the elevation-dependent hydrostatic pressure decrease. Such a conventional tensiometer used for depths greater than about 8 m will permit little to no useful measurements of matric potentials, because the upper section of the column exists at low absolute pressure and rapidly becomes filled with water vapor. Approaches used to circumvent this problem eliminate use of the water-filled column and rely instead on either burying pressure transducers at or near the depth of the tensiometer tip or use of an air-filled column. The latter option is not as suitable for monitoring transient flow conditions, since the air-filled cavity imparts low gauge sensitivity, hence a low response time.

Temperature variations in the field have long been known to make interpretation of tensiometer readings difficult. The phenomenon is complex and often sensitive to instrument design, since thermal effects on volume changes of tensiometer components and fluids can be substantial. This can be especially problematic when instrument sensitivity is high and the overall tensiometer–soil conductance is low. Keeping all instrument components below the soil surface (where maximum temperature variations occur) can diminish temperature-induced changes in tensiometer readings. When tensiometers are used in field locations exposed to freezing temperatures, protection from damage by ice formation is necessary. This can be accomplished to an extent by adding methanol or antifreeze to the water in the tensiometer, or by burying the instrument below the depth of soil freezing.

List of Technical Nomenclature

ρ_w	Density of water ($kg\,m^{-3}$)
τ	Response time (s)
ψ_m	Matric potential (Pa)
g	Acceleration due to gravity ($m\,s^{-2}$)
h	Head (m)
K	Conductance ($m^2\,s^{-1}$)
L	Length (vertical) (m)
P	Pressure (Pa) or (kPa)
R	Reading (tensiometer) (kPa)
S	Sensitivity (gauge) (m^{-2})
t	Time (s)
V	Volume (cm^3) or (m^3)

See also: **Hydrodynamics in Soils**; **Thermodynamics of Soil Water**; **Vadose Zone:** Hydrologic Processes

Further Reading

Cassel DK and Klute A (1986) Water potential: tensiometry. In: *Methods of Soil Analysis*, 2nd edn, pp. 563–596. Monograph No. 9. Madison, WI: American Society of Agronomy.

Jury WA, Gardner WR, and Gardner WH (1991) *Soil Physics*, 5th edn. New York: John Wiley.

Or D (2001) Who invented the tensiometery? *Soil Science Society of America* 65: 1–3.

Peck AJ and Rabbidge RM (1969) Design and performance of an osmotic tensiometer for measuring capillary potential. *Soil Science Society of America Proceedings* 33: 196–202.

Richards LA (1928) The usefulness of capillary potential to soil moisture and plant investigators. *Journal of Agricultural Research* 37: 719–742.

Richards LA (1949) Methods of measuring soil moisture tension. *Soil Science* 68: 95–112.

Sposito G (1981) *The Thermodynamics of Soil Solutions*. Oxford, UK: Oxford University Press.

Stannard DI (1990) Tensiometers – theory, construction, and use. In: Nielsen DM and Johnson AI (eds) *Ground Water and Vadose Zone Monitoring*, pp. 34–51. ASTM STP 1053. Philadelphia, PA: American Society for Testing and Materials.

Towner GD (1980) Theory of time response of tensiometers. *Journal of Soil Science* 31: 607–621.

Termites *See* **Fauna**

Terra Rossa *See* **Mediterranean Soils**

TERRACES AND TERRACING

G R Foster, Bryan, TX, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Terraces are runoff management structures used to control soil erosion by water. All terraces are topographic modifications of hillslopes. Terraces reduce rill and interrill erosion by shortening the overland flow path (slope) length or reducing effective landslope steepness. Some terraces reduce net erosion by depositing a portion of eroded sediment on the hillslope. Terraces capture runoff and convey it off the hillslope without causing excessive erosion. Captured runoff increases soil moisture and thus crop production. Terraces are designed, installed, and maintained to fit local climate, soil, topography, and crop management in accordance with available resources, preferred land use, and other socioeconomic factors.

Benefits and Limitations

Terrace systems are highly effective at preventing excessive rill erosion, eliminating ephemeral gully erosion, reducing sediment yield, conserving soil moisture, protecting landscape quality, and increasing land value.

Terraces are topographic modifications that require soil displacement to construct them. They work best on deep soils, such as loess soils. Terraces require a significant investment to build and maintain. Farming with terraces may be inconvenient, and they may limit the choice of farming practices. Terrace systems that do not fit local conditions can be worse than no terraces at all.

Runoff and Erosion on Hillslopes without Terraces

Runoff begins as overland flow spread uniformly around the hillslope. Overland flow is terminated by draws (concentrated flow areas) on the landscape where flow becomes channelized. These draws are formed during landscape evolution. Figure 1 illustrates these flow paths.

Soil erosion by water on the overland flow areas is known as rill and interrill erosion. Rill erosion is the detachment of soil particles by surface flow, and it occurs in a series of rills, which are small, parallel channels a few tens of millimeters wide. Location of rills is determined by soil surface microtopography. Interrill erosion occurs on interrill areas, which is the space between rills. Interrill erosion is detachment of soil particles by raindrop impact. Detachment is the separation of soil particles (sediment) from the soil mass. Deposition is the accumulation of sediment on the soil surface. Sediment load is the amount of the sediment that is transported downslope. Most downslope sediment transport is by surface runoff. Detachment increases sediment load in a downslope direction, while deposition decreases sediment load.

Erosion in concentrated flow areas is known as concentrated flow erosion. This erosion is also known as ephemeral gully erosion if tillage fills the channels and runoff reforms the channels during each cropping cycle. Rills are obliterated by tillage. Rills are typically reformed in different locations depending on the soil-surface microtopography left by tillage. Ephemeral gullies always occur in the same location because of the dominance of landscape macrotopography.

Types of Terraces

The major types of terraces are gradient, parallel-impoundment, bench, and ridge. Most terrace systems involve runoff interceptors that collect the overland flow and redirect it around the hillslope. Most terrace systems include water-conveyance structures that receive the intercepted flow and convey it to low points on the hillslope.

Gradient Terraces

Construction Gradient terraces use an embankment and associated channel to intercept overland flow and direct it around the hillslope. Flow from the terraces is discharged into a protected channel that conveys the flow downslope, as illustrated in Figure 2. These terraces are typically used on hillslopes less than 10% in steepness. Gradient terraces are constructed tens of meters apart and typically use low embankments and shallow channels so that the entire area can be farmed with tractors and wide farm

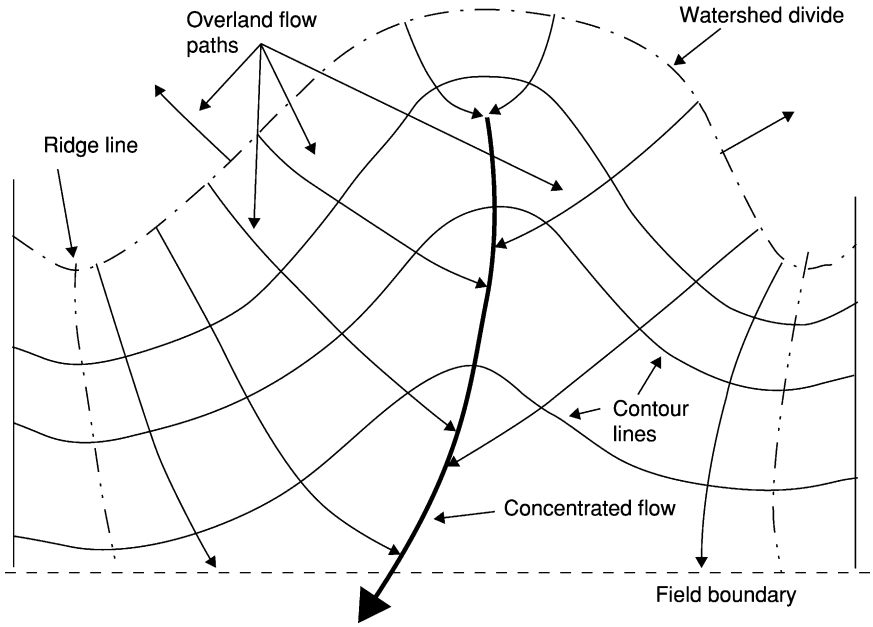


Figure 1 Overland flow and concentrated flow paths on a typical hillslope without terraces.

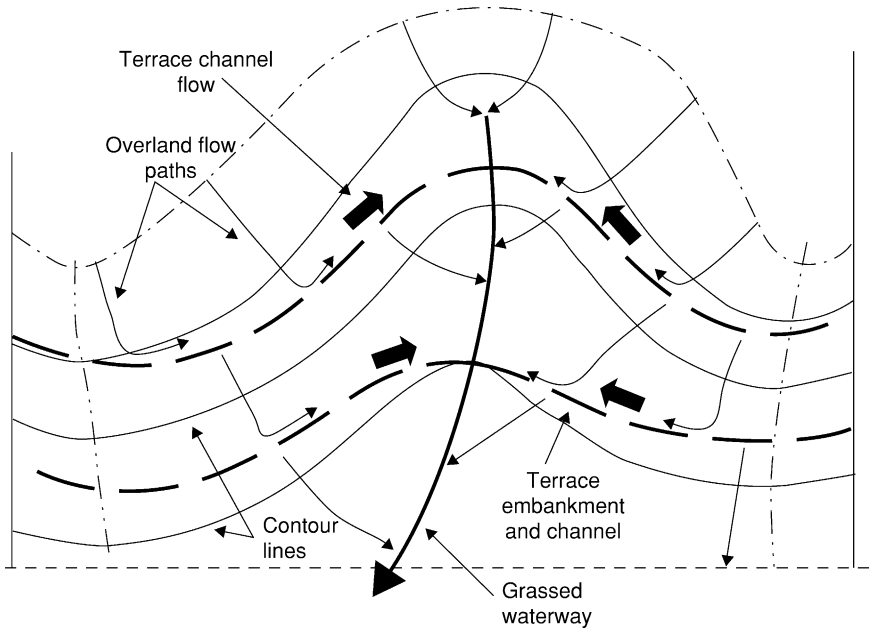


Figure 2 Overland flow and concentrated flow paths on a typical hillslope with gradient terraces on a uniform grade.

equipment. These terraces are referred to as broad-base terraces. Gradient terraces can be constructed with road graders, earthmovers, special terracing machines, and plows used in farming. Cuts and fills are balanced to minimize construction costs and the soil depth that is disturbed. Deep cuts expose subsoil and reduce crop production.

Gradient terraces shorten the 'slope length' of the overland flow path. Rill and interrill erosion vary approximately with the square root of slope length. Adding two gradient terraces to a 100 m slope length creates three slope lengths of 33 m minus the distance from the center of the terrace embankment to the upper edge of the terrace channel as illustrated in Figures 2 and 3. These terraces reduce rill and interrill erosion by ~40%.

Spacing Terraces are spaced to avoid excessive rill erosion. Equations for terrace spacing have been developed based on field experience. A rule of thumb is

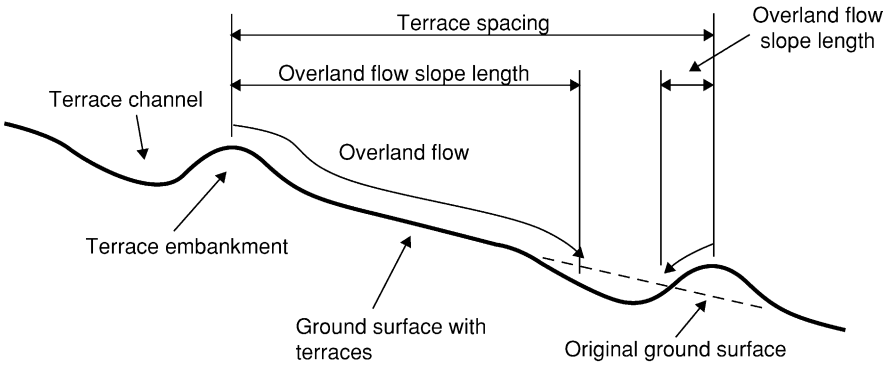


Figure 3 Terrace embankment and channel cross section for a broad base terrace.

that rill erosion is excessive when average annual rate for rill and interrill erosion exceeds 15 t ha^{-1} year^{-1}. Terrace spacing is decreased where rainfall amount and intensity are high and where the soil has a low infiltration rate, both of which produce high runoff rates. Terrace spacing is also decreased on steep hillslopes. High runoff rates and steep slopes increase runoff erosivity and the potential for rill erosion. Terrace spacing is decreased for soils that are susceptible to rill erosion.

Vegetative cover including plant stems, roots, and residue, and soil surface roughness reduce runoff erosivity. Terrace spacing is based on the most vulnerable period for rill erosion during a cropping cycle, which is often the seedbed preparation period, especially for clean-tilled row crops such as maize or soybean.

Terrace spacing can also be based on the average annual soil erosion rate estimated using erosion prediction technology such as the Revised Universal Soil Loss Equation (RUSLE). The terraces are spaced so that the estimated erosion rate on the inter-terrace interval is less than the soil loss tolerance value assigned to the site-specific soil. Soil loss tolerance values range from about 4 to 11 t ha^{-1} year^{-1}, based on the degree that erosion is judged to harm a particular soil.

Contouring, a practice where the direction of tillage and crop rows is on the contour around the hillslope, reduces rill and interrill erosion by about 50% on mild slopes. Contouring fails when runoff erosivity becomes too great. The failure location is known as the critical slope length, which is computed with RUSLE. Terrace spacing should not exceed the critical slope length.

Deposition and sediment delivery Gradient terraces also control erosion by deposition in the terrace channel. Deposition occurs when the sediment load reaching the channel from the inter-terrace area is greater than transport capacity of the flow in the channel. Transport capacity in the channel increases in direct proportion to increases in channel grade and channel flow rate. If transport capacity in the channel is greater than the incoming sediment load, no deposition occurs.

Sediment delivery ratio (SDR) is the ratio of the sediment load leaving the terrace channel to the incoming sediment load from the inter-terrace area. The sediment delivery ratio is about 0.2 and 0.4 for terraces on a 0.2% and 0.5% grade, respectively, in clean-tilled row crops. If terrace grade is greater than about 1.0%, no deposition occurs (as a rule of thumb).

Deposition also depends on sediment characteristics. Less very fine sediment is deposited than coarse sediment. Sediment eroded from most cropped soils is a mixture of primary particles and aggregates (conglomerates of primary particles). Silt-textured soils (e.g., 5% clay, 87% silt, 8% sand) are poorly aggregated and produce a high proportion of fine sediment particles. An SDR value for a silt soil might be 0.34. A silt loam soil (e.g., 15% clay, 65% silt, 20% sand) produces sediment having a balanced mixture of primary particles and aggregates. The SDR for that soil might be 0.26. A high sand soil (e.g., 4% clay, 6% silt, 90% sand) is eroded mainly as large, easily deposited primary particles. The SDR might be 0.12. In contrast, a high clay soil (e.g., 60% clay, 20% silt, 20% sand) erodes as sediment having a significant amount of very fine primary clay particles. The sediment also has a large fraction of large aggregates that are readily deposited. The easily deposited aggregates make up for the large fraction of primary clay in the sediment that is not readily deposited. The SDR for the clay soil is 0.24, about the same as for the silt loam soil. Clay is a bonding agent such that a large fraction of the sediment eroded from high clay soils is large aggregates.

Periodic maintenance is required to clear terrace channels of deposited sediment to maintain flow capacity. Otherwise, flow overtops the terrace and causes serious gully erosion that requires major

repairs. Deposited sediment removed from the terrace channel is spread over the field near the terrace to replace soil loss by erosion and soil displaced by the terrace construction. The ridges and channels are maintained so that farming over them can continue. Placing the deposited sediment on the terrace ridges makes them too high.

Only partial credit is given to deposition caused by terraces as soil saved relative to protecting the entire hillslope, because the deposition is on a relatively small area. Also, the deposition is remote from the location on the hillslope that produced the sediment. Erosion can degrade source areas even though much deposition occurs downslope. The credit given to deposition as soil saved decreases as terrace spacing increases because the fraction of the total area benefited by deposition decreases. No credit is given for deposition for terraces spacing greater than 100 m. A maximum credit of one half is given for spacing closer than 30 m. Very little credit is given for deposition that occurs in terraces located near the bottom of the hillslope.

Outlet channels Gradient terrace systems include outlet channels that collect flow discharged at the outlet of the terrace channels and convey the flow downslope. These channels must be well protected to prevent gully erosion. Flow in them is erosive because of a steep channel grade. Typically, these channels are grassed waterways. Rock can be used to line outlet channels where the flow is especially erosive. A major benefit of terrace systems is the elimination of ephemeral gully erosion, but this important benefit is not realized if stable outlet channels are not provided.

The capacity and stability of terrace and outlet channels are based on a design storm with a 24 h duration that occurs with a return period of 10 years, referred to as a 10 year-24 h storm. Runoff is computed based on the condition during the cropping cycle that produces the most runoff. Channel flow capacity is based on the condition that provides the greatest hydraulic resistance, e.g., unmowed grass in outlet channels and heavy residue cover in terrace channels. Channel stability is based on the condition when the channel is most vulnerable to concentrated flow erosion, e.g., freshly mowed grass in outlet channels and tilled seedbed in terrace channels.

Moisture conservation Gradient terraces are used in low precipitation regions to conserve soil moisture and increase crop production. A level terrace grade is used and the bottom of the terrace channel is widened to increase water capture and expand the benefited area (*See* **Water Harvesting**).

Parallel-Impoundment, Underground-Outlet Terraces

A terrace system suited for farming with wide equipment is one where the terraces are nearly parallel. Uniform-grade terraces are difficult to farm because the terraces are not parallel. Wide equipment does not work well when crossing terraces and tends to destroy them. Parallel terraces, illustrated in Figure 4,

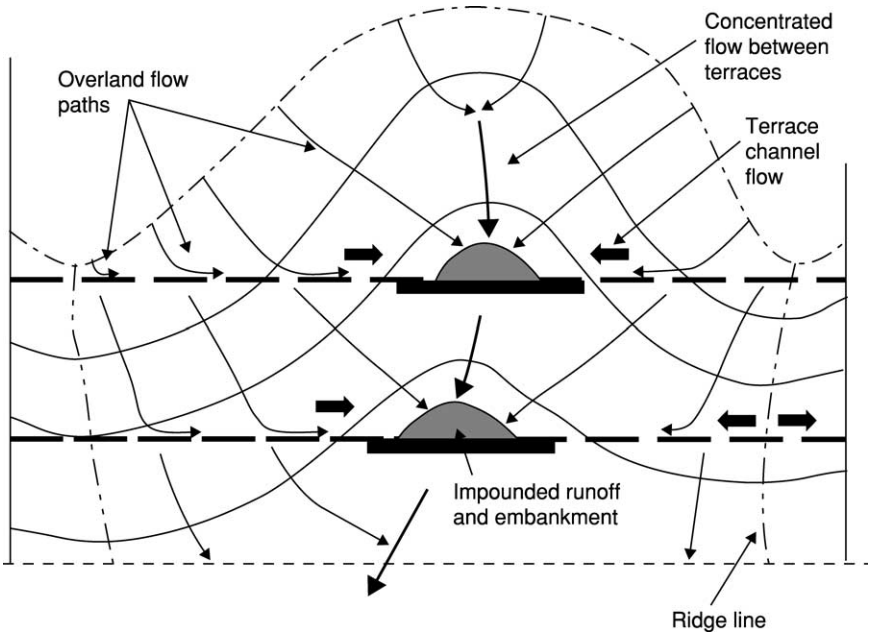


Figure 4 Overland flow and concentrated flow paths on a typical hillslope with parallel-impoundment, underground-outlet terraces.

require that grade varies along the terraces. Parallel-impoundment terraces are highly effective at controlling sediment yield and reducing ephemeral gully erosion, and are convenient to farm with large equipment. These terraces are less effective than gradient terraces at controlling rill and interrill erosion. Parallel-impoundment terrace systems include overland-flow interceptors, impoundments, and underground water-conveyance systems.

Overland-flow interceptors The overland-flow interceptors for parallel terraces are essentially the same as the embankments and channels used with gradient terraces. The slope lengths for the overland flow paths with a gradient terrace system are generally uniform along the terrace. However, the slope length of overland flow varies along parallel terraces. Slope lengths vary from essentially zero to the slope length of the natural landscape if the terraces are spaced far apart.

Channel grade varies along parallel terraces. The grade is zero where the channels cross a ridge and is steepest between the ridge and the impoundment in the draw. Deposition occurs uniformly along a uniform grade terrace channel, but deposition only occurs in parallel terrace channels where channel grade is sufficiently flat. Maximum channel grade is limited to ensure that no concentrated flow erosion occurs in the terrace channels. Terrace embankments and channels are relocated uphill to reduce terrace channel grade. The terrace spacing and hydraulic design procedures used for gradient terraces are also used to design parallel terraces.

Impoundments Impoundments located across the draws on the hillslope receive the flow from the overland-flow interceptors. These impoundments are sized to retain the runoff from a 10 year-24 h storm assumed to occur when runoff potential is greatest. The impoundment size is increased to store the sediment that will accumulate over the design life of the impoundment, usually 10 years. Maintenance can be performed periodically to remove and spread the deposited sediment over the hillslope to extend the life of the impoundments.

Runoff is retained in the impoundment to provide time for sediment to be deposited by settling to the bottom. Twenty-four hours is usually sufficiently long for most of the sediment to be deposited. The retention time is kept short to minimize inundation of crops and to speed drying so that farming operation can resume with minimal inconvenience.

Impoundment terraces very efficiently trap sediment. For example, an impoundment terrace will trap ~94% of the sediment eroded from a silt loam soil on the inter-terrace area for an SDR of 0.06. The SDR values for sediment eroded from sand, silt, and clay soils are 0.01, 0.07, and 0.14, respectively. The order of sediment delivery ratios among the soil textures differs from the SDR order for gradient terraces. This difference results from the sediment being continuously added along the gradient terrace channel, while the sediment is added to the impoundment at a point. Also, deposition mechanics in still water differ from deposition mechanics in channel flow.

The embankments for these impoundments are often so high that they cannot be crossed with farm equipment, which removes a portion of the field from crop production. The embankments are constructed with a steep backslope, illustrated in Figure 5, to minimize this loss. Dense grass is used on the backslope to prevent erosion. A bulldozer is used to construct the embankments.

Water-conveyance system Water is drained from the impoundments using a perforated riser pipe that projects above the floor of the impoundment. Water enters the riser and falls vertically into an underground pipe. Discharge rate is controlled using an orifice in the bottom of the riser to provide the desired retention time. Water flows from the riser into an underground pipe that runs up the draw and collects flow from all the terraces that cross that particular draw. A perforated underground pipe is sometimes used to provide subsurface drainage in the draw where seepage can collect.

Bench Terraces

Bench terraces reduce steep hillslopes to ones that are effectively flat. Bench terraces consist of a steep non-erodible section and a flat or nearly flat bench that is cultivated, as illustrated in Figure 6. The most impressive example of bench terraces is in China, where steep, highly erodible land is farmed. Reducing the effective landslope from 30% to 1% decreases erosion by ~99%.

Flat bench terraces are used in low rainfall areas to capture rainfall to increase soil moisture and crop production. Reduced runoff also reduces soil erosion (*See* **Water Harvesting**).

Flat benches with vertical backslopes Among bench terraces, flat benches with vertical backslopes provide the greatest land area for cultivation. However, constructing vertical walls for bench backslopes requires considerable resources. Vertical walls are easiest to construct if the wall is not high. However, a low wall results in a narrow flat bench area, which limits the cropping-management practices and size of equipment that can be used on the benches. A well-designed

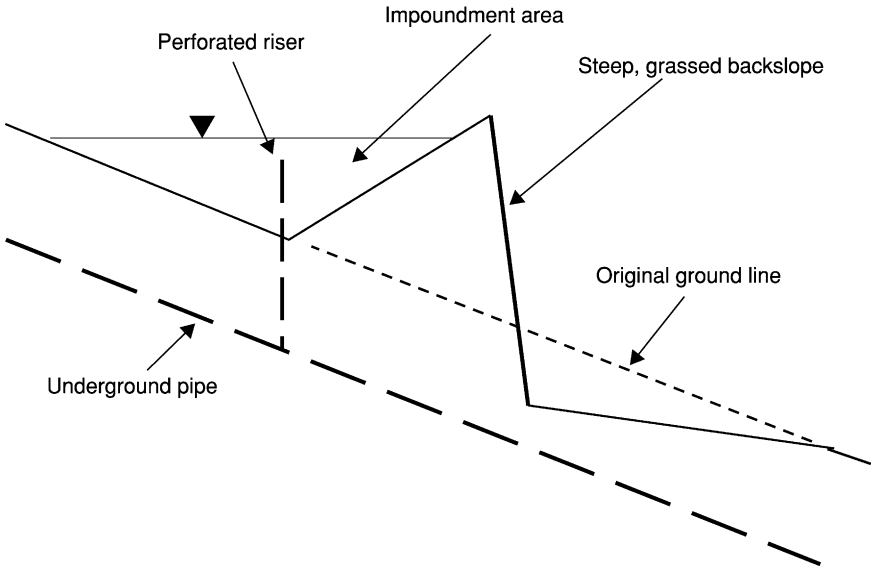


Figure 5 Impoundment embankment cross-section for an underground-outlet terrace.

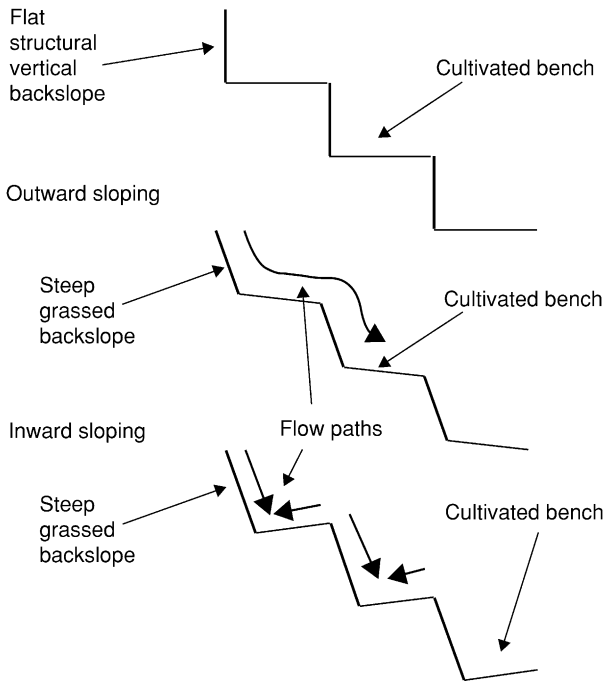


Figure 6 Types of bench terraces.

water-conveyance system is included with the terrace system to collect excess water and convey it downslope in stable channels.

Outward-sloping benches An alternative bench terrace system has a steep, grassed backslope (1:1 slope) and a relatively flat bench (less than 2%) wide enough (e.g., 3 m) to accommodate mechanically powered equipment. The bench can be outward or inward sloping. If the bench is outward sloping, as illustrated in Figure 6, overland flow originates at the top of the hillslope and cascades down the slope. Erosion is low because the grass essentially prevents erosion on the steep backslope, and erosion is low on the nearly flat bench.

While bench terrace systems theoretically perform well with water cascading from one terrace to the next, runoff must be carefully controlled to avoid concentrated flow and gully erosion. Maintaining bench terraces with the front edge perfectly on a level grade so that runoff flows uniformly over the bench edge is difficult. Also, the grass cover on the steep backslope must be well maintained to prevent areas where gully erosion can begin. Runoff should be collected at regular intervals downslope and brought to nonerodible channels that convey the runoff downslope.

Inward-sloping benches Inward-sloping benches, illustrated in Figure 6, eliminate the cascading flow down the slope. Each bench and its backslope are individual watersheds. The outer edge of the bench is left higher than the back edge so that runoff flows back toward the hillslope. The only runoff in each bench watershed is from the rainfall that falls directly on that watershed. The benches can be put on a slight grade around the hillslope so that runoff flows around the hillslope at the base of the backslopes to nonerodible collection channels that convey the runoff downslope.

Inward-sloping bench terraces have less rill and interrill erosion than do outward-sloping bench terraces. However, the difference is not great because at

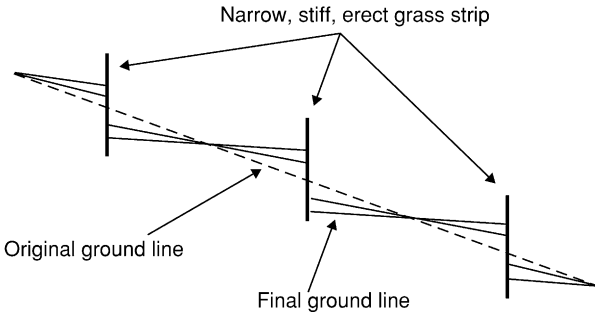


Figure 7 Narrow, stiff grass strips enable erosion and deposition to create bench terraces.

1% slope steepness and less, most of the erosion is caused by interrill erosion, which is not affected by slope length. For example, reducing slope length from 50 m to 3 m results in only a 20% reduction in erosion for a 1% steep slope.

Naturally formed benches Soil must be mechanically moved to construct the terraces described above. Permanent, narrow grass strips can be used to let nature build bench terraces. Narrow, approximately 0.5 m wide, strips of dense erect, stiff grass are planted very carefully on the contour. These dense strips slow the overland flow to cause deposition in backwater ponded on the upper side of the grass strip. Over time, the hillslope slowly evolves as illustrated in Figure 7. Reducing steepness from 10% to 2% for a 10 m hillslope segment by rotating the segment about its midpoint requires 0.35 m of erosion at the segment's upper end and 0.35 m of deposition at the segment's lower end. Approximately 100 years is required for this adjustment assuming that all sediment eroded on the segment's upper portion is deposited on the segment's lower portion, an average annual erosion rate of 50 t ha^{-1} $year^{-1}$, and a specific soil bulk density of 1.3. The process is often slow. However, an immediate benefit is that the grass spreads the runoff so that rill erosion is reduced for a few meters downslope from the strip.

Ridge Ridges and beds on the contour around the hillslope are actually a set of small terraces where the slope length is half the spacing of the ridges. The slope length of the overland flow is the distance from the top of the ridge to the middle of the furrow between the ridges. Almost all of the erosion on the ridges is interrill erosion because of short slope lengths.

The furrows between the ridges act like terrace channels. A flat furrow grade, less than 2% for clean-tilled row crops, deposits much of the sediment eroded on the ridges. Deposition in the furrows is given full credit, as soil saved protects the entire hillslope because the deposition is very near to the source of the sediment.

To function as terraces, ridges must be sufficiently high to avoid overtopping and be carefully constructed and maintained to ensure that they perform well. Erosion from failed ridges can be greater than if ridges had not been used. Overtopping in a local area is like a dam break. Failures cascade downslope, resulting in serious ephemeral gully erosion. Maintaining a sufficient height along the ridge and avoiding adverse furrow grades that pond water prevent overtopping in localized areas.

Ridges are most effective when arranged perfectly on the contouring. Maximum moisture is retained for crop utilization. When ridges overtop, the runoff flows over the ridges uniformly along the ridges if the ridges are perfectly level. The flow depth will be very shallow and much less erosive than if the runoff concentrates at low places and breaks over in a concentrated flow. Ridges perfectly on the contour reduce ephemeral gully erosion. When ridges are perfectly on the contour, the head of ephemeral gullies is further down the hillslope than if ridges and furrows are on a grade. Ridges and furrows on a slight grade concentrate overland flow far up the hillslope, which extends ephemeral gully erosion further up the hillslope than with ridges perfectly on the contour, as illustrated in Figures 8 and 9.

Straight-ridge alignment facilitates farming with large equipment. However, the furrow grade along straightened ridges is non-uniform, just as with parallel terraces. The grade must be sufficiently flat so that rill erosion does not occur in the furrows on areas adjacent to concentrated-flow areas. Runoff collects in these natural waterways and flows downslope. These concentrated-flow areas must be protected with a grass or another stable lining to prevent serious ephemeral gully erosion. If these concentrated-flow areas are not protected, ephemeral gully erosion can be greater with ridging slightly off the contour than with ridges directly up and downhill. However, wide grass waterways take land out of production, reduce farming convenience, and require significant maintenance.

Summary

Terraces are important structures used to safely convey runoff, control rill erosion, and prevent ephemeral gully erosion, especially on steep hillslopes with long slope lengths. Terraces are also used to conserve soil moisture and thus to increase crop production in low rainfall areas. Terraces intercept overland flow and redirect it around the hillslope where protected channels convey the runoff downslope.

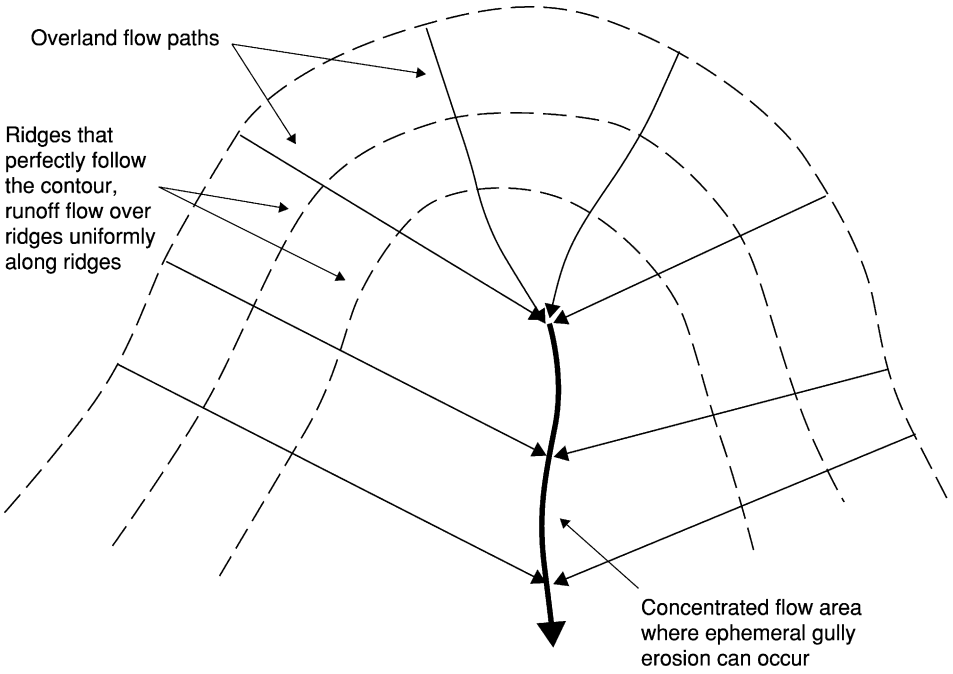


Figure 8 Overland flow paths and location of the head of ephemeral gullies when ridges are perfectly on the contour.

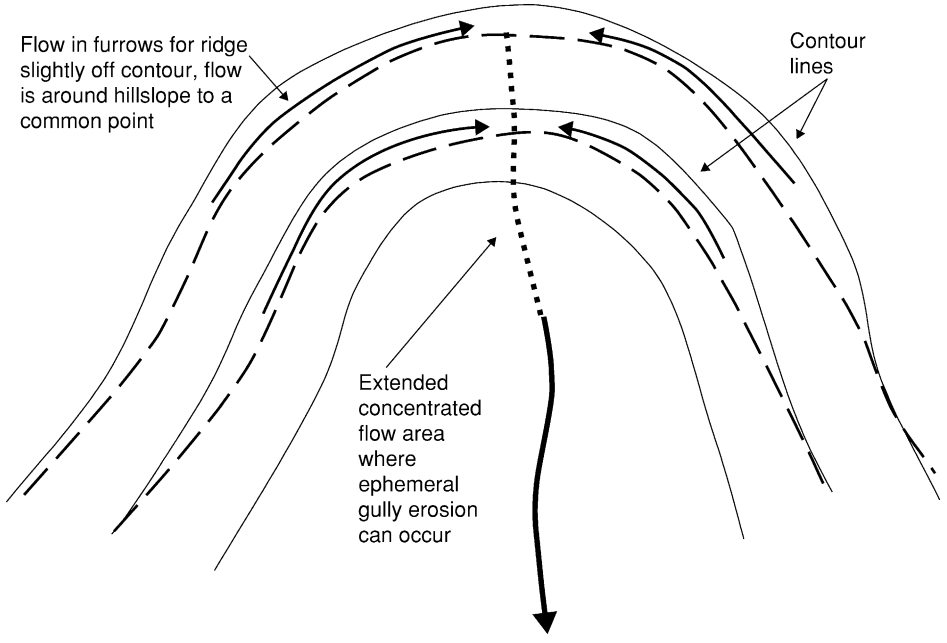


Figure 9 Overland flow paths and location of the head of ephemeral gullies when ridges are on a slight grade.

Terraces are used in a wide range of agroecosystems ranging from small plots farmed with human and animal power to large fields farmed with large machines. Terraces are best suited to deep soils where soil can be moved to create benches and ridges without seriously affecting the soil productivity. Terraces require significant resources of capital or labor to construct and must be maintained. Terraces on steep hillslopes in China graphically illustrate how terraces can be used to make maximum use of the landscape for crop production.

See also: **Aggregation:** Physical Aspects; **Conservation Tillage**; **Crop-Residue Management**;

Drainage, Surface and Subsurface; **Erosion:** Water-Induced; **Overland Flow**; **Precipitation, Watershed Analysis**; **Water Harvesting**; **Watershed Management**

Further Reading

Beasley RP, Gregory JM, and McCarty TR (1984) *Erosion and Sediment Control*, 2nd edn. Ames, Iowa: Iowa State University Press.

Foster GR and Highfill R (1983) Effect of terraces on soil loss: USLE P factor values for terraces. *Journal of Soil and Water Conservation* 38: 48–51.

Haan CT, Barfield BJ, and Hayes JC (eds) (1994) Hydrologic frequency analysis and rainfall-runoff estimation in stormwater computations. In: *Design Hydrology and Sedimentation for Small Catchments*, pp. 5–103. San Diego, California: Academic Press.

Hurni HA (1981) Nomograph for design of labour-intensive soil conservation in rain-fed cultivations. In: Morgan RPC (ed.) *Soil Conservation: Problems and Prospects*, pp. 185–210. New York: John Wiley.

Morgan RPC (1986) *Soil Erosion and Conservation*. New York: John Wiley.

Toy TE, Foster GR, and Renard KG (2002) *Soil Erosion: Processes, Prediction, Measurement, and Control*. New York: John Wiley.

Troeh FR, Hobbs JA, and Donahue RL (eds) (1991) Conservation structures. In: *Soil and Water Conservation*, 2nd edn, pp. 247–275. Engle Wood Cliffs, New Jersey: Prentice-Hall.

US Department of Agriculture-Natural Resources Conservation Service (2003) Conservation practice standards. In: *Field Office Technical Guide*. Available on Internet at www.usda.gov/technical/efotg/, Washington, DC.

Zhengsan F, Piehua Z, Qiande L, Baihe L, Letian R, and Hanxiong Z (1981) Terraces in the Loess Plateau of China. In: Morgan RPC (ed.) *Soil Conservation: Problems and Prospects*, pp. 481–513. New York: John Wiley.

Zingg AW (1940) Degree and length of land slope as it affects soil loss in runoff. *Agricultural Engineering* 2: 59–64.

TESTING OF SOILS

A P Mallarino, Iowa State University, Ames, IA, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Soil testing is an important diagnostic tool for determining the nutrient needs of plants and for environmental assessments. Some soils are inherently deficient in plant nutrients. Other soils had sufficient levels of nutrients in the past, but removal with crop harvest has depleted the reserves. Thus, soil testing is widely accepted and used in most advanced crop-production areas of the world to determine fertilization needs for crops. Soil testing can also be used to identify application rates of waste materials containing nutrients or other elements that could harm the environment. Waste materials such as animal manures and industry by-products may provide various plant nutrients. However, high application rates to soils designed to dispose of the material at a low cost may result in nutrient loads that are harmful to plant, animal, or human health. Nutrient management regulations are being developed to address land application of waste materials. Soil testing is required in many regulations and management guidelines to assess environmentally harmful levels of certain elements and to determine limits to application rates.

Soils are tested routinely for the primary nutrients phosphorus (P), potassium (K), and nitrogen (N). In some regions, soils are also routinely tested for other primary nutrients such as calcium (Ca), magnesium (Mg), and sulfur (S), and for other nutrients required in very small amounts by crops such as boron (B), copper (Cu), iron (Fe), manganese (Mn), molybdenum (Mo), and zinc (Zn). Soils receiving waste materials are also tested for elements such as arsenic (As), cadmium (Cd), nickel (Ni), lead (Pb), mercury (Hg), and selenium (Se) among others.

Two of the primary plant nutrients, N and P, may have harmful effects on the environment when applied to soils in excessive amounts. Excessive N and P applications to agricultural fields and ineffective nutrient, soil, and water conservation practices are increasing nutrient pollution in many regions of the world. However, the basic concepts of soil testing also apply to other elements. Phosphorus is included with the group of elements with relatively low mobility in soils (Ca, Cd, Cu, Fe, Mg, Mn, Mo, Ni, Pb, and Zn among others). N, especially in the nitrate form, is included with the group of elements with greater mobility (which include B, Cl, S, and others). Important concepts include the meaning of a soil-test value, soil-testing quality, use of soil testing to determine economically optimum nutrient application rates, and use of soil testing for environmental assessments.

Soil-Test Extractants and Methodologies

Phosphorus Soil Tests

Soil tests for P and other elements have been developed based on an understanding of the chemical forms in which the elements exist in the soil and empirical work. In most soils, P is primarily associated with aluminum (Al), Fe, and Ca. P precipitation and/or adsorption to Fe and Al oxides generally predominates in acid (pH < 7.0) to neutral soils, whereas reactions with Ca predominate in soils with significant amounts of calcium carbonate. The degree and strength to which P is bound in soils are largely determined by the amount and types of Fe and Ca compounds present and by other soil properties such as pH, organic matter, clay mineralogy, and the amount of P currently present in the soil. The influence of Fe increases and that of Ca decreases as rainfall, temperature, and weathering increase. These factors are also important in determining the plant availability and solubility of other immobile elements.

Phosphorus soil-test methods for agronomic use employ dilute strong acids, dilute weak acids, complexing ions, and/or buffered alkaline solutions. Most tests have been developed to reflect the soil properties related to P sorption that predominate in a region. For example, the Bray P-1 test was developed for use in the acid-to-neutral soils of the north-central region of the USA; the Olsen (or sodium bicarbonate) test was developed primarily for use on calcareous soils of the same region; and the Mehlich-1 test was developed for the low cation-exchange capacity and highly weathered soils common to southeastern regions. The Mehlich-3 P extractant was developed more recently. It has been adopted because it is suitable for a wider range of soil properties than other P tests and also for other nutrients (such as K, Ca, Mg, Cu, Fe, and Zn).

Increasing concerns about nonpoint-source P pollution from agriculture has prompted questions about the suitability of agronomic soil P tests for environmental purposes. Several soil tests have been proposed as environmental soil P tests because they may provide better estimates of dissolved P delivery to water resources. Two of these tests are being evaluated extensively. One test based on FeO-impregnated paper uses a sink approach to extract soil-bound P that is most likely available to aquatic organisms. The other test is based on weaker desorption reactions and uses either deionized water or a diluted calcium-chloride solution to extract soil P. These tests are environmentally sound for laboratory use because few hazardous chemicals are needed.

Although the results of the different soil P tests are often highly correlated for soils with similar properties, the actual quantities extracted can differ greatly. For example, the Mehlich-3 test extracts 1.5 to 2 times the amount of P as the Mehlich-1 and Olsen tests. The FeO-impregnated paper environmental P test usually measures 4–10 times the amount of P measured by the water extractant. These observations highlight the importance of understanding the meaning of a soil-test value.

Nitrogen Soil Tests

The cycling of N in soils is different from other nutrients and requires a different soil-testing approach. More than 97% of N in soils is present in organic forms and is converted to inorganic forms available to plants mostly through microbial activity. The rate of N release from organic to inorganic forms is difficult to predict because it depends on temperature, moisture, aeration, organic matter type, soil pH, and many other factors. The predominant inorganic end-result of mineralization is nitrate-N. This N form is highly mobile in soils and is subject to loss by leaching and to gaseous losses through denitrification. Because of these factors, N fertilizer recommendations for crops are frequently made on the basis of crop yield potential coupled with an N-credit system and less frequently on the basis of soil testing. The N-credit system is based on empirical information of the effects of previous crops, several soil properties, and climate on crop response to N fertilization.

Soil testing for nitrate-N has been used for a long time in low-rainfall regions and is becoming more common in humid areas of the USA, mainly for maize production. In low-rainfall areas, where nitrate leaching potential is minimal, the amount of nitrate-N in the soil profile before planting crops can be credited directly against the N needs of the crop. In many humid regions, measurements of soil profile nitrate or ammonium before planting have been unreliable to predict N needs of maize. However, nitrate-N testing early in the growing season has been shown to provide an index of the soil N-supplying capability for maize. Soil samples are collected deeper than for other nutrients (usually from a 0- to 30-cm depth) after planting the crop and before plants reach a height of 15–30 cm.

The nitrate-N soil-test interpretation is usually complemented by consideration of the previous crop, rainfall during the 2- to 4-week period before collecting soil samples, and soil properties that influence nitrate leaching. As work on this type of testing for nitrate in humid regions continues, it is becoming a useful diagnostic tool to decrease N fertilizer rates and reduce potential N pollution of water resources. As with all soil tests, local or

regional field calibrations are being conducted to determine critical concentration ranges and fertilizer recommendations.

Calibration of Soil Tests for Crop Production

Soil tests seldom extract the total amount of nutrients or elements in a soil sample. Soil tests have been developed to measure a fraction of the total soil nutrient concentration that correlates with plant growth. Interpreting a soil-test value requires an understanding of the impacts on test results of the extractant used, method of soil sampling, sample handling, and the intended use for the result. The amount of nutrient measured by various soil tests can vary widely depending on the extractant used. Important extractant properties include the concentration of the chemical compounds and the reaction time with the soil. The method used to measure the nutrient after extraction may be important for some nutrients. For example, colorimetric methods usually measure only orthophosphate P, while other methods may also measure other P forms. Also, the same soil test may measure widely different amounts of nutrients in soils with contrastingly different chemical and (or) mineralogical properties. Extractants used and interpretations of results often vary depending on soil properties and the purpose of soil testing (for example, measurement of total, soluble, or plant-available concentrations).

A soil-test method useful for predicting crop response to fertilization should produce values that are well correlated with plant nutrient uptake and growth. Greenhouse and field research is conducted to determine which soil-test extractant is best suited for a given combination of soil, crop, and growing conditions. Accurate interpretations of soil-test results and appropriate fertilizer recommendations require that the relationship between the amount of a nutrient measured by a given soil test and the crop response to the added nutrient must be known. The process of determining the probability of crop response at a given soil-test value is known as soil-test calibration and must be determined by field experimentation. The calibration procedure usually involves growing a crop or various crops in soils representative of the region on which the test will be used and the application of various fertilization rates. The soils should have a broad range of soil-test values that include deficient, optimum, and above-optimum values.

The crop yield response can be expressed as an absolute value or as a value relative to the yield achieved without nutrient addition. As an example, Figure 1 shows the relationships between the relative or absolute yield increase of maize and the amount of P extracted by a soil test. Although the specific shape of the relationship between soil-test values and crop growth or yield can differ, the general response is fairly consistent. At low soil-test values, crop yield is limited by nutrient deficiency. As soil-test values increase, yield increases at a reduced rate until a maximum yield value. At higher levels, there is no longer a relationship between the soil-test values and yield. The maximum yield value can remain approximately constant for a wide range of soil-test values or can decrease at excessively high nutrient levels.

The soil-test value at which crop growth or yield reaches a maximum is called the soil-test critical concentration. This is the soil-test value that best separates soils where a positive crop yield response to an added nutrient is likely from those soils where a response is not likely. Various mathematical models can calculate different critical concentration values and some models are asymptotic to a maximum. There is no reasonable agronomic consideration that justifies a single critical concentration. Moreover, economic

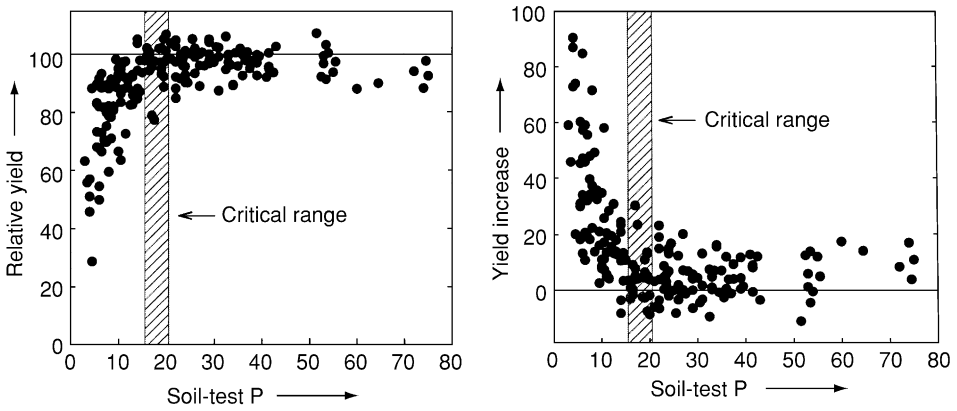


Figure 1 Example of the relationships between relative or absolute crop yield response and soil-test P values used for soil-test calibration.

considerations suggest that maximum economic yield perhaps should be the basis for determining critical concentrations instead of a physical maximum. Thus, critical concentration ranges are usually determined by using a combination of models and a variety of both agronomic and economic considerations. The soil-test response curve is often used to divide soil-test levels into several categories such as very low, low, optimum, high, or excessive.

A thorough understanding of crop response to nutrients and factors such as sampling date (season of the year), sampling method (mainly sampling depth), method of nutrient application (for example, band or broadcast), and application timing is needed to interpret soil-test results correctly. Also, other factors can influence crop growth (climatic factors, plant population, levels of other nutrients in the soil, and crop cultivars, among others). These factors may influence crop growth without affecting the amount of nutrient needed to produce a certain yield level or they may interact with soil nutrient levels.

Climatic factors are especially important for nutrients that are highly mobile in the soil, such as nitrate-N, B, Cl, and S. High nutrient mobility results from limited retention of nutrient chemical forms by soil organic or mineral constituents. High nutrient mobility results in more difficult soil-test calibrations and introduces more uncertainty when interpreting soil-test results. Levels of these nutrients in soil change markedly during the year as a result of changes in factors such as soil temperature and rainfall. Soil samples for relatively immobile nutrients (such as P, K, Ca, Mg, Cu, Zn) are often collected from the top 15–20 cm of soil, from where most of the nutrient uptake takes place. Other nutrient measurements, such as subsoil testing, are sometimes useful for the more immobile nutrients such as P or K but are more useful for mobile nutrients.

Calibration of Soil Tests for Environmental Purposes

The general concepts for soil-test calibration described for agronomic soil tests also apply to soil tests used for environmental purposes. With agronomic tests, the most important consideration is related to the meaning of a soil-test value for crop growth or yield. With environmental tests, the most important consideration relates to the meaning of a soil-test value for predicting environmental impacts. For several reasons, calibration and interpretation of soil tests used to predict potential environmental impacts are more complex than for agronomic uses. The most important reason is that a certain concentration of an element in the soil may have different impacts on the environment depending on the amounts absorbed by plants, amount retained by the soil, and potential delivery to environmentally critical areas. For example, soils with high clay content have a greater sorption capacity for P (and for other immobile elements) than coarse-textured soils, and a lesser amount of P is delivered as the distance from the source to channeled water flow increases. Establishment of nutrient loads that can result in unacceptable water-quality degradation will depend on factors such as the distance to a sensitive water body, the effectiveness of transport pathways for water and either total or dissolved nutrient fractions, and socioeconomic factors involving land or water use. A single, critical soil-test concentration range is unlikely to be found for nutrient impacts on the environment.

Not all areas of a field or landscape contribute equally to nutrient loss. For loss to occur, both a source of the nutrient and a mechanism for transporting it to water bodies or groundwater are needed. A key concept in effective management of nutrient pollution is to focus on the critical areas in which these two factors maximize the potential for loss. The P Index is an example of an assessment tool that considers soil-test values with other site-specific factors (such as hydrology, nutrient management, water-body proximity, and sensitivity to nutrient inputs) to determine the potential environmental risk of P levels in a field or parts of fields. The P Index includes source factors (soil-test values, as well as fertilizer and manure application rates, methods, and timing) and transport factors (such as soil erosion, surface runoff, subsurface drainage, and distance to streams or water bodies). Indexes are also being developed for N.

Agronomic soil-test methods designed to assess plant-available nutrients are sometimes used for environmental purposes. An effective environmental soil test should measure the nutrient forms most important to eutrophication or other negative environmental impacts. Eutrophication is a process whereby water bodies such as lakes, estuaries, or slow-moving streams receive excess nutrients that stimulate excessive aquatic plant or algae growth. This enhanced growth reduces dissolved oxygen in the water when dead plant or algae material decomposes, which causes ecosystem imbalances and may promote growth of toxic bacteria and unpleasant odor. Current research indicates that there are usually similar relationships between either agronomic or environmental soil P tests and dissolved P in surface runoff or subsurface drainage, except immediately after P additions. However, extending the concept of agronomic critical soil-test ranges to declare that nutrient concentrations above those needed for

optimum crop production will result in significant environmental impacts is not appropriate, because this ignores important aspects of nonpoint-source pollution. If soil tests are to be interpreted to predict the probability of nutrient pollution, calibrations relating soil-test results to specific measurements of environmental response (such as the concentration or load of P in surface runoff, for example) are necessary.

Quality of Soil Testing

Laboratory Quality Control

The quality of a laboratory soil-test result can be assessed on the basis of bias and precision. Bias refers to the deviation of the analytical result from the true value and measures accuracy of the result. Precision refers to the reproducibility of a given test value. A soil-testing laboratory could have good precision for a specific procedure but have bias in its results. A primary means of evaluating soil-test quality procedures as well as laboratory performance is through comparisons of tests performed on well-mixed subsamples of the same soil at different laboratories. In North America, a voluntary proficiency-testing program involving state and private laboratories was launched in the early 2000s to improve the quality of soil testing for agronomic purposes (the North American Proficiency Testing Program, NAPT). Results from this program indicate that, for example, the average variability of P tests is approximately 10–15% and that 65–70% of participating laboratories produced results within this level of variability. Variability is lower within any given laboratory. Data from the NAPT program indicate that precision levels for the P tests discussed are approximately 5–10% within each laboratory.

Soil Sampling for Soil Testing

For a soil-testing program to be effective, besides proper soil-test calibration and laboratory quality control, soil samples should be collected in a cost-effective manner and should accurately represent the nutrient levels in the area of interest. Sampling is a critical component of the soil-testing process because it usually represents the largest single source of error in soil testing. Many factors varying both spatially and temporally influence nutrient concentrations in soils. Sampling protocols must account for the great diversity in magnitude, structure, and spatial scale of nutrient variability present in agricultural fields and other ecosystems.

Soil nutrient variability within a crop field may be due to soil formation or management factors. Factors such as landscape position and soil parent material can cause great changes in soil texture, organic matter, drainage, and other properties. These properties affect nutrient levels directly through their influence on the amount of plant-available nutrient or indirectly though crop yield potential and, thus, the nutrient removal by crops. Variability caused by long-term history of management and land-use practices overlays that associated with soil-formation factors. The cumulative effects of nonuniform manure or fertilizer application are sources of potentially high soil-test variability. Proximity to livestock confinement areas, feed storage areas, and field boundaries are additional examples of historical factors that cause great variability in many fields. Small-scale variability usually predominates in fields with long histories of cropping and fertilizer or manure applications, especially when nutrients are applied using band methods. The challenge in these situations is to determine effective methods to delineate sampling areas within a field and the number of cores or borings needed for each composite sample to account for small-scale variability appropriately.

A variety of systematic and zone sampling approaches have been developed in different regions. The development of accurate and reliable global positioning technology, affordable geographic information systems, and variable-rate application equipment has led to widespread use of site-specific soil sampling approaches in North America. These sampling approaches are typically used to generate a soil fertility map to serve as an input to computer-controlled equipment for applying varying rates of one or more materials. One such approach is zone sampling, by which field subregions with more homogeneous properties than the field as a whole are delineated. Landscape position, soil color, soil mapping unit, and crop growth differences are examples of factors often used to help define management zones. Another site-specific approach involves systematic grid sampling, where soil-test patterns in a field are determined by means of a dense, systematic sample collection. A grid size of approximately 1 ha is most common today in the USA. Small-scale variability of P and K is so extreme in some fields that accurate within-field soil fertility mapping is impossible. Much uncertainty still exists regarding how best to perform site-specific soil sampling and generate accurate soil fertility maps.

Soil-test values may also change markedly with depth. This results from a combination of soil-forming and management factors and from the differential mobility of nutrients in soils. The elements with lower mobility tend to accumulate near the application point (often the most shallow soil layers), because they are more strongly retained by soil constituents than the more mobile nutrients. The tillage

system and the application method greatly influence vertical nutrient stratification. For example, the significant stratification of both P and K in no-till soils is well known, but deep nutrient application methods can reduce their concentration near the soil surface. Thus, the proper depth for soil sampling and its consistency are important considerations. Soil-test variability can also be the result of sampling-depth variability. Soil samples should be collected at the same depth used in the calibration research that serves as the quantitative basis for the soil-test interpretation and nutrient recommendations. Typically, this is a depth of 15–20 cm for the relatively immobile nutrients (such as P, K, Ca, Mg, Cu, Fe, Mn, Zn) and 30–60 cm for the more mobile nutrients (such as nitrate-N). However, the optimum soil-sampling depth varies with the objective of the soil testing. For example, soil sampling for predicting element loss with surface runoff may need to be shallower than for predicting uptake by plants or loss with subsurface drainage.

Use of Soil Tests to Determine Nutrient Loading Rates

Soil testing is becoming an established practice to determine whether environmentally unacceptable concentrations of plant nutrients or other elements are present in soils as a consequence of fertilization and waste applications. The term 'loading rate' implies that when excessive amounts of an element are applied to soils, it could be transported to environmentally critical areas and that the result may be harmful. Soils can be overloaded with a variety of nutrients, nonessential heavy metals, organic chemicals, and pathogenic organisms; and overloaded soils can affect the environment in various ways. For farmers and nutrient-management planners to make confident decisions about nutrient loading, soil tests must be able to predict loading rates and environmentally critical levels over a range of crops, soils, and environmental conditions.

In the USA and many European countries, regulatory initiatives are being established to mitigate water-quality problems associated with N and P water pollution. Other environmental efforts to which soil testing contributes include preventing nutrient toxicity to plants and protecting the food chain from accumulations of nutrients or elements harmful to animal and human health. These environmental problems usually occur at soil-test levels in excess of those required for optimal plant growth or as a result of clearly undesirable events (such as nutrient transport in floods or attached to eroded soil). Regarding phytotoxicity and feed or forage quality, soil tests have the potential to predict negative responses of plants to soils overloaded with a nutrient or nonessential elements.

Measurements and protocols for calibrating soil tests for environmental purposes are complex and are not well established. Soils can contribute to nonpoint-source pollution of water resources (mainly through erosion and surface runoff) even when soil-test levels are optimum for crop production. Nonincorporated surface applications of nutrient sources may result in high nutrient losses not correlated with soil-test levels. In this instance, nutrient transport may occur before any significant interaction between the nutrient source and the soil. A soil test alone does not permit accurate characterization of the risk of nutrient loss from soil to water. Estimates of the amount of nutrient directly lost from surface-applied materials are also needed. What happens between the field's edge and the surface water body, including the possibility for long-term retention of nutrients in buffer zones and any channel processes resulting in retention or release of nutrients, is a very important component of predicting loading rates. Complex diagnostic tools such as N and P risk indexes should be developed and tested for their efficacy to reduce nutrient pollution.

Summary

Soil testing is, and should be, an important component of more comprehensive environmental diagnostic tools. Researchers are attempting to define measurable parameters useful to improve estimates of the relationship between soil-test concentrations and loading rates to water resources. An example of such work is the ongoing international effort to develop better soil-testing procedures (including both methods of analysis and their interpretations) for N and P to identify watershed or field areas likely to contribute significant amounts of N and P to surface water and groundwater through erosion, surface runoff, or leaching. Growing evidence from field research demonstrates a strong relationship between soil-test values and concentrations of elements in surface runoff and subsurface drainage. However, local or regional calibrations are needed, because the relationships vary greatly across soil-test methods and soil properties. Moreover, the effective impact of a certain soil-test value on environmentally sensitive areas is strongly affected by the site hydrology and various transport mechanisms. Thus, if loading rate decisions regarding fertilizers or waste materials are made solely on the basis of soil-test concentrations, erroneous decisions will probably be made regarding potential element losses from fields. However, soil testing plays an important role

in optimizing crop production and protecting the environment.

See also: **Fertility**; **Fertilizers and Fertilization**; **Quality of Soil**

Further Reading

Carter MR (ed.) (1993) *Soil Sampling and Methods for Soil Analysis*. Boca Raton, FL: CRC Press.

Kamprath EG, Beegle DB, Fixen PE *et al.* (2000) *Relevance of Soil Testing to Agriculture and the Environment*. CAST Issue Paper 15. Ames, IA: Council for Agricultural Science and Technology.

Power JF and Dick W (eds) (2000) *Beneficial Uses of Agricultural, Municipal, and Industrial By-Products*. Madison, WI: Soil Science Society of America.

Steele K (ed.) (1995) *Animal Waste and the Land–Water Interface*. Boca Raton, FL: Lewis Publishers.

Tiessen H (ed.) (1995) *Phosphorus in the Global Environment: Transfers, Cycles, and Management*. Scientific Committee on Problems of the Environment (SCOPE), International Council of Scientific Unions (ICSU), United Nations Environment Programme (UNEP). Chichester, UK: John Wiley.

Westerman RL (ed.) (1990) *Soil Testing and Plant Analysis*, 3rd edn. Madison, WI: Soil Science Society of America.

TEXTURE

G W Gee, Pacific Northwest National Laboratory, Richland, WA, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Soil consists of an assemblage of particles that differ widely in size and shape. Some particles are coarse enough to be seen easily with the naked eye, while others are so tiny they can only be seen with powerful electron microscopes. The term 'soil texture' is an expression of the dominant particle sizes found in soil and has both qualitative and quantitative connotations.

Qualitatively, 'soil texture' refers to the feel of soil material. A soil when rubbed by hand can have a feeling that ranges from coarse and gritty to fine and smooth. Experienced soil classifiers can press soil into their hands and tell its texture, coarse or fine, or gradations in between. The expressions 'light soil' and 'heavy soil' are frequently used to characterize the general physical behavior of different soils. A coarse-grained, sandy soil tends to be loose, well-aerated, and easy to cultivate, so it is called a 'light soil.' A fine-textured soil tends to absorb water easily and becomes plastic and sticky when wet, and tight, compact, and cohesive when dry, so it is called a 'heavy soil.' These terms are misleading, since coarse-textured soils are generally denser, having lower porosity, than fine-textured soils, and thus are heavier rather than lighter per unit mass. Table 1 shows the range of densities and porosities that can be found in soils of various textures.

Quantitatively, soil texture refers to the relative proportions of various sizes of soil particles, sometimes called 'size separates.'

Soil Particle Size: Measurements and Classification

The size of soil particles can vary over 6 orders of magnitude, ranging from large stones and rocks (greater than 0.25 m in size) to tiny, submicron clays (less than 0.0001 mm). The shape of soil particles also varies widely. Particle shape is often an indication of the mineral assemblage and the degree of soil weathering that has taken place. For beach sand, composed of quartz, particles are uniformly rounded and almost spherical. In contrast, most clay materials are far from spherical, many being platelets, looking more like playing cards (e.g., kaolinite) or tubes (e.g., halloysite) than marbles. In spite of the variation in shape, soil particles are sized according to their 'effective' particle diameter. Because of the huge size range, no single method can be used to measure particle diameters. Measurements are typically made with tape or ruler for large particles (boulders or cobbles) of more than a few centimeters in diameter, with sieves for particles as small as approximately

Table 1 Range of values found in bulk densities and associated properties for soils of various textures

Description	*Bulk density*[a] *($g\,cm^{-3}$)*	*Porosity*[b] *($cm\,cm^{-3}$)*	*Void ratio*[c] *($cm\,cm^{-3}$)*
Clay	1.20	0.54	1.17
Loam	1.40	0.49	0.96
Sand	1.60	0.39	0.64
Sandy loam[d]	1.90	0.28	0.39
Sandstone	2.12	0.20	0.25

[a]Bulk density, mass of solids per total volume.
[b]Porosity, total volume of pores (liquid and gas phase) per total volume.
[c]Void ratio, total volume of pores (liquid and gas phase) per solid volume.
[d]Compacted subsoil.

0.05 mm and with a variety of sedimentation methods (e.g., pipet, hydrometer, laser light-scattering) for particles less than 0.05 mm.

Size Measurements

Size measurements depend on separating the soil into primary particles. Methods for soil analysis consist of disaggregating by washing, vibrating, or otherwise dispersing the soil into single units. Typically the larger particles are measured directly by tape, ruler, or calipers. Particles from 100 mm to 0.05 mm are sieved. Materials less than 2 mm are generally washed through a nest of screens for a given period of time after treating with some kind of soil dispersant such as sodium hexametaphosphate, which swamps the clay surfaces with sodium, causing dispersion of the clay materials and accelerating disaggregation. For tropical soils or soils with high organic, iron, or aluminum contents, special treatments are deployed to remove these binding agents and release the primary particles. Soils of volcanic ash origin are notorious for aggregation, and pretreatment by chemical or mechanical dispersion can significantly affect the amount of clay produced. In one test, researchers have demonstrated that the clay (less than 0.002 mm) content of a volcanic ash soil varies from 1% to 56% depending on the amount of pretreatment imposed on the sample. Some tropical soils consist largely of clay-size particles yet behave much like sands, in terms of their water-storage and related physical properties, because of the high degree of natural aggregation.

Fine Soil Measurements

Fine soil measurements rely on sedimentation of a soil suspension. The sample is stirred in water and dispersing agent, then the suspension is allowed to settle. Sampling is done at preselected times at a fixed depth in the suspension using either a pipet or hydrometer. Pipet techniques take a physical sample after a given settlement period, while hydrometer methods rely on estimates of effective densities of the settling particles. Knowing settlement depths allows for a straightforward calculation of the effective particle diameter. The Stokes law for viscous drag on a settling spherical object is combined with buoyant and gravitational forces to obtain the settlement rate. Combining forces and solving for the particle velocity yields:

$$v = d^2 g\,(\rho_p - \rho_w)/(18\eta) \qquad [1]$$

where d is equivalent particle diameter, g is gravity acceleration, ρ_p is particle density, ρ_w, is solution density, and η is solution viscosity.

Figure 1 shows the pipet method and Figure 2 shows a typical hydrometer used for soil-solution density measurements. This equipment is relatively inexpensive and readily available and, for routine analysis in soils, sedimentation, and engineering laboratories, the pipet and hydrometer are the standard tools for fine-fraction analysis. In recent years, the need for diagnostic tools in soil genesis and mineralogy have led to more sophisticated techniques being developed and tested. Laser light-scattering techniques are now being used in soil science more routinely to obtain size distributions of soil materials. Figure 3 shows the primary components of a laser light-scattering device.

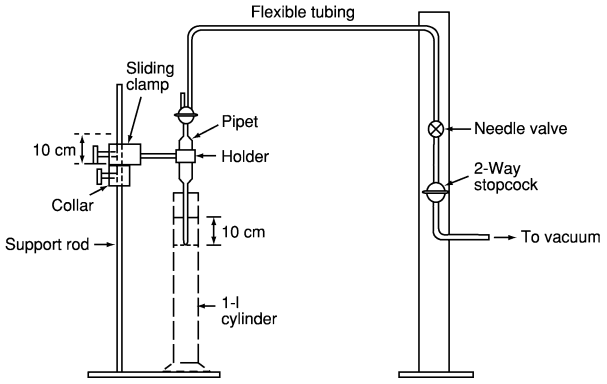


Figure 1 Schematic of pipet apparatus and stand for sedimentation analysis of soil-particle size.

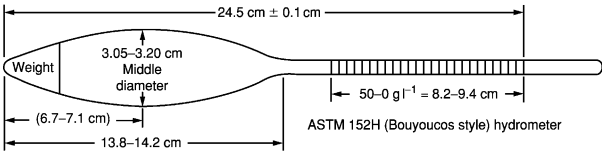


Figure 2 Schematic of ASTM 152H-type hydrometer used for soil particle-size analysis.

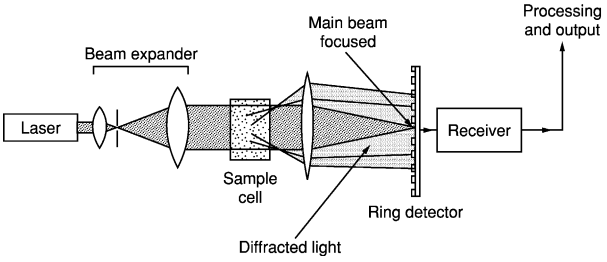


Figure 3 Laser diffraction particle-size analyzer showing the primary components: light sources, sample, focusing lens, detector, and processing system. (Reproduced with permission from Syvitski JPM (ed.) (1991) *Principles, Methods, and Application of Particle Size Analysis*. New York: Cambridge University Press.)

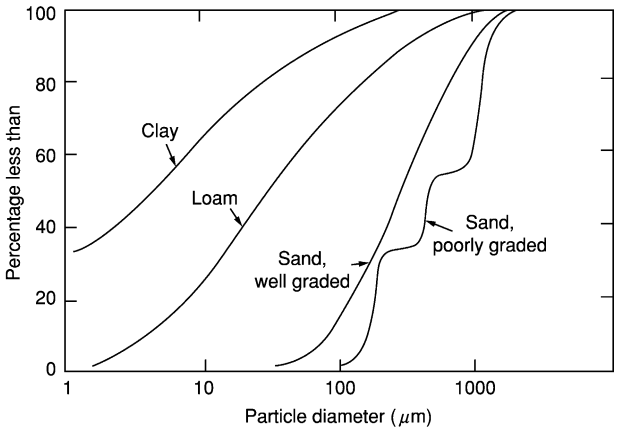


Figure 4 Particle-size distribution curves for several soil. (Adapted from Hillel D (1980) *Fundamentals of Soil Physics*. New York: Academic Press.)

Figure 4 shows particles-size distribution for a range of soils. All but the poorly sorted sand exhibit log-normal characteristics and might be considered fractal in nature. Much has been made about the fractal nature of particle-size distributions of soils and the possibility of relating the fractal dimension of the particle size to an equivalent pore-size distribution and then subsequently obtaining estimates of hydraulic properties from these parameters. In spite of this temptation, recent analysis suggests that many soils are actually multifractal, that is, they do not have the same fractal dimension over the entire particle-size range, so caution must be exercised in extrapolating fractal parameters beyond the range over which they might apply. Nevertheless, particle-size distribution curves are very useful and can help diagnose the origins of a soil as well as estimate numerous physical and chemical properties.

Figure 5 shows how various methods (sieving, sedimentation, and laser methods) are combined to obtain the size distribution of a soil sample. It appears that, with proper calibration and similar pretreatments (i.e., chemical and physical pretreatment), different methods can yield comparable results.

Size and Textural Classification Schemes

There are more than a dozen different schemes for classifying size separates. Figure 6 shows four of the more common schemes, including the size limits used in each scheme and the mesh numbers for the screens used in sieving the particles. Note that the size separates have arbitrary size limits, and for the most part the differences between schemes are relatively small, with division between the International Society of

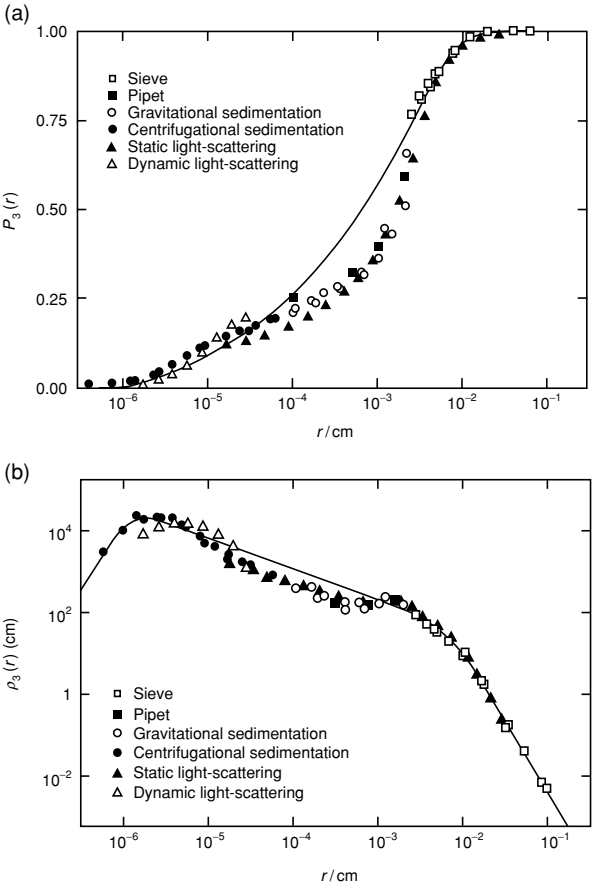


Figure 5 Cumulative mass distribution $P_3(r)$ (a) and number distribution density $\rho_3(r)$ for Buchberg soil as a function of particle radius for different measurement methods (solid lines represent empirical fit of three power-law functions). (Reproduced with permission from Wu Q, Borkovec M, and Sticher H (1993) On particle size distributions in soils. *Soil Science Society of America Journal* 57: 883–890.)

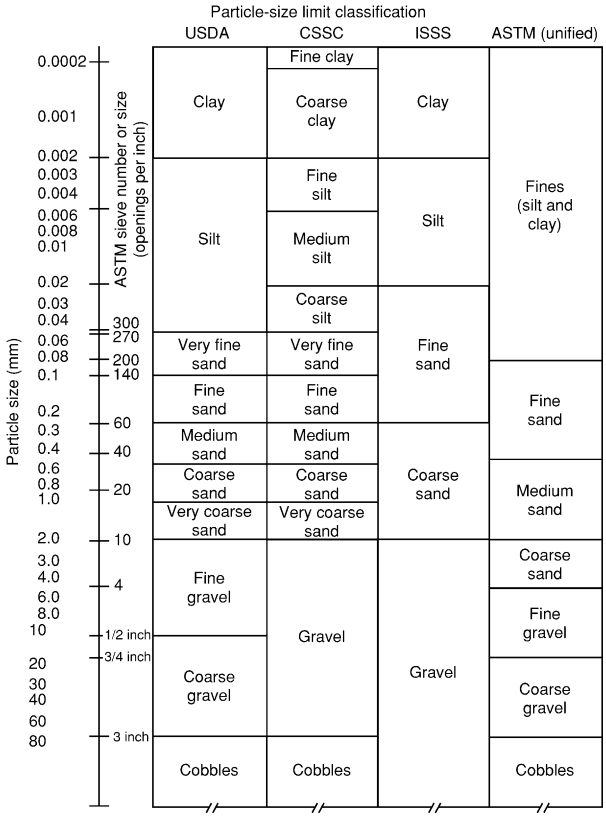


Figure 6 Particle-size limits according to several current classification schemes. USDA, US Department of Agriculture; CSSC, Canada Soil Survey Committee; ISSS, International Society of Soil Science; ASTM, American Society of Testing and Materials. (Adapted from Soil Survey Staff (1975); and McKeague JA (ed.) (1978) *Manual on Soil Sampling and Methods of Analysis*. Ottawa, Canada: Canadian Society of Soil Science.)

Soil Science (ISSS) and the US Department of Agriculture (USDA) classification being primarily the separation between the sand and clay fractions, with ISSS defining the uppermost silt limit at 0.02 mm and the USDA upper silt limit defined at 0.05 mm. Conversions from one system to another have been made and incorporated into user-friendly computer programs that translate one system readily into the another.

Textural Classification

Since soils rarely consist of one size range, textural classification is based on different combinations of size separates. The classification schemes that have been developed are as diverse as the applications for which the textural data are used. Two primary textural classifications are discussed, the universal classification scheme, used in engineering and adopted by the American Society of Testing and Materials (ASTM), and the USDA scheme that is widely used in agronomic science.

Universal classification (engineering) scheme Table 2 shows the engineering classification scheme used by the ASTM. In order to use this textural scheme, particle-size distribution is needed and certain engineering properties are required. From the size-distribution curve, the coefficient of curvature, C_c, can be obtained as:

$$C_c = (D_{30})^2/(D_{10} \times D_{60}) \quad [2]$$

where D_{60}, D_{30}, and D_{10} are the particle sizes corresponding to 60, 30, and 10% finer on the cumulative particle size distribution curve (Figure 7) and C_u, the coefficient of uniformity, is obtained as:

$$C_u = D_{60}/D_{10} \quad [3]$$

These properties are helpful in defining the types of gravel or sand that are present. C_u and C_c are shown in Figure 7 for a coarse soil. In this case, $C_u = 200$ and $C_c = 5.6$. If less than 5% of the sample passes the 200-mesh (0.075 mm) sieve, this soil is classed as a poorly graded gravel, GP. The other engineering properties are measures of the amount and type of fine materials present in the soil and are obtained from the liquid limit (LL) and plasticity index (PI). These properties (the so-called Atterberg limits) are obtained from standard methods and generally plotted on a PI versus LL diagram (Table 2). High LLs imply high clay content, and high PI implies the presence of swelling-type clays (e.g., smectites). Swelling clays such as smectites are typically more plastic than nonswelling clays (e.g., kaolinites). Kaolinitic clays are not plastic and grade out as materials that fall below the 'A' line on the PI versus LL curve (Table 2). In total, 16 different textural classes are used for engineering purposes, ranging from well-graded gravel (WG) to silts and clay (CH) and organic soils (PT).

USDA classification For agronomic and other soil-science applications, the textural designations are quite straightforward. It is assumed that the organic fraction is minimal and that only three size separates, sand, silt, and clay fractions, are being considered. In the USDA system, the fractions of the three size separates, sand (2–0.05 mm), silt (0.05–0.002 mm), and clay (0.002 mm) are plotted and 12 distinct textures are assigned according to the mix of separates (Figure 8). For the textural diagram, the assumption is:

$$m_d + m_t + m_y = 1 \quad [4]$$

where m_d is size fraction of sand, m_t is size fraction of silt, and m_y is size fraction of clay. Only two fractions are required to obtain texture from the textural triangle.

Table 2 Unified soil classification system, including plasticity chart

Criteria for assigning group symbols and names using laboratory tests[a]				*Soil classification*	
				Group symbol	*Group name*[b]
Coarse-grained soils more than 50% retained on No. 200 sieve	Gravels more than 50% of coarse fraction retained on No. 4 sieve	Clean gravels less than 5% fines[c]	$C_u > 4$ and $1 < C_c < 3$[e]	GW	Well-graded gravel[f]
			$C_u < 4$ and/or $1 > C_c > 3$[e]	GP	Poorly graded gravel[f]
		Gravels with fines more than 12% fines[c]	Fines classify as ML or MH	GM	Silty gravel[f,g,h]
			Fines classify as CL or CH	GC	Clayey gravel[f,g,h]
	Sands 50% or more of coarse fraction passes, No. 4 sieve	Clean sands less than 5%	$C_u > 6$ and $1 < C_c < 3$[e]	SW	Well-graded sand[i]
			$C_u < 6$ and/or $1 > C_c > 3$[e]	SP	Poorly graded sand[i]
		Sands with fines more than 12% fines[d]	Fines classify as ML or MH	SM	Silty sand[g,h,i]
			Fines classify as CL or CH	SC	Clayey sand[g,h,i]
Fine-grained soils 50% or more passes, No. 200 sieve	Silts and clays liquid limit less than 50	Inorganic	PI > 7 and plots on or above 'A' line[j]	CL	Lean clay[k,l,m]
			PI < 4 or plots below 'A' line[j]	ML	Silt[k,l,m]
		Organic	Liquid limit – oven-dried < 0.75 Liquid limit – not dried	OL	Organic clay[k,l,m,n] Organic silt[k,l,m,o]
	Silts and clays liquid limit 50 or more	Inorganic	PI plots on or above 'A' line	CH	Fat clay[k,l,m]
			PI plots below 'A' line	MH	Elastic silt[k,l,m]
		Organic	Liquid limit – oven-dried < 0.75 Liquid limit – not dried	OH	Organic clay[k,l,m,p] Organic silt[k,l,m,q]
Highly organic soils		Primarily organic matter, dark in color, and organic odor		PT	Peat

[a]Based on the material passing the 75-mm sieve.
[b]If field sample contained cobbles or boulders, or both, add 'with cobbles or boulders, or both' to group name.
[c]Gravels with 5–12% fines require dual symbols: GW-GM well-graded gravel with silt; GW-GC well-graded gravel with clay; GP-GM poorly graded gravel with silt; GP-GC poorly graded gravel with clay.
[d]Sands with 5–12% fines require dual symbols: SW-SM well-graded sand with silt; SW-SC well-graded sand with clay; SP-SM poorly graded sand with silt; SP-SC poorly graded sand with clay.
[e]$C_u = D_{60}/D_{10}$ and $C_c = (D_{30})^2/(D_{10} \times D_{60})$.
[f]If soil contains 15% or more sand, add 'with sand' to group name.
[g]If fines classify as CL-ML, use dual symbol GC-GM, or SC-SM.
[h]If fines are organic, add 'with organic fines' to group name.
[i]If soil contains ≥15% gravel, add 'with gravel' to group name.
[j]If Atterberg limits plot in hatched area, soil is a CL-ML, silty clay.
[k]If soil contains 15–29% plus No. 200, add 'with sand' or 'with gravel,' whichever is predominant.
[l]If soil contains ≥30% plus No. 200, predominantly sand, add 'sandy' to group name.
[m]If soil contains ≥30% plus No. 200, predominantly gravel, add 'gravelly' to group name.
[n]PI ≥4 and plots on or above 'A' line.
[o]PI <4 or plots below 'A' line.
[p]PI plots on or above 'A' line.
[q]PI plots below 'A' line.
(Reproduced with permission from ASTM (2000) *Standard Test for Classification of Soils for Engineering Purposes*. D 2487–98. 2000 Annual Book of ASTM Standards, 04.08: 238–247. Philadelphia, PA: ASTM.)

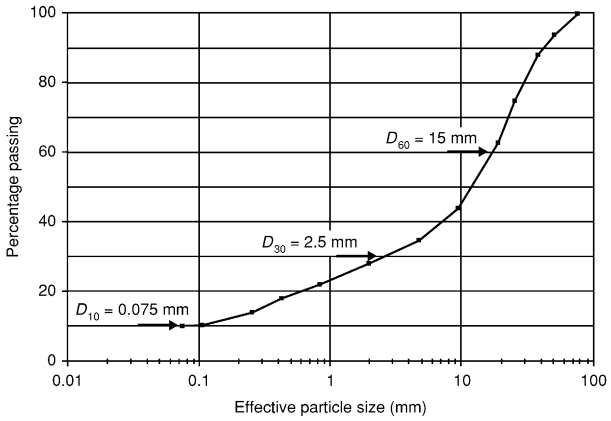


Figure 7 Particle-size distribution curve for a coarse soil, using sieve analysis. Also shown are the D_{60}, D_{30}, and D_{10} values required to compute the coefficient of curvature, C_c, and coefficient of uniformity, C_u. D_{60}, D_{30}, and D_{10} are the particle sizes corresponding to 60, 30, and 10% finer (percent passing).

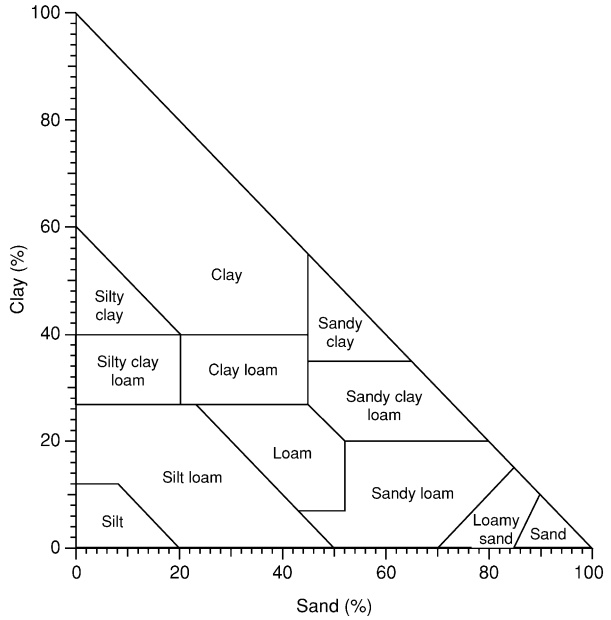


Figure 8 Textural triangle for USDA textural classification scheme.

Use of Texture Data for Estimating Hydraulic Properties

Soil scientists use soil-texture information to make qualitative judgments about a number of physical properties, but, until recently, there have been limited attempts to use texture for quantitative assessments of soil properties. Because the division of particles into size classes is entirely arbitrary, it is not surprising that texture data do not generally relate directly to fundamental physical properties. For example, it is not easy to determine how much silt it takes to equal a unit of clay, or if silt and sand fractions have any effect on a particular soil property of interest. In addition, there are underlying issues of soil mineralogy, organic versus inorganic fraction, particle-shape factors, and aggregation that complicate matters, so that the use of particle size data often requires some degree of empiricism in relating to such things as water-retention characteristics or effective surface area.

A more fundamental approach for describing soil texture is to use a statistical description for the particle-size distribution. A basic property of soil is the mean effective particle diameter. Particle-size distributions tend to be log-normal or fractal, so the appropriate means and standard deviations are log means and log standard deviations. These can be computed as:

$$d_g = \exp\left(\sum_i m_i \ln d_i\right) \quad [5]$$

$$\sigma_g = \exp\left(\sqrt{\sum_i m_i (\ln d_i)^2 - (\ln d_g)^2}\right) \quad [6]$$

where d_g is geometric mean diameter (in micrometers), m_i is mass fraction of component i (in grams), d_i is mean diameter of component i (in micrometers), and σ_g is geometric standard deviation (in micrometers).

When sand, silt, and clay fractions are known, the geometric mean diameter and the geometric standard deviation can be estimated in terms of the silt and clay fractions by using the following expressions:

$$d_g = \exp(5.756 - 3.454\, m_t - 7.712\, m_y) \quad [7]$$

$$\sigma_g = \exp[33.14 - 27.84\, m_t - 29.13\, m_y - (\ln d_g)^2]^{1/2} \quad [8]$$

where m_t is mass fraction of silt and m_y is mass fraction of clay.

Table 3 shows typical silt and clay fractions, geometric mean particle diameter, and geometric standard deviation for the 12 textural classes of the USDA classification scheme as identified in Figure 7. The standard textural triangle (Figure 7) can be transformed into a diagram based on the geometric mean particle size and the geometric mean standard deviation. From such a construct, textures can be interpreted in terms of more quantitative relationships. Coarser-soil soils can be easily incorporated into such a scheme. The standard textural triangle ignores the soil fraction greater than 2 mm. Frequently there is a need to incorporate this fraction into the size distribution, particularly when hydrologic properties such as water retention and unsaturated hydraulic conductivity are of interest.

Table 3 Typical silt and clay fractions and geometric mean particle diameter and geometric standard deviation for the 12 US Department of Agriculture textural classes

Texture	*Silt fraction*	*Clay fraction*	d_g (μm)	σ_g (μm)
Sand	0.05	0.03	212	4.4
Loamy sand	0.12	0.07	122	8.7
Sandy loam	0.25	0.10	62	12.2
Loam	0.40	0.18	20	16.3
Silt loam	0.65	0.15	11	9.6
Silt	0.87	0.07	10	4.1
Sandy clay loam	0.13	0.27	25	28.4
Clay loam	0.34	0.34	7	23.1
Silty clay loam	0.58	0.34	3	10.9
Sandy clay	0.07	0.40	11	39.7
Silty clay	0.45	0.45	2	13.8
Clay	0.20	0.60	1.5	22.8

Summary

Much effort has been expended in recent years to develop methods that relate particle-size distribution to pore size and subsequently to a host of other soil properties such as water retention and unsaturated hydraulic conductivity. While this is appealing, particularly when particle-size distribution data are available and other data are absent, this approach must be viewed with caution, because the particle-size distribution data obtained in most routine applications are at best empirical in nature. It needs to be pointed out that while particle-size distributions are important and offer what is often considered the most fundamental properties of a soil, the actual size distribution obtained from a laboratory analysis is almost always dependent upon the method used to obtain the particle-size distribution. As an example, researchers have found that some soils of volcanic ash origin have clay-sized particles that show great resistance to dispersion when subjected to gentle shaking but break down under high-energy source treatment. The chemical pretreatment and amount of mechanical work done on the soil are dictated by arbitrary decisions, so there is no absolute size distribution for a given soil. Because we are dealing with natural systems, we must rely on empirical methods and arbitrary limits and thus are constrained by the methods deployed to obtain size distributions and, resulting textural descriptions for soils.

Further Reading

Allen T (1981) *Particle Size Measurement*. London, UK: Chapman and Hall.

American Society for Testing and Materials (2000) *Standard Test for Classification of Soils for Engineering Purposes*, pp. 238–247. D 2487-98. 2000 Annual Book of ASTM Standards, vol. 04.08. Philadelphia, PA: ASTM.

American Society for Testing and Materials (2000) *Standard Test Method for Liquid Limit, Plastic Limit, and Plasticity Index of Soils*, pp. 546–558. D 4318-98. 2000 Annual Book of ASTM Standards, vol. 04.08. Philadelphia, PA: ASTM.

American Society for Testing and Materials (2000) *Standard Test Method for Particle-Size Analysis of Soils*, pp. 10–17. D 422-63 (1998). 2000 Annual Book of ASTM Standards 04.08. Philadelphia, PA: ASTM.

Christopher TBS and Mokhtaruddin AM (1996) A computer program to determine soil textural class in 1-2-3 for Windows and Excel. *Communications in Soil Science Plant Analysis* 27: 2315–2319.

Folk RL (1980) *Petrology of Sedimentary Rocks*, 2nd edn. Austin, TX: Hemphill.

Gee GW and Or D (2002) Particle size analysis. In: Dane JH and Topp GC (eds) *Methods of Soil Analysis*, part 4, pp. 255–293. *Physical Methods*. SSSA Book Series No. 5. Madison, WI: Soil Science Society of America.

Hillel D (1980) *Fundamentals of Soil Physics*. New York: Academic Press.

Kubota T (1972) Aggregate-formation of allophanic soils: effects of drying on the dispersion of the soils. *Soil Science and Plant Nutrition* 18: 79–87.

Loveland PJ and Whalley WR (2001) Particle size analysis. In: Smith KA and Mullins CE (eds) *Soil and Environmental Analysis: Physical Methods*, 2nd edn. New York: Marcel Dekker.

McKeague JA (ed.) (1978) *Manual on Soil Sampling and Methods of Analysis*. Ottawa, Canada: Canadian Society of Soil Science.

Muggler CC, Pape T, and Buurman P (1996) Laser grain-size determination in soil genetic studies. 2. Clay content, clay formation, and aggregation in some Brazilian oxisols. *Soil Science* 162: 219–228.

Shiozawa S and Campbell GS (1991) On the calculation of mean particle diameter and standard deviation from sand, silt, and clay fractions. *Soil Science* 152: 427–431.

Syvitski JPM (ed.) (1991) *Principles, Methods, and Application of Particle Size Analysis*. New York: Cambridge University Press.

US Department of Agriculture (1982) *Procedures for Collecting Soil Samples and Methods of Analysis for Soil Survey*. Soil Survey Investigations Report No. 1. Washington, DC: Soil Conservation Service.

Warrick AW (2002) *Soil Physics Companion*. Boca Raton, FL: CRC Press.

Yong RN and Warkentin BP (1996) *Introduction to Soil Behavior*. New York: Macmillan.

THERMAL PROPERTIES AND PROCESSES

D Hillel, Columbia University, New York, NY, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

The soil temperature, as it varies in time and space, is a factor of primary importance in determining the rates and directions of soil physical processes and of energy and mass exchange with the atmosphere. Temperature governs evaporation and aeration, as well as the types and rates of chemical reactions that take place in the soil. Finally, soil temperature strongly influences biological processes such as seed germination, seedling emergence and growth, root development, and microbial activity.

Soil temperature varies in response to changes in the radiant, thermal, and latent energy exchange processes that take place primarily through the soil surface. The effects of these phenomena are propagated into the soil profile via a complex series of transport processes, the rates of which are affected by time-variable and space-variable soil properties.

Modes of Energy Transfer

In general, there are three principal modes of energy transfer: radiation, convection, and conduction. By radiation, we refer to the emission of energy in the form of electromagnetic waves from all bodies above 0 K. According to the Stefan–Boltzmann law, the total energy emitted by a body, J_t, integrated over all wavelengths, is proportional to the fourth power of the absolute temperature T of the body's surface. This law is usually formulated:

$$J_t = \epsilon\sigma T^4 \qquad [1]$$

where σ is a constant and ϵ is the emissivity coefficient, which equals unity for a perfect emitter (generally called a 'black body'). The absolute temperature also determines the wavelength distribution of the emitted energy. The Wien law states that the wavelength of maximal radiation intensity λ_m is inversely proportional to the absolute temperature:

$$\lambda_m = 2900/T \qquad [2]$$

where λ_m is in micrometers. The actual intensity distribution as a function of wavelength and temperature is given by Planck's law:

$$E_\lambda = C_1/\lambda^5[\exp(C_2/\lambda T) - 1] \qquad [3]$$

where E_λ is energy flux emitted in a given wavelength range and C_1 and C_2 are constants.

Since the temperature of the soil surface is generally of the order of 300 K (though it can range from below 273 K, the freezing point, to 330 K or higher), the radiation emitted by the soil surface has its peak intensity at a wavelength of approximately 10 μm and its wavelength distribution over the range of 3–50 μm. This is in the realm of infrared, or heat, radiation.

A very different spectrum is emitted by the sun, which acts as a black body at an effective surface temperature of approximately 6000 K. The sun's radiation includes the visible light range of 0.3–0.7 μm, as well as some infrared radiation of greater wavelength (up to approx. 3 μm) and some ultraviolet radiation ($\lambda < 0.3$ μm). Since there is very little overlap between the two spectra, it is customary to distinguish between them by calling the incoming solar spectrum 'short-wave' radiation, and the spectrum emitted by the Earth 'long-wave' radiation.

The second mode of energy transfer, called 'convection,' involves the movement of a heat-carrying mass, as in the case of ocean currents or atmospheric winds. An example more pertinent to soil physics would be the infiltration of warm waste water (from, e.g., a power plant) into an initially cold soil.

Conduction, the third mode of energy transfer, is the propagation of heat within a body by internal molecular motion. Since temperature is an expression of the kinetic energy of a body's molecules, the existence of a temperature difference within a body will normally cause the transfer of kinetic energy by the numerous collisions of rapidly moving molecules from the warmer region of the body to their neighbors in the colder region. The process of heat conduction is thus analogous to diffusion and, in the same way that diffusion tends in time to equilibrate a mixture's composition throughout, heat conduction tends to equilibrate a body's internal distribution of molecular kinetic energy – that is, its temperature.

In addition to the three modes of energy transfer described, there is a composite phenomenon which one may recognize as a fourth mode, namely latent heat transfer. A prime example is the process of distillation, which includes the heat-absorbing stage of evaporation, followed by the convective or diffusive movement of the vapor, and ending with the heat-releasing stage of condensation. A similar catenary process can also occur in transition back and forth

from ice to liquid water in soils subject to freezing and thawing.

Energy Balance for a Bare Soil

The radiation balance of a bare soil surface can be written:

$$J_n = (J_s + J_a)(1 - \alpha) + J_{li} - J_{lo} \quad [4]$$

Here J_n is the net radiation, that is, the sum of all incoming-minus-outgoing radiant energy fluxes; J_s the incoming flux of short-wave radiation directly from the sun and J_a the short-wave diffuse radiation from the atmosphere (sky); J_{li} the incoming long-wave radiation flux from the sky and J_{lo} the outgoing long-wave radiation emitted by the soil; and, finally, α is the albedo, or reflectivity coefficient, which is the fraction of incoming short-wave radiation reflected by the soil surface rather than absorbed by it. In the present context, all terms that do not pertain to the soil, namely J_s, J_a, and J_{li}, are disregarded.

The albedo α is an important characteristic of soil surfaces, and it can vary widely in the range of 0.1–0.4, depending upon the soil's basic color (whether dark or light), the surface's roughness, and the inclination of the incident radiation relative to the surface. In the short term, the albedo also depends on the changing wetness of the exposed soil. The drier the soil, the smoother its surface, and the brighter its color, the higher its albedo. To a certain extent, the albedo can be modified by various surface treatments such as tillage and mulching.

Apart from the reflected short-wave radiation, the soil also emits long-wave radiation. In accordance with eqn [1], the emitted flux J_{lo} depends primarily on soil surface temperature but is also affected by the soil's emissivity. This parameter, in turn, depends on soil wetness, but its range of variation is generally small, i.e., between 0.9 and 1.0.

The net radiation received by the soil surface is transformed into heat, which warms the soil and air and vaporizes water. We can write the surface energy balance as follows:

$$J_n = S + A + LE \quad [5]$$

where S is the soil heat flux (the rate at which heat is transferred from the surface downward into the soil profile), A is the 'sensible' heat flux transmitted from the surface to the air above, and LE is the evaporative heat flux, a product of the evaporative rate E and the latent heat per unit quantity of water evaporated, L.

The total surface energy balance (combining eqns [4] and [5]) is therefore:

$$(J_s + J_a)(1 - \alpha) + J_{li} - J_{lo} - S - A - LE = 0 \quad [6]$$

Conventionally, all components of the energy balance are taken as positive if directed toward the surface, and negative otherwise.

Conduction of Heat in Soil

The conduction of heat in solid bodies was analyzed as long ago as 1822 by Fourier, whose name is associated with the linear transport equations that have been used ever since to describe heat conduction. The first law of heat conduction, known as the Fourier law, states that the flux of heat in a homogeneous body is in the direction of, and proportional to, the temperature gradient:

$$q_h = -\kappa \nabla T \quad [7]$$

Here q_h is the thermal flux (i.e., the amount of heat conducted across a unit cross-sectional area in unit time), κ is thermal conductivity, and ∇T the spatial gradient of temperature T. In one-dimensional form, this law is written:

$$q_h = -\kappa_x \partial T/\partial x \quad \text{or} \quad q_h = -\kappa_z \partial T/\partial_z \quad [8]$$

Here $\partial T/\partial x$ is the temperature gradient in any arbitrary direction designated x, and $\partial T/\partial z$ is, specifically, the gradient in the vertical direction representing soil depth ($z = 0$ being the soil surface). The subscripts attached to the thermal conductivity term are meant to account for the possibility that this parameter may have different values in different directions (i.e., that it may be nonisotropic). The negative sign in these equations is due to the fact that heat flows from a higher to a lower temperature (i.e., in the direction of, and in proportion to, a negative temperature gradient).

Equation [7] is sufficient to describe heat conduction under steady-state conditions, that is, where the temperature at each point in the conducting medium is invariant and the flux is constant in time and space. To account for nonsteady (transient) conditions, we need a second law analogous to Fick's second law of diffusion. To obtain the second law of heat conduction, the principle of energy conservation in the form of the continuity equation is invoked, which states that, in the absence of any sources or sinks of heat, the time rate of change in heat content of a volume element of the conducting medium must equal the change of flux with distance:

$$\rho c_m \partial T/\partial t = -\nabla \cdot q_h \quad [9]$$

where ρ is mass density and c_m specific heat capacity per unit mass (called simply 'specific heat' and

defined as the change in heat content of a unit mass of the body per unit change in temperature). The product ρc_m (often designated C) is the specific heat capacity per unit volume, and $\partial T/\partial t$ is the time rate of temperature change. Note that ρ represents the total mass per unit volume, including the mass of water in the case of a moist soil. The symbol ∇ ('del') is the shorthand representation of the three-dimensional gradient. An equivalent form of eqn [9] is:

$$\rho c_m \partial T/\partial t = -(\partial q_x/\partial x + \partial q_y/\partial y + \partial q_z/\partial z)$$

where x, y, z are the orthogonal direction coordinates.

Combining eqn [9] with [7] gives the second law of heat conduction:

$$\rho c_m \partial T/\partial t = -\nabla \cdot (\kappa \nabla T) \quad [10]$$

which, in one-dimensional form, is:

$$\rho c_m \partial T/\partial t = \partial/\partial x(\kappa \partial T/\partial x) \quad [11]$$

Sometimes there is need to account for the possible occurrence of heat sources or sinks in the realm where heat flow takes place. Heat sources include such phenomena as organic matter decomposition, wetting of initially dry soil material, and condensation of water vapor. Heat sinks are generally associated with evaporation. Combining all these sources and sinks into a single term S, we can rewrite the last equation:

$$\rho c_m \partial T/\partial t = \partial/\partial x(\kappa \partial T/\partial x) \pm S(x,t) \quad [12]$$

in which the source–sink term is shown as a function of both space and time.

Volumetric Heat Capacity of Soils

The volumetric heat capacity C of a soil is defined as the change in heat content of a unit bulk volume of soil per unit change in temperature. Its units are calories per cubic centimeter per degree (Kelvin) or joules per cubic meter per degree. As such, C depends on the composition of the solid phase (mineral and organic constituents) of the soil, on bulk density, and on soil wetness (Table 1).

The value of C can be estimated by summing the heat capacities of the various constituents, weighted according to their volume fractions:

$$C = \sum f_{si} C_{si} + f_w C_w + f_a C_a \quad [13]$$

Here, f denotes the volume fraction of each phase: solid (subscripted 's'), water ('w'), and air ('a'). The solid phase includes a number of components subscripted 'i,' such as various minerals and organic matter, and the symbol Σ indicates the summation of the products of their respective volume fractions and heat capacities. The C value for water, air, and each component of the solid phase is the product of the particular density and the specific heat per unit mass (i.e., $C_w = \rho_w c_{mw}$; $C_a = \rho_a c_{ma}$; $C_{si} = \rho_{si} c_{msi}$).

Table 1 Densities and volumetric heat capacities of soil constituents (at 10°C) and of ice (at 0°C)

	Density ρ		*Heat capacity C*	
Constituent	*g cm^{-3}*	*kg m^{-3}*	*cal cm^{-3} K*	*J m^{-3} K*
Quartz	2.66	2.66×10^3	0.48	2.0×10^6
Other minerals (mean)	2.65	2.65×10^3	0.48	2.0×10^6
Organic matter	1.3	1.3×10^3	0.6	2.5×10^6
Water (liquid)	1.0	1.0×10^3	1.0	4.2×10^6
Ice	0.92	0.92×10^3	0.45	1.9×10^6
Air	0.001 25	1.25	0.003	1.25×10^3

Reproduced with permission from Hillel D (1998) *Environmental Soil Physics*. San Diego, CA: Academic Press.

Most of the minerals composing soils have nearly the same values of density (approx. 2.65 g cm^{-3} or 2.65×10^3 kg m^{-3}) and of heat capacity (0.48 cal cm^{-3} K or 2.0×10^6 J m^{-3} K). Since it is difficult to separate the different kinds of organic matter present in soils, it is tempting to group them all into a single constituent (with mean density of approximately 1.3 g cm^{-3} or 1.3×10^3 kg m^{-3}, and a mean heat capacity of approximately 0.6 cal cm^{-3} K or 2.5×10^6 J m^{-3} K).

The density of water is less than half that of mineral matter (approx. 1 g cm^{-3} or 1.0×10^3 kg m^{-3}); its specific heat is more than twice as large (1 cal cm^{-3} K, or 4.2×10^6 J m^{-3} K). Finally, since the density of air is only approximately 1/1000 that of water, its contribution to the specific heat of the composite soil can be neglected generally.

Thermal Conductivity of Soils

Thermal conductivity, designated κ, is defined as the quantity of heat transferred through a unit area of the conducting body in unit time under a unit temperature gradient. As shown in Table 2, the thermal conductivities of specific soil constituents differ very markedly (see also Table 3). Hence the space-averaged (macroscopic) thermal conductivity of a soil depends upon its mineral composition and organic matter content, as well as on the volume fractions of water and air.

Since the thermal conductivity of air is very much smaller than that of water or solid matter, a high air content (or low water content) corresponds to a low

thermal conductivity. Moreover, since the proportions of water and air vary continuously, κ is also time-variable. Soil composition is seldom uniform in depth, hence κ is generally a function of depth as well as of time. Unlike heat capacity, thermal conductivity is sensitive not merely to the volume composition of a soil but also to the sizes, shapes, and spatial arrangements of the soil particles.

The relationship between the overall thermal conductivity of a soil and the specific conductivities and volume fractions of the soil's constituents is very intricate, as it involves the internal geometry or structure of the soil matrix and the mode of transmission of heat from particle to particle and from phase to phase.

The dependence of thermal conductivity and diffusivity on soil wetness is illustrated in **Figure 1**. The influence of latent heat transfer by the diffusion of water vapor in the air-filled pores is proportional to the temperature gradient in these pores. It can be taken into account by adding to the thermal conductivity of air an apparent conductivity due to evaporation, transport, and condensation of water vapor (the so-called vapor-enhancement factor). This value is strongly temperature-dependent and rises rapidly with increasing temperature.

Because soil water potential depends on temperature, the development of a temperature gradient generally induces the movement of water as well as of heat. Hence techniques for measuring heat transfer through a soil sample based on steady-state heat flow between two planes maintained at a constant temperature differential involve the risk of changing the sample's internal moisture distribution and therefore its thermal properties. During the process of measurement, the soil near the warmer plane becomes drier, while that near the cooler plane becomes wetter. Early attempts to measure thermal conductivity failed to recognize this pitfall as they purported to maintain constant soil moisture conditions during prolonged, steady-state heat flow. Hence their results can only be considered approximations at best. While steady-state methods may be sufficiently accurate for measuring thermal conductivity of dry soils, short-term, transient heat-flow techniques are preferable, in principle, for moist soils.

Table 2 Thermal conductivities of soil constituents (at 10°C) and of ice (at 0°C)

Constituent	*mcal cm^{-1} sK*	*Wm^{-1} K*
Quartz	21	8.8
Other minerals (average)	7	2.9
Organic matter	0.6	0.25
Water (liquid)	1.37	0.57
Ice	5.2	2.2
Air	0.06	0.025

Reproduced with permission from Hillel D (1998) *Environmental Soil Physics*. San Diego, CA: Academic Press.

Simultaneous Transport of Heat and Moisture

The flows of water and of thermal energy under nonisothermal conditions in the soil are interactive phenomena: the one entails the other. Temperature gradients affect the moisture potential field and induce both liquid and vapor movement. Reciprocally, moisture gradients move water, which carries heat. The simultaneous occurrence of temperature gradients and of moisture potential gradients in the soil therefore brings about the combined transport of heat and moisture.

Two separate approaches to the combined transfer of heat and moisture have been attempted: (1) a

Table 3 Average thermal properties of soils and snow[a]

Soil type	*Porosity*	*Volumetric wetness*	*Thermal conductivity (10^{-3} cal cm^{-1} s^{-1}°C)*	*Volumetric heat capacity (cal cm^{-1} s^{-1}°C)*	*Damping depth (diurnal) (cm)*
Sand	0.4	0.0	0.7	0.3	8.0
	0.4	0.2	4.2	0.5	15.2
	0.4	0.4	5.2	0.7	14.3
Clay	0.4	0.0	0.6	0.3	7.4
	0.4	0.2	2.8	0.5	12.4
	0.4	0.4	3.8	0.7	12.2
Peat	0.8	0.0	0.14	0.35	3.3
	0.8	0.4	0.7	0.75	5.1
	0.8	0.8	1.2	1.15	5.4
Snow	0.95	0.05	0.15	0.05	9.1
	0.8	0.2	0.32	0.2	6.6
	0.5	0.5	1.7	0.5	9.7

[a]After van Wijk and de Vries (1963).
Reproduced with permission from Hillel D (1998) *Environmental Soil Physics*. San Diego, CA: Academic Press.

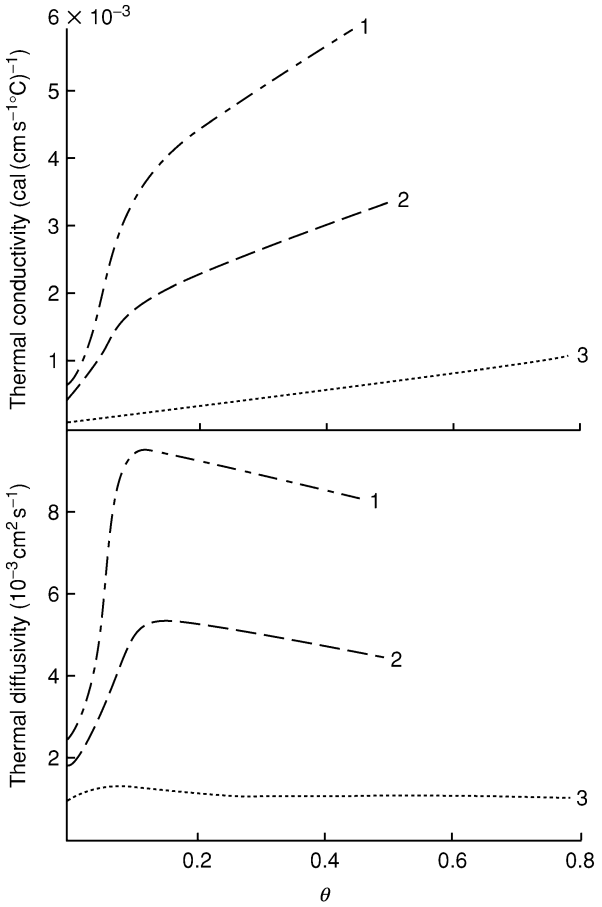


Figure 1 Thermal conductivity and thermal diffusivity as functions of volume wetness (volume fraction of water) for: (1) sand (bulk density 1460 kg m^{-3}, volume fraction of solids 0.55); (2) loam (bulk density 1330 kg m^{-3}, volume fraction of solids 0.5); and (3) peat (volume fraction of solids 0.2). (After de Vries, 1975.) (Reproduced with permission from Hillel D (1998) *Environmental Soil Physics*. San Diego, CA: Academic Press.)

mechanistic approach, based on a physical model of the soil system, and (2) a thermodynamic approach, based on the phenomenology of irreversible processes in terms of coupled forces and fluxes. Though starting from different points of view, the two approaches have been shown to be related and, properly formulated, can be cast into an equivalent mold.

The mechanistic approach is based on the concept of viscous flow of liquid water under the influence of gravity and of capillary and adsorptive forces, and on the concept of vapor movement by diffusion. Local 'microscopic-scale' thermodynamic equilibrium between liquid and vapor is assumed to exist at all times and at each point within the soil. The general differential equation describing moisture movement in a porous system under combined temperature and moisture gradients for unidimensional vertical flow is, accordingly:

$$\partial\theta/\partial t = \nabla \cdot (D_T \nabla T) + \nabla \cdot (D_w \nabla\theta) - \partial K/\partial z \quad [14]$$

where θ is volumetric wetness, t time, T absolute temperature, D_T the water diffusivity under a temperature gradient (the sum of the liquid and vapor diffusivities), D_w the water diffusivity under a moisture gradient, K the hydraulic conductivity, and z the vertical space coordinate. The last term on the right-hand side is due to the gravity gradient and becomes positive if z is taken to be increasing downwards.

The heat transfer equation is, similarly:

$$C_v \partial T/\partial t = \nabla \cdot (\kappa \nabla T) - L\nabla \cdot (D_{w,vap} \nabla\theta) \quad [15]$$

Here C_v is volumetric heat capacity, κ apparent thermal conductivity of the soil, L latent heat of vaporization of water, and $D_{w,vap}$ diffusivity for heat conveyed by water movement (mostly vapor). Equations [14] and [15] are of the diffusion type, involving θ- and T-dependent diffusivities as well as gradients of both θ and T.

Taken together, eqns [14] and [15] describe the coupled transport of moisture and heat in soils. The assumption of local thermodynamic equilibrium links the vapor pressure p_v to the matric potential ψ by the following relation: $p_v = p_{vs}h = p_{vs}\exp(Mg\psi/RT)$, where p_{vs} is the saturated vapor pressure at the particular temperature T, h relative humidity, M molar mass, g acceleration due to gravity, and R the universal gas constant.

The approach based on the thermodynamics of irreversible processes formulates a pair of phenomenological equations in which the fluxes of moisture q_w and heat q_h are expressed as linear functions of the moisture potential (e.g., pressure) gradient dp/dz and the temperature gradient dT/dz:

$$\begin{aligned} q_w &= -L_{ww}(1/T)dp/dz - L_{wh}(1/T^2)(dT/dz) \\ q_h &= -L_{hw}(1/T)dp/dz - L_{hh}(1/T^2)(dT/dz) \end{aligned} \quad [16]$$

The four phenomenological coefficients occurring in these equations (L_{ww}, L_{wh}, L_{hw}, L_{hh}, relating water flow to the water potential gradient, water flow to the thermal potential gradient, heat flow to the water potential gradient, and heat flow to the thermal potential gradient, respectively) are unknown functions of p (or θ) and T. According to the Onsager theorem, the cross-coupling coefficients L_{wh} and L_{hw} are equal when the fluxes and forces are properly formulated. Thus, the number of coefficients that must be measured is reduced.

An apparent advantage of the irreversible thermodynamics approach is that it makes no *a priori* assumptions regarding the mechanisms of the transport phenomena formulated. Hence it would seem to be less restrictive than a physical theory whose validity

is constrained at the outset by its mechanistic assumptions. The disadvantage of this approach, however, is precisely its failure to address itself to, and provide insight into, the nature and internal workings of the phenomena considered.

Thermal Regime of Soil Profiles

In nature, soil temperature varies continuously in response to the ever-changing meteorological regime acting on the soil–atmosphere interface. That regime is governed by a regular periodic succession of days and nights, and of summers and winters. Yet the regular diurnal and annual cycles are perturbed by such irregular episodic phenomena as cloudiness, cold waves, heat waves, rainstorms or snowstorms, and periods of drought. Add to these external influences the soil's own changing properties (i.e., temporal changes in reflectivity, heat capacity, and thermal conductivity as the soil alternately wets and dries, and the variation of all these properties with depth), as well as the influences of geographic location, vegetative cover, and, finally human management, and one can expect the thermal regime of soil profiles to be complex indeed.

The simplest mathematical representation of nature's fluctuating thermal regime is to assume that at all depths in the soil the temperature oscillates as a pure harmonic (sinusoidal) function of time around an average value. Assume that, although soil temperature varies differently at different depths in the soil, the average temperature is the same for all depths. A starting time ($t = 0$) is chosen such that the surface is at the average temperature. The temperature at the surface can then be expressed as a function of time (Figure 2):

$$T(0,t) = T_{ave} + A_0 \sin \omega t \qquad [17]$$

where $T(0,t)$ is the temperature at $z = 0$ (the soil surface) as a function of time t, T_{ave} is the average temperature of the surface (as well as of the profile), and A_0 is the amplitude of the surface temperature fluctuation (the range from maximum, or from minimum, to the average temperature). Finally, ω is the radial frequency, which is 2π times the actual frequency. In the case of diurnal variation, the period is 86 400 s (24 h), so $\omega = 2\pi/86\,400 = 7.27 \times 10^{-5}\,s^{-1}$. Note that the argument of the sine function is expressed in radians rather than in degrees.

The last equation is the boundary condition for $z = 0$. For the sake of convenience, let us assume that at infinite depth ($z = \infty$) the temperature is constant and equal to T_{ave}. Under these circumstances, the temperature at any depth z is also a sine function of time, as shown in Figure 3:

$$T(z,t) = T_{ave} + A_z \sin[\omega t + \phi(z)] \qquad [18]$$

in which A_z is the amplitude at depth z. Both A_z and $\phi(z)$ are functions of z but not of t. They can be determined by substituting the solution of eqn [18]

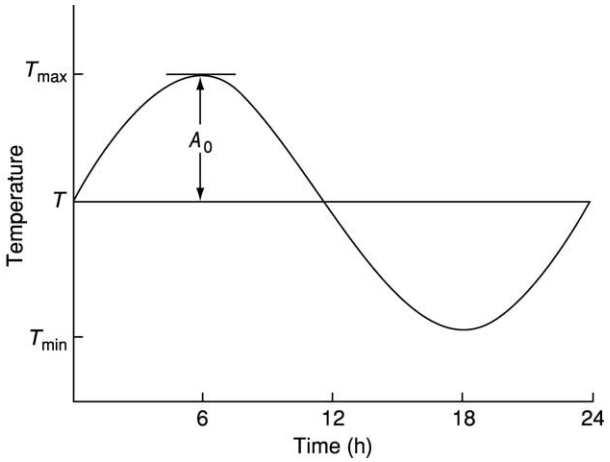


Figure 2 Idealized daily fluctuation of surface soil temperature, according to the equation: $T = T_{ave} + A_0 \sin(\omega t/p)$, where T is temperature, T_{ave} average temperature, A_0 amplitude, t time, and p period of the oscillation (in this case, p refers to the diurnal 24 h). (Reproduced with permission from Hillel D (1998) *Environmental Soil Physics*. San Diego, CA: Academic Press.)

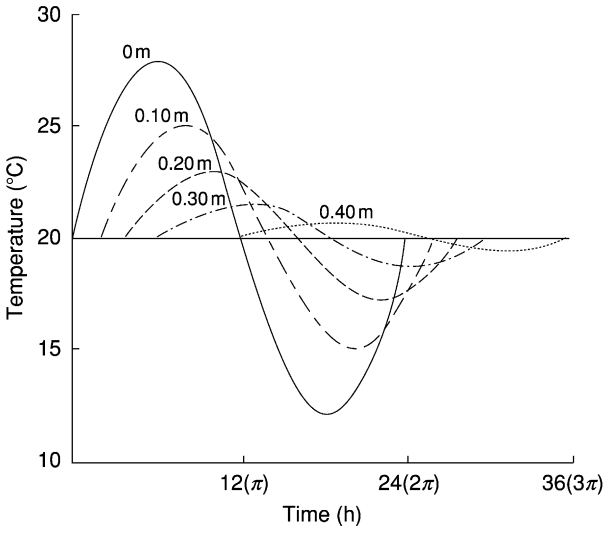


Figure 3 Idealized variation of soil temperature with time for various depths. Note that at each succeeding depth the peak temperature is damped and shifted progressively in time. Thus, the peak at a depth of 0.4 m lags about 12 h behind the temperature peak at the surface and is only about 1/16 of the latter. In this hypothetical case, a uniform soil was assumed, with a thermal conductivity of 1.68 $Jm^{-1}s^{-1}deg^{-1}$ (or 4×10^{-3} $cal\,cm^{-1}s^{-1}deg^{-1}$) and a volumetric heat capacity of 2.1×10^6 $Jm^{-3}deg^{-1}$ (0.5 $cal\,cm^{-3}deg^{-1}$). (Reproduced with permission from Hillel D (1998) *Environmental Soil Physics*. San Diego, CA: Academic Press.)

in the differential equation $\partial T/\partial t = \kappa(\partial^2 T/\partial z^2)$. This leads to the solution:

$$T(z,t) = T_{ave} + A_0[\sin(\omega t - z/d)]/e^{z/d} \qquad [19]$$

The constant d is a characteristic depth, called the 'damping depth,' at which the temperature amplitude decreases to the fraction $1/e$ ($1/2.718 = 0.37$) of the amplitude at the soil surface A_0. The damping depth is related to the thermal properties of the soil and the frequency of the temperature fluctuation as follows:

$$d = (2\kappa/C\omega)^{1/2} = (2D_h/\omega)^{1/2} \qquad [20]$$

At any depth the amplitude of the temperature fluctuation A_z is smaller than A_0 by a factor $e^{z/d}$, and there is a phase shift (i.e., a time delay of the temperature peak) equal to $-z/d$. The decrease in amplitude and increase in phase lag with depth are typical phenomena in the propagation of a periodic temperature wave in the soil (Figure 4).

The physical reason for the damping and retarding of the temperature waves with depth is that a certain amount of heat is absorbed or released along the path of heat propagation when the temperature of the conducting soil increases or decreases, respectively. The damping depth is related inversely to the frequency, as can be seen from eqn [20]. Hence it depends directly on the period of the temperature fluctuation considered. The damping depth is $(365)^{1/2} = 19$ times larger for the annual variation than for the diurnal variation in the same soil.

The annual variation of soil temperature down to considerable depth causes deviations from the simplistic assumption that the daily average temperature is the same for all depths in the profile. The combined effect of the annual and diurnal variation of soil temperature can be expressed by:

$$T(z,t) = T_{ave,y} + A_y[\sin(\omega_y t + \phi_y - z/d_y)]/e^{z/d_y} + A_d[\sin(\omega_d t + \phi_d - z/d_d)]/e^{z/d_d} \qquad [21]$$

where the subscripted indices y and d refer to the yearly and daily temperature waves, respectively. Thus $T_{ave,y}$ is the annual mean temperature. The daily cycles are now seen to be short-term perturbations superimposed upon the annual cycle. Vagaries of weather (e.g., spells of cloudiness or rain) can cause considerable deviations from simple harmonic fluctuations, particularly for the daily cycles. Longer-term climatic irregularities can also affect the annual cycle. The soil temperature profile as it varies seasonally is shown in Figure 4.

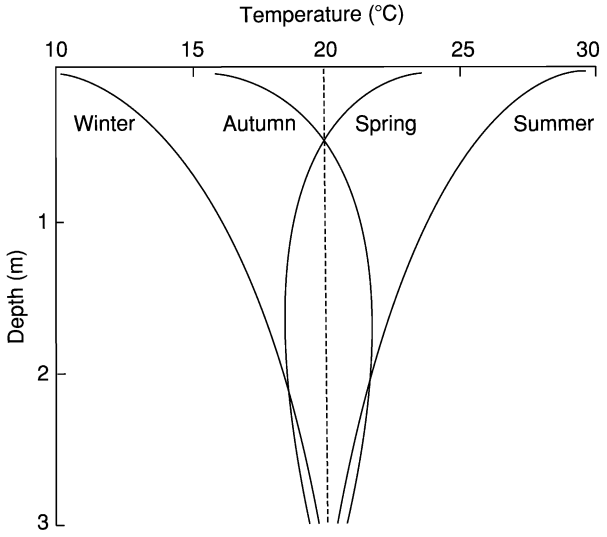


Figure 4 Soil temperature profile as it varies from season to season in a frost-free region. (Reproduced with permission from Hillel D (1998) *Environmental Soil Physics*. San Diego, CA: Academic Press.)

An alternative approach is possible, with fewer constraining assumptions. It is based on numerical rather than analytical methods for solving the differential equations of heat conduction. Mathematical simulation models relying on digital computers now allow soil thermal properties to vary in time and space (e.g., in response to periodic changes in soil wetness), so as to account for alternating surface saturation and desiccation and for profile-layering. They also allow various climatic inputs to follow more realistic and irregular patterns. The surface amplitude of temperature need no longer be taken to be an independent variable, but one that depends on the surface energy balance and thus is affected by both soil properties and above-soil conditions.

Other innovations of practical importance involve the development of techniques for monitoring the soil thermal regime more accurately and precisely than was possible previously. One such technique is the infrared radiation thermometer for scanning or remote sensing of surface temperature for both fallow and vegetated soils without disturbance of the measured surface. Knowledge of the surface temperature and its variation in time is important in assessing energy exchange between the soil and the overlying atmosphere, as well as in determining boundary conditions for within-soil heat transfer.

See also: **Energy Balance**; **Evaporation of Water from Bare Soil**; **Radiation Balance**

Further Reading

Carslaw HS and Jaeger JC (1959) *Conduction of Heat in Solids*. Oxford, UK: Oxford University Press.

de Vries DA (1975) Heat transfer in soils. In: de Vries DA and Afgan NH (eds) *Heat and Mass Transfer in the Biosphere*, pp. 5–28. Washington, DC: Scripta Book Co.

Groenevelt PH and Bolt GH (1969) Non-equilibrium thermodynamics of the soil–water system. *Journal of Hydrology* 7: 358–388.

Hillel D (1977) *Computer Simulation of Soil–Water Dynamics*. Ottawa, Canada: International Development Research Centre.

Hillel D (1998) *Environmental Soil Physics*. San Diego, CA: Academic Press.

Jackson RD and Taylor SA (1986) Heat transfer. In: *Methods of Soil Analysis*. Monograph No. 9, pp. 349–360. Madison, WI: American Society of Agronomy.

Kutilek M and Nielsen DR (1994) *Soil Hydrology*. Cremingen, Germany: Catena-Verlag.

Philip JR and de Vries DA (1957) Moisture movement in porous materials under temperature gradients. *Transactions of the American Geophysical Union* 38: 222–228.

van Bavel CHM and Hillel D (1976) Calculating potential and actual evaporation from a bare soil surface by simulation of concurrent flow of water and heat. *Agricultural Meteorology* 17: 453–476.

THERMODYNAMICS OF SOIL WATER

P H Groenevelt, University of Guelph, Guelph, ON, Canada

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

The word 'thermodynamics' is derived from $\vartheta\epsilon\rho\mu$os (heat) and $\delta\upsilon\nu\alpha\mu\iota$s (force, power). It presents the science of all forms of energy and mass, including entropic ('waste') heat contained in the mass at ambient temperature. The origin of the word suggests that this branch of science deals with both the statics (equilibrium) and the dynamics (nonequilibrium) of energy and mass. Statics (classical thermodynamics) deals with the state of the system in which no heat or mass transfer occurs (this state is extremely rare in natural soils). Dynamics (nonequilibrium thermodynamics) deals with transport processes of mass and heat.

'Soil water' is often used interchangeably with the term 'soil moisture.' Here we distinguish between the two terms: 'soil water' indicating the chemical component H_2O in the soil and 'soil moisture' meaning the soil solution. Thermodynamics distinguishes the different chemical components in the system. It deals with the interaction between and transfer of these components and heat.

Classical (Equilibrium) Thermodynamics of Soil Water

The birth of thermodynamics is rather confusing. The founders tried to deal with such things as the effectiveness of steam engines (S. Carnot). R. Meyer, who was the first (in 1842) to publish the equivalence of work and heat, based his conclusion on his observations of the degree of 'redness' of human blood. No wonder that thermodynamics had a shaky start. The first mathematical formulation of the 'First Law of Thermodynamics' (the law of conservation of energy) came from Helmholtz and Joule in 1847. These scientists were dealing with systems that were 'on the move.' Yet the system had to be at equilibrium and the 'changes' had to be infinitesimally small and 'reversible' (meaning that the changes can be reversed without any loss of useful energy at the ambient temperature, or entropy production).

One would expect the First Law of Thermodynamics to equate integral values of the different forms of energy. Yet the equations always come in differential form first. At first only 'closed' systems were considered, implying that no mass could move into or out of the system. It was not until J. Willard Gibbs published his two great treatises, *The Equilibrium of Heterogeneous Substances* (in the *Transactions of the Connecticut Academy*, 1876 and 1878), that the 'system' was opened up and many of the mysteries of the previous 50 years were clarified. The *Transactions of the Connecticut Academy* were not widely read in Europe, and it was not until C.N. Lewis published his famous book in 1923, 20 years after Gibbs' death, that Gibbs' works became widely known. In the year of his retirement, Gibbs remarked that, during the 30 years of his teaching at Yale University, he estimated that only half a dozen of his students had benefited from his lectures.

We start with the integrated Gibbs equation:

$$E = U + \Psi = TS - PV + \sum_{i} \mu_i m_i + \sum_{i} \psi m_i \qquad [1]$$

where E is the energy of the system, U is the 'internal' energy, and Ψ is the energy derived from external force fields.

U consists of three 'blobs' of different forms of energy: the first one (TS) being the entropic energy (heat in the system at the ambient temperature T); the second one ($-PV$) being the energy the system has lost in order to create a volume, V, for itself to exist at the ambient pressure P; the third one being the sum of the chemical energy of each component (i) in the system, with m_i being the mass of component i and μ_i being the chemical potential of component i.

Ψ is derived from external force fields such as gravity or an electrical or magnetic field (in the gravity field $\Psi = gh$, g being the gravitational acceleration, and h the height above the reference level). If other forms of energy (such as kinetic energy or viscous stress energy) play a role in the transformations of energy, they should be added to eqn [1]. Although kinetic and viscous stress energy play a dominant role in energy transformations in the soil solution on the Navier–Stokes scale, on the Darcy scale they are almost always negligible and will therefore be ignored here.

The 'free' energy, G, was defined by Gibbs as:

$$G \equiv U - TS + PV = \sum_i \mu_i m_i \qquad [2]$$

Thus, μ_i is the specific (per unit mass) Gibbs free energy of constituent i.

The first expressions of the First Law were all in differential form, even after Gibbs 'opened up' the system:

$$\mathrm{d}U = T\mathrm{d}S - P\mathrm{d}V + \sum_i \mu_i \mathrm{d}m_i \qquad [3]$$

This is the differential form of the Gibbs equation. It is not trivial, because if one carries out the complete differentiation of eqn [1], one finds that, after subtraction of eqn [3] there is a 'leftover':

$$S\mathrm{d}T - V\mathrm{d}P + \sum_i m_i \mathrm{d}\mu_i = 0 \qquad [4]$$

This leftover is known as the Gibbs–Duhem equation.

From eqn [4] it follows that:

$$\mathrm{d}\mu_w = Sm_w^{-1}\mathrm{d}T - Vm_w^{-1}\mathrm{d}P + \sum_{(k\neq w)} m_k m_w^{-1}\mathrm{d}\mu_k \qquad [5]$$

This is the central equation in classical thermodynamics of soil water. It dictates how the chemical potential (specific Gibbs free energy) of water in the soil changes with the temperature, the pressure, and the composition of the soil solution.

The guiding principle is now that, when the system is in equilibrium, $\mathrm{d}\mu_w = 0$. The specific Gibbs free energy of water is the same everywhere, or, as far as water is concerned, there is no transport and no entropy production. In short, the water is at rest. The inverse statement, "when $\mathrm{d}\mu_w = 0$, the water is at rest," is not necessarily true and is not a fundamental principle of thermodynamics. It is possible that the last two terms of the right-hand side of eqn [5] (temporarily) balance each other, making $\mathrm{d}\mu_w = 0$, while both water and the solute are on the move and entropy is being produced. This is possible when 'leaky' semipermeable membranes are present. Heavy clay soils are excellent examples of leaky semipermeable membranes.

Therefore, the concept of the 'total' potential of the soil water (the sum of all the component potentials) is not very useful. When the whole system is at rest (complete equilibrium), then the total potential of the water is equal everywhere in the system.

But, when the total potential of the water is equal everywhere, this does not necessarily mean that the water is at rest. The gradient of the total potential is not the appropriate driving force in the flux equation of water. This will become clearer in the section on nonequilibrium thermodynamics, below.

The apprentice in thermodynamics often asks: "What is the practical use of all this?" Thermodynamics is still not a coherent, comprehensive science. Developments are still happening haphazardly. Some scientists, such as Clifford Truesdell, have complained bitterly about the inconsistencies and the lack of mathematical rigor, using words such as "thermodynamics was approached through detours in the fog of word-play." Here, only two of the practically useful results will be discussed.

The Tensiometer (and the Pressure Membrane Apparatus)

Early scientists such as L.A. Richards, G.H. Bolt, and M.J. Frissel, working on the thermodynamics of soil moisture, immediately realized that water in the soil is present in the soil solution and that the soil solution is never pure water. The soil solution contains dissolved salts and ions, as well as counterions, that is, a surplus of cations, necessary to counterbalance the surplus of electrons present in the lattices of clay particles. The electrical double layer acts as an osmometer and the counterions cause a volume element of the solution, close to a clay surface, to carry a volume charge. This gives the soil solution an additional opportunity to possess energy (in the presence of an electrical field). In addition, there are microforce fields acting on the solution that is in close proximity of solid surfaces (London–Van der Waals forces). All this implies that eqn [5] should contain a few more terms if the thermodynamic system were chosen on the Navier–Stokes scale.

The first requirement in soil physics is to formulate and assess the hydrostatics of soil water, i.e., establish

the characteristic relationships when the water is at rest. There are no 'membranes' in nature or in the laboratory that are semipermeable to heat. Thus, when a system is brought to equilibrium, the temperature will be the same everywhere. The energy level in the external force fields, such as gravity, can easily be established separately. In order to measure the energy level of soil water derived from internal force fields (menisci, electrical double layers, London–Van der Waals forces), it is now advantageous to bring the soil solution in equilibrium with either a compartment containing a solution that reflects the composition of the soil solution or one that contains pure water. There are membranes that are ('perfectly') semipermeable to all dissolved materials in the soil solution. In the latter case (using a perfect osmometer and measuring the hydrostatic pressure in the osmometer with respect to atmospheric pressure), the energy level of the soil water is measured at the ambient temperature and at the particular position in the external force field that results from all internal force fields plus all dissolved materials (the total osmotic pressure of the soil solution). In the former case, using a perfectly 'leaky' membrane such as the sintered ceramic cup (tensiometer cup) or a Visking membrane in a pressure plate apparatus, the solution inside the compartment (tensiometer cup) will reflect the composition of the 'free salts' in the soil solution and will show the 'equilibrium' solution. By measuring the hydrostatic pressure in a tensiometer with respect to atmospheric pressure, the energy level of all internal force fields is measured (including electrical double layers), excluding the free salt. This energy level is called the 'matric potential.' The counterions cannot equilibrate in the tensiometer cup. They act as internal miniosmometers and can be considered part of the matrix (the solid phase of the soil). Because most soils are quite 'leaky' as far as the exclusion of dissolved materials is concerned, the water in the soil moves largely as a solution. Therefore the gradient of the matric potential is the dominant driving force on the soil water (in addition to the force of the gravitational field). The imperfect semipermeability of a clay soil, causing osmotic pressure differences to drive the water, will be dealt with in the section Nonequilibrium Thermodynamics of Soil Water, below. The action of osmotic pressure gradients will then be formulated as 'capillary' osmosis and the imperfect semipermeability will be evaluated by the 'reflection coefficient.'

The Maxwell Relations and Their Use in the Hydrostatics of Swelling Soils

One of the many complaints about the lack of mathematical rigor lies in its origin: a closed pot on a bench in a chemistry laboratory. Temperature and pressure are allowed to change in steps, ΔT and ΔP, although only in small steps, δT and δP. But entropy production is not allowed and therefore the steps have to be infinitesimally small, dT and dP. They begin to look like a differential, but differential with respect to what? Time or space? Everything becomes cleaner when the denominator of the differential is specified. This first comes about, while still in the realm of classical thermodynamics, when a cross-differentiation is performed on the differential form of the Gibbs equation. This leads to the Maxwell relations, several of which can be constructed. Many more can be found after performing Legendre transforms on the Gibbs free energy. Most of the Maxwell relations are quite useless, but occasionally one finds a gem. One example is a Maxwell relation for shrinking soils. For many decades soil scientists have searched for an expression for the differential of the matric potential (tensiometer pressure), p, with respect to the load (or overburden pressure), P, at constant ratio of mass of water and mass of solids. Here, both the dependent and the independent variable are intensive variables. The appropriate Maxwell relation shows that the above differential is equal to the differential relationship between the void ratio (volume of voids per volume of solids), e, and the moisture ratio (volume of water per volume of solids) at constant load pressure, P. Here, both the dependent and the independent variable are extensive variables.

Thus:

$$\mathrm{d}p(\mathrm{d}P)^{-1} \text{ at constant } m_{\mathrm{w}}(m_{\mathrm{solids}})^{-1} = \mathrm{d}e(\mathrm{d}\vartheta)^{-1} \quad [6]$$

at constant P. This means that the behavior of the tensiometer reading can be predicted upon loading from the slopes of the shrinkage lines for different values of P. Because the shrinkage lines for different values of P are nearly parallel, one only needs to measure the unloaded shrinkage line in order to predict how the tensiometer reading changes when moving deeper into the soil profile or upon loading by machinery.

Nonequilibrium Thermodynamics of Soil Water

The thermodynamic 'system' is now chosen to be a unit volume element in the soil solution. All the extensive variables are expressed per unit volume. Subsequently, each of the terms of the differential form of the Gibbs equation (eqn [3]) are written as differentials with respect to time.

The mathematical sloppiness of classical thermodynamics disappears once the terms of the differential

form of the Gibbs equations are transformed into proper time differentials and the realm of nonequilibrium thermodynamics is entered.

Next, for each of the terms separately, an appropriate conservation (continuity) equation is constructed. The general form of such a conservation equation shows that the change of the content of the entity of concern, with time, is equal to the negative of the divergence of the flux of that entity (inflow minus outflow), plus or minus one or more source or sink terms. Thus, for the entropic energy term:

$$\mathrm{d}\left(SV^{-1}\right)\mathrm{d}t^{-1} = -\mathrm{div}\ j_s + \sigma \qquad [7]$$

where SV^{-1} is the entropy per unit volume of solution, j_s is the flux of entropy, and σ is the entropy production term. This latter term plays a central role in nonequilibrium thermodynamics. If all the processes and forms of energy are accounted for, nothing is counted twice, and all algebra is carried out correctly, then the entropy production term is nonnegative. As the absolute temperature is always positive, the energy dissipation term, $T\sigma$, must also be nonnegative. When σ and thus $T\sigma$ are zero, the system is at equilibrium and nothing moves. If anything moves, then entropy is produced, and σ and $T\sigma$ are positive. Energy is being dissipated; that is, energy is transformed from useful energy to waste energy (i.e., heat at the ambient temperature). This is the Second Law of Thermodynamics.

As the first term on the right-hand side of Eqn [3] is already multiplied by T, one finds, after replacing the time differentials of the different terms in Eqn [3] by the right-hand side of their appropriate conservation equations, the product $T\sigma$ as the source term of entropy conservation equation. After all the appropriate substitutions are made, the term $T\sigma$ is singled out (written explicitly) and the equation is then called 'the (energy) dissipation function.'

On the right-hand side of the dissipation function, one usually finds the sum of a number of products of fluxes and forces. One may reshuffle these forces and the fluxes, under the strict principles that the rules of algebra are carried out correctly and that nothing is forgotten or counted twice.

The resulting complete expression for the energy dissipation is then the sum of products of fluxes and conjugated forces:

$$T\sigma = -j_q(\mathrm{grad}\ T) - j^V(\mathrm{grad}\ H) - j^D(\mathrm{grad}\ \pi) - j^E(\mathrm{grad}\ E) \qquad [8]$$

where j_q is the caloric (Fourier) heat flux, j^V is the (Darcy) flux of the soil solution, j^D is the (Fick) diffusion flux, and j^E is the electric current (usually indicated by I). grad π is the gradient of the osmotic pressure, and grad E is the gradient of the electrostatic potential.

The primary forms of energy dissipation are the dissipation of heat, pressure, mixing, and electrical energy. If simple linear, homogeneous transport equations are constructed relating a flux to its conjugated flux (the one that is occurring in the same product), the transport equations of Fourier (1822), Darcy (1856), Fick (1855), and Ohm (1827) are obtained. The disciplines of physics and chemistry have long recognized that 'coupled' transport can occur. Examples of such coupled transport processes are osmosis, electro-osmosis, thermo-osmosis, the Peltier effect, etc. A coupled transport process takes place when a flux arises due to a nonconjugated force (that is, a force that occurs in a product other than the one in which the flux of concern occurs).

Nonequilibrium thermodynamics now postulates that each flux occurring in the energy dissipation function is a linear, homogeneous (no intercept) function of all forces occurring in the same equation for the total energy dissipation. Thus:

$$j_q = -L_{TT}(\mathrm{grad}\ T) - L_{TV}(\mathrm{grad}\ H) - L_{TD}(\mathrm{grad}\ \pi) - L_{TE}(\mathrm{grad}\ E) \qquad [9]$$

$$j^V = -L_{VT}(\mathrm{grad}\ T) - L_{VV}(\mathrm{grad}\ H) - L_{VD}(\mathrm{grad}\ \pi) - L_{VE}(\mathrm{grad}\ E) \qquad [10]$$

$$j^D = -L_{DT}(\mathrm{grad}\ T) - L_{DV}(\mathrm{grad}\ H) - L_{DD}(\mathrm{grad}\ \pi) - L_{DE}(\mathrm{grad}\ E) \qquad [11]$$

$$j^E = -L_{ET}(\mathrm{grad}\ T) - L_{EV}(\mathrm{grad}\ H) - L_{ED}(\mathrm{grad}\ \pi) - L_{EE}(\mathrm{grad}\ E) \qquad [12]$$

The 'phenomenological' coefficients in the transport equations [9–12] are indicated by the letter L with two subscripts, say K and M. The first one (K) indicates which entity is being transported. The second one (M) indicates which driving force causes the transport. The terms for which the coefficient has identical subscripts represent the 'straight,' well-known transport processes, the laws of Fourier (1822), indicated by L_{TT}, Darcy (1856), indicated by L_{VV}, Fick (1855), indicated by L_{DD}, and Ohm (1827), indicated by L_{EE}. All other coefficients are known as 'coupling' coefficients, representing 'coupling' processes. They come in pairs, a pair being indicated by L_{KM} and L_{MK}. The two coefficients of a pair are called 'twin' coefficients. They represent all possible coupling phenomena.

If the energy dissipation equation is complete and all algebra has been carried out correctly, then the two twin coefficients are equal:

$$L_{KM} = L_{MK} \quad [13]$$

Thus the matrix of coefficients in eqns [9–12] is symmetrical. In the above matrix there are six of these pairs of equal twins. These equalities are known as the Onsager reciprocal relations (ORRs).

The fundamental value of the procedure outlined above is that all possible forms of energy dissipation are accounted for, even if they have never been observed.

For soil physics these forms are extremely important. After the merciless denigration of the ORRs by Truesdell, soil physicists should rise up and accept the validity and the great practical usefulness of ORRs. The framework of the matrix of coefficients in eqns [9–12] is comparable, in nature but not quite in stature, to the periodic system as proposed by Mendeleyev. For his systematic framework, Onsager received the Nobel Prize in Chemistry (1968), but Mendeleyev never received this accolade. The first such prize in chemistry was awarded to Jacobus van't Hoff, even though Mendeleyev was still alive. The refusal by the Nobel Committee to award the prize to Mendeleyev continued until he died in 1907. The framework makes it possible, in case the measurement of a certain coefficient is difficult or expensive, to measure its twin, which often appears to be easier or cheaper to measure.

When building models for transport processes in clay soils, e.g., based on the electrical double-layer theory, these ORRs can be used to verify the correctness of the model: if the cross-coefficients are not equal, researchers can be assured that somewhere they have made a mistake.

The actual occurrence of a coupled transport process always relies on some kind of a selection mechanism, such as a mechanism that can select between molecules of different chemical nature, e.g., water and salts, or a mechanism that can select between 'hot' and 'cold' molecules, e.g., a liquid–gas interface. It should be noted that convection (advection) is never a selection mechanism.

The coupled transport processes are discussed here in pairs of twin phenomena and indicated by their (equal) coefficients:

- L_{TV} (thermofiltration) and L_{VT} (thermo-osmosis). Both these phenomena are very common in soils. The transport of heat due to a water potential gradient in the absence of a temperature gradient and the transport of water due to a temperature gradient in the absence of a water potential gradient are occurring constantly in the soil. In unsaturated soil they are due to evaporation and condensation (the selection mechanism here is the heat of vaporization–condensation). In saturated soils the magnitude of these phenomena is very small (the selection mechanism is now the heat of wetting). In frozen soils they are caused by freezing and melting (here the selection mechanism is the heat of freezing–melting and solidification–sublimation);
- L_{TD} and L_{DT} (thermodiffusive processes). These phenomena are known as, respectively, the Dufour effect and the Soret effect. The magnitude of these effects in soils is very small, but their occurrence is quite probable;
- L_{TE} and L_{ET} (thermoelectric processes). These phenomena are known as, respectively, the Peltier effect and the Seebeck effect. The magnitude of these phenomena in soils is small, but their occurrence is quite likely;
- L_{VD} (osmosis, also called capillary osmosis) and L_{DV} (reverse osmosis or salt sieving). These phenomena are very common in soils. The magnitude is directly related to the clay content of the soil. (The selection mechanism lies with the electrical double layer.) Clay particles expel negative ions, and therefore they expel dissociated salts (negative adsorption). The longstanding conflict as to whether the osmotic pressure (potential) should be added to the hydraulic potential of water to produce the 'total' potential of water, the gradient of which then serves as the driving force on the water, is here resolved. As the value of L_{VD} is almost always smaller then the value of L_{VV} (except for a perfectly semipermeable membrane), the concept of the total potential is useless. The ratio L_{VD} to L_{VV} is known as the 'reflection coefficient';
- L_{VE} (electro-osmosis) and L_{EV} (streaming current). These electrokinetic phenomena also find their cause in the existence of electrical double layers. The most commonly observed result of the effects is the streaming potential in clay soils;
- L_{DE} (electrophoresis) and L_{ED} (diffusion current). These phenomena, together with osmosis and reverse osmosis, electro-osmosis, and streaming current are extensively discussed in the literature. The magnitude of all six coefficients on the basis of electrical double-layer theory has been calculated.

Modern Development

The branch of science called 'nonequilibrium thermodynamics' is now split into two subbranches. The first one is now called 'linear nonequilibrium thermodynamics.' Indeed all transport equations presented

here are linear. This subbranch of science deals with processes that are 'not far from equilibrium.' The second subbranch is now called 'nonlinear nonequilibrium thermodynamics.' Its great proponent is Ilya Prigogine, who received the Nobel Prize in Chemistry for his work in 1977. This subbranch of science deals with 'violent' processes ('far away from equilibrium'), such as Liesegang rings and atom bombs. The natural transport processes in the soil are never violent, and, as long as the Darcy law (one of the primary energy-dissipating processes) holds, all coupling processes fall in the park of 'not far from equilibrium.' However, if there are cases where the Darcy law breaks down and turns into the nonlinear Forchheimer equation (that is, when the water in the soil becomes turbulent), soil physicists may have to turn their attention to this latest branch of thermodynamics.

See also: **Darcy's Law**; **Heat and Moisture Transport**; **Hydrodynamics in Soils**

Further Reading

Bolt GH and Frissel MJ (1960) Thermodynamics of soil moisture. *Netherlands Journal of Agricultural Science* 8: 57–78.

Gibbs JW (1961) *The Scientific Papers of J. Willard Gibbs,* vol. I, *Thermodynamics*. Mineola, NY: Dover Publications.

Glansdorff P and Prigogine I (1971) *Thermodynamic Theory of Structure, Stability and Fluctuations*. Chichester, UK: John Wiley.

Groenevelt PH (1971) Onsager's reciprocal relations. *Search* 2: 264–267.

Groenevelt PH (1979) Transport processes. In: Fairbridge RW and Finkl CW (eds) *The Encyclopedia of Soil Science*, part 1, pp. 578–581. Stroudsberg, PA: Daudon, Hutchison & Ross.

Groenevelt PH and Bolt GH (1969) Non-equilibrium thermodynamics of the soil-water system. *Journal of Hydrology* 7: 358–388.

Groenevelt PH and Bolt GH (1972) Water retention in soil. *Soil Science* 113: 238–245.

Groenevelt PH and Elrick DE (1976) Coupling phenomena in saturated homo-ionic montmorillonite. II. Theoretical. *Soil Science Society of America Journal* 40: 820–823.

Groenevelt PH and Grant CD (2001) Re-evaluation of the structural properties of some British swelling soils. *European Journal of Soil Science* 52: 469–477.

Groenevelt PH and Grant CD (2002) Curvature of shrinkage lines in relation to the consistency and structure of a Norwegian clay soil. *Geoderma* 106: 235–245.

Groenevelt PH and Kay BD (1974) On the interaction of water and heat transport in frozen and unfrozen soils. II. The liquid phase. *Soil Science Society of America Proceedings* 38: 400–404.

Groenevelt PH, Grant CD, and Semetsa S (2001) A new procedure to determine soil water availability. *Australian Journal of Soil Research* 39: 577–598.

Katchalsky A and Curran PF (1965) *Non-Equilibrium Thermodynamics in Biophysics*. Cambridge, MA: Harvard University Press.

Kay BD and Groenevelt PH (1974) On the interaction between water and heat transport in frozen and unfrozen soils. I. Basic theory: the vapor phase. *Soil Science Society of America Proceedings* 38: 395–400.

Pippard AB (1964) *The Elements of Classical Thermodynamics*. Cambridge, UK: Cambridge University Press.

Truesdell C (1969) *Rational Thermodynamics*. New York: McGraw-Hill.

Tillage *See* **Conservation Tillage**; **Cultivation and Tillage**; **Zone Tillage**

TILTH

D L Karlen, USDA Agricultural Research Service, Ames, IA, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Tilth is defined as the physical condition of soil as related to its ease of tillage, fitness as a seedbed, and its promotion of seedling emergence and root penetration. It is also an important indicator of soil quality, because it relates to soil structure in terms such as 'mellowness' or 'friability.' Soils with poor tilth are described as appearing lifeless, resembling brick or concrete, or being cloddy. If soils with poor tilth have a high percentage of sand, they will often disaggregate or separate into their primary sand-, silt-, and clay-size particles with little or no tendency to bind together. Soils with poor tilth can also be massive and difficult to till, work into a suitable seedbed or seed. Soils with medium tilth are characterized as being

somewhat cloddy, with a tendency to 'ball up' and be 'rough pulling' (i.e., having a high draft or requiring a greater energy input) when worked into a seedbed. Soils with good tilth are described as being crumb-like, easy to slice (like cutting butter), and spongy. Defining the descriptive terminology for soil tilth is a challenging task for those specializing in soil management. Tilth is difficult, if not impossible, to quantify. A common description of this dilemma is that tilth is what all farmers can measure through the soles of their boots, but no soil scientist can define.

History of Tilth

The concept of soil tilth predates modern agriculture, but a clear definition has never evolved because of differing perceptions. For example, an old childhood story explaining the concept of soil tilth describes how, in his dying moments, an elderly man called his sons to his side to tell them that if they would dig diligently in the garden, they would find a hidden treasure. The sons did so, but alas, they found no treasure of silver or gold. The subsequent harvest from this much-worked garden, however, was so large that the father's meaning gradually dawned upon his sons and they realized the wisdom of his words. A sixteenth-century book entitled *Boke of Husbandry*, written to provide instructions for growing peas and beans, tells the farmer how to determine whether the soil is ready for planting. It states that he should walk on the plowed ground and:

> if it synge or crye, or make any noyse under thy fete then it is to wete to sowe; and if it make no noyse, and wyll beare thy horses, thanne sowe in the name of God.

In the 1930s, several authors described how climatic conditions and various soil-, crop-, and animal-management practices affected soil tilth. By sieving soil samples and separating the material into different sizes, they showed how freezing, drying winds, rainfall, tillage, animal compaction, and other management factors affected tilth. Eventually, 'soil tilth' was used as a 'blanket' term, describing all the soil conditions that determine the degree of fitness of a soil as an environment for the growth and development of a crop plant. Soil structure was identified as a key factor in determining soil tilth, because structure is a soil property that can be altered rapidly through tillage operations and changes in environmental factors such as rainfall or temperature.

Soil tilth gradually came to be recognized as a dynamic condition. As virgin grassland or forest soils are cultivated, the organic matter concentrations, base (cation) saturation, and porosity decrease, while bulk density increases. These changes reduce the granulation or tendency of the soil to form stable aggregates and gradually degrade soil structure. Declining soil tilth is associated with increased runoff, erosion potential, need for tillage to prepare an adequate seedbed, and fertilizer to sustain reasonable and profitable crop yields.

Soil-management research through the mid twentieth century focused on identifying basic properties and processes that affect soil structure and tilth (i.e., flocculation, cementing by organic and inorganic colloids, wetting and drying cycles, freezing and thawing, organic matter cycling, biological activity, and tillage). One result of these efforts was that, in 1959, the 86th US Senate passed Senate Document No. 59, confirming the need:

> to conduct basic research on the relation of the physicochemical nature of soils, the role of organic matter, the activities of microbes, and the effects of mechanical manipulation upon the structural attributes of soils; [directed] toward predicting the effects of soil management practices [and] providing an optimum environment for root growth of crops in different kinds of soils.

However, 30 years elapsed before the US Department of Agriculture (USDA) Agricultural Research Service established the National Soil Tilth Laboratory in Ames, Iowa. Significant progress has been made in understanding many of the scientific aspects of soil structure, organic matter cycling, activities of microorganisms, and effects of various soil- and crop-management practices. However, with regard to adoption of improved soil-management practices, economic pressures and political forces beyond the control of most farmers often contribute to declining soil tilth.

Soil Organic Matter and Tilth

Organic matter is one of the main factors affecting soil tilth because of its role in aggregation (**Figure 1**). Studies with disturbed and undisturbed soils show that, as organic matter increases, compaction decreases and there is more space for air and water exchange. The constant addition of organic matter throughout the soil profile is one reason why virgin grassland soils generally have excellent aggregation and structure. Those soils, especially in humid regions, generally have a high organic matter content, a high percentage of water-stable aggregates, and a structure that accommodates rapid plant growth with unlimited root proliferation. These soil characteristics occur because in humid regions grasses replace most of their roots and top growth each year. Also, when the aboveground plant material dies, it falls on the soil surface, where it is decomposed or mixed into the upper part of the soil by earthworms and other soil organisms. The dead

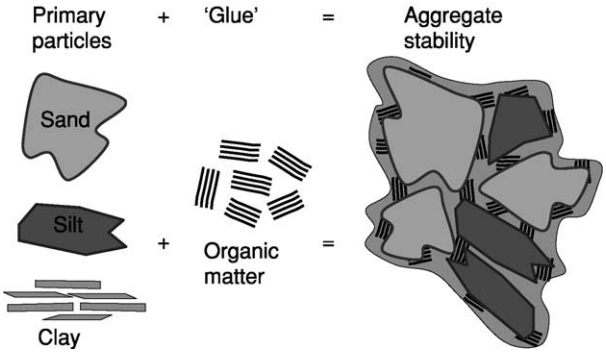


Figure 1 A conceptual illustration of soil aggregation, defined as the process whereby primary soil particles (sand, silt, and clay) are bound together, usually by natural forces and substances (glues) derived from root exudates, microbial activity, and other organic matter sources.

roots and plant residues provide a food source for fungal and bacterial decomposition processes, leading to the formation of soil aggregates, which modify effects of soil texture with regard to water and air relationships, and root penetration. Increased water-stable aggregation, annual proliferation of plant roots, and the associated wetting and drying cycles gradually develop a soil structure that is stable and, with good management, generally resists degradation.

The regular addition of organic materials can improve tilth by creating larger and more stable aggregates that are not easily eroded by wind or water. Organic matter, both as residue on the soil surface and as a binding agent for aggregates near the surface, strongly influences water infiltration, retention, and the potential for particle detachment and erosion. Surface residues intercept rainfall or irrigation water, and dissipate the energy before the drops hit the surface, detach soil particles through the splash, or begin to form a surface crust by filling the voids and cracks between the aggregates with fine particles. Organic materials on the soil surface also slow the water as it flows across the field, thus increasing opportunities for the water to infiltrate. Another characteristic of soils with good tilth is the presence of large channels or pore space between the water-stable aggregates. These channels greatly enhance the ability of soil to conduct water from the surface into the subsoil, where, because of enhanced root proliferation (another characteristic of good tilth), plants can subsequently use the water to support growth and development.

Soil-Management Effects on Tilth

Tilth, as an indicator of the physical condition of a soil, is strongly influenced by any soil management practice that quantitatively affects soil structure, aggregation processes, or soil organic matter cycling. Incorporating perennial crops into extended rotations, reducing tillage intensity and frequency, including cover crops, and optimizing nutrient and water management are among the most common practices that can be used to create the desired condition.

Surface mulch from crop residues, cover crops, animal manure, municipal sludge, leaves, or other carbonaceous sources provides both physical protection from raindrop impact and a food source for earthworms and other soil microorganisms. The surface mulch also moderates soil temperature and moisture extremes at the soil surface. If soils are left bare and exposed, surface temperatures can be very high, causing the soil to dry out. When this occurs, earthworms and other soil insects move deeper into the soil, leaving a surface zone that contains very few active organisms. Bacteria and fungi that live in thin films of water die or become inactive, slowing the natural process of organic matter cycling and aggregation.

Tillage Effects on Soil Tilth

Although tillage can improve tilth over the short term by loosening surface soil, disrupting crusts, and creating a more favorable soil–water–air environment for plants, it is also a primary cause for the long-term deterioration of tilth. The loosening process that often has positive short-term plant-production benefits has negative long-term effects, because it increases the rate of organic matter decomposition through chemical and microbial mineralization. This occurs because previously protected organic matter is exposed by the tillage and subsequently used as a food source by microorganisms. The loss of organic matter decreases aggregate stability, increasing the potential for surface crusting, poor seedling emergence, runoff, erosion, and other indicators of poor soil tilth (e.g., decreased infiltration of rainfall or irrigation water, decreased soil water retention, and decreased nutrient cycling). Tillage itself or the wheel traffic associated with the operations can further reduce the quality of tilth by increasing compaction. This reduces total pore space and frequently increases water-filled pore space, leading to reduced aeration, slower nutrient cycling, and a decreased volume of soil for plant root proliferation. These processes obviously occur in fields used for crop production, but they have similar negative effects on tilth at construction sites, in forests during logging operations, on campgrounds, athletic fields, or even desert areas where off-road recreational vehicles are allowed.

Tillage has been described as being 'addictive' for soils because of the vicious cycle that can be

established. Soils are tilled to disrupt surface crusts, compacted zones, or other real or perceived limitations to crop production. Doing so may provide short-term improvements in tilth and nutrient cycling, but these periods of improvement become shorter and shorter as more of the organic matter is lost through oxidation. Ultimately the soils may be 'burned out' and must either be abandoned until natural processes (e.g., wetting and drying; plant root proliferation, exudation, death, and decay; weathering) can restore the tilth or management practices (e.g., crop rotation, tillage, water management) are changed. As a result of these interactions, monitoring of soil tilth may provide a useful indicator of the sustainability of an agricultural system.

In situations where nitrogen is the most limiting factor, increasing nitrogen fertilization rates and therefore crop productivity has been shown to increase soil organic matter, even with moldboard plowing. However, without risking increased nitrogen leaching and potential water-quality degradation, development and adoption of reduced- or no-tillage agricultural practices can generally improve soil tilth even more by slowing organic matter decomposition, maintaining aggregate stability, and preserving soil structure. The critical management practice for these systems generally focuses on crop-residue management. Maintaining crop residues on the soil surface and the lack of loosening through tillage reduces dispersion of the surface aggregates by rainfall or runoff. By leaving crop stubble on untilled soil, the potential effects of wind erosion are minimized. Reducing or eliminating tillage also diminishes tillage erosion and keeps soil from being moved downhill by tillage tools.

Studies from around the world illustrate how tillage can affect soil tilth. In one study where direct drilling was compared with several methods of cultivation for wheat production, five soil series from southern England were examined. At each site, the tillage treatments had been in place for at least 4 years. The results showed that aggregates from the surface were more stable after direct drilling than after plowing. Most of the effects were attributed to a positive interaction between aggregate stability and organic matter content (i.e., two critical indicators of soil tilth). Direct drilling did decrease both aggregate stability and organic matter content in soils where grass had been grown before the treatments were imposed, but the decrease was less than that observed with plowing. Benefits of direct drilling were assumed to be greater root density in surface layers and less organic matter mineralization than in plowed soils. A similar response was observed in the USA when moldboard plowing and no-till crop-production practices were evaluated on land being taken out of the Conservation Reserve Program.

Extended Crop Rotations

Soil tilth is affected in two ways by incorporating forage or hay crops into an extended rotation. First, soil organic matter decomposition decreases because the soil is not being disturbed or tilled each year; second, grass and legume sods develop extensive root systems, part of which will naturally die each year, adding new organic matter to the soil. The dead roots provide a food source for fungal and bacterial decomposition processes, leading to the formation of soil aggregates and improved tilth.

To demonstrate the effect of perennial root systems on soil aggregation, researchers have compared aggregate size distributions for soils under continuous corn production with those where corn is grown in rotation with oat and meadow. The data show that, with continuous corn production, aggregate size is less than half that for corn grown in rotation. The studies also show that aggregation decreases slowly but consistently over a 7-year period of continuous corn production. In related studies, comparisons among crop sequences show that aggregation increases by switching to a corn–oat–meadow rotation after 11 years of continuous corn production. Also, it takes only 4 years of continuous corn production to decrease the aggregation established under either bluegrass or alfalfa to less than that found after 18 years of a corn–oat–meadow rotation. The highly significant correlation between crop yield and aggregate size confirms that soil tilth and productivity are both affected by crop rotation.

Cover Crops

Growing cover crops can improve soil tilth by decreasing erosion, increasing infiltration, and adding organic residues to the soil. These benefits are achieved because cover crops are often grown only during seasons when the soil is especially susceptible to erosion, although in orchards and vineyards the cover crops are frequently grown year-round. The leaves and stems of the cover crop intercept rainfall and dissipate its energy, while roots bind the soil and hold it in place. The amount of benefit from cover crops depends on the above- and belowground biomass and rooting structure of the plant being grown and the length of time before the soil is prepared for the next crop. For example, in the southeastern USA, crimson clover can be grown successfully as a cover crop prior to cotton. Initially planted in late autumn, the clover grows slowly throughout the winter and early spring, often reseeding itself for the next year

before the cotton crop is planted. In more northern areas such as the US Corn and Soybean Belt, where cold temperatures significantly shorten the growing season, warm-area crops such as crimson clover cannot survive. Oat, rye, and hairy vetch are some of the possible cover crops for the northern states of the USA, but the amount of protection against soil erosion and especially the amount of nitrogen supplied by the vetch are generally much less than in the southern USA. Also, in semiarid regions, where water is often the most limiting factor, growing cover crops may not be a viable management practice.

The quantity of organic residues returned to the soil by cover crops varies depending on plant species and the length of the growing season. Depending on temperature and moisture conditions, production can vary from less than 1000 kg ha^{-1} to more than 10 000 kg ha^{-1}. With low amounts of residue, there may be some detectable changes in active organic matter (e.g., particulate organic matter), but, because the material decomposes so rapidly, there generally will be no detectable change in total soil organic matter. However, a 5-year study in California, USA, has shown that growing a clover cover doubles total organic matter in the top 5 cm of soil (13–26 g kg^{-1}) and increases it by 2 g kg^{-1} at the 5–15-cm depth.

Additional benefits of cover crops include sequestering nutrients (especially nitrate N) so they cannot be leached below the crop root zone, supplying nutrients (usually N) to the following crop, suppressing weeds, breaking pest cycles, and providing habitat for beneficial insects and other fauna. However, before a farmer or land manager decides to grow a cover crop to maintain or improve soil tilth, it is important to determine the objectives. Is the main purpose to add available nitrogen to the soil or to provide large amounts of carbon? Is erosion control during the late autumn and early spring the primary goal? Does the soil have a compaction problem that needs to be alleviated? Will the combination of cover crop, climate, and water-holding capacity of the soil result in excess water depletion that adversely affects the primary crop? Once the primary purpose for growing a cover crop has been determined, decisions on plant species (i.e., legume or grass), the optimum date of planting, and when to eliminate the cover crop will be easier to make.

Evaluating Soil Tilth

Several factors can cause soil tilth to change, including tillage, compaction due to wheel traffic or grazing, reconsolidation due to rainfall or irrigation, and cultivation. Qualitative observations of surface crusts, cloddiness, poor seedling emergence, ponding, restricted plant rooting, soil organism diversity and population (especially earthworms), and increased runoff, erosion, or compaction have traditionally been used to evaluate soil tilth. Among the more quantitative measurements that have been used are aggregation (e.g., mean weight diameter, percentage of water-stable aggregates, or aggregate-size distributions), penetration resistance, infiltration rates, and crop performance.

Penetration resistance is measured most precisely with a penetrometer, although a stiff wire, spade, or tile probe can also be used. For precise measurements, the force required to push a rod with a cone-shaped tip through the soil to a known depth is recorded. To minimize the variation and provide the most quantitative data possible, measurements are taken 24–48 h after rainfall or irrigation so that the soil moisture content is relatively uniform throughout the measurement zone. Penetrometer data are also more reliable in nonstructured sandy or sandy loam soils. In general, if the resistance exceeds 2070 kPa (300 psi) the soil will be too hard for root penetration. Although it is not possible to be quantitative with a wire, tile probe, or spade, it is still possible to use these tools to observe soil structure and qualitatively describe soil tilth. If the soil has good tilth, it is fluffy and granular, and separates easily along the various structural planes of weakness. If the soil structure is massive or compacted, the tilth, by definition, is poor.

Measurements of infiltration rates with a rainfall simulator or other less-sophisticated tools (i.e., single- or double-ring infiltrometers) are also useful for evaluating soil tilth. These measurements are based on the time required for a known volume of water to enter the soil. If a soil has good tilth, water entry will be moderately fast for the first 5–10 cm of water. If water continues to enter at a very high rate, it could indicate that the soil has no ability to retain the water or that there are cracks, crevices, fractures, root channels, worm holes, or other macropores carrying the water away. If water enters the soil at an extremely slow rate or not at all, tilth is generally in a less-desirable condition, perhaps due to compaction or surface crusting. Under those conditions, rainfall and irrigation water usually runs off, increasing the potential for soil erosion and further degradation of soil tilth.

Tilth Indices

Efforts to develop a tilth index have been initiated to improve soil management and provide guidance for custom or prescribed tillage. The need for a tilth index was envisioned because even though an experienced person may be able to determine whether a soil has good tilth by sight and feel, there was no available method to measure or quantify it. Furthermore, if soil tilth can be quantified, it may be possible to measure

Table 1 Soil measurements and corresponding coefficients developed to compute a soil tilth index[a] for Mollisols and Alfisols in the Midwestern USA

Indicator	*Description*	*Tilth coefficient*
Bulk density	Mass per unit volume of dry soil	1.0, for $BD \leq 1.3\ g\ cm^{-3}$ $-1.5 + 3.87 \times BD - 1.5 \times BD^2$, for $1.3 \leq BD \leq 2.1\ g\ cm^{-3}$ 0.0, for $BD \geq 2.1\ g\ cm^{-3}$
Penetration resistance (cone index)	A measure of soil strength, indicating ease of root penetration, plant growth, and yield	1.0, for $CI \leq 1.0$ MPa $1.012 - 0.002 \times CI - 0.01 \times CI^2$, for $1.0 \leq CI \leq 10.0$ MPa 0.0, for $CI \geq 10.0$ MPa
Soil organic matter content	The organic fraction of soil excluding nondecayed plant and animal residue	1.0 for $OM \geq 50\ g\ kg^{-1}$ $0.59 + 0.122 \times OM - 0.008 \times OM^2$, for $10\ g\ kg^{-1} \leq OM \leq 50\ g\ kg^{-1}$ 0.70, for $OM \leq 10\ g\ kg^{-1}$
Aggregate uniformity coefficient	Shape of the grain-size (aggregate) distribution curve (all of equal size gives an AUC 1.0)	1.0, for $AUC \geq 5$ $0.348 + 0.245 \times AUC - 0.023 \times AUC^2$, for $2 \leq AUC \leq 5$ 0.75, for $AUC \leq 2$
Plasticity index	The difference in water content between the liquid limit and the plastic limit of a soil (cohesiveness)	1.0, for $PI \leq 150\ g\ kg^{-1}$ $1.02 + 0.0009 \times PI - 0.00016 \times PI^2$, for $150\ g\ kg^{-1} \leq PI \leq 400\ g\ kg^{-1}$ 0.80, for $PI \geq 400\ g\ kg^{-1}$

BD, bulk density; CI, cone index; UC, uniformity coefficient; OM, soil organic matter; PI, plasticity index; AUC, aggregate uniformity coefficient.
[a]Tilth index $= CF_{(BD)} \times CF_{(CI)} \times CF_{(UC)} \times CF_{(OM)} \times CF_{(PI)}$.

seasonal variation and to understand the subtle effects of soil- and crop-management practices such as conservation tillage or extended rotations.

The initial effort to develop an index assumed that tilth of a mineral soil could be characterized by examining bulk density, strength, aggregate characteristics, organic matter content, and consistency of a soil simultaneously. Bulk density, cone index (penetration resistance or strength), and an aggregate uniformity coefficient were chosen to reflect soil properties that respond to short-term manageable; organic matter content was chosen to reflect long-term management effects. The plasticity index was used as a measure of consistency, which reflects the relative ease with which a soil can be deformed or ruptured. This index relates to soil water properties and is strongly influenced by inherent soil characteristics (predominantly clay type and amount). These five measurements or indicators were also chosen because they can be measured easily in the field or with a routine soil test (e.g., for organic matter; see Table 1).

The search for a better tilth index continued with the addition of soil-water content. Concurrently, and with an even greater emphasis on the biological processes that affect tilth, the concept of soil quality has emerged as a strategy to quantify the effects of various soil-management practices on physical, chemical, and biological properties and processes occurring within the soil (*See* **Quality of Soil**). In the USA, the concept evolved during the 1990s, with two distinct areas of emphasis, education and assessment, both based on principles of soil science. Information sheets about the various indicators (or measurements) that can be used to evaluate soil quality, a *Soil Biology Primer* and *Guidelines for Soil Quality Assessment in Conservation Planning*, were among the educational materials developed for those with minimal knowledge of soil resources. For qualitative evaluation, scorecards were developed to provide farmers with a self-assessment of their current soil and crop management practices. The scoring is relatively simple (e.g., poor, fair, good) and generally based on observations of tilth, earthworms, runoff or ponding of water, plant vigor, ease of tillage, crop growth and development, and yield. Soil-quality test kits have been developed to provide semiquantitative indicator data and guidelines have been written to help interpret the data. A more comprehensive framework that uses scoring curves to interpret a full suite of potential indicators has been developed to provide a more comprehensive evaluation of soil tilth or soil quality. The most important aspect of these efforts is that quantitative methods for characterizing soil tilth do appear to be feasible and with further refinement may be useful for guiding soil and crop management decisions toward more sustainable land-use practices.

Summary

The concept of soil tilth is often easier for a farmer to recognize than for a scientist to describe. Traditionally, tilth has been described in many ways,

usually focusing on the physical condition of a soil and how it responds to tillage, functions as a seedbed, or affects seedling emergence and plant root growth and development. Soil tilth is strongly influenced by soil organic matter and soil structure. Soil management, especially tillage practices, cover crops, and crop rotations, often affects soil tilth because of its effect on soil organic matter. Soil scientists, ecologists, and agricultural engineers are continuing to evaluate different biological, chemical, and physical measurements or indicators that can be used individually or combined into index values to describe tilth quantitatively. That information may then help to improve soil-management practices through technologies such as no-tillage agriculture, the use of cover crops, and more diversified and extended crop rotations.

See also: **Compaction**; **Conservation Tillage**; **Infiltration**; **Organic Matter:** Principles and Processes; **Quality of Soil**; **Structure**; **Texture**

Further Reading

Foth HD and Ellis BG (1997) *Soil Fertility*, 2nd edn. Boca Raton, FL: CRC Press.

Karlen DL, Erbach DC, Kaspar TC *et al.* (1990) Soil tilth: a review of past perceptions and future needs. *Soil Science Society of America Journal* 54: 153–161.

Karlen DL, Ditzler CA, and Andrews SS (2003) Soil quality: why and how? *Geoderma* 114: 145–156.

Magdoff F and van Es H (2000) *Building Soils for Better Crops*. Burlington, VT: Sustainable Agriculture Publications.

Schjønning P, Elmholt S, and Christensen BT (eds) (2004) *Managing Soil Quality – Challenges in Modern Agriculture*. Wallingford, UK: CABI.

Schwab GO, Fangmeier DD, Elliot WJ, and Frevert RK (1993) *Soil and Water Conservation Engineering*, 4th edn. New York: John Wiley.

Singh KK, Colvin TS, Erbach DC, and Mughal AQ (1992) Tilth index: an approach to quantifying soil tilth. *Transactions of the American Society of Agricultural Engineers* 35(6): 1777–1785.

TIME-DOMAIN REFLECTOMETRY

G C Topp, Agriculture and Agri-Food Canada, Ottawa, ON, Canada
T P A Ferré, University of Arizona, Tucson, AZ, USA

Canadian Crown Copyright © 2005, Published by Elsevier Ltd. All Rights Reserved.

Introduction

The vital role of water in maintaining the life of the landscape requires the development of techniques to monitor and sustain water supply and its quality. Time-domain reflectometry (TDR) is a recently developed technology for use in soil and landscape processes, which has become widely used in highly diversified applications. Despite the importance of water and ionic solutes in the mass and energy balances of the soil profile, it has been only in the last 20 years that rapid, *in situ*, nondestructive measurement of soil water content and ionic solute concentration has become possible in the form of TDR. This is an electromagnetic (EM) technique using radar principles at radio frequency where estimates of water content and electrical conductivity of soil can be made separately from the same wave. The effect of the soil and water on the propagation velocity of the EM wave is analyzed to provide a reliable measure of the water content. The bulk electrical conductivity (EC) is determined from an analysis of the rate of decrease in amplitude during the propagation of the wave in soil. The principles of EM wave propagation are also used for the design of soil probes and to indicate the precision of measurements. The strong interaction between water content and electrical conductivity usually hinders the independent and separate determination of these properties in soil. This difficulty is overcome with TDR where both measurements are from the same sampled region and with the same EM wave.

Water molecules have unique electrical properties that determine the electrical properties of soil. Because of the strong dependence of soil electrical properties on the amount of water in a soil and on the quality of that water, electrical and electromagnetic measurements can be very useful for characterizing the soil water content. One method with which these properties can be measured is TDR. TDR can be thought of as one-dimensional radar using radio waves that propagate along a wave-guide. This wave-guide is constructed from parallel metal rods that are inserted in the soil. As radio waves travel through any material, the characteristics of the wave are continuously changed by the medium through which the wave is traveling. These changes in wave properties are measured by TDR and a wave equation analysis is used to estimate two electrical parameters

of the soil: dielectric permittivity (dielectric constant) and electrical conductivity. These properties are then used to estimate the water content and the electrical conductivity of the soil, respectively.

As the name implies, time-domain reflectometry relies on the measurement of the travel time of reflected waves. In this way, it is similar to other radar applications. In TDR, the wave velocity is measured by recording the time difference between the entry of a wave into the soil and the return of the reflected wave from the end of the TDR probe. Currently available TDR instruments operate in the range of radio frequencies (10 to 1000 MHz). In this frequency range, the dependence of the wave velocity on the water content is not strongly influenced by the electrical conductivity of the pore water or the electrical properties of the soil solids. This is supported by direct measurements made since the mid-1970s that show that, while other soil factors such as density, texture, temperature, and soluble salts affect the TDR wave velocity, their effects are less pronounced than are the effects of water content. That is, the soil water content is the most important soil factor affecting the travel time measured by TDR. This property of TDR makes the method very effective and efficient for the measurement of soil water content.

Soil salinity is an issue of great importance in regard to irrigation and agricultural production in arid and semiarid regions. It has long been realized that the electrical conductivity of pore water can be related to pore-water salinity. However, the bulk soil electrical conductivity is not solely dependent upon the concentration and composition of salts in solution. This property is also highly dependent on the soil texture, the temperature, and the water content. TDR can be used to infer the bulk electrical conductivity of a soil. The energy loss of the transmitted signal is related to the bulk electrical conductivity of the soil. This energy loss is determined from the relative amplitudes of the signal source and that of the wave reflected from the end of the probe. TDR has the advantage of simultaneously measuring both the water content and the bulk electrical conductivity in approximately the same volume of porous medium. Under some conditions, this information can be used to infer the bulk electrical conductivity, and thereby the salinity, of the pore water.

In the 20-plus years since TDR was first applied to soil measurements, the method has become a *de facto* standard for soil water content measurement and monitoring. The popularity of the method for environmental monitoring and research arises from a combination of its accuracy in a wide range of soils and its relative ease of use compared with many other available techniques. TDR provides real-time, *in-situ* soil water content measurements. Multiple soil probes can be incorporated into a switching network connected to a single instrument and data logger allowing for remote automated monitoring. For most soils, the accuracy of measurements of volumetric water content change is within $\pm 0.02\,m^3\,m^{-3}$ without the need for soil-specific calibration, and similar absolute water contents can be achieved with calibration. There is considerable flexibility in the design and placement of TDR probes, allowing users to modify water content measurement networks to conform to the requirements of any specific study. Finally, because TDR measures the volumetric water content, the data are directly applicable to hydrologic water balance analyses with no need for the measurement of supporting soil parameters such as bulk density.

Measurement of Soil Water Content Using TDR

Measuring the volumetric water content of a soil is difficult. As a result, methods have been developed that infer the volumetric water content based on the water potential, bulk electrical conductivity, or dielectric permittivity. The water potential–volumetric water content relationship is nonlinear and hysteretic. The bulk electrical conductivity depends on many factors in addition to the water content. In contrast, the bulk dielectric permittivity of a soil is highly correlated with the volumetric water content. For these reasons, recent advances in the development of indirect methods of measuring the volumetric water content have focused on methods that rely on the measurement of dielectric permittivity. TDR has long been used as a laboratory method to measure the dielectric permittivity of liquids. In an unrelated field, TDR has been widely applied to the location of faults in buried cables. However, it was not until it was demonstrated that a single correlation existed between the bulk dielectric permittivity and the volumetric water content for a wide range of porous materials that these two distinct uses came together to give rise to the TDR method for soil water content measurement. Over the past 30 years, TDR has been applied at many scales and under a broad range of conditions spanning agricultural and forestry environments and seasonal conditions from summer to winter. In addition, the method has been improved and modified.

Most TDR instruments launch a fast-rise voltage step (rise time <200 ps) along a transmission line that is connected to a probe that is inserted in the soil or medium of interest. The voltage pulse propagates as a planar EM wave, traveling in the soil and guided by the conductors that form the probe. The properties of

the soil that govern the propagation of the TDR pulse are described collectively by the propagation constant of the soil, from which velocity of propagation, v ($m\,s^{-1}$), and attenuation coefficient (α) are derived. The TDR instrument records and displays the voltage returning to the instrument as a function of elapsed time following the launch of the voltage pulse. This gives rise to a characteristic plot known as the waveform.

The waveform is a record of the voltage that arrives back at the instrument after being reflected from the end of the probe. Commonly, the returning voltage, V, is displayed as a normalized voltage, known as the reflection coefficient (ρ), as a function of time (t) (Figure 1, where $V/V_0 = 1 - \rho$), where V_0 is the voltage of the transmitted pulse. From the TDR waveform it is possible to determine both v and α. As discussed above, v is used to infer the volumetric water content while α is used to determine the electrical conductivity. A portion of the TDR step pulse travels a distance, L_{cable} (m), along the connecting cables to the probe and then reflects back to the instrument. A portion of the energy that enters the probe travels an additional distance, L (m), along the probe, and then travels the total distance back to the instrument. Taking the difference in these travel times to be t(s), the velocity of the wave can be determined from the known probe length, L, as:

$$v = \frac{2L}{t} \qquad [1]$$

From electromagnetic wave propagation theory, v is given in terms of relative dielectric permittivity and electrical conductivity of soil, which can be expressed as:

$$v = \frac{c}{\sqrt{\epsilon_{ra}}} \qquad [2]$$

where c ($= 3 \times 10^8\,m\,s^{-1}$) is the velocity of light and other EM waves in vacuum, and ϵ_{ra} is the apparent relative dielectric permittivity or apparent dielectric constant that controls the rate of propagation of the step pulse. Combining eqns [1] and [2] gives:

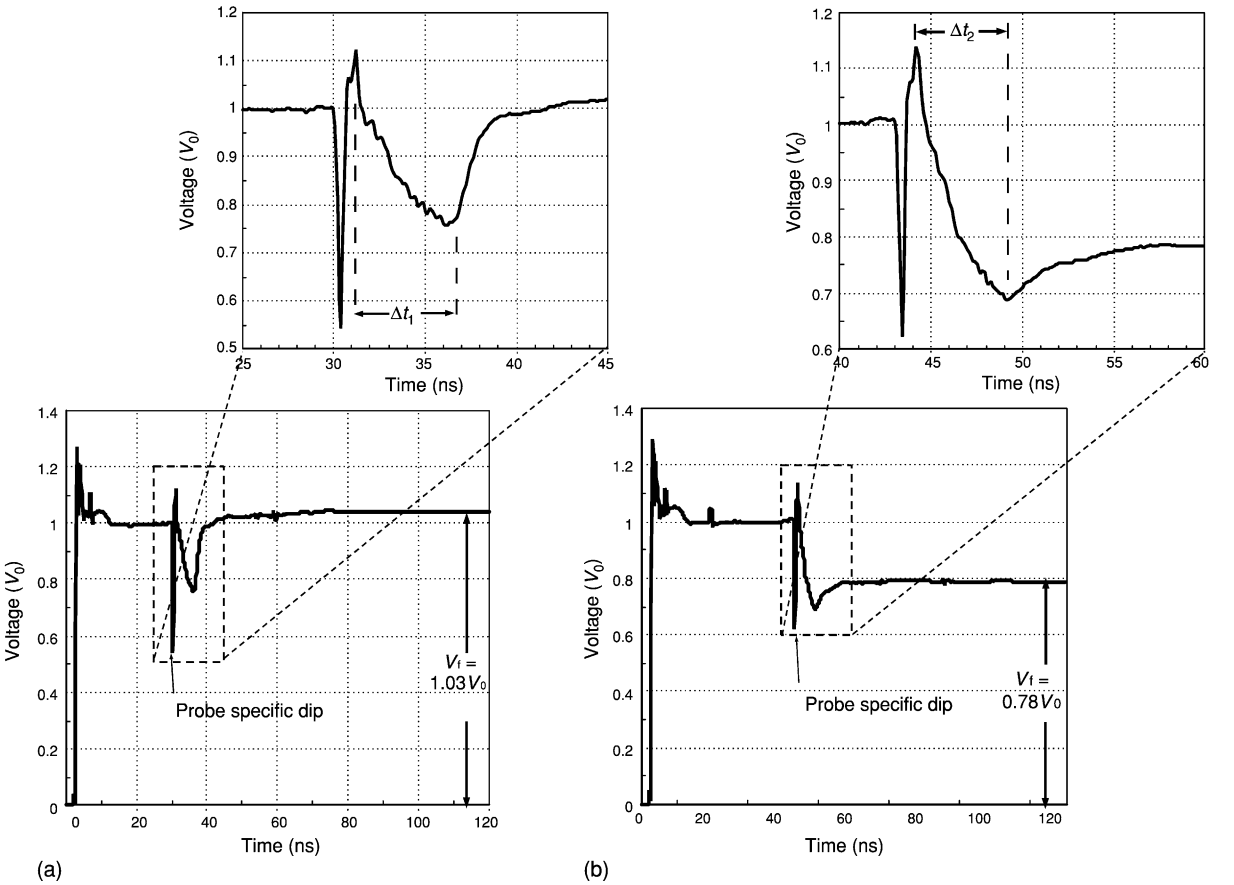


Figure 1 Two time-domain reflectometry curves from 20-cm probes in silty clay loam soil. The soils are at similar water content but the soil solution is more conductive in (b), giving a smaller return reflection and resulting lower V_f. In (a) $\theta_v = 0.304\,m^3\,m^{-3}$ and $\sigma_0 = 57\,mS\,m^{-1}$ and in (b) $\theta_v = 0.271\,m^3\,m^{-3}$ and $\sigma_0 = 95\,mS\,m^{-1}$.

$$\epsilon_{ra} = \left(\frac{ct}{2L}\right)^2 \quad [3]$$

More complete analyses of the EM wave equations show that ϵ_{ra} is primarily the result of the relative dielectric permittivity (real component) or dielectric constant of the soil, ϵ'_r. Also, ϵ_{ra} is affected to some extent by the electrical conductivity, σ_0 ($S\,m^{-1}$), the frequency of the signal, ω ($2\pi\,s^{-1}$), and the relative dielectric loss or imaginary component of relative dielectric permittivity, ϵ''_r. The effect of these factors on EM waves is expressed as:

$$\epsilon_{ra} = \frac{\epsilon'_r}{2}\left[1 + \sqrt{1 + \left(\frac{\epsilon''_r + \frac{\sigma_0}{\omega\epsilon_0}}{\epsilon'_r}\right)^2}\right] \quad [4]$$

where ϵ_0 is the dielectric permittivity of free space ($= 8.85 \times 10^{-12}\,F\,m^{-1}$). Most TDR applications in soil have assumed that $(\epsilon''_r + \sigma_0/\omega\epsilon_0)/\epsilon'_r << 1$, which leads to $\epsilon_{ra} \approx \epsilon'_r$. This assumption allows for the determination of the volumetric water content without prior knowledge of the bulk electrical conductivity.

The high relative dielectric permittivity of water imparts to wet soil a very strong dependence of ϵ_{ra} on the water content. The early laboratory work in TDR adopted an empirical approach, making no assumption about the 'state' of water in porous materials. Instead it related the TDR-measured ϵ_{ra} to the volumetric soil water content, θ_v ($m^3\,m^{-3}$), based on oven drying to 105°C. This early work gave a consistent empirical relationship between relative dielectric permittivity and volumetric water content for a wide range of soils and became a model for calibration of the TDR as a method for measurement of water content. Furthermore, it was found that the relationship is essentially independent of soil bulk density, texture, ambient temperature, and salt content, within the range of these parameters most often encountered. The equation now used widely for calibration is:

$$\theta_v = -5.3 \times 10^{-2} + 2.92 \times 10^{-2}\epsilon_{ra} - 5.5 \times 10^{-4}\epsilon_{ra}^2 + 4.3 \times 10^{-6}\epsilon_{ra}^3 \quad [5]$$

This relationship has extremely broad applicability, providing good descriptions for the water content of soils of widely varying properties, such as iron-rich volcanic soils, the unfrozen water content in frozen soils, soils with up to 50% gravel, oil shale waste, and crushed concrete. The widely obtained agreement between observations and eqn [5] has not been true for observations made in low-density soils, mineral soils with high organic or clay contents and artificial soils, such as glass beads. The widespread use of TDR has resulted in a number of applications in which eqn [5] cannot be used, and some effort has been devoted to finding alternative relationships between ϵ_{ra} and θ_v. However no relationship has emerged that shows more general applicability without the requirement for other supporting measurements (e.g., clay content, temperature).

Alternatives to empirically derived calibrations are often based on dielectric mixing models. The application and use of mixing models require varying levels of prior knowledge about the soil, such as porosity, density, and the relative dielectric permittivity of soil components. The application of mixing models for water content determinations has shown that 'square root' mixing models are most common. Such models are equivalent to a linear relationship between $\sqrt{\epsilon_{ra}}$ and the TDR travel time, t in eqn [3], which leads to a convenient linear calibration for water content. The linear relationship between θ_v and $\sqrt{\epsilon_{ra}}$ is supported by calibration data from numerous sources and is very similar to the polynomial expression in eqn [5] (Figure 2). Fitting a linear relationship where possible presents a significant improvement over fitting a

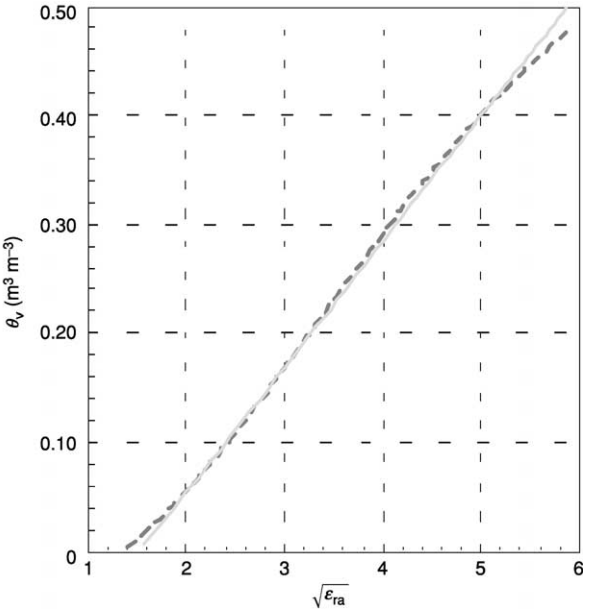


Figure 2 The similarity between eqn [5] (dashed line) and eqn [6] (solid line) over the water content range (0.05 to 0.45 $m^3\,m^{-3}$) applicable for most soils. Note the *x*-axis is $\sqrt{\epsilon_{ra}}$ and shows the linearity of eqn [6].

$$\theta_v = -5.3 \times 10^{-2} + 2.92 \times 10^{-2}\epsilon_{ra} - 5.5 \times 10^{-4}\epsilon_{ra}^2 + 4.3 \times 10^{-6}\epsilon_{ra}^3 \quad [5]$$

$$\theta_v = 0.115\sqrt{\epsilon_{ra}} - 0.176 \quad [6]$$

Reproduced with permission from White I, Knight JH, Zegelin SJ, and Topp GC (1994) Comments on ''Consideration of the use of time-domain reflectrometry (TDR) for measuring soil water content'' by Whalley WR. *European Journal of Soil Science* 45: 503–508.

polynomial calibration curve because there are only two parameters to define, making the fitting procedure much simpler. A linear regression of $\sqrt{\epsilon_{ra}}$ on θ_v gives rise to (Figure 2)

$$\theta_v = 0.115\sqrt{\epsilon_{ra}} - 0.176 \quad [6]$$

The constant term includes the electrical properties of the dry solids, primarily the bulk density and the relative permittivity of the constituent material. The slope coefficient in the first term on the right hand side embodies the effects of dissolved solutes and clay surfaces on the electrical properties of the water phase as seen by TDR. This slope changes relatively slowly as salt and clay contents increase. Clay and salt content do introduce some curvature to the relationship between $\sqrt{\epsilon_{ra}}$ and θ_v, indicating the importance of calibration checks to assure an appropriate calibration for high-resolution measurements and for unique soil conditions. Precise calibration equations become more important where absolute measures of water content are required as contrasted with differences or changes in water where more general calibrations usually suffice.

Instrumentation and Wave-Guides

The original and still widely used TDR instrument for soil was the portable cable-tester. This instrument displays the TDR waveform on a screen, allowing analyses to be performed and recorded manually. Alternatively, the data can be recorded and analyzed digitally on a personal computer (PC). Currently, a number of companies offer TDR instrumentation aimed directly at soil and environment studies. Most commercial instruments now offer automated analysis of the TDR waveform as a part of the basic instrument with network switching for multiple probes as an option. Most of these instruments make use of general calibration relationships that closely approximate eqns [5] or [6]. In addition, they allow for the calibration relationships to be customized for the desired soil characteristics.

One of the most important components of a TDR system is the soil probe or transmission line. The basic elements are the conductive components, often parallel metallic rods, which act as wave-guides. Currently the most common soil probes are of the parallel rod type, consisting of two or three parallel rods. These may vary in length from 0.1 to 1.0 m and with probe separations from 0.01 to 0.1 m. Probes that use bare metallic rods measure total water content along the path of the TDR pulse regardless of the distribution of water along this path. However, the distribution of electromagnetic energy is not uniform in the directions perpendicular to the rods. Furthermore, the distribution of energy depends on the number of rods, their diameters, and their separations. A rod separation to rod diameter ratio less than 10 is a practical rule of thumb for probe design. Further, the rod diameter should be large in relation to the dominant pore size, e.g., 10 times the dominant particle or aggregate size. We have found that 6-mm rods spaced at 50 mm have worked well in a variety of studies in tilled and untilled agricultural soil.

Many other probe configurations have come into current use, including profiling probes with the conductors mounted on opposite sides of a dielectric central core. The parallel wires mounted on a central core have been diversified to include a helical-wrapped pair on a single shaft probe and serpentine pathways on a plastic surface giving a surface probe. In all cases the energy density of the propagating wave decreases rapidly with distance from the conductors. Spatial weighting of the measurements in the cross-section perpendicular to the direction of wave propagation is a very important consideration, especially for probes that have nonmetallic components near the metal rods. Numerical analyses have been applied to the wave-equation response from parallel rod probes embedded in a porous medium. One aspect of this work was the definition of sample areas in the transverse plane that are not limited to predefined shapes, allowing for a more realistic description of the lateral sensitivity of different probe configurations. In general, TDR probes should be installed in a manner that minimizes variability of the water content within their sampling volume, especially in the direction that is transverse to the rods.

An important application of TDR relates to the measurement of shallow soil water content profiles. Three general approaches to water content profiling are depicted in Figure 3 using vertical and horizontal probes. In general, vertical rods (Figure 3a) give the most accurate measurement of cumulative water from the surface to the ends of the rods, regardless of the water content distribution with depth. Horizontal rods (Figure 3b) give higher resolution of the water content profile, with less accurate measurements of the total water stored to a given depth. It is possible to place probes depicted in Figure 3a and 3c at an angle off the vertical. Angled rods at say 45° give an optimal compromise between horizontal and vertical configurations. Another consideration is the potential for causing cracks in the soil or gaps around the rods. These features can have a significant impact on TDR measurements and on water movement, and vertical emplacements are most susceptible to these influences. In addition, water content profiles that are determined by the differencing of adjacent measurements (e.g., Figure 3a) introduce the error from both measurements, which can lead to a cumulative error

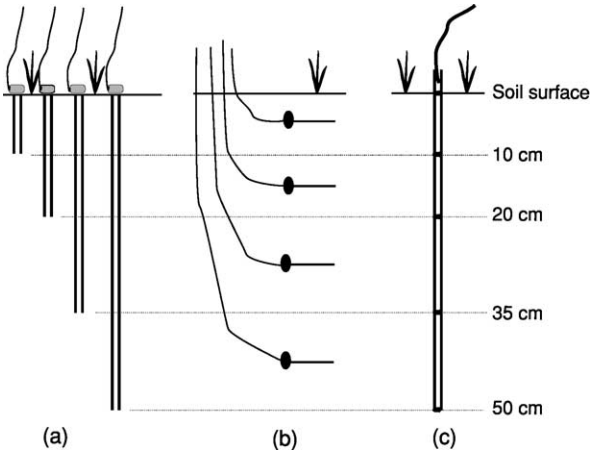


Figure 3 Three options for measurement of water content profiles. (a) Adjacent vertical; (b) horizontal; (c) segmented vertical. Probe lengths and intervals can be adjusted to the monitoring requirements. Installing probes in (a) and (c) off vertical results in a compromise between vertical and horizontal installation but optimizes the advantages from each. Reproduced with permission from Ferré PA and Topp GC (2002) Time domain reflectometry. In: Dane JH and Topp GC (eds.) *Methods of Soil Analysis Part 4 Physical Methods*. SSSA Book Series No. 5, Soil. Sci. Soc. Amer., Madison, WI.

as large as $\pm 0.03\,m^3\,m^{-3}$. On the other hand, horizontal rods require excavation for insertion at depth, which can also affect water movement. In addition, because the head of a horizontal probe is placed beneath the ground surface, it is important that it be watertight to avoid the influence of water entry. Practically, vertical rods are the simplest to install. However, their tendency to move vertically out of the soil during winter by the processes of frost-heave can be countered somewhat by the use of angled rods.

An alternative to the continuous-rod probes is the segmented probe depicted in Figure 3c. This probe configuration makes use of diode-shorting to achieve probe segmentation electronically and is available commercially. This probe is essentially a series of vertical probes placed in the same hole. As a result of its design, these probes provide both the water content profile and the total profile stored water from a single installation. Installation of this type of probe is of critical importance as any disturbance of the soil adjacent to the probe is in the most sensitive part of the measured region. Similarly, air gaps or preferential water flow along the probe will also affect measurements.

Impact of Salinity on Water Content Measurement

The explicit effect of the bulk electrical conductivity on TDR water content determinations has not been fully evaluated. Research is continuing with varied EM techniques to define conditions when the direct effect of bulk electrical conductivity (EC) results in unacceptable error in water content estimates. It appears, however, that the influence is very small for mineral soils having a clay content less than 50% and pore-water electrical conductivities less than $5\,dS\,m^{-1}$. For example, measurements in a soil of 55% clay gave an overestimate of water content of $0.014\,m^3\,m^{-3}$ due to the assumption that $\epsilon_{ra} = \epsilon'_r$. Typically, high EC conditions lead to difficulty in measuring the water content due to total extinction of the waveform through energy loss. This places a practical limit on the maximum length of TDR rods. However, to maintain acceptable water content accuracy, probe length should be at least 7.5 cm. If the EC is too high to use probes at least 7.5 cm in length, nonconductive rod coatings can be used to minimize the effects of signal loss. These coatings change the water content calibration and they should only be used when absolutely necessary.

Some of the limitations of TDR analysis and application may be addressed by a new approach to TDR waveform analysis. Standard analyses involve interpretation of the time of arrival of reflections. However, in addition to this time-based analysis, a TDR waveform can be analysed in the frequency domain using Fourier analysis. Conversion of TDR data into the frequency domain provides additional frequency-dependent information about the electrical properties of the soil and water. Frequency domain analyses of measurements made with 3-cm length probes may provide high-precision measurement of relative dielectric permittivity in saline soils where the time-domain analysis is not feasible. Further developments of this approach may extend the range of applicability of TDR.

Measurement of Bulk Electrical Conductivity Using TDR

As the TDR wave propagates through soil guided by the probe rods, its velocity is decreased by the presence of water. In addition, the magnitude of the wave is decreased by electrical losses associated with the bulk electrical conductivity of the soil, σ_0. The decrease in magnitude of the wave is expressed in the attenuation coefficient, α. It is possible to express σ_0 in terms of the values obtained from the TDR waveform (Figure 1):

$$\sigma_0 = \left(\frac{1}{120\pi L}\right)\left(\frac{Z_0}{Z_i}\right)\left(\frac{2V_0}{V_f} - 1\right) \qquad [7]$$

where Z_0 (ohms) is the characteristic impedance of the TDR probe (a factor of the probe geometry), Z_i is the output impedance of the TDR instrument, V_0 (volts) is the initial step voltage output and V_f is the

final voltage remaining after all multiple reflections from the probe, respectively. In practice, the quantity $(Z_0/120\pi LZ_i)$ is often treated as a calibration constant for the instrument, cable, and probe combination, which can be obtained from measurements in solutions of known EC. After the probe and TDR instrument have been characterized, the bulk EC is calculated from the value of V_f as the only independent variable at each measurement site or measurement time. Although the measurement is straightforward, it is important that V_f be determined well after all multiple reflections have returned and the TDR waveform has reached a final constant value.

The bulk EC measured by TDR has applications in research into solute transfer processes, particularly ionic solute breakthrough experiments. At a heterogeneous field site, using vertically installed TDR rods, solute mass flux measured with TDR was in good agreement with that obtained from solution samplers. The TDR-measured EC has been used to monitor nitrate levels in the field with a view to improving crop nutrient management and minimizing leaching to groundwater.

Impact of Water Content on Pore-Water Salinity Measurement

The EC of the soil solution is often desired for soil and environment studies, particularly in relation to soil salinity assessment or monitoring migration of polluting chemicals. The relationship between bulk EC and solution EC is strongly dependent on both the soil water content and the clay content of the soil. TDR offers the distinct advantage of measuring both water content and EC in the same sample and under identical conditions. Thus, TDR can be very useful for establishing both the water content and the pore-water EC simultaneously. TDR has been used in both direct and indirect approaches and to elucidate the relationship among water content, bulk EC, and solute EC. However, it is not always possible to determine the pore water EC if the water content varies along the TDR probes, as this will give a gradient of pore water EC along the probe.

Conclusion

In the 25 years since the early applications in soils, TDR has revolutionized the field measurement of soil water content. In 1991, a survey of soil physicists in four countries reported TDR was used one-third as often as gravimetric sampling and one-half as often as neutron moderation. It is most likely that TDR use had surpassed both these methods by the end of the 1990s, making it more ubiquitous than any other technique. The capability to log TDR data and the relatively nondestructive nature of the installations have allowed *in situ* monitoring, contributing to the much wider use of TDR. For water balance and water-use efficiency studies, TDR is particularly applicable because the probe designs and orientations can be chosen to meet a variety of conditions, such as agricultural row crops, drip irrigation patterns, and forested landscapes. Through the application of TDR in soil much has been learned about the soil's electromagnetic properties in the radio frequency range. This knowledge has assisted in the development of other electromagnetic techniques, such as ground-penetrating radar and remote sensing using either active or passive microwave radars. TDR is specifically valuable for 'ground-truthing' radar applications because both are interpreting measured EM properties of soil. As a consequence, TDR can be applied to defining the sampling regions of radar applications by appropriate choice of probe geometries for the TDR. Water content measurement by capacitance techniques has been accelerated recently, in part because appropriate frequencies and probe geometries were identified through TDR studies. Although capacitance devices are less expensive, they have not yet offered the ability to measure EC as can be done by TDR.

TDR is prominent among several emerging technologies being used to enable the long-neglected investigation of roots as the major agents of water and chemical transfer in soil. TDR water content measurements also improve the value and significance of other environmental parameters. For example, frequent measurements of temperature, water content by TDR, and oxygen concentrations in the rooting zone of maize can help to demonstrate when tillage and trafficking limit both O_2 supply and crop growth. Prior to TDR, measurements of O_2 supply for roots and other biota in the soil had not been possible. Time-domain transmissiometry (TDT), which is similar to TDR, has been combined with a soil penetrometer to measure both soil strength (compaction) and water content. The combination of sensors greatly enhances the value of the soil penetrometer as a diagnostic tool for soil compaction.

Regulations now play a leading role in the need to monitor soil contamination and harmful infiltration of contaminated solutions into groundwater. Cities in the USA such as Los Angeles, Tucson, Las Vegas, and Glendale are using TDR for such monitoring. Agencies from around the world have reported using TDR in environmental monitoring applications as diverse as nonaqueous liquid-phase concentrations (Sweden), subsurface pollutant detection (United Arab Emirates), changes in groundwater level and

crude oil thickness (USA), characterization of diesel-contaminated soil (Canada), and measuring water-table elevations within sidewalk systems (USA).

The application of TDR in soil is continuing to develop and diversify in ways that will enhance the value of TDR for soil and environmental monitoring. The analyses of spatial sensitivity within probes have facilitated designs for specific uses, often allowing combinations of TDR with other measurements such as with tensiometers and the cone penetrometer.

List of Technical Nomenclature

α	Attenuation coefficient
ϵ_0	Dielectric permittivity of free space ($F\,m^{-1}$)
ϵ_r	Relative dielectric permittivity (dielectric constant)
ϵ_r'	Real component of relative dielectric permittivity (dielectric constant)
ϵ_r''	Imaginary component of relative dielectric permittivity (relative dielectric loss)
ϵ_{ra}	Apparent relative dielectric permittivity (dielectric constant)
θ_v	Volumetric soil water content ($m^3\,m^{-3}$)
ρ	Reflection coefficient
σ_0	Electrical conductivity ($S\,m^{-1}$)
ω	Frequency of signal ($2\pi\,s^{-1}$)
c	Velocity of light ($m\,s^{-1}$)
L	Length (m)
t	Time(s)
V	Voltage (volt)
v	Propagation velocity of electromagnetic waves ($m\,s^{-1}$)
Z_0	Characteristic impedance of probe (ohms)
Z_i	Output impedance of instrument (ohms)

See also: **Neutron Scattering**; **Salination Processes**; **Water Content and Potential, Measurement**; **Water Potential**

Further Reading

Clothier BE and Green SR (1997) Roots: the big movers of water and chemical in soil. *Soil Science* 162: 534–543.

Dowding CH (ed.) (2001) *Second International Symposium and Workshop on Time Domain Reflectometry for Innovative Geotechnical Applications*. IL, USA: Infrastructure Technology Institute at Northwestern University Evanston.

Ferré PA and Topp GC (2000) Time-domain reflectometry sensor techniques for soil water content measurements and electrical conductivity measurements. In: Baltes H, Gopel W, and Hesse J (eds) *Sensors Update*, vol. 7, pp. 277–300. Weinheim: Wiley-VCH.

Ferré PA and Topp GC (2002) Time domain reflectometry. In: Dane J and Topp GC (eds) *Methods of Soil Analysis*, Part 4. 3rd edn, pp. 434–446, 534–545. Madison, WI: Soil Science Society of America.

O'Connor KM and Dowding CH (1999) *GeoMeasurements by Pulsing TDR Cables and Probes*. Boca Raton: CRC Press.

Topp GC and Reynolds WD (1998) Time domain reflectometry: a seminal technique for measuring mass and energy in soil. *Soil and Tillage Research* 47: 125–132.

Wraith JM (2002) Time domain reflectometry. In: Dane J and Topp GC (eds) *Methods of Soil Analysis*, Part 4. 3rd edn, pp. 1289–1297, 1311–1321. Madison, WI: Soil Science Society of America.

TROPICAL SOILS

Contents
Arid and Semiarid
Humid Tropical

Arid and Semiarid

H C Monger, New Mexico State University, Las Cruces, New Mexico, USA
J J Martinez-Rios, Universidad Juarez del Estado de Durango, Mexico
S A Khresat, Jordan University of Science and Technology, Irbid, Jordan

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Arid and semiarid soils occupy about one-third of the Earth's ice-free land surface. They are sources and sinks of atmospheric CO_2, sources and sinks (mainly sources) of global dust, and substrate that support high biodiversity of plants and animals. Arid and semiarid soils uniquely accumulate secondary minerals, such as calcite and gypsum, as the result of low rainfall and limited leaching. Because of sparse vegetative cover and high susceptibility to wind and water erosion, many arid and semiarid soils have low resistance and low resilience to disturbance. Hence land degradation (i.e., desertification) is common on most continents with arid and semiarid soils.

These dryland soils have been a factor of primary importance in human history. The oldest hominid fossils are found in east Africa in sediments with paleosols containing pedogenic carbonate indicating an arid or semiarid climate. The transition from hunting-and-gathering to agriculture took place in arid and semiarid Mesopotamia about 10 000 years ago. The production of surplus food on the floodplain soils of the Euphrates and Nile gave rise to early civilizations in Sumeria, Babylonia, and Egypt. Today several large urban centers are located on arid and semiarid soils around the world that have adequate groundwater or river water supplies. If irrigated, arid and semiarid soils are an important source of local and global food production. Still, most arid and semiarid land is sparsely populated, open, and often wilderness land.

Climatic Controls

The terms arid, deserts, semiarid, and steppes are used variously to describe dryland conditions. Arid (an adjective) describes a climatic condition of low rainfall – commonly taken to be less than 250 mm (10 in.) of mean annual precipitation (Figure 1a). Desert (a noun) is a region of the Earth's land surface within an arid climate. Likewise, semiarid is a climatic condition characterized by a mean annual precipitation between 250 and 500 mm (Figure 1a). A region of the Earth's surface within a semiarid climate is a steppe.

Closely linked to annual precipitation, and especially to soil moisture, is vegetation. The driest deserts are often barren of plant cover, but most deserts have scattered shrubs, cacti, forbs, and grasses. Though some steppe vegetation can occur in areas like the Badia of Jordan that receive as little as 100 mm of rainfall, generally steppe vegetation is characterized by higher amounts of rainfall in which short-grass prairie is bordered by desert vegetation on the arid side and tall-grass prairie, savanna, or woodlands on the subhumid side. Thus, by these definitions arid soils are synonymous with desert soils and semiarid soils are synonymous with steppe soils.

But to define deserts and steppes by precipitation alone is to ignore other important climatic variables, mainly temperature. One expression of the combined influences of both precipitation and temperature is the de Martonne aridity index based on the formula:

$$I_a = P_{mm}/(T^{\circ}C + 10) \qquad [1]$$

where the aridity index (I_a) is equal to the mean annual precipitation in millimeters (P_{mm}) divided by the mean annual temperature in degrees Celsius ($T^{\circ}C$) plus 10. According to this index, values below 5 characterize true deserts, values of approximately 10 demarcate dry steppes, values of about 20 represent prairies, and values above 30 typify forest. The boundary for the Chihuahuan desert of Mexico and the USA, for example, is based on an aridity index of 10 or lower.

The Köppen system also demarcates deserts and steppes as a function of both precipitation and temperature (Figure 1b). For example, some cold regions in the high latitudes of North America and Eurasia that get semiarid-amounts of precipitation are coniferous forest instead of steppes, and many cold

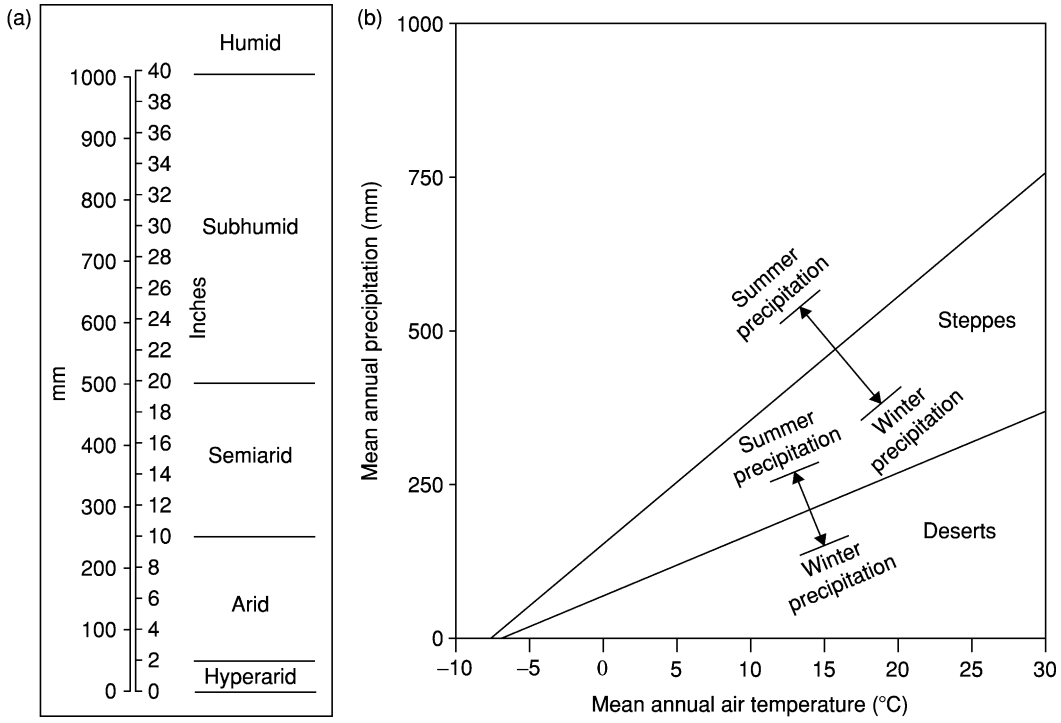


Figure 1 (a) Climate categories based on mean annual precipitation. (b) Desert and steppe boundaries as a function of both mean annual precipitation and temperature according to the Köppen system. Arrows show how boundaries shift according to whether precipitation falls mainly in the summer or winter.

areas that receive arid-amounts of precipitation are steppes instead of deserts.

Seasonality of precipitation is another climatic factor that affects desert and steppe boundaries. For a given mean annual temperature, the boundary of a steppe will extend into wetter climates if its precipitation falls mainly in the summer (Figure 1b). In other words, if precipitation falls in summer, the area of a steppe will be larger because summer evapotranspiration depletes soil moisture more thoroughly than winter evapotranspiration. Similarly, the size of a desert will be larger if its precipitation falls in the summer rather than in the winter.

Soil water potential is another way of defining boundaries of arid and semiarid soils. Soil moisture measured as soil water potential, which includes the influence of particle size and salts, is more important to vegetation than annual precipitation alone. The Soil Taxonomy system, for example, uses soil water potential to define moisture regimes as a criterion for classifying soils. The aridic moisture regime, for example, includes soils too dry to support nonirrigated crops, and is defined as soil that is moist (i.e., water held at tensions greater than −1500 kPa) for no more than 90 consecutive days when the soil temperature at a depth of 50 cm is above 8°C. Soils with the ustic and xeric moisture regimes are transitional between the aridic moisture regime and soils of humid climates that have the udic moisture regime. Semiarid soils occur within ustic and xeric moisture regimes, their drier subdivisions, and wetter subdivisions of the aridic moisture regime, namely, the aridic ustic, aridic xeric, xeric aridic, and ustic aridic regimes.

Processes of Soil Formation in Arid and Semiarid Soils

Water – its amount and depth-of-wetting – is the major driver that gives rise to differences between arid and humid soils. In humid soils >50% of the water entering the soil drains downward through the profile to groundwater. In arid soils, <10% flushes through the soil profile to the groundwater. A humid soil receiving 1300 mm of rain, for example, would have about 650 mm that percolated through its profile in a year. An arid soil receiving 200 mm would have <20 mm that percolated through its profile. Thus, nearly all water in arid soils and much water in semiarid soils enters and leaves via the soil surface. Notable exceptions are low-lying areas that receive runoff water (e.g., playas). In these topographic lows, soils are nonsaline because of deep leaching, if the water table is deep and the soils have high permeability. However, if the water table is shallow or soils have slow permeability, these topographic lows are zones of salt accumulation

As with humid soils, the processes of soil genesis (i.e., gains, transfers, transformations, and losses) operate in arid and semiarid soils, but the magnitude and direction of these processes are different. The shallow depth-of-wetting and incomplete leaching have a major impact on gains because authigenic minerals, such as calcite, silica, and gypsum, accumulate in the profile and give rise to the formation of calcic, petrocalcic, duripans, gypsic, and petrogypic horizons. Gains of dust are also important. Silicate clay dust, for example, contributes to the formation of argillic horizons, carbonate dust to the formation of calcic horizons, and gypsum dust to the formation of gypsic horizons. Gains of photosynthetic carbon in the form of soil organic matter are lower than in humid soils. But gains of photosynthetic carbon released as respired soil CO_2 that leads to HCO_3^- and $CaCO_3$ formation are higher than in humid soils.

Transfers of material down the profiles of arid and semiarid soils include illuvial clay, carbonate, and salts. Arid soils typically display a chromatographic pattern of an argillic horizon overlying a calcic horizon. If gypsum and soluble salts are also present, the profile can contain an argillic overlying a calcic overlying a gypsic overlying a salic horizon. Transfers of materials also occur up the profiles of some arid and semiarid soils. These include capillary rise of soluble material and particles moved upward by ants and termites. In some cases, desert pavements are formed by the upward movement of coarse fragments lifted by silts and fine sands that accumulate beneath them.

Important transformations in arid and semiarid soils include rock disintegration resulting from the crystallization of salts, thermal fluctuations, and chemical weathering. Other transformations involve the decomposition of organic matter and formation of clay minerals, such as palygorskites and sepiolites. The formation of kaolinite is common in parent materials containing feldspars. Losses are mainly in the form of erosional truncation of soil horizons.

Factors of Soil Formation in Arid and Semiarid Soils

The five soil-forming factors interact in assorted ways to produce arid and semiarid soils (Figure 2). Climate, especially water supply, is the defining factor. Water is not only the agent for mineralogical gains, transformations, and transfers, it is also essential for nutrient supply and direct use by vegetation. Vegetation, in turn, has a feedback link to soil by adding organic matter, translocating ions via bioaccumulation, and providing ground cover that protects the soil from erosion. Vegetation is also linked to animals by being their food supply, while animals are linked to vegetation by herbivory and seed dispersal (Figure 2). Animals affect soil by bioturbation – humans alone move about 40 GT of soil per year. Soil, in turn, affects animals by providing habitat, which is important for nematodes, gastropods, earthworms, crustaceans, mites, spiders, ants, termites, mice, moles, rabbits, gophers, birds, foxes, badgers, deer, bear, and even humans in some places of the arid world.

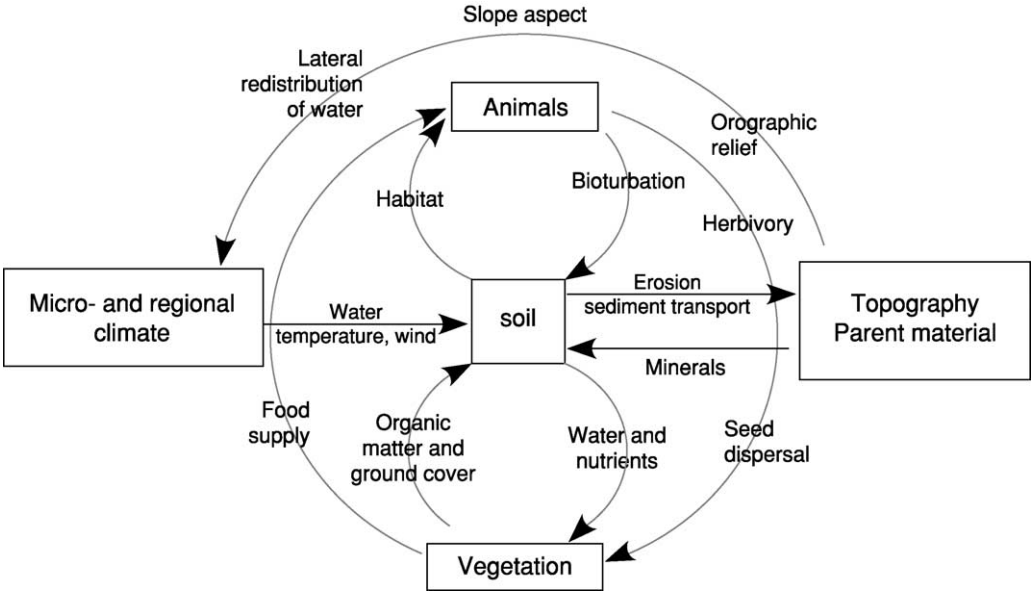


Figure 2 Illustration of the links among the soil-forming factors of climate, biota (vegetation and animals), topography, and parent material. The fifth factor, time, affects all of the illustrated links in proportion to the duration that the factors operate.

Parent material and topography are the main factors of the geologic setting that affect arid and semiarid soils. Parent material has a direct link to soil as a supplier of mineral detritus. Topography has its link to soil by its influence on regional climate at the orographic scale and by its influence on microclimate (resulting from slope aspect and laterally redistributed water) at the local scale. Over the long term, climate through its effects on soils alters the configuration of the landscape by erosion and sediment transport (Figure 2).

Properties of Arid Soils

Arid soils have surface horizons with several unique characteristics. Many arid soils, for example, are covered by desert pavement that overlies vesicular A and E horizons. Other arid soils are covered by salt efflorescence in areas where shallow groundwater has risen by capillarity and evaporated at the surface. Still other arid soils are covered by microbiotic crusts or by blankets of aeolian sand or silt. Nearly all arid soils have lower amounts of organic matter than their more humid counterparts. For classification purposes, the surface horizon (i.e., epipedon) that is ubiquitous for arid soils is the ochric epipedon. Other epipedons of arid soils with much smaller occurrences are the mollic, anthropic, and in very rare cases of grass sod over shallow basalt, the histic.

Subsurface horizons of arid soils are uniquely different from subsurface horizons of humid soils in some instances, yet similar in other instances. Subsurface horizons in arid soils that are uniquely different include horizons dominated by calcium carbonate, secondary silica, gypsum, and soluble salts, while horizons common to both arid and humid soils include those with weak structural and color development and accumulations of illuvial clay and sesquioxides. Diagnostic horizons unique to arid soils include the calcic, petrocalcic, duripan, gypsic, petrogypsic, natric, and salic horizons (Figure 3). Diagnostic horizons found in both arid and humid soils include the cambic, argillic, and in rare cases, the oxic.

In addition to diagnostic horizons, other soil properties, such as vertic, andic, lithic, climatic, anthropogenic, depth to groundwater, and particle-size characteristics are used to make taxonomic subdivisions in arid soils. At the highest taxonomic levels, arid soils include Leptosols, Gypsisols, Durisols, Calcisols, and Solonchaks in the World Reference Base (WRB) system; Desert gray brown, Desert takyr-like, Desert sandy, and Meadow desert soils in the Russian system; and Halosols and Aridisols in the Chinese system. In the Soil Taxonomy system, arid soils are classified as, in order of abundance, Aridisols, Entisols, Vertisols, Oxisols, and Andisols (Table 1). Moving dune fields also occupy large areas of deserts, especially in North Africa.

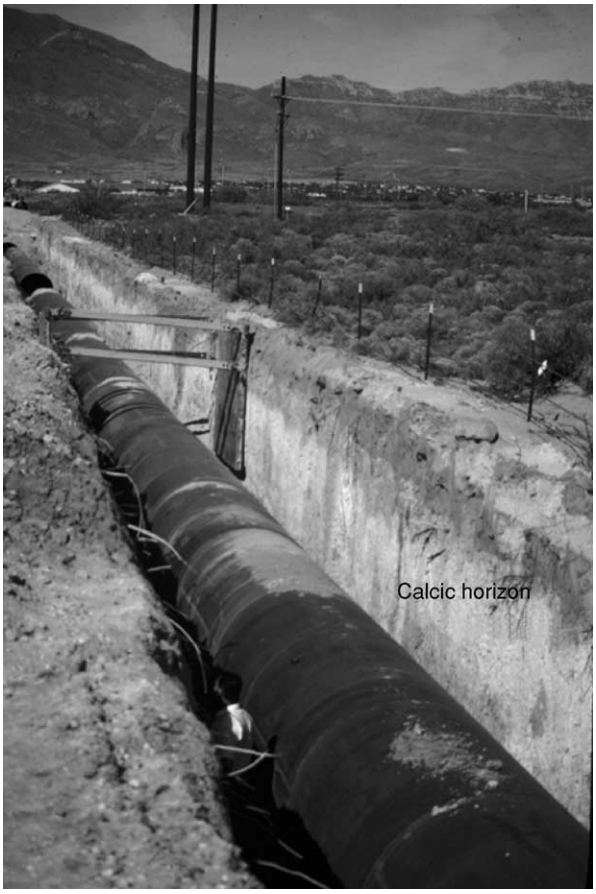


Figure 3 Calcic horizon in an arid soil in the Chihuahuan desert of North America. This site has an annual rainfall of 250 mm and an annual temperature of 16°C.

Properties of Semiarid Soils

Semiarid soils, owing to more rainfall and the homogenizing effects of greater vegetative cover, have surface horizons with more organic matter than arid soils and fewer unique features like desert pavements, vesicular horizons, and efflorescence. Although the ochric epipedon is still a common horizon of semiarid soils, the mollic epipedon is widespread and dominates the steppe areas that border tall-grass prairies.

Being transitional between arid and humid soils, semiarid soils have a large variety of subsurface horizons. As with arid soils, semiarid soils might have calcic, petrocalcic, duripan, gypsic, natric, and salic horizons. Yet, as with humid soils, semiarid soils might have albic, argillic, and kandic horizons as well as fragipans and plinthite.

Table 1 Global extent of arid and semiarid soils (km^2) based on the Soil Taxonomy system[a]

Soil order	*Suborder*	*Africa*	*Asia*	*Australia / Oceania*	*Europe*	*South America*	*Central America*	*North America*	*Global*
Arid soils									
Aridisols									15 798 100
	Cryids	653	417 587	–	103	165 909	–	499 351	1 083 601
	Salids	95 249	595 400	919	130	59 756	–	17 698	769 152
	Durids	–	–	–	–	–	–	–	–
	Gypsids	347 458	326 137	–	3454	–	–	–	677 050
	Argids	450 884	1 635 130	1 710 271	1218	520 175	1111	1 093 291	5 412 080
	Calcids	1 818 639	2 204 270	613 547	1013	98 680	–	202 235	4 938 384
	Cambids	842 681	1 125 259	346 253	5573	422 429	2259	173 378	2 917 832
Vertisols	Torrerts	196 570	65 210	602 630	–	6233	1754	42 435	914 832
Oxisols	Torroxes	9346	–	4247	–	16 550	–	–	30 143
Andisols	Torrands	834	–	–	–	95	–	150	1078
Entisols[b]	In aridic (torric) moisture regimes								12 600 308
Total arid soils									29 344 460
Semiarid soils[c]									
Mollisols									
	Ustolls	4371	1 587 526	21 689	303 635	409 796	–	1 744 438	4 071 455
	Xerolls	72 125	403 161	55 502	236 656	769	–	165 132	933 346
Vertisols									
	Usterts	722 630	600 544	14 863	29 001	38 107	18 247	114 387	1 763 776
	Xererts	11 692	47 348	20 258	17 105	1050	–	969	98 423
Oxisols	Ustoxes	1 718 701	18 457	39 249	–	1 338 805	1295	360	3 116 866
Andisols									
	Ustands	11 235	12 308	–	1975	11 414	11 004	7750	55 686
	Xerands	–	–	–	8937	2098	–	17 840	28 876
Alfisols									
	Ustalfs	2 470 581	1 084 231	501 676	253 839	998 870	22 555	346 868	5 678 621
	Xeralfs	80 071	209 377	270 520	118 833	25 340	–	176 034	880 174
Ultisols									
	Ustults	1 649 310	824 575	101 301	–	1 091 366	58 903	131 048	3 856 502
	Xerults	933	2672	–	–	22	–	15 958	19 586
Inceptisols									
	Ustepts	1 675 118	832 200	257 463	377 338	611 443	83 255	331 262	4 168 080
	Xerepts	163 793	178 312	264	310 697	8342	–	9246	670 654
Entisols[d]	In ustic moisture regimes								4 620 113
Entisols	In xeric moisture regimes								840 021
Total semiarid soils									30 802 179
Total ice-free land area									130 268 185

[a] Courtesy of USDA–Natural Resources Conservation Service, Soil Survey Division, World Soil Resources, 2002.
[b] Entisols with the aridic (torric) moisture regime are designated at the great group level (e.g., Torripsamments and Torriorthents).
[c] The semiarid soils category contains some soils of subhumid climates (500–1000 mm annual precipitation) in cold regions and along coasts.
[d] Entisols with the ustic and xeric moisture regimes are designated at the great group level (e.g., Ustorthents and Xerofluvents).

Other diagnostic properties used to subdivide semiarid soils include redox, petroferric, vertic, andic, lithic, climatic, anthropogenic, particle-size, and groundwater characteristics. Classification of semiarid soils includes the Kastanozems, Chernozems, and Phaeozems in the WRB system; Meadow Chernozem-like, Meadow Chestnut, Semidesert Brown, and Semidesert Meadow Brown soils in the Russian system; and Isohumisols and Ferrallisols in the Chinese system. In the Soil Taxonomy system, semiarid soils are classified as, in order of abundance, Alfisols, Entisols, Mollisols, Inceptisols, Ultisols, Oxisols, Vertisols, and Andisols (Table 1).

Human Land Use of Arid and Semiarid Soils

Early civilizations arose in arid and semiarid Sumeria in the fourth millennium BC as irrigated agriculture on floodplain soils encouraged stable settlements, led to surplus food, and freed people to pursue specialized trades and to develop social order and cultural creativity. Similar cultural developments arose along rivers in other arid and semiarid climates, such as the Indus of ancient India and Hoang-Ho of ancient China. In the western hemisphere as well, societies developed in arid and semiarid climates, such as the Inca, Aztec, and Hohokan cultures.

Grazing of domesticated cattle, sheep, and goats has been the traditional land use of arid and semiarid soil away from the irrigated floodplains. Today, in addition to grazing, arid and semiarid soils have ecological and global biogeochemical properties important to humans. Arid and semiarid ecosystems have some of the highest biodiversity on Earth as grasses, cacti, shrubs, reptiles, birds, and mammals have adapted to wide and rapid shifts in temperature and moisture. Much of the carbon in the global carbon cycle, at least 800×10^{15} g, is stored as soil carbonate in arid and semiarid soils. The source of much of the local, regional, and global dust is from arid and semiarid soils.

Desertification, which is defined by the United Nations Convention on Desertification as 'land degradation resulting from climatic and human activities,' is a major issue of importance in arid and semiarid climates because desertification directly affects about one-sixth of the world's population in both developing and developed countries. Increasing human population and livestock pressures have accelerated desertification and caused shrubs to invade grasslands with subsequent erosion of exposed topsoil. Wind erosion is a major agent for desertification in arid and semiarid regions because it removes organic matter and fine mineral particles in A horizons that have important water and nutrient storage properties. Water erosion is also important, especially at the arid–semiarid boundary of about 250 mm of precipitation where erosion rates are commonly at their greatest.

Limited water supply is of paramount importance to expanding populations in arid and semiarid regions of North and South America, Australia, Africa, Asia, and the Middle East. In the western USA, for example, the drought did not move to the people, the people moved to the drought. While the lack of water threatens population growth in arid and semiarid regions, the lack of water is the natural state, and the state that imparted the unique properties to arid and semiarid soils. It is also the lack of water that makes arid and semiarid places some of the most open, least developed places on Earth, and some of the last 'wild' places for future generations to enjoy and study.

See also: **Classification of Soils**; **Desertification**; **Factors of Soil Formation:** Climate

Further Reading

Birkeland PW (1999) *Soils and Geomorphology*. New York: Oxford University Press.

Brandt J and Thornes JB (1996) *Mediterranean Desertification and Land Use*. New York: John Wiley.

Cooke R, Warren A, and Goudie A (1993) *Desert Geomorphology*. London: UCL Press.

Dregne HE (1983) *Desertification of Arid Lands*. New York: Harwood Academic Press.

Durant W (1935) *Our Oriental Heritage*. New York: Simon and Schuster.

Eswaran H, Reich PF, Kimble JM *et al.* (2000) Global carbon stocks. In: Lal R, Kimble JM, Eswaran H, and Stewart BA (eds) *Global Climate Change and Pedogenic Carbonates*, pp. 15–61. Boca Raton, FL: CRC Press.

Gile LH, Peterson FF, and Grossman RB (1966) Morphology and genetic sequences of carbonate accumulation in desert soils. *Soil Science* 101: 347–360.

Harris D (1996) *The Origins and Spread of Agriculture and Pastoralism in Eurasia*. Washington, DC: Smithsonian Institution Press.

Krogh L (2002) Major classification systems. In: Lal R (ed.) *Encyclopedia of Soil Science*, pp. 176–182. New York: Marcel Dekker.

Schlesinger WH, Reynolds JF, Cunningham GL *et al.* (1990) Biological feedbacks in global desertification. *Science* 247: 1043–1048.

Schmidt RH (1979) A climatic delineation of the 'real' Chihuahuan desert. *Journal of Arid Environments* 2: 243–250.

Smith GD (1986) *The Guy Smith Interviews: Rationale for Concepts in Soil Taxonomy*. Soil Management Support Service monograph no. 11. Washington, DC: US Government Printing Office.

Soil Survey Staff (1999) *Soil Taxonomy – A Basic System of Soil Classification for Making and Interpreting Soil Surveys*. USDA Agriculture Handbook number 436. Washington, DC: US Government Printing Office.

Spaargaren OC (2000) Other systems of soil classification. In: Sumner ME (ed.) *Handbook of Soil Science*, pp. E137–E174. Boca Raton, FL: CRC Press.

Strahler AN and Strahler AH (1987) *Modern Physical Geography*. New York: Wiley.

Velasco MH (1983) *Uso y Manejo del Suelo*. Mexico: Editorial Limusa.

Wilding LP (2000) Introduction: general characteristics of soil orders and global distribution. In: Sumner ME (ed.) *Handbook of Soil Science*, pp. E175–E182. Boca Raton, FL: CRC Press.

Humid Tropical

S W Buol, North Carolina State University, Raleigh, NC, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Soils are the physical, chemical, and biological media at the upper surface of the Earth's land areas capable of accepting plant roots. A wide range of geologic materials, soil moisture, and temperature conditions

that differ regionally and among adjacent soils on local landscape positions provide for diverse suites of contrasting soils that interact differently with biological communities and human attempts to sustain food production within the humid tropics.

Humid Tropical Setting

Many climatic classifications have been published. The parameters of soil temperature and moisture regimes used to classify soils do not conform to general concepts of humid tropics. The following criteria, adopted by the National Cooperative Soil Survey in the USA, are most universally used when classifying soil.

Soil-Temperature Regimes of the Tropics

Mean annual soil temperatures are 2–4°C warmer than mean annual air temperatures. Soil-temperature regimes (STRs) are defined by two criteria, mean annual soil temperature and seasonal temperature difference (Figure 1). Seasonal soil-temperature difference is determined as the mean soil temperature of June, July, and August compared with the mean soil temperature of December, January, and February. Almost all of the soils within the geographic tropics have soil temperatures that seasonally differ by less than 6°C and are identified by the prefix 'iso' placed before the name of the mean annual soil-temperature regime. Except for a few soils near the northern and southern extremities of the geographic tropics in Africa, this one characteristic is the only soil property that is nearly universal with the concept of tropical soil. The practical aspect of this soil property is that seasonal soil temperatures seldom have to be considered when planting food crops.

Soils with mean annual soil temperatures of 22°C or higher are classified as isohyperthermic; soils with mean annual soil temperatures of 15–22°C are identified as isothermic. Freezing conditions are seldom a problem in the isothermic and isohyperthermic STRs. Higher elevations in tropical areas have isomesic STRs, with mean annual soil temperatures less than 15°C; crop growth is slow, night-time freezing is common, and even cold-tolerant crops such as potatoes seldom grow well where mean annual soil temperature is below 10°C.

Soil-Moisture Regimes of the Tropics

Soil-moisture regimes (SMRs) are defined to classify a soil's ability to supply water to plants without irrigation (Figure 2). In soils where the groundwater table is not reached by the roots of most crop plants, the SMR is determined by the seasonal distribution of rainfall in 'normal' years. Normal years are ±1 standard deviation of long-term means. Most food crops require a reliable supply of water for at least 90 consecutive days. To calculate SMRs, the mean monthly precipitation is compared with the calculated potential evapotranspiration. A soil-water balance is then constructed for the mean rainfall in the area. The duration of time during which water is normally available either from average rainfall or as stored available water in the soil during a period when soil temperatures are warm enough to grow the crop determines the SMR.

The perudic SMR has precipitation that exceeds potential evapotranspiration every month of normal years. Although this may seem desirable, these areas present weed, insect, and disease problems, and the constantly humid conditions make it difficult to harvest mature grain crops.

The udic SMR has fewer than 90 cumulative days when water is not available in the rooting zone in normal years. It is possible to grow food crops any time of the year without irrigation when the temperature is warm enough for that crop. Available water is less reliable during some part of the year, and farmers often select more drought-tolerant crops or may choose not to plant during that period, but perennial plants are adequately supplied with water throughout most years in most of the udic SMR areas.

The ustic SMR is borderline to the common concept of humid. In normal years, soils with an ustic SMR have at least 90 consecutive days when moisture is available, but more than 90 cumulative days when water is not available in the rooting zone. Natural vegetation is either seasonal rain forest or savanna. At least one crop can be reliably grown each year, and it is possible to grow two crops per year on some soils, but there is a seasonal dry period of 90 days or more when crop production is not possible without irrigation. The reliable dry season of the ustic SMR is a distinct advantage for weed and disease control and grain-crop harvest in isohyperthermic and isothermic STRs.

Soils that have a reliable moisture supply for fewer than 90 consecutive days in normal years have an aridic SMR and are excluded from the concept of humid tropics.

The above SMRs are used to classify only the well-drained upland soils. Groundwater often saturates the rooting zone of some soils within all areas of the humid tropics. These soils are identified as having an 'aquic soil moisture condition' and are commonly referred to as poorly drained soils. Soils with aquic soil-moisture conditions most often occur in areas adjacent to rivers and lakes or in broad, level

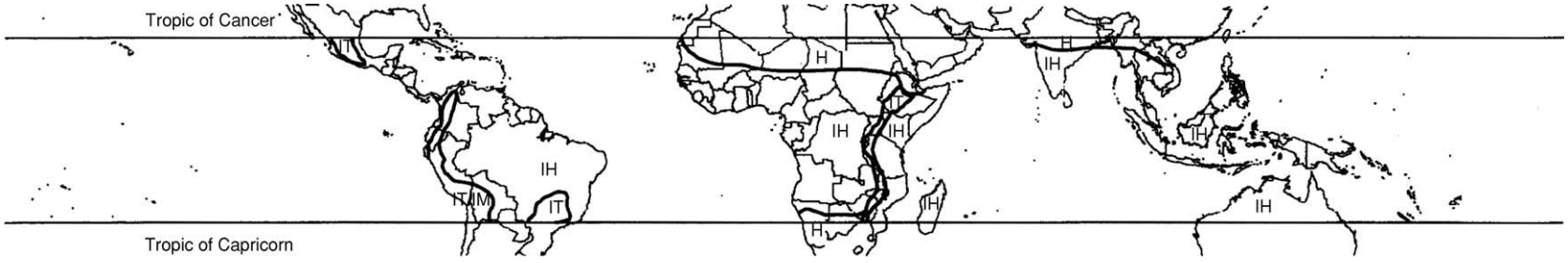


Figure 1 Generalized map of soil-temperature regimes in the tropics. IH, isohyperthermic; IT, isothermic; IT/IM, isohyperthermic and isothermic; H, hyperthermic.

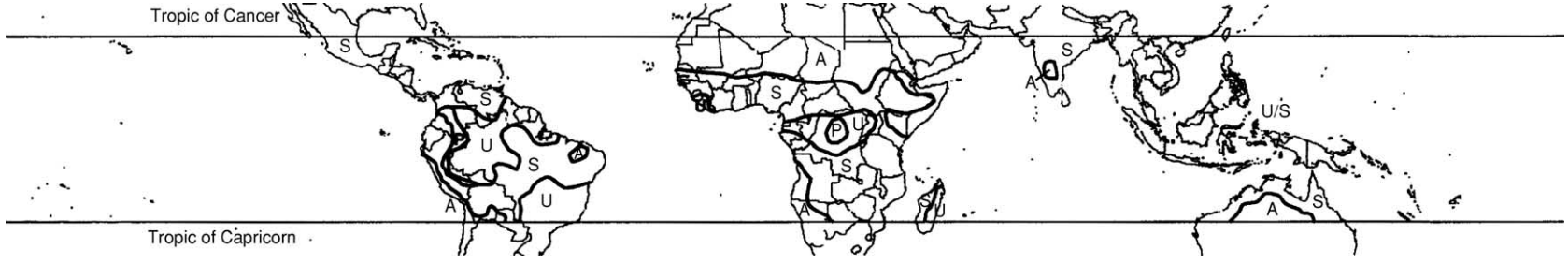


Figure 2 Generalized map of soil-moisture regimes in the tropics. P, perudic; U, udic; S, ustic; U/S, udic and ustic; A, aridic.

landscapes. Soils with aquic soil-moisture conditions are saturated with water only during some part of the year. They are often seasonally utilized for subsistence food crops, but most require engineered drainage systems before commercial agricultural production is possible.

Chemical and Mineralogical Composition of Soils

Of the chemical elements needed for plant and animal physiology, only carbon, oxygen, hydrogen, nitrogen, and to some extent sulfur are derived from air and water. The other essential elements are obtained from the minerals in the soil. Mineralogical properties of soils are derived from the geologic material within which the soil is formed. An inadequate supply of any essential element limits plant growth. The most frequent limitations result from insufficient plant-available nitrogen, phosphorus, potassium, calcium, or magnesium.

Practically no nitrogen is present in soil minerals. Nitrogen enters the soil as ammonium and nitrate dissolved in rainwater or via fixation from the air by nitrogen-fixing microbes. Some nitrogen-fixing microbes in the soil are symbiotic and the nitrogen they extract from the air is incorporated into their legume plant host. Other nitrogen-fixing microbes are not symbiotic, and the nitrogen they extract from the air is incorporated into their cells. Nitrogen is concentrated in organic residues in the surface layers of soil. As organic residues decompose, inorganic forms of nitrogen are released into the soil solution and become available to growing plants, leach into the groundwater during periods of excessive rainfall, or return to the air as nitrogen gas during periods when the soil is saturated with water. Plant-available nitrogen contents in soil are transient and closely related to supplies of organic residue.

Phosphorus is present in only a few minerals. Iron and aluminum phosphates are extremely insoluble and do not release phosphorus rapidly enough for rapid plant growth. The release rate is so slow that soils with high iron and aluminum contents tend to absorb phosphate applied as fertilizer and decrease its availability to plants. Apatite, a more soluble calcium phosphate mineral capable of supplying plant-available phosphorus, is a common source of phosphorus and often present in limestone.

Potassium is present in mica and feldspar minerals. These minerals are rather easily decomposed in the soil environment and consequently are seldom present in materials that have been repeatedly transported and deposited on the land surface.

Calcium and magnesium are most abundant in carbonate minerals associated with limestone and some sandstone. Carbonate minerals are also relatively unstable when subjected to weathering and therefore are present only in recent geologic sediments, limestone and some sandstone.

Soil pH is a measure of the acidity or alkalinity of water in the soil and has a direct effect on how rapidly many of the essential elements are available to growing plants. In the absence of carbonate minerals, soils in the humid tropics are acid in reaction and only limited quantities of essential elements present in the soil are available for plant growth. Acid soils with pH values less than approximately 5.2 also have a concentration of aluminum ions that is toxic to some but not all crop plants. Additions of lime (finely ground calcium and calcium:magnesium carbonates) are desirable and often necessary to reduce or eliminate aluminum toxicity and increase the availability of the essential elements to most crop plants growing in acid soils.

The rate at which essential elements in the soil are available to plants is critical to understanding soil fertility. Plants extract the elements they need from the soil as inorganic ions in the soil solution. The amount of each essential element in the soil that is available to plants changes rapidly as the moisture content of the soil changes and also depends on the rate at which organic compounds decompose to release organically bound elements as available inorganic ions. Less than approximately 1% of the total amount of most essential elements in the soil is present in an available form. Plant species differ greatly in the rate at which they need to acquire essential elements for adequate growth. The rate at which nutrients become available influences natural plant communities and is directly related to human food production. Most human food crops require 90–120 days to mature. Food crops must have a rate of nutrient availability many times faster than required by native ecosystems. A high-yielding grain crop of rice, wheat, or corn must acquire approximately as much phosphorus in 90 days as trees acquire from the same area of land in more than 20 years. In addition, tree roots usually penetrate more deeply and exploit a larger volume of soil than food crops. Therefore, the concentration of available nutrient elements near the soil surface must be considerably greater to supply adequately the needs of a food crop than to support tree growth.

Humans harvest and transport their food crops to a domicile some distance from the site where the crop was grown. Often the seed portion of the plant is consumed and only the less nutrient-rich stems, leaves, and roots of the crop plant are returned to the soil as organic residue. Considerable amounts of

organic residues are required to fertilize a crop plant, because these residues decompose slowly to release inorganic ions for crop growth. The common practice of burning residues facilitates rapid crop growth by releasing organically bound nutrients.

Historically humans have populated areas of soil with high levels of mineral fertility. These are commonly igneous or volcanic materials of basic mineral composition, sedimentary rocks such as limestone rich in calcium, magnesium, and phosphorus, and recent flood plains frequently renewed by depositions of material derived from fertile geologic materials and eroded surface soil. Where the mineral composition of the soils contains only small amounts of essential elements and a large amount of slow-growing natural biomass is present, a system of food production known as slash and burn is practiced. Although some essential elements are volatilized and lost, fire is the primary method of rapidly decomposing organic material and creating a short time period when nutrients contained are rapidly available as inorganic ions. If there is enough biomass, at least one crop can be successfully grown in the 90 days after burning. If correctly done, burning also assures that the surface temperature of the soil becomes high enough to reduce weed competition by killing most weed seeds near the soil surface. A second and third crop is often possible before the available supply of essential elements is exported from the field as human food and the rate of nutrient availability is reduced to a point where crop yields are low and/or weeds become a major problem. After the farmer abandons that land, a succession of native communities that are able to grow with lower rates of nutrient flow from the soil invade the site. After some years, the slow-growing native vegetation acquires enough nutrients in its biomass that it can again be cut, dried, and burned to obtain a site for another brief sequence of crop plants. This method of nutrient-availability management has numerous variations among different indigenous cultures. Only low human population densities are sustained by slash-and-burn agriculture because of the long periods of time (usually between 10 and 30 years and inversely related to the mineral fertility of the soil) that must be allowed for natural vegetation to accumulate sufficient quantities of essential elements needed to fertilize a food crop after burning. Where domestic animals are allowed to graze large areas of native vegetation, essential elements concentrated in their excrements are often collected and used to fertilize small areas of food crops. In areas where infrastructure enables crops to be exported and essential nutrients imported as concentrated fertilizer, continuous food-crop production is practiced on even the most chemically infertile soils. Many combinations and variations of these strategies presently exist throughout the humid tropics.

Mineral and Chemical Grouping of Soils

Broad geographic areas of soils with similar mineral and chemical composition are outlined in Figure 3. Many localized areas of soils that differ significantly from those described in the following discussions are present within each identified area.

Reworked Sediments of Low Fertility

Unlike many areas in temperate latitudes, where soils form from material that has been loosened and deposited by massive glaciations in rather recent geologic times, many areas in the tropics have land surfaces that have been exposed to weathering, soil formation, and erosion for many geologic ages. In these areas the present soils have formed from material almost devoid of minerals that contain essential elements, and soil acidity renders the low quantities present only slightly available to plants. Soils formed from such materials are not only deficient in available forms of the major plant essential elements, but also often deficient in plant-available forms of copper, zinc, boron, and sulfur. The nutrient-poor natural biomass growing on such areas often contains inadequate amounts of essential nutrients for slash-and-burn agriculture. Without the importation of essential nutrients, human habitation is often nomadic and limited to small, isolated sites of more fertile soil.

The watershed uplands between the Amazon and Parana rivers in Brazil are representative of infertile reworked material. The soils form in thick deposits of material that has been weathered through several cycles of erosion and deposition. Quartz, kaolinite, gibbsite, and iron oxides are the predominant minerals. Content of essential elements is very low, and the majority of the soils are Oxisols (Ferralsols) (*Nomenclature of Soil Taxonomy* is used to identify soils and, where applicable, the corresponding units of the *Soil Map of the World* follow in parentheses) and, in very sandy sediments, equally infertile Quartzipsamments (Ferralic Arenosols). The most infertile of these soils, Acrustox (Acric Ferrosols), support sparse native vegetation of grass and low shrubs called 'cerrado' in Brazil. Each year wildfires burn extensive areas during the dry season of the ustic SMR. Attempts to graze the native vegetation often result in nutrient deficiencies in cattle. Extensive areas of these soils are now being utilized for mechanized agriculture with external sources of lime and fertilizer. In addition to wheat, soybeans, and other food crops, sugar cane for ethanol fuel production is grown.

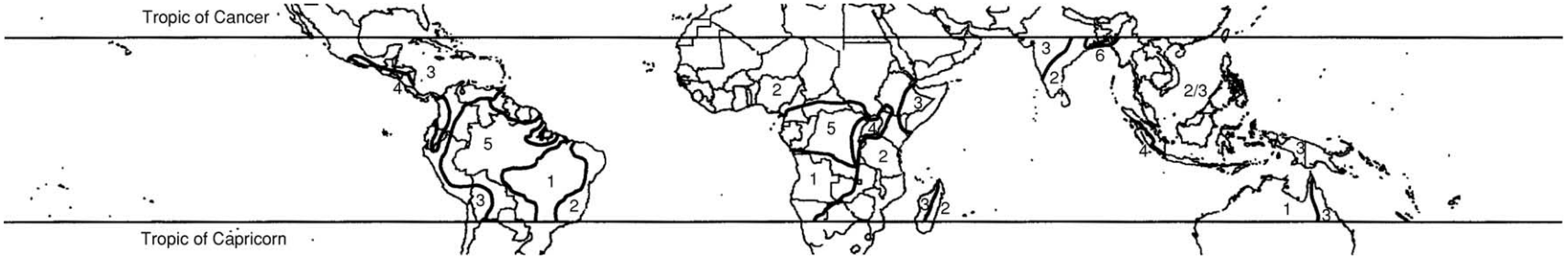


Figure 3 Generalized map of mineral and chemical soil regions in the tropics. 1, low-fertility sediments; 2, acid igneous rock; 3, limestone and other base-rich rock; 4, volcanic materials; 5, major river basins; 6, flood plains.

On inclusions of limestone, basalt, and sediments from these more fertile materials, Eutrustox (Rhodic Ferralsols) are present. These more fertile soils were naturally forested with seasonal deciduous forests that have been removed as farmers seek the more fertile soils for food-crop production.

Acid Igneous Rock

These areas are most extensive in Southeast Asia, parts of the eastern highlands in South America, West Africa, the central highlands of East Africa, and Eastern India. Ultisols (Acrisols) and Inceptisols (Cambisols) formed in these materials can be considered of moderate fertility but most commonly deficient in P and Ca. In most areas there are inclusions of more base-rich sedimentary rock where Alfisols (Luvisols) have formed. Most agriculture is limited to slash and burn unless infrastructure for markets and fertilizer exist.

Limestone and Other Base-Rich Rock

Where tectonic upheavals have exposed areas rich in carbonate and other nutrient-bearing rock, the soils are relatively rich in essential elements. Soils on the Deccan plateau of western India are fertile Vertisols (Vertisols) and Mollisols (Cambisols) formed in base-rich rock. The Andean chain along the west coast of South America and in Central America was formed by tectonic uplift and associated volcanic activity. Some of the uplifted materials are limestone, rich in carbonate, while other portions of the region are granite and schist, relatively poor in essential elements. Tectonic instability, steep slopes, and glaciers at the highest elevations have resulted in rapid erosion of surface materials and deposits of relatively fertile sediments in intermountain valleys of fertile Eutrudepts and Haplustepts (Eutric Cambisols).

Volcanic Material

A small proportion of the soils in humid tropical areas are formed in relatively nutrient-rich material derived from volcanic activity. Volcanic ash and basalt materials of various ages, including some recent volcanic activity, form Andisols (Andosols). Soils derived from volcanic ash and basalt historically have supported intense human settlement. Java, southern Sumatra, parts of the Philippines and other Pacific islands, Central America, the Andean mountains in South America and the Rift Valley in Africa are major areas of volcanic material. Many volcanic areas are mountainous and the higher elevations have isothermic or isomesic STRs where crop growth is reliable but cool temperatures slow crop growth at the higher elevations. Development and maintenance of roads are expensive due to the rugged relief, earthquakes, and volcanic activity. Transportation infrastructure is seldom reliable beyond local markets.

Major River Basins

The major river basins include the Amazon, Orinoco, and Paraguay river basins in South America, and the Congo basin in Africa (Figure 3). The basins are composed of sedimentary materials derived from surrounding uplands and mountains. The fertility of the soils reflects the mineral composition of the upland areas from which the sediments are derived.

The western portion of the Amazon basin and the central portion of the Congo basin have udic and perudic SMRs, while the eastern portion of the Amazon basin has an ustic SMR. Oxisols (Ferralsols) predominate in the Congo basin and eastern portion of the Amazon basin, where the sediments are derived from the Guyana and Brazilian shields. Ultisols (Nitosols) predominate in the western part of the Amazon basin, where the sediments are derived from the Andes and are slightly more nutrient-rich than those to the east. Both Ultisols and Oxisols are nutrient-poor, but inclusions of more nutrient-rich material are present. Alfisols (Luvisols and Eutric Nitosols) are formed in these more base-rich sediments in western Brazil and on the eastern boarders of the Congo basin. Many areas of these more chemically fertile soils are being cleared of forest vegetation for crop production and cattle grazing.

Some of the sediments in the basins are sandy. Soils formed in the sandy sediments are very infertile Quartzipsamments (Arenosols), where the water table is deep, or Spodosols (Podzols), formed where the water table is near the soil surface. It has been estimated that perhaps 10% of the Amazon basin has these kinds of soils. One major area of sandy soils is in the headwaters of the Rio Negro in the northwestern part of the basin, where it merges with the Orinoco watershed. Smaller areas of sandy soils are scattered throughout the basins, often detectable by coffee-colored 'black' waters of the rivers that drain these sandy watersheds.

Floodplains

Soils formed in river floodplain sediments and deltas are often fertile. The fertility is derived from eroded topsoil from the surrounding uplands. Where the mineral composition of the soil is relatively rich in essential elements, people are able to grow and harvest food with little or no attention to chemical fertilization. Crop yields are usually low but reliable. Most soils formed in floodplains are Entisols and Inceptisols, with aquic soil conditions (Gleysols). Despite the

physical dangers from flooding, human settlements are concentrated in these areas, attracted by the soil fertility and access to transportation afforded by the rivers. Road maintenance is extremely difficult where broad floodplains are prevalent. Most agriculture is subsistent unless the flooding hazards and periodic saturation can be controlled. Narrow areas of flood plains are present along coastlines and rivers throughout the humid tropics, but only a few areas are shown in Figure 3.

Organic Deposits

Organic remains accumulate in areas that remain saturated for most of the year, forming organic soils (Histosols). Histosols are present in many floodplain areas but contiguous areas are too small to outline in Figure 3. Cultivation is problematic and drainage systems are needed to aerate the soil and provide stable trafficking.

Human Utilization

The predominant soil limitation in humid tropical regions is low chemical fertility. Major areas of soil are formed from geologic material that contains very limited quantities of life-essential phosphorus, potassium, calcium, and magnesium. The most chemically infertile soils are present as uplands in the interiors of Africa and South America. Most of these soils are physically deep and have reliable moisture for one or two food-crops each year. The most infertile of these soils do not support enough natural biomass to sustain even slash-and-burn subsistence agriculture. Where rapid and reliable infrastructure has been developed, sustained commercial agriculture is now practiced. Reliable markets, fertilizer, lime, and fuel supplies are essential. A sufficient amount of phosphorus must be applied and mixed into the soil to saturate the iron and aluminum oxide surfaces to the extent that sufficient phosphorus becomes available before crops can be grown. Carbonate, in the form of crushed limestone, must be applied to raise the pH of the soil and inactivate the extractable aluminum. Nitrogen fertilizer is needed for nonlegume crops. Potassium is required for high yields, and small amounts of copper, zinc, boron, and molybdenum are needed in many areas. After an initial investment is made in altering the chemical conditions, fertilizer requirements are annually no greater than in other grain-growing soils around the world. Modern soil-testing technology is utilized to determine annual fertilizer formulations and rates. The physical stability of the inert soil minerals, a paucity of river systems that must be bridged, and the gentle topography aid in economical road construction in most parts of the region.

Somewhat less infertile soils with dense forest vegetation can be and are being used for slash-and-burn subsistence agriculture. The rate at which essential nutrients are released from the minerals is too slow to replenish those removed in rapidly growing food crops. The slow-growing forest vegetation that regenerates in abandoned fields can accumulate enough nutrients in its biomass that when cut and burned one or two food crops can be grown only every 10–30 years. These soils have the potential for more intense food crop production if chemically fertilized, which is seldom possible without a market infrastructure.

Human populations in the humid tropics are concentrated on soils formed from materials relatively rich in phosphorus, calcium, magnesium, and potassium. Such soils are primarily located in floodplains and deltas, in areas of volcanic activity, and in upland areas underlaid by limestone, base-rich igneous, and metamorphic rock.

Importing essential elements to compensate for those harvested in human food can increase crop yields from all soils. Applying animal and human wastes wherein the essential elements have been gathered from sites distant from the food-growing area is a partial solution in many cultures. However, the concentration of essential elements in organic residue and waste materials is low, transportation is difficult, and the rate at which the residues decompose to release inorganic ions for crop uptake is slow. Where commercial infrastructure is available, concentrated inorganic fertilizers have proven to be the most economical method of supplying nutrients harvested in human food.

Further Reading

Buol SW and Eswaran H (2000) Oxisols. *Advances in Agronomy* 68: 151–195.

Buol SW and Sanchez PA (1988) Soil characteristics and agronomic practices for sustainable dryland farming. In: Unger PW, Jordan WR, Snead TV, and Jensen RW (eds) *Challenges in Dryland Agriculture: A Global Perspective*. Proceedings of the International Conference on Dryland Farming, Bushland, TX: Amaillo.

Cochrane TT, Sanchez LG, de Azevedo LG, Porras JA, and Garver CL (1985) *Land in Tropical America*. Cali, Colombia: Centro Internacional de Agricultura Tropical (CIAT).

FAO–UNESCO (1971–1981) *Soil Map of the World*, vols I–IX. Paris, France: UNESCO.

Lepsch IF, Buol SW, and Daniels RB (1977) Soils–landscape relationships in the Occidental Plateau of Sao Paulo State, Brazil. I. Geomorphic surfaces and soil mapping units. *Soil Science Society of America Journal* 41: 104–109.

Lopes AS and Cox FR (1977) A survey of the fertility status of surface soils under "cerrado" vegetation in Brazil. *Soil Science Society of America Journal* 41: 742–747.

Lopes AS and Cox FR (1977) Cerrado vegetation in Brazil: an edaphic gradient. *Agronomy Journal* 69: 828–831.

Osher LJ and Buol SW (1998) Relationship of soil properties to parent material and landscape position in eastern Madre de Dios, Peru. *Geoderma* 83: 143–166.

Sanchez PA (1976) Properties and management of soils in the tropics. New York: John Wiley.

Soil Survey Staff (1999) *Soil Taxonomy: A Basic System of Soil Classification for Making and Interpreting Soil Surveys*, 2nd edn. Agricultural Handbook No. 436. Washington, DC: USDA–NRCS.

Tyler EJ, Buol SW, and Sanchez PA (1978) Genetic association of soils in the Upper Amazon Basin of Peru. *Soil Science Society of America Journal* 42: 771–776.

Van Wambeke A (1981) *Calculated Soil Moisture and Temperature Regimes: South America*. SMSS Technical Monograph No. 2. Ithaca, NY: Cornell University Press.

U

UNSTABLE FLOW

T S Steenhuis, J-Y Parlange, and Y-J Kim, Cornell University, Ithaca, NY, USA
D A DiCarlo, USDA Agricultural Research Service, Oxford, MS, USA
J S Selker, Oregon State University, Corvallis, OR, USA
P A Nektarios, Agricultural University of Athens, Athens, Greece
D A Barry, University of Edinburgh, Edinburgh, UK
F Stagnitti, Deakin University, Warrnambool, VIC, Australia

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

The importance of understanding unstable preferential flow processes cannot be overstated considering its relevance to crop irrigation, groundwater recharge, and vadose-zone transport of nutrients and contaminants. An important aspect of preferential infiltration is the residence time of water in the vadose zone, where contaminants and nutrients are degraded or exchanged in the media. The residence time is shorter for soils with preferential flow than without. Preferential flow can also be caused by macropores. Both types of preferential flow are similar in that the gravity force dominates sorptive processes. This article is concerned mainly with the unstable finger flow formation.

Conditions for Unstable Flow

Although the existence of fingered flow has been known since 1945, the practical importance of instability in stratified soils was not recognized prior to 1972, when Parlange and coworkers demonstrated that fingering can occur in a fine-over coarse-textured profile. In the early years, investigators were concerned with instabilities caused by gradients of temperature or solute concentration in miscible liquids. A few noticed that 'narrow tongues' formed in hydrophobic sands.

Several attempts were made in the 1950s and 1960s by Saffman and other researchers to relate the fingering in petroleum engineering to vadose-zone fingers; but this was problematic, because the first laboratory experiments used a viscous oil, glycerin, which was displaced by compressed air in Hele–Shaw cells. This is not an obvious analog to water displacing air in a stratified medium. As a consequence, viscosity became initially a dominant feature in many theories, although this is now known to be incorrect. This does not mean that some of the equations derived by Hele–Shaw cells cannot be properly reinterpreted. A necessary condition for instability, for water and air in a porous media, neglecting both viscosity and density of air, becomes:

$$K_s > Q \qquad [1]$$

where Q is the water flux imposed by the upper layer of fine material, and K_s is the saturated conductivity of water in the coarse layer underneath. The condition in eqn [1] is obviously necessary, because if Q was greater than K_s, the whole area would have to be saturated to carry the water and no finger would be present. It is a fundamental contribution of Hillel and coworkers to have noticed that eqn [1] is not constraining enough and that water entry in the coarse layer is associated with a 'water entry' suction for the water to penetrate the coarse layer. They further suggest that, upon rewetting, the water entry suction will be higher, resulting in drier fingers, leading to interesting hysteresis phenomena. In all cases, hysteresis is crucial by limiting lateral capillary diffusion of water which would otherwise remove the presence of fingers, as often happens in Hele–Shaw cells.

Finger Diameter and Structure

In early work in Hele–Shaw cells, the optimal width of the fingers was found by balancing the destabilizing effects of gravity and the stabilizing effect of surface tension. Finger width for soils were initially derived 'Green and Ampt soils,' with discontinuous wetting fronts. In the mid-1970s, Parlange and coworkers derived an expression for the finger

width, d, based on the analysis of Richards' equation, i.e., a diffuse front, yielding:

$$d = \pi \frac{S^2}{K(\theta - \theta_i)} \frac{1}{1 - Q/K} \quad [2]$$

where S is the sorptivity given by:

$$S_c^2 = \int_{\theta_i}^{\theta} D\left[\theta + \bar{\theta} - 2\theta_i\right] d\theta \quad [3]$$

where D is the soil-water diffusivity and θ_i is the initial water content, assumed small enough that the soil-water conductivity, K, at θ_i is negligible compared with its value at θ_c. The value of θ_c is a strong function of the initial water content and is not predicted by the theory. Equation [2] is valid for a two-dimensional finger in a slab chamber. In the field, where the fingers are three-dimensional π is replaced by 4.8.

Initially, for a soil initially dry, Parlange and coworkers assumed that θ_i in eqn [2] corresponds to saturation. However, fingers are rarely saturated in soils (as they are in a Hele–Shaw cells) and water content varies along the fingers. As shown in **Figure 1** the finger tip is the wettest (the red color) and it dries behind the finger. The moisture content, θ, varies with depth (measured from the interface between layers) according to the equation:

$$z = \int_{\theta_o}^{\theta} \frac{D d\bar{\theta}}{K - v(\bar{\theta} - \theta_i)} \quad [4]$$

where v is the constant downward speed of the fingers obtained after a short time, i.e., after all mergers have taken place and the fingers have reached a steady configuration. θ_o is the value of θ at $z = 0$ and, if we assume that the maximum value of θ corresponds to a water entry value θ_e, then eqn [4] gives:

$$vt = \int_{\theta_o}^{\theta_e} \frac{D d\theta}{K - v(\theta - \theta_i)} \quad [5]$$

which gives $\theta_o(t)$ when v and θ_e are known. In particular, when $t \to \infty$, θ_o approaches an asymptotic value $\theta_{o\infty}$ with:

$$K(\theta = \theta_{o\infty}) = v(\theta_{o\infty} - \theta_i) \quad [6]$$

When Q/K_c is negligible in eqn [2], we obtain a simpler equation for d:

$$d = \pi S_c^2 / K_c(\theta_c - \theta_i) \quad [7]$$

Note that both S_c^2 and K_c are inversely proportional to the viscosity; accordingly d is independent of viscosity. The influence of viscosity can be felt only through $[1 - Q/K_c]$ in eqn [2] and, thus, is irrelevant when fingering is important and Q/K_c is small.

Finally, for coarse sands, and as long as θ is not too small where the description of the Gardner–Ritsema soil-water conductivity is realistic:

$$K = K_c \exp \alpha(h - h_c) \quad [8]$$

where α is more or less constant and h is the matric potential. Then, eqn [3] shows that for a coarse sand:

$$S^2 \simeq 2(\theta - \theta_i) \frac{K_c}{\alpha} \quad [9]$$

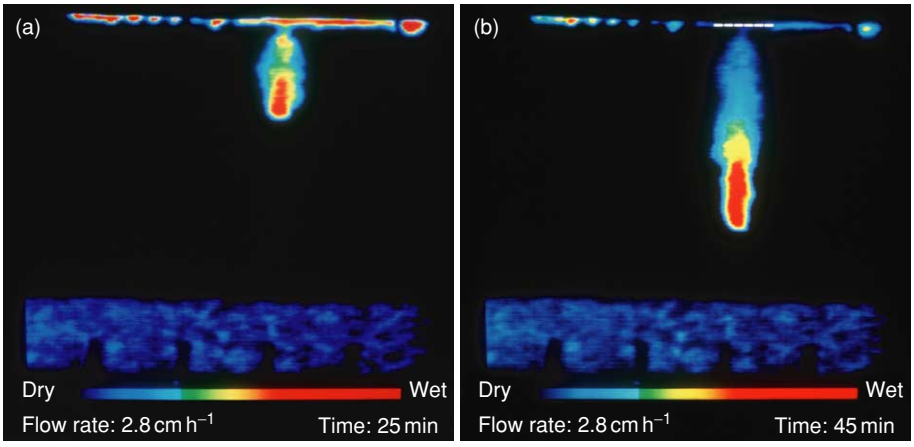


Figure 1 Finger formation in a coarse sand. Note that near the soil surface the distribution zone is visible. This zone carries the water from the rainfall that is uniformly applied at a rate of 2.8 cm h^{-1} to the one finger. Color indicates the moisture contents, with red the closest to saturation. Then the colors for decreasing moisture contents are: yellow, light green, light blue, dark blue, and black: (a) 25 min after water application; (b) 45 min after water application. Reproduced from Nektarios PA, Steenhuis TS, Petronic AM, and Parlange J-Y (1999) Fingered flow in laboratory golf putting greens. *Journal of Turfgrass Management* 3: 53–67.

Hence, we find the remarkably simple result:

$$d \simeq 2\pi/\alpha \qquad [10]$$

i.e., the dependence on θ_o and θ_i has disappeared. This explains why d is essentially constant in time and space and is inversely proportional to Gardner's α.

Note that, for water contents between θ_e and θ_o, all properties are measured on a drying curve of the matric potential. However, as the finger moves downwards within the sand, there is a very narrow zone at the finger tip where the water content increases rapidly, thus operating on a wetting curve, but reliable matric potential data are almost impossible to get in that region. Insight in this region was obtained by measuring reliable water contents down to the pore scale with neutron radiography. A 'dynamic' pressure, $h(\theta)$, was calculated based on Darcy's law by measuring the velocity of the finger, the distributed moisture content, and the unsaturated conductivity. The shape of the curve is similar to that which can be derived from earlier fingered flow experiments depicted in Figure 2 and is very much different from expected soil-water pressure relationships for the same soil. Although these effects are sometimes attributed to nonequilibrium effects, applying continuum results to the pore scale is problematic.

In fingered flow experiments where initial moisture content is varied, the finger size initially decreases when the moisture content increases from 0 to 0.005 $cm^3\,cm^{-3}$ and then later increases with increasing moisture content when it reaches the width of the chamber at 0.04 $cm^3\,cm^{-3}$ (Figure 3; note that the

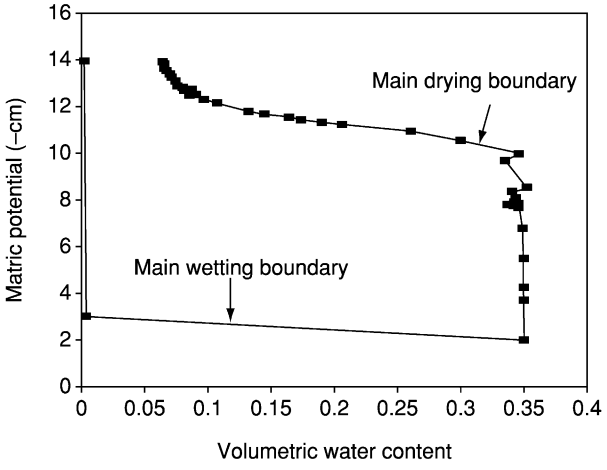


Figure 2 Nonequilibrium soil moisture curve obtained from fingered flow experiments. *h*, matric potential. Reproduced from Lin, Y, Steenhuis TS, and Parlange J-Y (1994) Closed-firm solution for finger width in sandy soils at different water contents. *Water Resource Research* 30: 951.

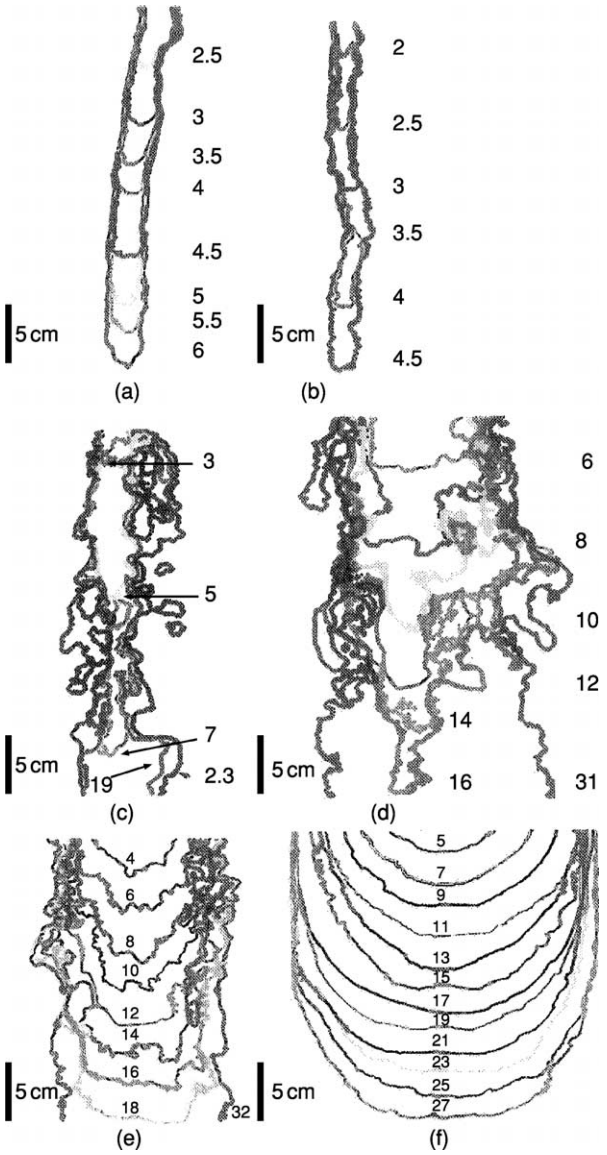


Figure 3 (a) Tracings of the totally dry sand. The sequence of tracings were taken every 30 s. The numbers accompanying the tracings are the minutes after infiltration started. A scale indicates 5 cm; (b) the advancing wetting front in the 0.01 $cm^3\,cm^{-3}$ moist sand. The infiltration front has a smaller width than in the totally dry case; (c) the advancing wetting front in the 0.02 $cm^3\,cm^{-3}$ initial moist sand pack. The tracings are depicted every 2 min. In addition, the latest tracing (after 23 min of infiltration) is shown; (d) advancing wetting fronts in the 0.03 $cm^3\,cm^{-3}$ moist sand. Two-minute tracings are shown until the front reached the bottom of the visible area. In addition, the tracing of the front is depicted when the infiltration was stopped (after 31 min); (e) advancing wetting fronts in the 0.04 $cm^3\,cm^{-3}$ initial moisture case. Two-minute tracings are shown until the front reached the bottom of the visible area. In addition, the front is depicted when the infiltration was stopped (after 32 min.); (f) the tracings of the wetting front when the chamber was imbibed from the bottom, or with a 0.04 $cm^3\,cm^{-3}$ (measured) initial moisture content. Tracings are depicted every 2 min. Reproduced with permission from Bauters TWJ, Dicarlo DA, Steenhuis TS, and Parlange J-Y (2000) Soil water content dependent wetting front characteristics in sands. *Journal of Hydrology* 231–232: 244–254.

surface induction zone is not shown). As θ_i increases, θ_c decreases rapidly and Q/K_c is not negligible in eqn [2] any more. Eqn [10] does not hold and d increases rapidly with θ_i.

The theory above applies only for a liquid, e.g., water or oil, and a gas, e.g., air. To extend the formula to the case of oil and water, it is necessary to obtain the dependence of the front velocity on its curvature. To obtain the dependence of the front velocity on curvature, the flow in the narrow diffuse region, ahead of it, must be analyzed, and to do so we must know the flow further ahead (here the pure oil). In the case of a gas, this presents no problem as it is assumed that the air can move freely ahead and does not affect the flow of water. Usually this will not be the case when oil is displaced by water, although it represents a limiting case.

The discussion has been limited, so far, to the movement of liquids. To include the transport of solutes, it is necessary to recall the physical configuration of the flow paths (**Figure 4**, where for clarity **Figure 1** has been redrawn schematically). The soil near the surface is more or less uniformly wet. This is the distribution zone in which water and solutes are funneled into the fingers of the conveyance zone below. The thickness of the distribution zone is that of the fine layer for layered soils and is of the order of the finger size for soils without such a layer. In the conveyance zone, all fingers move down with a velocity v. The number of fingers depends on the flow rate. For example, in **Figure 1** the flow rate is relatively low and only one finger is formed. In other experiments, where the flow rate is higher, several fingers form and, as a consequence, are spaced closer.

For the case when the initial concentration in the distribution zone is C_o, and the rainfall is solute-free, the concentration in the percolating water out of the distribution zone can be described as similar to a linear reservoir:

$$C = C_o \exp(-\lambda t) \quad [11]$$

where λ is the coefficient equal to q/w, q is the steady-state flow rate, w is the apparent water content of the distribution zone and equals $L(\rho k_d + \theta_s)$, d is the depth of the distribution zone, θ_s is the saturated moisture content, ρ is the bulk density of the soil, and k_d is the desorption partition coefficient. In case water is added with a solute concentration, C_o, the concentration in the water leaving the distribution zone is:

$$C = C_o(1 - \exp(-\lambda t)) \quad [12]$$

Equations [11] and [12] are equivalent to those used for sludge by the US Environmental Protection Agency in predicting the loss of metals from the incorporation zone.

It is reasonable to assume that the transport in the preferential flow paths of the conveyance zone can be described with the convective–dispersive equation, viz:

$$D\frac{\partial^2 C}{\partial x^2} - v\frac{\partial C}{\partial x} = \frac{\partial C}{\partial t} \quad [13]$$

where D is the dispersion coefficient, and v is the velocity of the solute and equals approximately $q/(\beta(\rho k_d + \theta))$, where β is the wetted fraction in the conveyance zone by fingers and θ is the moisture content. Using Laplace transforms, eqn [13] can be integrated subject to the boundary condition described in eqn [11], with no solutes initially present in the column, and for $4D\lambda/v^2 < 1$ as:

$$C = \frac{1}{2}C_o \exp(-\lambda t)\left[\exp\left\{\frac{vx}{2D}(1-a)\right\}\operatorname{erfc}\left(\frac{x - vta}{2\sqrt{Dt}}\right) + \exp\left\{\frac{vx}{2D}(1+a)\right\}\operatorname{erfc}\left(\frac{x + vta}{2\sqrt{Dt}}\right)\right] \quad [14]$$

where

$$a = \sqrt{1 - \frac{4D\lambda}{v^2}}$$

The last term can usually be neglected when x or t is sufficiently large, i.e., $(x + vta)/(4Dt)_{1/2} > 3$. For the boundary condition (eqn [12]), we can find the solution by superposition.

Note that in eqn [14], as discussed above, the velocity is a function of the soil properties and not of the flow rate. The number of fingers in the conveyance zone adjusts itself to facilitate the different flow rates. This has been demonstrated with a fingered

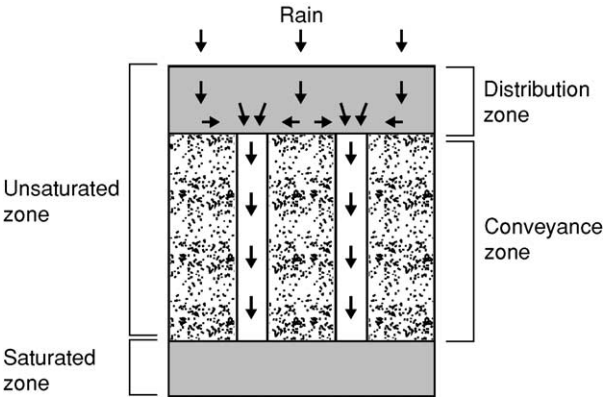


Figure 4 Schematic of the fingered flow process in the soil with preferential flow paths.

flow experiment that was carried out in the laboratory with 50-cm-long, 14-cm-diameter duplicate columns filled with coarse sand subjected to different intensities of steady-state rainfall. Chloride was applied at the surface as a pulse during each steady-state rainfall using different columns for different flow rates. The results are given as a function of cumulative flow (Figure 5a) rate and time (Figure 5b). In Figure 5a, the initial breakthrough of chloride takes more water for the high flow rate of 2 cm h^{-1} than for the low flow rate (0.3 cm h^{-1}). This is, obviously, a direct consequence of having a greater portion of the column wetted for the high flow rate. In Figure 5b, when chloride concentrations are plotted as a function of time, initial breakthrough for all three flow rates is approximately at the same time, clearly establishing that water and solute velocity is the same for the different flow rates. When the results are fitted to eqn [14], the solute velocity is, in all cases, approximately 35 cm h^{-1}, with no clear trend among the different flow rates, as expected.

In conclusion, it has become obvious in the last 25 years that, for soils in the field, preferential flow is the rule rather than the exception. Fingered flow is one of these types of preferential flow and can often be found above the primary coastal aquifers that are overlain by coarse-grained soils. Recognizing that these fingered flow paths can rapidly carry chemicals down to groundwater is a first step in keeping our aquifers clean for generations to come.

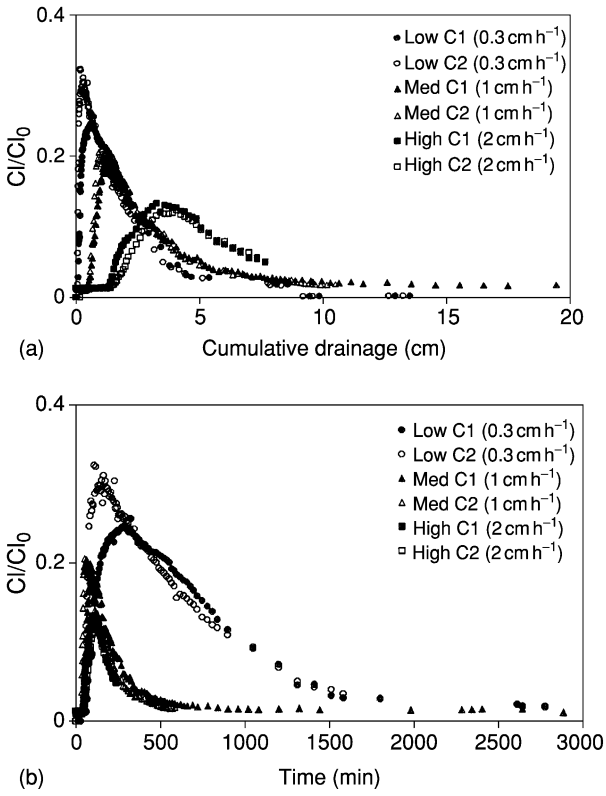


Figure 5 Chloride breakthrough curves (CBTs) for sand columns with a pulse application for three different flow rates of 0.3, 1, and 2 cm h^{-1} plotted: (a) as a function of cumulative drainage after application of the chloride pulse; (b) as a function of time after application of the chloride pulse. The experiment was carried out in duplicate with C1 and C2 indicating the two columns.

Further Reading

Bauters TWJ, DiCarlo DA, Steenhuis TS, and Parlange J-Y (2000) Soil water content dependent wetting front characteristics in sands. *Journal of Hydrology* 231–232: 244–254.

DeBano LF and Dekker LW (2000) Water repellency bibliography. *Journal of Hydrology* 231–232: 409–432.

Glass RJ, Steenhuis TS, and Parlange J-Y (1989) Mechanism for finger persistence in homogeneous, unsaturated, porous media: theory and verification. *Soil Science* 148: 60–70.

Hill DE and Parlange J-Y (1972) Wetting front instability in homogeneous soils. *Soil Science Society of America Proceedings* 36: 697–702.

Hillel D (1987) Unstable flow in layered soils: a review. *Hydrological Processes* 1: 143–147.

Parlange J-Y and Hill DE (1976) Theoretical analysis of wetting front instability in soils. *Soil Science* 122: 236–239.

Philip JR (1975) Stability analysis of infiltration. *Soil Science Society of America Proceedings* 39: 1042–1049.

Raats PAC (1973) Unstable wetting fronts in uniform and nonuniform soils. *Soil Science Society of America Proceedings* 37: 681–685.

Saffman PG and Taylor GI (1958) The penetration of a fluid into a porous medium or Hele–Shaw cell containing a more viscous liquid. *Proceedings of the Royal Society of London. Series A: Mathematical and Physical Sciences* 245: 312–331.

Selker J, Leclerq P, Parlange J-Y, and Steenhuis TS (1992a) Fingered flow in two dimensions. 1. Measurement of matric potential. *Water Resources Research* 28: 2513–2521.

Selker J, Parlange J-Y, and Steenhuis TS (1992b) Fingered flow in two dimensions. 2. Predicting finger moisture profile. *Water Resources Research* 28: 2523–2528.

Steenhuis TS, Boll J, Shalit G, and Merwin IA (1994) A simple equation for predicting preferential flow solute concentration. *Journal of Environmental Quality* 23: 1058–1064.

URBAN SOILS

J L Morel, C Schwartz, and L Florentin, Laboratoire Sols et Environnement ENSAIA-INPL/INRA, France
C de Kimpe, Agriculture and Agri-Food Canada, Ottawa, ON, Canada

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Soils in urban and suburban areas are transformed by human activities. For several decades, soil surveys and research were largely focused on agricultural and forest lands, and intensively managed and disturbed soils were not much investigated as revealed by the white (nonmapped) areas representing cities on most soil maps. Urban soils are used for many purposes, including urban and industrial activities, forestry, and agriculture. They are characterized by a strong spatial heterogeneity resulting from the various inputs of exogenous materials and the mixing of original soil material. The basic functions of natural and extensively modified soils are essentially the same. The evolution of urban soils is controlled by the same factors as natural soils, but the human factor imposes extremely rapid transformation cycles in comparison with those dominant under natural conditions. They often hold pollutants that may be a threat to human health. Anthropogenic soils can be investigated at least in part with traditional soil survey approaches; however these methods must be properly adapted and new methodology must still be developed. It is only through a multidisciplinary approach that urban soils will be better understood and their use optimized to protect human health and the quality of natural resources, e.g., preventing groundwater contamination.

Definition

'Urban soils' is a class of Anthropic soils, a term already used in several classification systems. Urban soils are soils extensively influenced by human activities, found mostly but not only in urban areas. They include: (1) soils that are composed of a mixture of materials differing from those in adjacent agricultural or forest areas, and that may present a surface layer greater than 50 cm, highly transformed by human activity through mixing, importing, and exporting material, and by contamination; (2) soils in parks and gardens that are closer to agricultural soils but offer different composition, use, and management than agricultural soils; and (3) soils that result from various construction activities in urban areas and that are often sealed. According to this definition, urban soils are essentially under strong human influence in urban and suburban environments; they may exert a strong effect on human health, on plants and soil organisms, and on water infiltration. They are differentiated from other strongly influenced soils such as those found in quarries, mines, and mine tailings, and airfields away from cities. However, it is sometimes difficult to set a clear boundary between urban soils and agricultural soils (Figure 1).

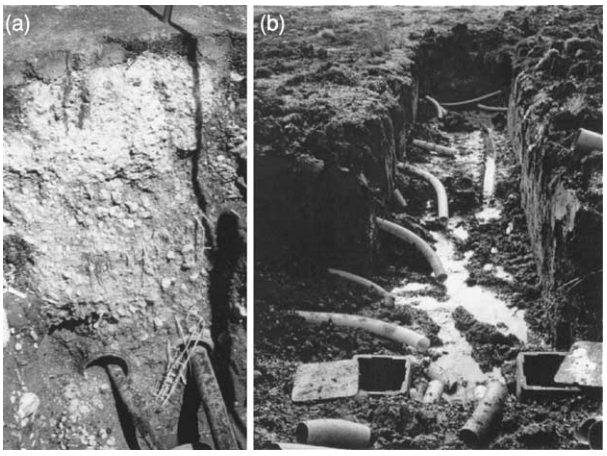


Figure 1 Urban soils and agricultural soils: (a) urban soil profile showing a sealed surface, imported soil materials, and electric wires; (b) agricultural soil profile showing strong perturbation induced by the incorporation of plastic drains.

Use of Soils in Urban Areas

In cities, soils provide support for infrastructure (buildings, roads, railways, parking lots, bridges), shelter for cables (electricity, telephone, television) and pipes of various size and composition (drinking water, wastewater, gas), and substrate for plants (isolated trees along streets, trees in public parks, and ornamental and edible plants in public and private gardens). They are also used for agricultural (horticulture, suburban agriculture, gardening), and industrial (buildings, mining, industrial waste disposal) production activities, and for recreation (stadiums, playgrounds). For centuries, the regions surrounding residential areas have been used for provision of construction materials and domestic waste disposal.

There are numerous soil types in urban areas as a result of these various uses, which differ in the intensity and duration of human impact. Human impact on urban soils may be light (e.g., urban forests) or, in contrast, maximal in developed built

areas (e.g., sealed and artificial soils). Therefore, soils in a city differ according to the degree of human transformation. In general, artificiality increases from the periphery to the center of the city, where original soils are often removed and replaced by anthropogenic materials. Older cities show the most modified soils, as they are generally constructed on their own waste materials that have accumulated over the years. For example, cities like Paris and Moscow are built on several meters of anthropogenic materials that hold remains of former human activities and materials.

Another main feature of urban soils is the high frequency of usage change with time. An example is the conversion of former industrial sites (now termed 'brownfield' sites) to new activities, including residential, public, and recreational activities. Such intensive changes are often dramatic and may cause problems for residents as a result of possible soil contamination. Landfills located at the periphery of cities are, knowingly or not, frequently converted to residential development as a result of the pressure of increasing urban population. This may result in increasing risks for residents from contamination through direct contact, inhalation, and/or consumption of garden produce.

Evolution of Soils in Urban Areas

As known from traditional soil science, a soil is the outcome of the evolution of a parent mineral material under the influence of climate, vegetation, topography, and time. Soils generally display a sequence of layers with an organically enriched surface horizon. Soil formation is generally a slow process that involves typically three main processes: weathering, transport, and accumulation. Differences in climatic conditions, parent rocks, and topography are at the origin of a wide range of soil types developed under natural conditions and, in general, the soil type can be deduced from knowledge of the conditions that have controlled its formation.

The three fundamental processes of formation (climate, parent rock, topography) and evolution of soils also apply to urban soils. However, urban soil formation is strongly influenced by the 'human factor' (Figure 2), which often creates a new 'parent rock' (e.g., debris from former buildings or industrial wastes), and modifies the conditions of its evolution through its influence on water circulation and composition in urban areas. The influence of the 'human factor' can also be interpreted in terms of soil formation and evolution: (1) weathering: transformation of the original material by mixing, compaction, or aeration of material layers; (2) transport: excavation of soil layers, leading to the partial or total elimination

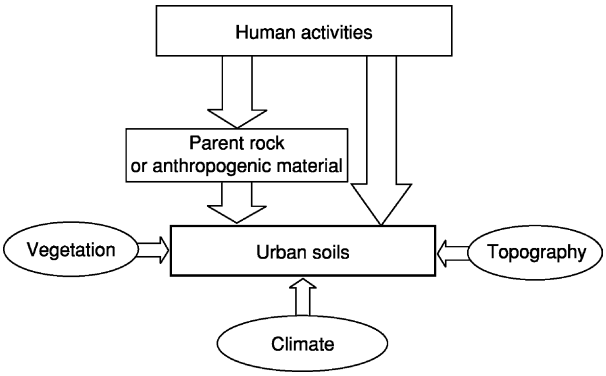


Figure 2 Urban soil formation and evolution.

of the original soil profile; and (3) accumulation: addition of exogenous materials from various origins (soil materials, minerals, technological compounds, and inert, organic, or toxic wastes). The kinetics of these processes are very rapid in comparison to natural processes, as a result of the increasing use of modern equipment, like tractors and bulldozers. A week or even a day is often the time scale required to modify completely the urban landscape. Heterogeneity in the landscape and in the soil may therefore change rapidly. Digging for new buildings, bringing in material from large distances for landscaping, disposing of rubbish, debris and topsoil for leveling and preparing the land for a new use often prevent any relation with the original parent material in the vicinity of the site being examined. Under such conditions, the natural weathering mechanism of soil formation does not play a great role if any, whereas transport and accumulation are predominant.

Another main feature of urban soils is the construction of barriers, e.g., sidewalk, which notably reduces the infiltration of water in soil, and increases its transport to streams, often causing flooding problems. Oxidizing conditions generally dominate in urban soils as a result of lack of water in the soil profile. However, soil compaction and leakage of drinking or wastewater from pipes may induce locally strong changes in redox and water flux conditions.

Composition and Heterogeneity of Urban Soils

The composition of soils in the cities depends largely on the nature of the materials in which they are developed. Coarse textures, with mostly sand and coarser material (rubble and gravel), generate great horizontal as well as vertical heterogeneity. This is a frequent characteristic of urban soils. Bulk density is generally low (<0.5) or very high (>1.60) as a function of the parent material. As an example, Figure 3 shows a soil

Figure 3 Soil profile in a twelfth-century urban area (Nancy, France).

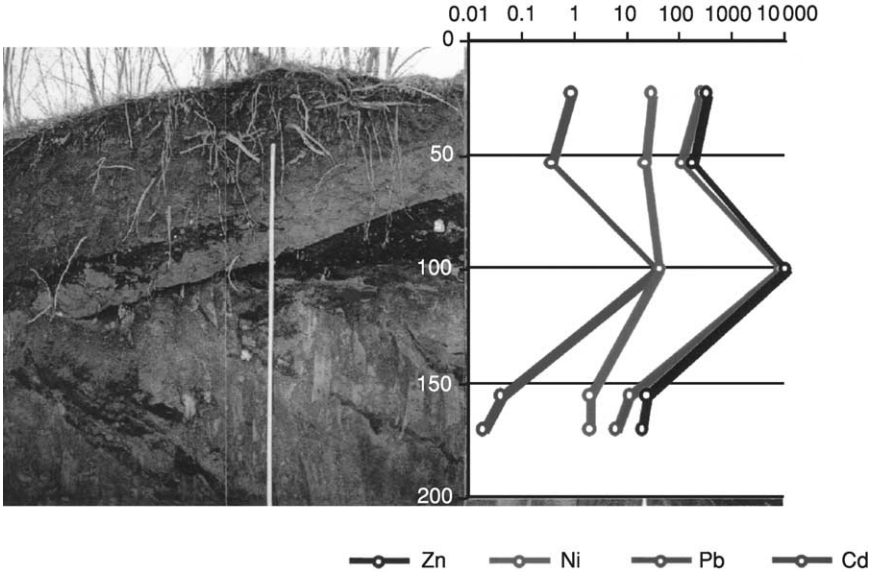


Figure 4 Soil profile and metal concentrations (mg kg^{-1}) in a former industrial site.

profile observed from a pit excavated for building construction in the city of Nancy, France. The soil exhibited a depth of more than 7 m and consisted of six distinct horizons. The three horizons at the bottom were the relic of a former agricultural soil. It presented a clear Ap horizon (15 cm) colored by organic matter, very compacted and weakly calcareous as a result of former agricultural practices or addition of calcareous construction material. The second horizon was not calcareous and offered a higher porosity with galleries of earthworms. The three upper horizons (0 to 1 m) were successive depositions of raw material. They were very heterogeneous and made of earthy material of sandy clay texture dominated by more than 80% of fine material. They were rich in organic matter and presented a brownish color. Numerous tree roots were present. The second horizon was strongly compacted and the third was characterized by the presence of building material of lighter density with fragments of dark gray schists coated with clayey sands.

Another example is given by a soil developed on a former industrial site (Figure 4). With the exception of the two upper horizons made of agricultural soil material brought in to facilitate the revegetation of the site, most of the material was of anthropogenic

origin, deposited during industrial activity and mixed and compacted at the closing of the factory. Many types of debris are present, including concrete, metal wires, slag, and industrial sand, each offering specific physical (density, porosity, mechanical resistance) and chemical properties (nutrients, metals, organic pollutants). In general, the analysis shows the high pH that is a common feature in anthropic soils, often as the result of alkalizing products mixed in the soil. Organic matter content may be high, and the C/N ratio is generally high because of the presence of organic material low in N, especially in industrial soil material; these high values are often due to contamination with oily wastes.

Soil Contamination in Urban Areas

Because of their diversified origin, urban soils may contain pollutants, the location, characteristics (including availability), and potential evolution of which must be established properly for future land use. Urban and suburban soils prompt a strong interest because of the growing public concern about the environment and human health. For example, urban horticulture provides a nonnegligible percentage of the food supply to large populations, especially in developing countries, and there is a need to improve this type of production while ensuring its safety, as well as addressing the issues related to a wide range of land uses. Also, in urban areas children are often in direct contact with soil material, and soil quality (pollutants, pathogens) in playgrounds may affect their health. Some urban soils, in particular those observed on brownfields where they were previously used as support for industrial production, may contain large amounts of mineral and organic pollutants that accumulated over time. The profile of the industrial site described above contains elevated amounts of some metals (Cd, Pb, and Zn), especially in the horizons made of industrial material, e.g., slag. Also organic contaminants (e.g., hydrocarbons, polychlorobiphenyls, herbicides) are commonly found in urban soils; polycyclic aromatic hydrocarbons (PAHs), a special class of hydrocarbons, some of which exhibit carcinogenic properties, are found in some former industrial sites (e.g., gasworks, coking plants). The soil shown in Figure 4 was also contaminated by lead and zinc (Pb and Zn), two metals widely present in urban areas (Table 1). An urban soil profile studied in the city of Nancy showed the influence of the material in place or brought in. These materials contributed to modifying the content of metals in the profile. The metal concentrations in the urban soil were lower than in the industrial site, but metals were present at significant concentrations in the upper horizon as a result of the various urban activities, including traffic and water runoff from zinc roofs.

Table 1 Concentration of metals in the profile of an urban soil (see Figure 3)

	Cd	Cu	Cr	Hg	Ni	Pb	Zn
Depth (cm)	($mg\,kg^{-1}$)						
20	0.67	66.2	65.0	2.00	28.3	415.1	325.7
175	0.08	11.8	39.5	0.04	24.5	28.4	50.4
199	0.75	41.0	59.7	0.06	50.0	21.4	119.3
215	0.26	20.2	44.7	0.04	28.6	26.6	70.4
294	0.11	15.0	54.0	0.02	32.6	24.3	63.4
725	0.03	5.8	25.7	0.02	13.4	26.0	32.1

Urban soils have different retention capacities for organic components, but their alkaline reaction often limits the mobility of heavy metals. In most urban soils, a pH shift to alkalinity constitutes an alkaline (carbonate) geochemical barrier in the topsoil. This barrier hinders the mobilization of heavy metals. Methods developed for agricultural soils (such as plant tests, selective extraction, microbiological and enzymatic tests) may be adapted for assessing the risk of transfer of pollutants to the food chain.

Garden soils

Gardens are a place of strong interactions between soils and human activities. In general, a very intensive form of agriculture is conducted, resulting in soils with high fertility and great diversity, deriving from the multiplicity of gardening practices. Soil quality (i.e., nutrient and pollutant content) in gardens is related to the quality of the parent materials, but through various inputs and modifications of the soil profile, the gardener is the most important factor in soil quality. Inputs range from traditional agricultural amendments and chemicals (manure, lime, pesticides, fertilizers) to domestic wastes and industrial amendments that may contain several inorganic and organic contaminants. In general, rates of amendment application are far higher than in traditional agriculture production, and garden soils tend to exhibit a deep upper horizon with a high concentration of organic matter and mineral nutrients, e.g., nitrogen and available phosphorus. In general, the organic matter content is directly related to the age of the garden. Therefore, the soil quality, hence fertility, of garden soils is as variable as gardening practices. This may be reflected by the biomass production of lettuce grown on various soil samples collected from a set of gardens (Figure 5). But pollutants may also accumulate in garden soils and be transferred to the food chain by direct consumption of vegetables.

Heavy metals tend to accumulate in garden soils and, in general, their concentration is on average twice that in agricultural soils, probably because of the input of various amendments to the garden soils (Figure 6). As for fertility, contamination is highly variable from one garden to another. This can be observed from the analysis of a set of soil samples collected from family gardens located in the Lorraine region (France) and in the nearby Saarland region (Germany) that showed a wide range of values, probably due to gardener practices and the proximity of industries. Soil physical properties and accumulation of heavy metals in the Ap horizon depend on parent material, substrates, and anthropogenic modifications. In Saarbrücken, only 18% of the garden soils were natural soils. Also metal concentration correlated well with soil fertility, as expressed by the total and available phosphorus contents (Figure 7). Finally, the content of heavy metals can be explained in decreasing order of importance by the age of the garden, previous use of the garden (including practices of the gardeners), natural concentration of metals in the soils, and atmospheric deposition.

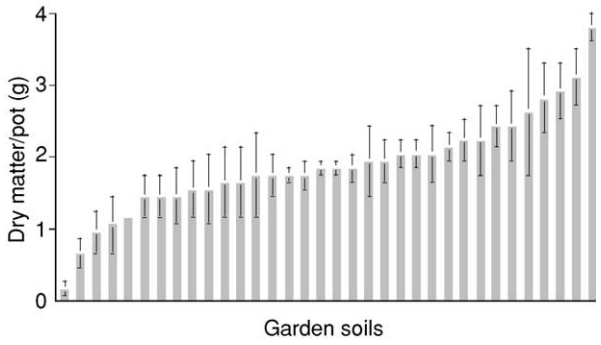


Figure 5 Biomass production of lettuce (*Lactuca sativa*) grown on a set of soil samples collected in family gardens in the Lorraine region (France).

Functions of Urban Soils

The link between soil functions and various soil properties such as substrate, texture, and humus content for soils at urban, industrial, and mining sites can be made using a set of indicators of soil quality such as those developed in Germany: rooting depth, wetting and aeration, nutrient status, and acid neutralization capacity. The high degree of surface sealing in urban soils limits the water partitioning that normally exists under natural conditions. Reduced seepage towards the subsurface causes high surface runoff and floods. Attempts are made to reduce this effect by increasing rainwater infiltration and groundwater recharge, and by reducing discharge by increasing water retention. Water and air supply in urban soils, an important feature for plant and tree growth, are controlled by bulk density, amount of medium, coarse, and very coarse mottles, organic matter content, and texture expressed as contents of clay, silt, and sand.

Urban soils are characterized by a great ecological heterogeneity, and show special distinctness of vegetation and fauna. They are habitats for plants and soil organisms, and for their filtering, buffering, and transforming of organic and inorganic pollutants. The root depth is, however, often limited due to abrupt horizon transitions, especially in the presence of a large percentage of coarse material (>2 mm). As a medium for plant growth, urban soil supports a large population of amenity vegetation in diversified habitats, including parks, gardens, roadsides, and turf areas. Much of the urban vegetation is cultivated, but there are also relics from natural vegetation or spontaneous infestation by opportunistic species.

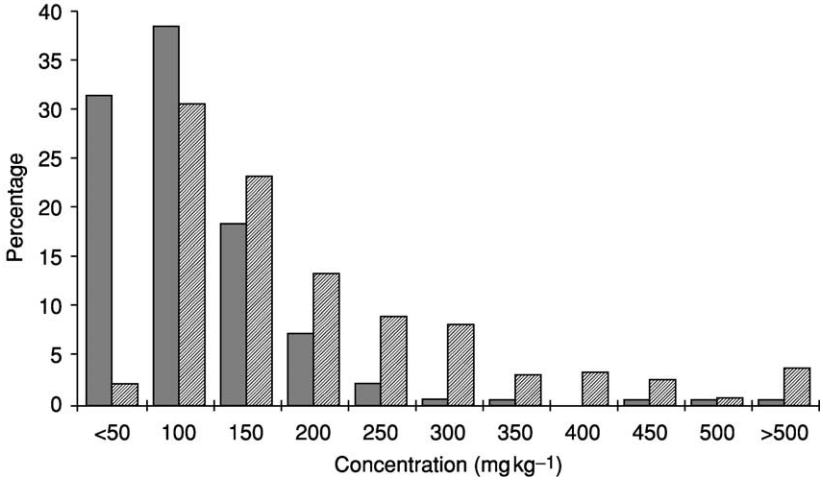


Figure 6 Frequency distribution of copper concentration in garden soils. ■ Regional soils ($n = 185$); ▨ garden soils ($n = 233$).

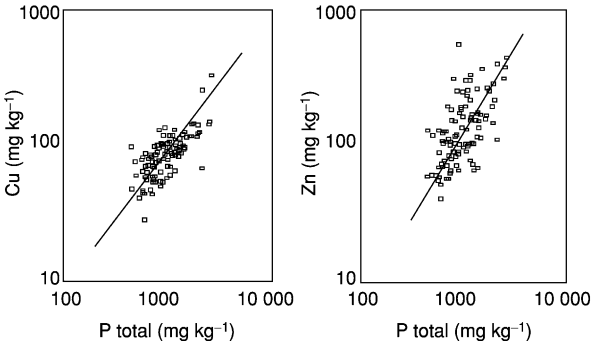


Figure 7 Relation between phosphorus and zinc and copper contents in garden soils.

The habitats may create stresses to vegetation survival: whereas normal root spread requires a circular soil disk preferably with a diameter equivalent to the crown spread and one meter deep, most urban tree roots are closeted in a narrow and shallow strip of substandard soil. In urbanized areas, roots of isolated trees are often sequestered in a restricted space; lateral room space is hindered, with a narrow strip, often less than 2 m, of poor-quality soil sandwiched between building foundations and highly compacted road material. The subsurface material quality is also often ignored. The presence of mortar, concrete, and asphalt attests to the widespread contamination of sites by construction debris that poses a physical hindrance to root development. Also the volume for water and nutrient storage is largely diminished.

Pedology and Archeology

Contrary to the situation in natural and slightly modified soils, e.g., agricultural or forest soils that most often display distinguishable horizons, a major challenge in urban soils, not only for their characterization, but also for determining their potential uses, is the heterogeneity of the layers associated with the fact that the latter are not always horizontal (Figure 3). Heterogeneity can also be very large within a single layer. This prompts a number of questions: what is an urban soil profile? What is an urban pedon? What area, both horizontally and vertically, is required for its description? What volume of soil is necessary for its characterization? How to describe the texture of urban soil, in particular the material coarser than sand? The presence of rubble may indeed confer special properties, and prevent some future uses of the soil.

Concepts must be properly developed in order to develop a suitable and useful classification scheme. Soil description involves the characterization of the successive layers with a specific composition. Urban soils are thus polycyclic soils composed by

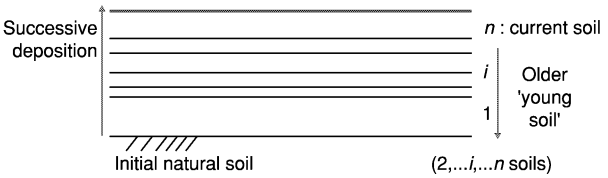


Figure 8 Urban soils in old cities are developed on waste material.

the superposition of several young soils (Figure 8). Using this approach, and as long as the human factor is clearly established, an urban soil survey is not strikingly different from a traditional soil survey: urban soils are the ultimate members of a continuum characterized by increasing human influence, and ranging from soils slightly or not affected by human influence (forests, rangeland) to agricultural lands and to urban soils where the human imprint is maximum, and where most natural features have often disappeared. For example, during early city development in previous centuries, when large machinery and transportation facilities were not available, rubble and debris were often disposed of at the site or at a short distance: cities were actually built on their own wastes. Therefore, the soils may contain layers or strata reflecting the processes that ruled the accumulation and such layers are the 'memory' of past activities: each layer corresponds to sedimentary materials representing an urban soil, generally poorly developed but significantly influenced by people.

Soils contain a large array of historical information, which has been proved very useful in understanding ecological and anthropological evolution. The materials brought about by human activities may also undergo pedological evolution leading to their significant transformation. In general, archeologists and pedologists have had only a few interactions in the past. Combining pedology and archeology by sharing research tools will increase our understanding of human evolution and help to predict better the problems connected to urban soils. Each layer of an urban soil contains products that are characteristics of the technological evolution of human societies. Combining archeology and pedology enables prediction of the location of some pollutants, such as heavy metals (Pb, Cu, Zn) and their fate in the profile.

Management of Urban Soils

The world is in the midst of a massive urban transition unlike that of any other time in history. In 1975, approximately one-third of the world's people lived in urban areas. By 2025, the proportion will have risen to almost two-thirds, which corresponds to more

than 5 billion people, mostly in developing countries. In developed countries, the population shift involves migration away from concentrated urban zones to large, sprawling metropolitan regions. Since the beginning of the twentieth century, the French population has increased considerably in cities, and it amounts to up to 70% of the population. In these large cities, as well as in smaller ones, green and open spaces play an important role as recreation areas, and they are the 'lungs' of urbanites. In the above context, characterizing the land that provides the infrastructure for these growing cities becomes particularly important. It is also a challenge: industrial and service businesses that are often located at the outskirts of the cities or in the suburbs near workers and employees are not particularly welcome amidst the new urban developments, and they are forced to move to a greater distance. This is favoring the emergence of brownfield sites, lands and buildings contaminated by previous industrial activities. Such areas are then recycled and often used for other human activities, that may present various risks for health. Waste disposal, especially the accumulation in landfills outside cities, creates similar concerns.

Within urban and suburban areas, soils are used as physical support for diverse anthropocentric activities and needs: infrastructure both at the soil surface and underground (e.g., roads, parking lots, railroads, pipes, sewers, ducts) and for buildings, sources (e.g., topsoil, landfill material from excavation) and sinks (e.g., dredged sediments, compost of organic urban wastes) of raw materials, food production, recreational activities (e.g., parks, landscaping), and memory and cultural heritage (e.g., cemeteries). Soil scientists have the expertise, as well as the social responsibility, to address issues related to these uses.

See also: **Applications of Soils Data**; **Land-Use Classification**

Further Reading

Arbeitskreis Stadtböden (1988) Substrate und Substratmerkmale von Böden der Stadt- und Industriegebiete. *Mitteilungen der Deutschen Bodenkundlichen Gesellschaft* 56: 311–316.

Bullock P and Gregory PJ (1991) *Soils in the Urban Environment*. Oxford: Blackwell Scientific Publications.

Burghardt W (1994) Soils in urban and industrial environments. *Zeitschrift für Pflanzenernährug und Bodenkunde* 157: 205–214.

Burghardt W, Zuzok A, and Heinen P (1987) Untersuchungen zur Kennzeichnung der Anreicherung und Verteilung von Schwermetallen in urbanen Böden. *Landschaft + Stadt* 30–38.

Craul PJ (1992) *Urban Soil in Landscape Design*. New York: John Wiley.

De Kimpe C and Morel JL (2000) Urban soils: a growing concern. *Soil Science* 165: 31–40.

FAO-UNESCO (1990) *Soil Map of the World – Revised Legend*. Rome: FAO.

Guillerme A (1999) Histoire du sol urbain. *Comptes Rendus de l'Académie d'Agriculture de France* 85: 129–140.

Hiller DA and Burghardt W (1993) Neues Leben im toten Boden. *Die Geowissenschaften* 1: 10–16.

Hollis JM (1992) Proposals for the classification, description and mapping of soils in urban areas. *English Nature Science* 2(4): 40.

United Nations (UN) Population Division (1995) *World Urbanization Prospects: The 1994 Revision*, pp. 132–139. New York: UN.

World Resources Institute (1996) Cities and the environment. In: *World Resources 1996–97*, pp. 1–30. New York: Oxford University Press.

V

VADOSE ZONE

Contents
Hydrologic Processes
Microbial Ecology

Hydrologic Processes

J W Hopmans, University of California–Davis, Davis, CA, USA
M Th van Genuchten, US Salinity Laboratory, Riverside, CA, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

The distinction between groundwater and the unsaturated zone is usually made within a hydrologic context that views water as the agent of change of the subsurface and the main driver for transport of chemicals between the atmosphere and groundwater. This region between the soil surface and groundwater table known as the vadose zone, involves a complex array of time-dependent, nonlinear physical, chemical, and biological processes, including interactions with groundwater and the atmosphere. The soil is the uppermost part of the vadose zone, subject to fluctuations in water and chemical content by infiltration and leaching, water uptake by plant roots, and evaporation from the soil surface. It is the most dynamic part of the subsurface, as changes occur at increasingly smaller time and spatial scales when moving from the groundwater toward the soil surface. Soil depth is usually controlled by the maximum rooting depth of plants (generally within a few meters of the soil surface). By contrast, the vadose zone may extend much deeper than the surficial soil layer and includes unsaturated rock formations and alluvial materials to depths of 100 m or more.

In the last few decades of the twentieth century, research interests in the deeper vadose zone have increased dramatically, instigated by a need to sustain quality of groundwater and maintain adequate resources for drinking water and ecologic purposes. Unquestionably, our society has negatively impacted the quantity and quality of its soil, water, and air resources. Chemical pollution generated by past agricultural, industrial, and municipal activities has contaminated soil and groundwater and surface-water systems worldwide. Unfortunately it continues to do so up to the present. Scientists and others are now increasingly aware that soil is a critically important component of the Earth's biosphere, not only because of its food-production function, but also as a safekeeper of local, regional, and global environmental quality. For example, management strategies in the vadose zone will offer the best opportunities for preventing or limiting pollution, and for remediation of ongoing pollution problems. Because chemical residence times in groundwater aquifers can range from a few years to thousands of years, their pollution is often essentially irreversible. Prevention or remediation of soil and groundwater contamination starts, therefore, with proper management of the vadose zone. Understanding the intricate processes in the vadose zone is a challenge because of the many complex nonlinear physical, chemical, and biological interactions that simultaneously control the transfer of heat, mass, and momentum between the atmosphere and the groundwater table.

Physical Processes

The physical characteristics of the vadose zone control such processes as natural or artificial recharge to deep groundwater and surface water–groundwater interactions. When considering the water budget of the vadose zone as a whole, many terms are similar and equally as important as those of a soil-water budget, with differences in measurement techniques mostly

predicated by differences in the spatial scale. Whereas typical soil-infiltration measurements are conducted over areas smaller than 1 m^2, estimates of infiltration rates over much larger areas and depths are required in the vadose zone, for which the assumption of one-dimensional flow is generally inadequate. Whereas the soil-water budget requires quantification of drainage, i.e., flow of water beneath the rooting zone (*See* **Drainage, Surface and Subsurface**), the water budget of the vadose zone includes net infiltration (*See* **Infiltration**), percolation, and recharge rates at large depths. 'Net infiltration' is generally defined as the water flux below the root zone that is not further influenced by evaporation or plant transpiration. 'Percolation rate' is defined as the net infiltration rate at any depth within the vadose zone, whereas 'recharge rate' defines the water flux into the groundwater across the groundwater table. Depth variations in percolation rates are caused by lateral flow and temporal variations in precipitation and evapotranspiration (*See* **Evapotranspiration**), whereas estimated rates are affected by measurement type. Specific vadose zone methods to estimate percolation and recharge rates include physically based methods that assume darcian water flow (*See* **Darcy's Law**) throughout the vadose zone, and tracer methods. The darcian methods generally require dedicated and highly specialized instrumention for large-depth measurements, but neglect fractured flow. Environmental tracer methods estimate percolation rate or age of water at a given depth based on *in situ* concentrations of natural tracers such as tritium, chloride, chlorine-36, and nonradioactive isotopes such as deuterium and oxygen-18, assuming that these are mass-conservative.

Accurate estimation of vadose zone water and solute fluxes is especially challenging in arid climates, where they are orders-of-magnitude smaller than in humid agricultural settings. However, because of these small flow rates, the relevant time scales of recharge rate can be orders-of-magnitude larger. A variety of methods have been reviewed, including detailed numerical modeling, to estimate percolation and recharge rates at Yucca Mountain, Nevada. The different methods were compared to study the potential and performance of Yucca Mountain as a repository site for high-level radioactive waste over time scales of 1000 years and longer (Figure 1). The mean water table depth is approximately 500 m, whereas the potential repository location is at a mean depth of 300 m below the land surface, within a densely welded and fractured tuff horizon. In February 2002, President George W. Bush endorsed a formal recommendation by the Department of Energy (DOE) for the Yucca Mountain to accept a total of approximately 85 000 metric tons of radioactive wastes.

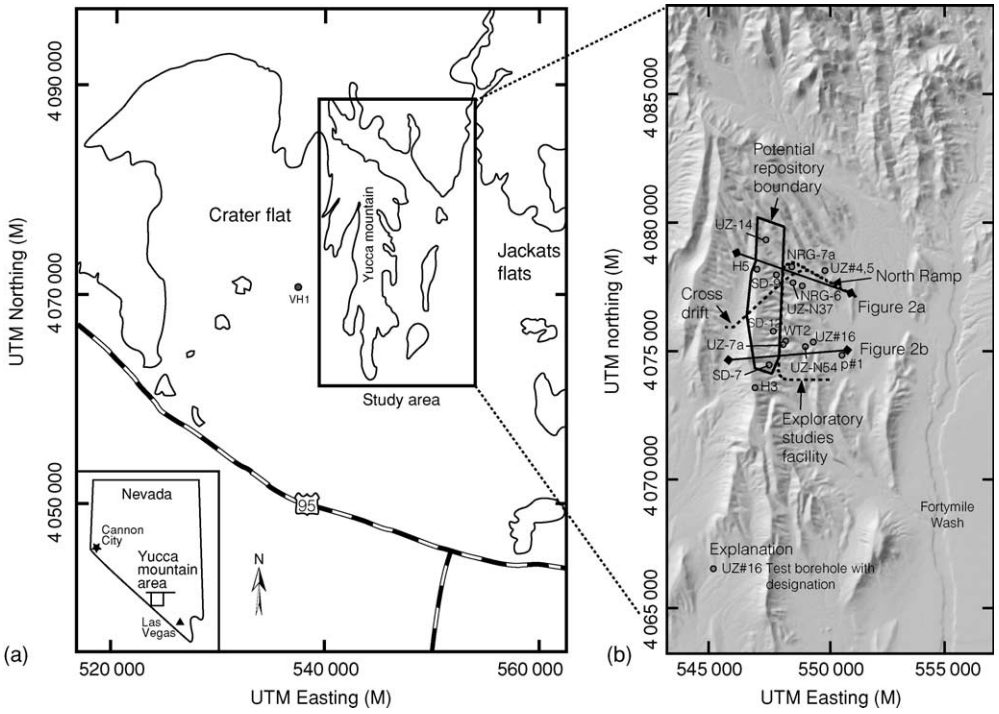


Figure 1 Location of (a) the Yucca Mountain area, Nevada, and (b) the study area. (Reproduced with permission from Flint AL, Flint LE, Kwicklis EM, Faryka-Martin JT, and Bodvatsson GS (2002) Estimating recharge at Yucca Mountain, Nevada, USA: comparison of methods. *Hydrogeology Journal* 10: 180–204.)

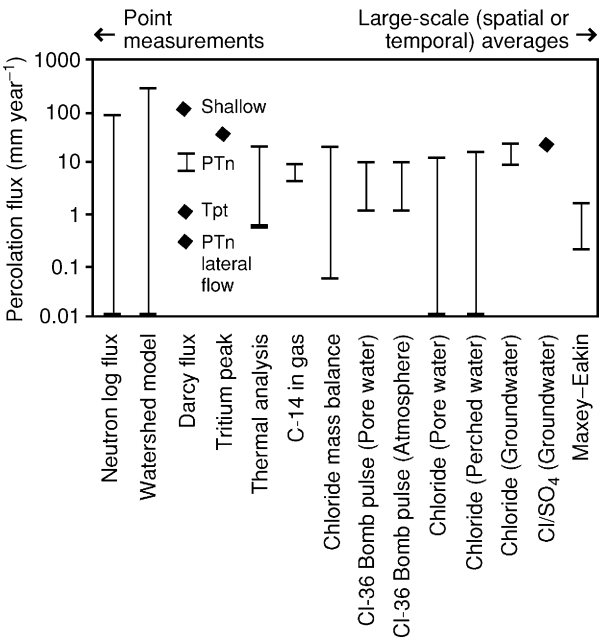


Figure 2 Comparison of percolation fluxes estimated by various methods. Bars represent ranges of estimates; points represent single estimates. PTn and Tpt indicate different hydrogeologic units of the Yucca Mountain site. (Reproduced with permission from Flint AL, Flint LE, Kwicklis EM, Faryka-Martin JT, and Bodvarsson GS (2002) Estimating recharge at Yucca Mountain, Nevada, USA: comparison of methods. *Hydrogeology Journal* 10: 180–204.)

Estimated percolation rates at Yucca Mountain vary between 0.5 and 20 mm year^{-1} (Figure 2), with differences in magnitudes controlled by variable precipitation, topography, soil depth, and highly variable, physical vadose zone properties, including fractures and faults. The different methods each yield percolation values typical of their specific space and time scales, with results complementing and partially overlapping each other. For example, shallow methods are representative for relatively small time scales, whereas rates estimated from the deep vadose zone and recharge measurement techniques are an integration of spatial mixing over much longer time scales.

Chemical Processes

The justification of vadose zone characterization and monitoring of contaminant transport arises from the simple observation that, for contaminants and microorganisms to reach the groundwater table, they must pass through the vadose zone first. Therefore, monitoring of the vadose zone and appropriate management practices using vadose zone concepts is a prerequisite for understanding and successfully preventing groundwater contamination. Much of the *Handbook of Vadose Zone Characterization and Monitoring* is devoted to current and emerging techniques for vadose zone measurement and monitoring of chemical and microbial pollutants. The general theories and concepts of chemical fate and transport in the vadose zone are largely identical to those occurring in soils. Therefore, the focus here is on the added complexities to which constituents are subjected as they move through the vadose zone. Most additional complications in vadose zone transport arise because of the much wider range in relevant space and time scales and the presence of preferential flow mechanisms (fractures, sediment, and rock beddings) that make it difficult to delineate transport networks for the larger vadose zone. Moreover, the increasing physical heterogeneity (*See* **Spatial Variation, Soil Properties**) and associated spatially variable soil moisture conditions and preferential flow mechanisms of the vadose zone affect geochemical and microbial reactions, resulting in spatial and temporal variations of concentrations and transport of which the implications have yet to be fully understood.

As dissolved solutes move through the vadose zone, various physical, chemical, and biological properties control their fate. In addition to diffusion and dispersion, the fate and transport of chemicals in the subsurface are influenced by sorption on to the solid phase and biological transformations. Diffusion and dispersion of the transported chemical are a function of both pore size distribution and water content. Mechanical or hydrodynamic dispersion is the result of water mixing within and between pores as a result of variations in the pore-water velocity. Increasing dispersivity values causes greater spreading of chemicals, thereby often also decreasing their peak concentration. Sorbed chemicals move through the vadose zone slower than noninteracting chemicals, with the degree of sorption largely depending on mineral type, specific surface area of the solid phase, and organic matter fraction. In addition, biogeochemical processes and radioactive decay affect contaminant concentration, for example by cation exchange, mineral precipitation and dissolution (*See* **Precipitation–Dissolution Processes**), complexation, oxidation–reduction reactions (*See* **Oxidation–Reduction of Contaminants**), and microbial biodegradation and transformations. Moreover, all of these processes depend on such environmental conditions as temperature, pH, water saturation, and redox status (*See* **Redox Potential**), including, their spatial variations.

The fate and transport of microorganisms (including pathogenic bacteria, viruses, and protozoan parasites) follow many similar processes to chemical compounds. For example, many microbes exibit

enhanced transport relative to the water solution as a result of electrostatic repulsion from negatively charged solid surfaces, as well as owing to size exclusion from smaller areas of the pore space. There is also evidence that biologically reactive solutes and bacteria preferentially adsorb to gas–liquid interfaces and may be subject to inferfacial adsorption–desorption processes. In addition, transport of microorganisms is controlled by their retention (such as filtration) in the porous matrix, which is a function of the size of the specific microorganisms, the water-filled pore size distribution, and pore-water velocity as well as concentration and ionic composition of the aqueous phase. As such, microbial biomass accumulation may reduce the porosity and alter the hydraulic properties of vadose zone soils. Additionally, inorganic, organic, and microbiologically active colloids with diameters between 0.01 and 10 μm can strongly adsorb otherwise immobile chemicals or microbes and thus facilitate their transport by acting as a mobile solid phase.

No better example of the complexities of vadose-zone transport can be illustrated than by highlighting progress made at the US DOE Hanford site in Washington State, toward measurement, characterization, modeling, and remediation of vadose zone contamination from leaking, high-level radioactive waste storage tanks. The subsurface fate of the nuclear waste is increasingly complex because of various potential preferential flow mechanisms occurring, such as fingered flow, tunnel flow, and flow associated with poorly sealed boreholes (**Figure 3**). Soon after World War II, the US Atomic Energy Commission built many 2.0- to 3.8-million-liter, carbon-steel single-shell tanks at the Hanford Site to store high-level waste generated from production of plutonium and uranium. Inventory studies show that more than 4 million liters of waste leaked from these tank farms into the surrounding vadose zone. The vadose zone at Hanford consists of permeable and poorly consolidated sands and gravels that are bedded but highly discontinuous in the lateral direction. Also, the vadose zone includes many sharp contrasts in texture between layers. The regional groundwater table at depths of 60–90 m below the surface discharges to the Columbia River, which flows adjacent to the Hanford site. Gamma logging data from approximately 140 dry wells (boreholes) surrounding the tanks shows contamination to depths of at least 40 m, probably mostly caused by preferential flow. The different preferential flow processes have caused widespread contamination of ^{137}Cs and other radionuclides, which otherwise would have been highly retarded. Fingered flow occurs because of density-driven flow, whereas tunnel flow is caused by zones of increased permeability around the tanks. The geologic complexities of the Hanford site, combined with man-made changes, have prevented adequate prediction of the transport of the various radioactive nuclides, even with the use of sophisticated multidimensional flow and transport models.

Whereas this example applies to point-source pollution, even more complications arise when responding to questions on nonpoint pollution of groundwater such as occurs through crop production by application of fertilizers (e.g., nitrates) and irrigation (e.g., salinity and toxic trace elements). Typically, using distributed modeling techniques, flow and transport

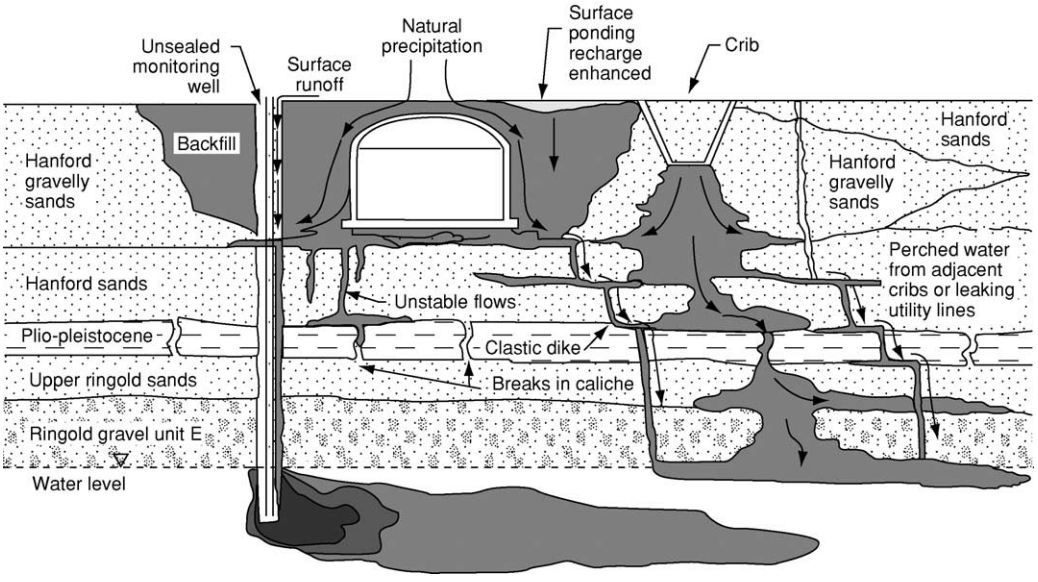


Figure 3 Conceptual model of fluid blow beneath single-shell tanks at the US Department of Energy's Hanford site.

are simulated for major characteristic units within a watershed or landscape (e.g., using soil map units, vegetation type, and slope). However, the connection and integration of the resulting simulation units within the watershed beg many questions of scale and associated nonlinearity effects. These nonlinear effects may be exacerberated when including chemical processes that are highly variable in space and time because of local soil variations in soil chemical composition, saturation, aeration, pH, redox status, and other factors. There is a need for interdisciplinary and multifaceted research approaches to improve the understanding of biological effects on soil chemical reactions and processes.

Biological Processes

In the soil environment, bioremediation and phytoremediation can be used to reduce vadose zone and groundwater contamination, by means of biological processes sustained by selected plants or microbes. Favorable environmental conditions of plant and microbial growth in soils and the vadose zone allow for bioremediation of contamination by inorganic and organic chemicals. Microbial processes mainly transform the parent chemical, thereby reducing its concentration, and by the formation of less-toxic metabolites, whereas plants can accumulate specific chemical species, e.g., through bioaccumulation of heavy metals in plant tissues. In addition, root water uptake by plants and trees can reduce leaching (*See* **Leaching Processes**) of water and salts below the rooting zone, thereby controlling percolation rates and contaminant transport through the vadose zone.

Differences between the microbiology of the vadose zone and the saturated zone are mostly caused by differences in water saturation, which affects the prevailing fluid fluxes and the availability of oxygen and other nutrients. With oxygen being the oxidizing agent for aerobic microorganisms, its supply may be limited in the deep vadose zones and/or in locally saturated (parched) areas. For most other nutrients, fluxes are controlled by the degree of water saturation and associated percolation rates and by organic matter supply rates. Consequently, nutrient availability is relatively abundant in the root zone and in humid climates, while their fluxes generally are limited in the vadose zones of arid climates. Although the vadose zone is generally unsaturated, local water-saturated inclusions may occur, promoting anaerobic microbiological processes, using electron acceptors other than oxygen, such as nitrate, Mn^{4+}, Fe^{3+}, or sulfate, depending on the redox status of the system. Irrespective of climate, the local nutrient supply in the vadose zone is at times predominantly controlled by diffusion to or from regions where transport is mainly by preferential flow through macropores (*See* **Macropores and Macropore Flow, Kinematic Wave Approach**) and cracks. Whereas, in general, microbial population density and activities are low in the vadose zone below the soil-rooting zone, their densities and activities can be orders-of-magnitudes higher at contaminated sites that facilitate microbial colonization. Microbes are generally located at air–water and water–solid interfaces; however, microbial heterogeneity is generally unpredictable, as it is conditioned by local variations in nutrient availability.

Enhanced microbial degradation (*See* **Pollutants:** Biodegradation) is generally achieved by gaseous nutrient deliveries of electron donors and acceptors to the contaminated sites, e.g., by bioventing through injection of air to simulate aerobic biodegradation of petroleum hydrocarbons (*See* **Hydrocarbons**). Microbial degradation of recalcitrant organic contaminants may require a secondary carbon supply, whereas other nutrients such as N and P are sometimes required. Inorganic contaminant concentrations of heavy metals or radionuclides can be reduced by microbes through their transformation (*See* **Metals and Metalloids, Transformation by Microorganisms**) to less-toxic states or by changing their mobility by using them as electron acceptors. A specific form of bioremediation is phytoremediation (defined as remediation through plants), which is mostly effective in near-surface soils. Many processes may contribute to phytoremediation (**Figure 4**), including specific root uptake followed by bioaccumulation and/or volatilization, and biodegradation in the rhizosphere (*See* **Rhizosphere**) sustained by root exudates and organic matter of decaying roots.

Certain plants have been adapted genetically to grow in soils containing toxic levels of metals. A classic example of this is the use of specific plants for the bioaccumulation of selenium, for example crop and grassland species. Research in selenium remediation methods accelerated in the 1980s after the discovery of high concentrations of Se in agricultural drainage water, followed by high mortality rates of grazing waterfowl at Kesterson Reservoir in Merced County, California. The specific advantage of crop or grassland plant species to remediate Se-laden soils is that these can be harvested and subsequently used as a Se supplement for Se-deficient forage or as an amendment for Se-deficient rangelands. Studies have demonstrated that grasses such as tall fescue can effectively take up Se if soils are supplemented with organic matter. Other successful examples include bioaccumulation of lead and other toxic metals, and radionuclides such as Cs, U, Cd, and Cr.

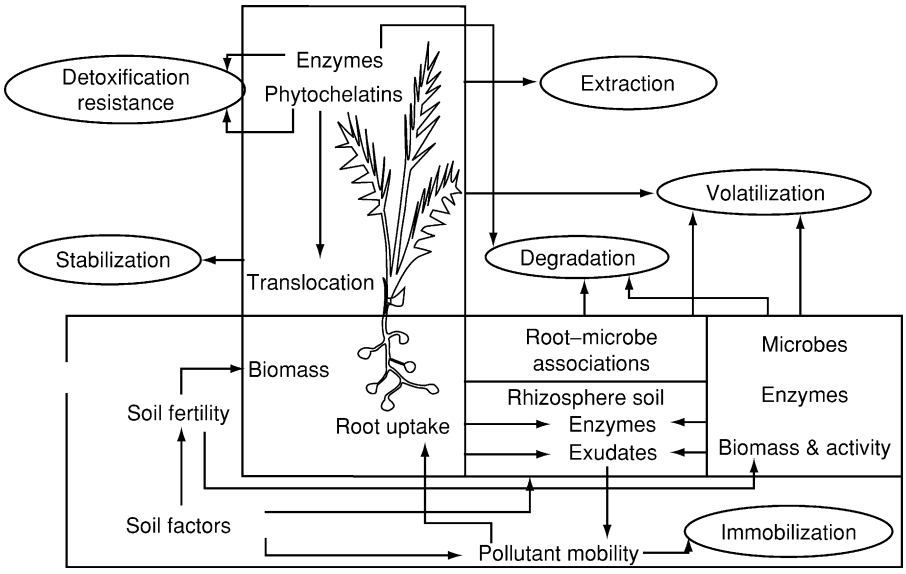


Figure 4 Simplified overview of soil–plant–microbe processes and interactions involved in phytoremediation. (Reproduced from Wenzel WW, Adriano DC, Salt DI, and Smith R (1999) Phytoremediation: a plant–microbe-based remediation system. In: Adriano DC, Bollag J-M, Frankenberger WT Jr, and Sims RC (eds) *Bioremediation of Contaminated Soils*, pp. 457–508. Agronomy Series No. 37. Madison, WI: American Society of Agronomy, Inc.)

There are many questions that need to be resolved before bioremediation techniques can be applied universally to a wide range of contaminant classes. Major issues revolve around the complex interactions between microbiological, geochemical, and hydrogeologic processes, the unpredictable local distribution of microsites, and the difficulty in supplying nutrients to these spatially distributed, contaminant-transforming microorganisms (*See* **Spatial Patterns**). In addition, successful bioremediation is controlled by the physical (e.g., soil moisture, temperature, oxygen diffusion) and chemical (e.g., organic matter type and content, soil adsorptive properties, availability of micronutrients) soil environment.

An additional example of mostly unknown territory occurs in the mixing zone between surface and subsurface water underneath streams, known as the hyporheic zone (HZ). This region of increased biochemical activity within the upper few centimeters of stream sediment affects the type and rate of material transformation as water moves downstream, thereby significantly changing stream-water composition and the stream ecosystem, as well as groundwater chemistry. Specific examples include the influence of oxygen supply to fish eggs buried in the HZ, the influence of particulate and dissolved organic matter on microbial activity and resulting stream-water chemistry, and the function of the HZ to denitrify, thereby ameliorating high-N stream load. There is a definite need to understand better the hydrology of the HZ and the coupling of hydrologic with biogeochemical processes, through dedicated interdisciplinary experiments and numerical modeling.

Scale Issues

For the past few decades, soil scientists have applied soil physical data to characterize flow and transport processes in large-scale, heterogeneous vadose zones, using measurement scales that are typically much smaller. For example, prediction of soil-water dynamics at the field scale is usually derived from the measurement of soil hydraulic properties from laboratory cores, collected from a limited number of sampling sites across large spatial extents. Soil parameters obtained from these small-scale measurements are included in numerical models with a grid or element size that is many times larger, with the numerical results extrapolated to predict large-scale flow and transport behavior. Because of the typical nonlinearity of physical properties, their use across spatial scales is inherently problematic. Specifically, the averaging of processes determined from discrete, small-scale samples may not describe the true soil behavior involving larger spatial structures. Moreover, the dominant physical flow processes may vary between spatial scales. Considering that soil physical, chemical, and biological measurements are typically conducted for small measurement volumes and that the natural variability of soils is enormous, the main question asked is how

small-scale measurements can provide information about large-scale flow and transport behavior. Answers to this question may require the estimation of appropriate, effective soil parameters for use in describing the behavior of pollutant plumes at the field or landscape scales.

A conceptual solution to the problem of scale issues of vadose-zone modeling might lie in considering the controlling effect of small-scale processes on larger-scale flow behavior. Hence, vadose zone properties are nonunique and scale-dependent, resulting in effective properties that vary across spatial scales and merely serve as calibration parameters in simulation models. Therefore, their accurate prediction in heterogeneous materials can only be accomplished using scale-appropriate measurements, including those that measure at the landscape scale.

Opportunities and Challenges

There are a number of opportunities that have come about in the past decade through experimental innovations and increasing environmental awareness. These opportunities include improved physical characterization of the vadose zone at larger spatial scales, which needs to be extended to chemical and biological measurements also.

First, it is becoming increasingly clear that there is a pressing need for subsurface observations and property-measurement techniques at spatial scales much larger than the usual laboratory or field-plot scale. The scale problem is extremely complex because of the general presence of large spatial and temporal variabilities of the soil physical, chemical, and biological properties in question. For example, at the heart of many hydrologic projects lies the need to understand better the flow and transport of water and associated chemical constituents into and through the vadose zone above an aquifer, and within a watershed. Hence, developmental work is needed on fundamental concepts and measurement technologies to establish appropriate soil parameters for use in theories or models describing the behavior of vadose zone water flow and pollutant plumes across spatial scales. In addition, appropriate measurement techniques and field experiments are needed to characterize effective field-scape and landscape-scale soil properties. In particular, we note the potential of using inverse methodologies perhaps in combination with rapidly improving invasive and noninvasive geophysical techniques, to infer *in situ* dynamic soil physical characteristics and the development of instruments that combine multiple measurements within a single device, thereby minimizing soil heterogeneity effects.

Second, although it is evident that large-scale characterization is needed, there is also increasing awareness within the scientific community that the physical, chemical, and biological processes in the vadose zone are controlled by mechanisms operating at the pore-size scale. Improved predictions of subsurface flow and transport will probably be a function of the development and application of innovative pore-scale modeling approaches (e.g., Lattice–Boltzmann, percolation, and related methods), and associated measurement techniques that operate at the microscopic level. Examples of the latter are nuclear magnetic resonance (NMR), computed tomography (CT), and spectromicroscopy.

Third, improved characterization and interpretation of subsurface processes will require increasing efforts toward an interdisciplinary partnership, integrating physical with chemical and biological processes. The soil physicist and/or vadose zone hydrologist must seek collaborations in other disciplines to ensure that different measurement types are collected and integrated to study more effectively and determine relationships between flow and transport processes at the microscale and in the laboratory, with ultimate application to the watershed scale. To achieve this goal, soil physics must be taught within the broader context of hydrology and the environmental sciences.

See also: **Darcy's Law**; **Drainage, Surface and Subsurface**; **Evapotranspiration**; **Hydrocarbons**; **Infiltration**; **Leaching Processes**; **Macropores and Macropore Flow, Kinematic Wave Approach**; **Metals and Metalloids, Transformation by Microorganisms**; **Oxidation–Reduction of Contaminants**; **Pollutants:** Biodegradation; **Precipitation–Dissolution Processes**; **Redox Potential**; **Rhizosphere**; **Spatial Patterns**; **Spatial Variation, Soil Properties**

Further Reading

Adriano DC, Bollag J-M, Frankenberger WT Jr, and Sims RC (1999) *Bioremediation of Contaminated Soils*. Agronomy Series No. 37. Madison, WI: American Society of Agronomy.

Ainsworth CC, Brockman FJ, and Jardine PM (2000) Biogeochemical considerations and complexities. In: Looney BB and Falta RW (eds) *Vadose Zone Science and Technology Solutions*, pp. 829–947. Columbus, OH: Battelle Press.

Cheng HH and Mulla DJ (1999) The soil environment. In: Adriano DC, Bollag J-M, Frankenberger WT Jr, and Sims RC (eds) *Bioremediation of Contaminated Soils*, pp. 1–13. Agronomy Series Number 37. Madison, WI: American Society of Agronomy.

Dane JH and Topp GC (2002) *Methods of Soil Analysis,* part 4, *Physical Methods*. SSSA Book Series No. 5. Madison, WI: Soil Science Society of America, Inc.

Ferré TPA and Kluitenberg GJ (eds) (2003) Advances in measurement and monitoring methods, special section. *Vadose Zone Journal* 2: 443–654.

Findlay S (1995) Importance of surface–subsurface exchange in stream ecosystems: the hyporheic zone. *Limnology and Oceanography* 40: 159–164.

Flint AL, Flint LE, Kwicklis EM, Faryka-Martin JT, and Bodvarsson GS (2002) Estimating recharge at Yucca Mountain, Nevada, USA: comparison of methods. *Hydrogeology Journal* 10: 180–204.

Hopmans JW and Bristow KL (2001) Current capabilities and future needs of root water and nutrient uptake modeling. *Advances in Agronomy* 77: 103–183.

Hopmans JW, Nielsen DR, and Bristow KL (2002) How useful are small-scale soil hydraulic property measurements for large-scale vadose zone modeling. In: Smiles D, Raats PAC, and Warrick A (eds) *Heat and Mass Transfer in the Natural Environment*, pp. 247–258. Geophysical Monograph Series no. 129. Washington, DC: American Geophysical Union Publications.

Parlange MB and Hopmans JW (1999) *Vadose Zone Hydrology: Cutting across Disciplines*. Oxford, UK: Oxford University Press.

Shafer DS, Bertsch JF, Koizumi CJ, and Fredenburg EA (2000) Gamma borehole logging for vadose zone characterization around the Hanford high-level waste tanks. In: Looney BB and Falta RW (eds) *Vadose Zone Science and Technology Solutions*, pp. 445–457. Columbus, OH: Battelle Press.

Sparks DL (2001) Elucidating the fundamental chemistry of soils: past and recent achievements and future frontiers. *Geoderma* 100: 303–319.

Stephens DB (1996) *Vadose Zone Hydrology*. Boca Raton, FL: CRC Press.

Tyler SW, Scanlon BR, Gee GW, and Allison GB (1998) Water and solute transport in arid vadose zones. In: Parlange MB and Hopmans JW (eds) *Vadose Zone Hydrology: Cutting Across Disciplines*, pp. 334–373. Oxford, UK: Oxford University Press.

Wenzel WW, Adriano DC, Salt DI, and Smith R (1999) Phytoremediation: a plant–microbe-based remediation system. In: Adriano DC, Bollag J-M, Frankenberger WT Jr, and Sims RC (eds) *Bioremediation of Contaminated Soils*, pp. 457–508. Agronomy Series No. 37. Madison, WI: American Society of Agronomy, Inc.

Wilson LG, Everett LG, and Cullen SJ (1995) *Handbook of Vadose Zone Characterization and Monitoring*. Boca Raton, FL: Lewis.

Wood TI and Faybishenko B (2000) Large-scale field investigations in fractured basalt in Idaho: lessons learned. In: Looney BB and Falta RW (eds) *Vadose Zone Science and Technology Solutions*, pp. 396–405. Columbus, OH: Battelle Press.

Wu LW (2000) Selenium accumulation and uptake by crop and grassland plant species. In: Frankenberger WT Jr and Engberg RA (eds) *Environmental Chemistry of Selenium*, pp. 657–686. New York: Marcel Dekker.

Microbial Ecology

P A Holden and N Fierer, University of California–Santa Barbara, Santa Barbara, CA, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

The vadose zone consists of unsaturated porous media and rock extending from the surface soil to the groundwater table. This zone is characterized by a lack of water relative to other major Earth compartments such as the ocean, fresh water, aquatic sediments, and groundwater. Low nutrient availability, low water content, and low potential energy per unit volume of water, termed 'water potential,' constrain life in the deeper vadose-zone relative to the surface. However, the vadose zone harbors numerous and diverse microbes, including bacteria, fungi, protozoa, and viruses. Vadose-zone microbial ecology refers to the study of interactions between vadose-zone microbes and their environment. Generally, microbial population density, diversity, and total activity, as well as available nutrients, and moisture and temperature fluctuations, decline sharply with depth below surface soil. For example, relative to surface soil, culturable heterotrophic bacteria in the unsaturated subsurface are only one-tenth as numerous. Regardless, the integral mass of microbes along the depth profile is large, which accounts for the overall importance of vadose-zone microbes as catalysts in terrestrial nutrient-cycling, including pollutant biodegradation.

Definition of the Vadose Zone

The vadose zone is the Earth's terrestrial subsurface that extends from the surface to the regional groundwater table. As shown in Figure 1, the vadose zone includes surface soil, unsaturated subsurface materials, and a transiently inundated capillary fringe. The subsurface materials include partially weathered soils and unweathered parent material. The vadose zone may be very shallow (less than 1 m) or very deep (extending hundreds of meters or more), depending on the depth to the water table.

The vadose zone has low water content relative to the saturated zone below the water table and is therefore commonly referred to as the unsaturated zone. Above the capillary fringe, vadose-zone pore spaces are generally air-filled, with thin water films coating solid particles. Pore spaces become water-filled when rainfall percolates, followed by drainage and gradual drying. It is the relative lack of water and its

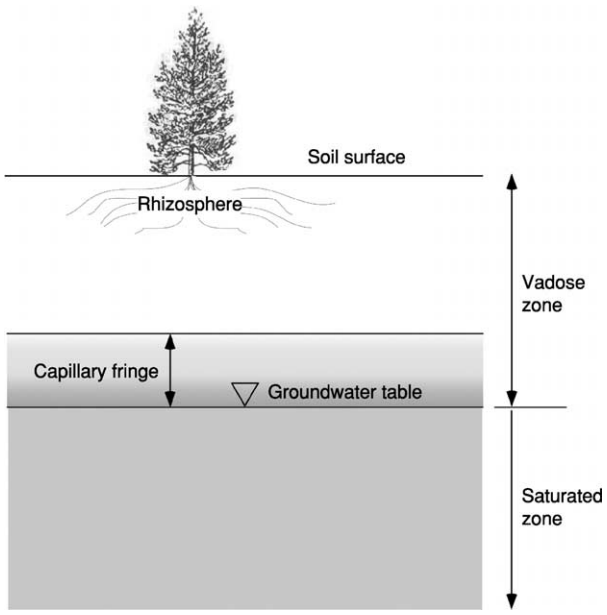


Figure 1 The vadose-zone profile.

transience, the complete lack of sunlight, and complex physical and chemical gradients that make the vadose zone an interesting and unique compartment for microbes.

Vadose-zone microbial ecology is the study of the microbes residing in the vadose zone, their interactions with each other, and their interactions with their surrounding environment. Compared with soil microbiology, a field for which a significant amount of agriculturally oriented research has been conducted, vadose-zone microbial ecology is a relatively young science. However, questions regarding the presence of microbes in the vadose zone, where they live, what they do, and how they do it are integral to the work of biogeochemists, environmental scientists, and engineers concerned with the fate of nutrients and pollutants.

Abundance and Distribution of Microbes in the Vadose Zone

Quantity Along a Depth Gradient

The abundance and distribution of microbes in the vadose zone are key parameters that need to be measured in order to understand vadose-zone microbial ecology. Abundance is studied by sampling (typically by aseptically coring through the face of a trench or vertically from the surface) and laboratory quantification at each depth interval sampled. Laboratory quantification of microbial abundance is performed using a variety of methods, including: directly counting eluted bacteria, culturing, measuring substrate uptake and/or mineralization, measuring total phospholipid fatty acids (PLFAs), measuring biomass carbon (C), quantifying extractable deoxyribonucleic acid (DNA), or counting using a most-probable-number (MPN) technique. Specific phylogenetic or functional groups of microbes can also be sensitively quantified using molecular biological approaches such as either fluorescence *in situ* hybridization (FISH) coupled with directly counting cells, or quantitative polymerase chain reaction (Q-PCR). Typically, each of these methods provides estimates of microbial abundance that are in general agreement with one another. Along a depth gradient (Figure 2), microbes tend to be most numerous in the surface soil, with their abundance declining sharply below the surface then reaching a nonzero plateau in the subsurface. The initial rate of decline with depth is more rapid in some vadose-zone systems than in others. Microbial population densities tend to increase in the capillary fringe and at the water table. In shallow vadose zones, the majority of the microbial biomass is in the surface material. In deeper vadose zones, microbial biomass densities are highest in the surface regions, but the total microbial biomass in deeper materials can be high due to the integration of sparser population densities across a large spatial area.

Generally, the number of culturable cells decreases more rapidly with depth than the number of total extractable cells. For example, by culturable counts, subsurface populations of heterotrophic bacteria may be one-tenth the size of surface populations, but, by direct counts, subsurface populations may be one-fifth the size of surface populations. In general, the

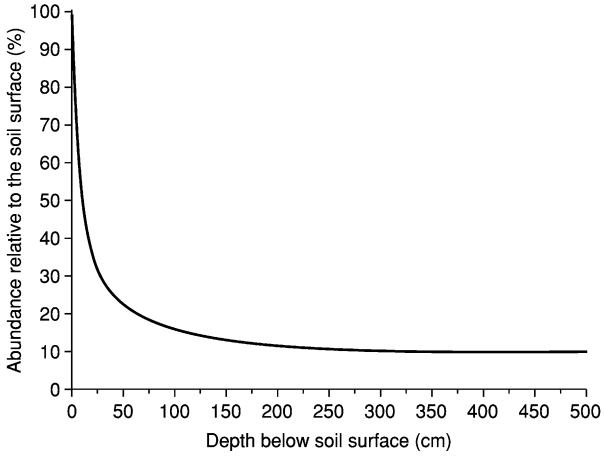


Figure 2 Typical depth trend of the vadose-zone constituents that are relevant to microbial ecology.

total numbers of culturable bacteria are higher in polluted vadose materials.

Protozoan population densities decline rapidly with depth below surface soil, which may be attributable to the concomitant decline in abundance of bacteria, their primary food source. Direct counts of fungal populations also decline with depth, but at a more dramatic rate than bacteria. Overall, total nonbacterial microbial biomass, including protozoa, algae, diatoms, and fungi, is lower than bacterial biomass in surface soils and decreases more dramatically with depth.

While the majority of soil microbial biomass is found in the surface horizons (Figure 1), where nutrients are most abundant (Figure 2), these horizons also experience extreme fluctuations in moisture and temperature which may result in significant temporal variability in population sizes. Microbial population sizes in the subsurface tend to experience a minimal degree of seasonal variation.

Distribution at the Meso- and Microscales

The distribution of microbes with depth through the vadose zone, as described above, is useful to know if we want to build large-scale mathematical models of microbially mediated processes in the vadose zone. However, there is also a secondary nature to microbial distribution which occurs at the meso-, and micro-, or perhaps 'microbe-' scales in soil. As depicted in Figure 3a, mesoscale variation in microbial abundances depends on the spatial distribution of bulk soil properties, including flow paths that preferentially conduct water through both surface and subsurface soils. Different microbial communities have been found along preferential flow paths as compared to the surrounding vadose material. In surface soils, preferential flow paths harbor more carbon and therefore more microbes; in the subsurface, this phenomenon is attenuated. At the microscale, the scale of individual pores, microbes are less active in the smallest pores where the diffusional resupply of nutrients is restricted. In contrast, larger, interconnected pores tend to contain more metabolically active microbes.

At the scale of individual microbes, which in some cases may be at the pore scale, there are at least three known possible configurations for microbial growth habits (Figure 3b). These configurations are mainly applicable to bacteria, the most abundant microbes in the vadose zone, but may also apply to other microbial particles, including fungal spores, viral particles, and protozoa. The distribution of filamentous fungi is generally a mesoscale phenomenon, because fungal hyphae can structurally, and physicochemically, bridge pores and particles. As shown in Figure 3b,

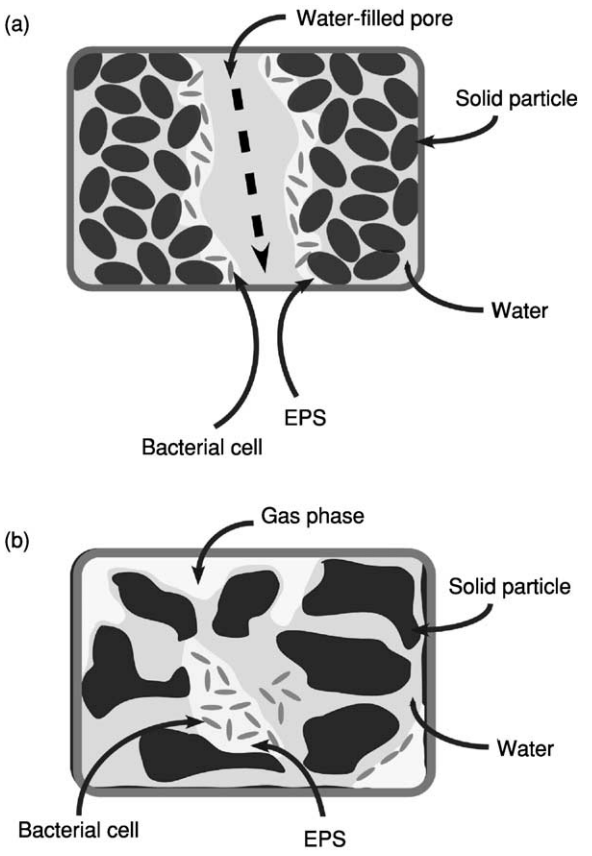


Figure 3 (a) Mesoscale and (b) microscale orientations of vadose-zone bacteria relative to other constituents, including extracellular polymeric substances (EPS) surrounding bacterial cells.

microbes can colonize the air–water interface, exist planktonically in water, or attach to surfaces and grow in biofilms consisting of cells and their associated expolymeric substances (EPS). The growth habit dictates movement: microbes at the air–water interface and in bulk water are mobile, but microbes embedded in EPS are relatively immobile. The growth habit also affects the availability of nutrients and the susceptibility to predation: microbes in EPS tend to be more protected from predation but often have reduced access to nutrients. Compared with what we know regarding gross microbial distributions at the macro- or depth-gradient scale, we know very little about microbial distributions at smaller scales and how specific configurations affect larger-scale vadose zone processes.

Diversity of Vadose Microorganisms

The microbes inhabiting the vadose zone are diverse taxonomically, metabolically, and morphologically. There are thousands of bacterial species (species richness) in a gram of surface soil. The proportional

representation of each taxonomic group (evenness) varies with depth within a vadose zone and between different types of vadose zones. For example, the overall abundance of Gram-positive bacteria has been shown to increase relative to Gram-negative bacteria along a vadose-zone depth profile, and there are lower fungal-to-bacterial ratios at depth than at the surface.

Taxonomic diversity is assessed with a range of techniques, including: the identification of culturable microbes, PLFA fingerprinting, fingerprinting based on 16S ribosomal DNA (rDNA), and the enumeration of different 16S rDNA sequences in a clone library. Metabolic diversity can be assessed by exposing serial enrichment cultures or microcosms to a variety of substrates and measuring the potential for substrate metabolism. Both taxonomic and metabolic diversity decline with depth through the vadose-zone profile, following the general trend depicted in Figure 2. In a related pattern, the number of dead and viable-but-nonculturable cells (VBNC) tends to increase with vadose-zone depth. Morphological diversity can be assessed by the microscopy of cultivated cells, but, as with any diversity assessment based on conventional laboratory culturing, the method may yield biased estimates of diversity because so few soil microbes (less than 1%) can be cultivated.

A fundamental question in vadose-zone microbial ecology is: How did the microbes, particularly those found at greater depths, get there? In part, diversity studies address this question by looking for similarities between the microbial community composition of the surface soil and that of the deeper vadose zone. If identical taxonomic groups are found at the surface and at depth, this suggests that microbes are transported downward through the vadose zone. If there are microbial groups found at depth that are unique from those at the surface, this suggests that microbes may have resided in the initial parent materials before the formation of the vadose zone. Generally, while there is overlap in the taxonomies of surface and deep vadose microbial communities, there are also unique groups.

Water is the main agent responsible for the inoculation of the subsurface by the surface. Thin water films in the vadose zone typically limit the transport of microbes through the profile. However, water and gas transiently advect and diffuse through soil pores. Preferential flow of water through certain pores causes these paths to have differing chemical and biological characteristics when compared with surrounding soil. Further, recharge of water to the vadose zone varies and, in turn, this leads to additional variations in pore chemistry and microbial communities along preferential flow paths. Overall, site hydrology is a key determinant of the abundance and types of microbes inhabiting the vadose-zone profile. However, at any given depth, the variability in microbial community composition and abundances can be high, and this heterogeneous distribution contributes greatly to microsite-level variation in microbial processes.

Activity of Vadose-Zone Microbes

Microbial metabolic activity declines with depth in the vadose zone, roughly paralleling the curve in Figure 2. The observed decline in activity with depth is related to the previously mentioned declines in microbial viability and abundance with depth. However, *in situ* activity is strongly affected by available nutrients. For example, *in situ* denitrification rates at depth in the vadose zone are generally much lower than the rates measured for surface soils (top 10 cm). The decrease in denitrification rates is partly a result of the low abundance of denitrifying bacteria, but is primarily a result of the low carbon availability in the deeper depths of the vadose zone. When subsurface vadose-zone samples are amended with nutrients, the activity per viable microbe in the subsurface is approximately the same as that in the surface.

Physiological markers of stress indicate that vadose-zone microbes are subject to starvation with increasing depth. Starvation is enhanced when water availability is low, because both the intrinsic physiology of microbes and the substrate availability to microbes are affected. Microbial biodegradation activity declines with depth, as shown with simple analysis of CO_2 evolution patterns. However, degradation rates can be as high or higher when either water or aqueous-phase nutrients are added to the vadose zone.

Measurements of enzyme activity are used to assess enzyme abundance in vadose-zone samples. Dehydrogenase, a useful measure of oxidative activity in all microbes, positively correlates with bacterial abundance, with dehydrogenase activity generally decreasing with depth below the surface. Other enzymes related to carbon, nitrogen, phosphorus, and sulfur metabolism show similar patterns. The low enzyme activity of deeper soils appears to be related to the low numbers of microbes and the low quantities of available nutrients.

Physical and Chemical Characteristics Affecting Vadose-Zone Microbial Ecology

The vadose zone is an open, multiphase system (Figure 3b) containing dissolved and undissolved solids (organomineral complexes and biomass), liquids

(water and nonaqueous-phase liquid pollutants), and gases (also known as soil gases). Open systems are inherently dynamic which, in the vadose zone, means that water, gases, and solutes are continuously being redistributed. The concept of chemical equilibrium in the vadose zone is only valid over short distances and short time scales. The ramification for microbes is that their local environment is in a state of continuous flux, due to either microbial metabolism or external environmental factors. Six key environmental factors are commonly cited as important to vadose zone microbial ecology: water availability, temperature, nutrient availability, oxygen, redox potential, and pH. Nutrient availability, redox potential, and oxygenation change with alterations in soil-water content: drying soil increases air-filled soil pore space, while wetting inhibits oxygenation and lowers redox potentials. Water is therefore considered to be a unifying environmental factor in the vadose zone.

Vadose-zone microbial ecology is affected by both soil-water potential and soil-water content (θ). Total soil-water potential (ψ) is defined as the potential energy per unit volume of soil and has two main components in the near subsurface: matric potential (ψ_m) and solute potential (ψ_s). In nonsaline soils, ψ_m is the dominant component of ψ. Total soil-water potential is related to the relative humidity of surrounding soil gas by:

$$\psi = \frac{RT}{V_w} \ln\left(\frac{\text{rh}}{100}\right) \qquad [1]$$

where R is the gas constant ($P - L^3/\text{mol} - T$), T is temperature (T), V_w is molal volume of water (moles per cubic liter), and rh is relative humidity (percentage).

As per eqn [1], ψ is negative when the rh is less than 100%. Alternatively, ψ is zero when the rh is equal to 100% (i.e., when water activity or $a_w = 1$). The vadose-zone environment may have a ψ equal to zero over a range of water contents that qualitatively span moist to fully saturated. For all $\psi = 0$ conditions, soil-water content indicates the degree to which pores are air-filled and thus oxygen-bearing. As soil dries, both θ and ψ decrease, but ψ is most relevant to microbial ecology. At approximately -0.25 MPa, vadose-zone bacteria become slightly nutrient-stressed, because the diffusional resupply of solutes is restricted. A total soil-water potential of -1.5 MPa is approximately the lower limit of what is physiologically tolerable to many surface soil bacteria, but filamentous fungi and actinomycetes are often more resilient to even lower water potential conditions. Vadose-zone microbes can adapt to be extremely desiccation-tolerant, as evidenced by the presence of viable microbes in deep, arid vadose zones where the total water potential is *c.*-50 MPa or less.

Water is an important physical environmental factor governing vadose-zone microbial ecology; temperature is another. For both water and temperature, average conditions vary from surface to subsurface. However, perhaps more importantly, both water content (and potential) and temperature are more variable in surface soils as compared to deep in the vadose zone. For example, temperature extremes are greatest at the soil surface but either increase or decrease (depending on the season) with depth to a constant temperature of around 20°C (*See* Figure 4 in **Thermal Properties and Processes**). Similarly, surface soils wet and dry frequently depending on the climate and season, yet subsurface soils are more stably moist or dry. In general, more stable temperature and moisture regimes at depth will select for vadose-zone microbial communities with narrower temperature and moisture optima and with lower tolerance for extreme or rapid fluctuations in environmental conditions.

The availability of nutrients to microbes in the vadose zone declines steeply with depth below the surface to low equilibrium values (**Figure 2**). In many vadose-zone environments, nutrients are more abundant in the surface soils because of plants and their contributions to soil nutrition. However, the textural composition of the vadose zone also changes with depth, affecting nutrient availability. For example, the percentage of clay, which controls the availability of exchangeable ions, typically decreases with depth. Further, the percentage of organic matter, which controls the abundance and availability of dissolved and available sorbed organic nutrients, also decreases with depth.

Soil gas composition affects redox status and aeration, and it often changes with depth through the vadose zone. For example, the oxygen content of soil gas in the subsurface is generally below the atmospheric concentration of 21%, often dropping to 15% or less. Also, CO_2 concentrations often range from 0.033% at the surface to as high as 4% or more at soil depths below 2 m. The CO_2 increase is due to the combined effects of microbial respiration and low rates of diffusional exchange between gases above- and below-ground. The high rates of CO_2 production and limited diffusional transport tend to increase alkalinity at depth, resulting in higher pHs in the subsurface compared with the soil surface. The unique gas conditions of the deeper soil depths facilitate the selection of microaerophiles and carboxyphiles.

The physical morphology of the vadose-zone particles is complex near the soil surface and more homogeneous in the subsurface (**Figure 4**). Undoubtedly,

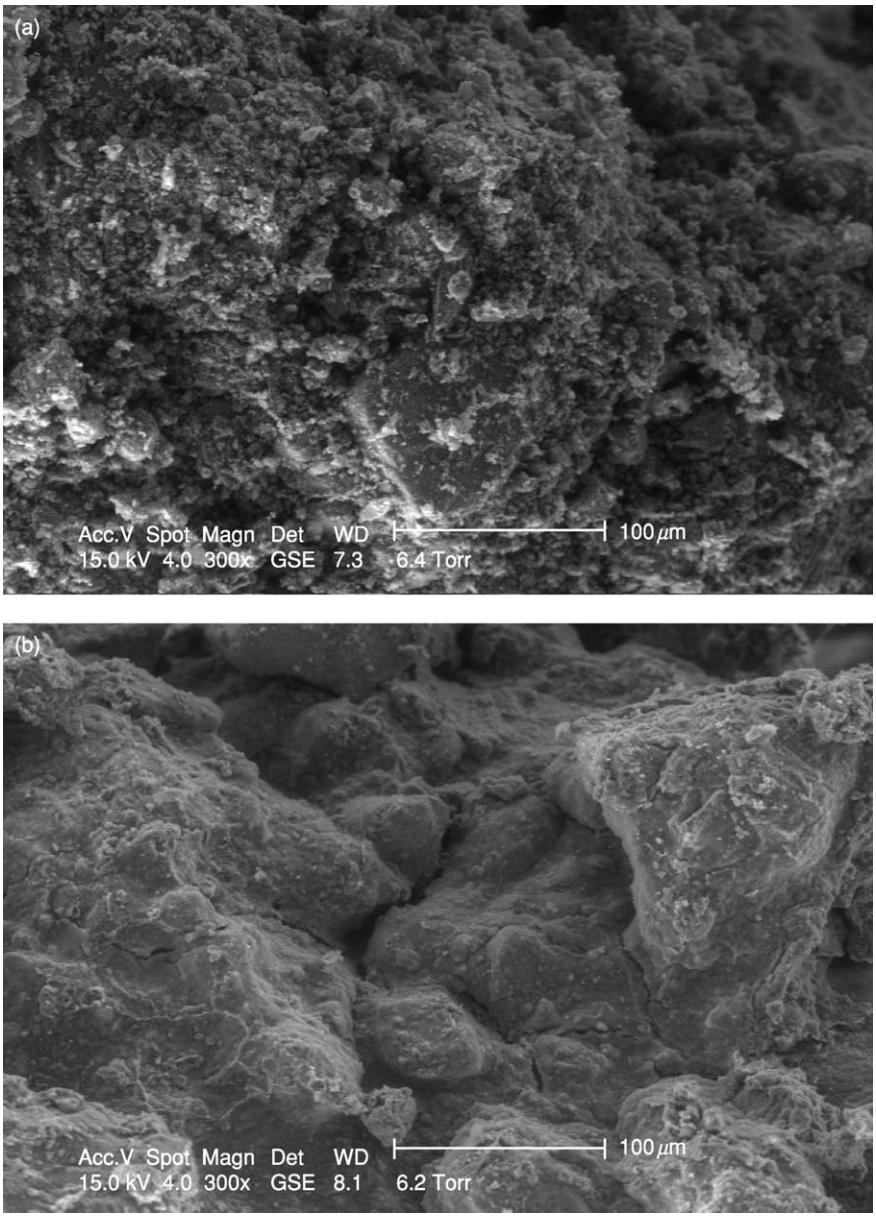


Figure 4 Environmental scanning electron micrographs of (a) surface and (b) subsurface vadose-zone materials at 300× magnification.

microbial growth and activity shape the physical appearance of vadose-zone particles and surfaces, but the relationship between gross surface textural morphology and microbial ecology is not yet quantified. Certainly, the physical composition of the vadose zone can influence the distribution and abundance of microbes: low-porosity rock such as basalt and higher-porosity sands and silts with low carbon content harbor fewer microbes as compared to buried soils (paleosols), which contain more organic carbon. Soil texture also influences the distribution of microbes, since higher clay contents are often associated with greater substrate availability due to the sorption of nutrients on to clay surfaces. However, at a given water content, soils with higher clay contents have lower soil-water potentials, restricting the rates of nutrient diffusion.

Modeling Vadose-Zone Microbial Processes

Mathematical models are useful for predicting the fate of nutrients and pollutants in the vadose zone. Although the vadose zone is not a well-mixed system, because of its heterogeneity it is sometimes modeled as a complete mix reactor where microbes are not

limited by diffusional or advective resupply of nutrients. In such cases, X, the concentration of growing microbes (n per cubic liter) increases according to a simple first-order relationship with respect to time:

$$\frac{dX}{dt} = \mu X \qquad [2]$$

where μ is specific growth rate (t^{-1}), where t is true.

Specific growth rate, for a particular microbe and for a given set of environmental conditions, varies with substrate concentration, as per the Monod equation:

$$\mu = \frac{\mu_{max} S}{K_s + S} \qquad [3]$$

where μ_{max} is maximum specific growth rate (t^{-1}), S is substrate concentration (per cubic milliliter), and K_s is substrate concentration at one-half μ_{max} (per cubic milliliter).

Equation [2] can be related to a change in substrate concentration over time by a yield coefficient, Y, where Y is mass of substrate utilized per mass of microbes. In that case, the rate of substrate depletion is:

$$\frac{dS}{dt} = -\mu X Y \qquad [4]$$

An equation that is analogous to eqn [3] but more applicable to substrate depletion by nongrowing microbes is the Michaelis–Menten equation:

$$v = \frac{v_{max} S}{K_m + S} \qquad [5]$$

where v is substrate depletion rate ($ml^{-3}\,t^{-1}$), v_{max} is maximum substrate depletion rate ($ml^{-3}\,t^{-1}$), S is substrate concentration (per cubic milliliter), and K_m is substrate concentration at one-half v_{max} (per cubic milliliter). Note that eqn [5] describes the variable v, whose units are different from the Monod specific growth rate, μ, because v describes the mass of substrate depleted per unit volume of catalysts (free enzymes or nongrowing microbes). When cells are nongrowing and do not experience mass transfer limitations:

$$\frac{dS}{dt} = -\frac{vX}{\rho_c} \qquad [6]$$

where ρ_c is microbial cellular density (per cubic milliliter).

Transport processes may limit the observed rate of microbial reaction in the vadose zone. At the scale of microbial biofilms (**Figure 5a**), the composite biofilm EPS and cells (**Figure 3a**) can significantly impede the diffusion of nutrients to biofilm bacteria. If we assume a steady state at that spatial scale, the

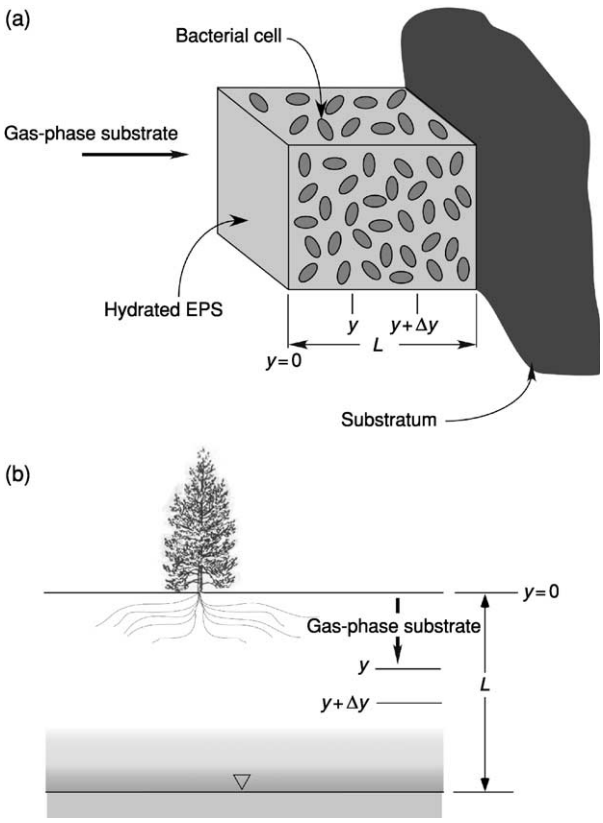


Figure 5 Idealized control volumes depicting (a) biodegrading bacterial biofilm and (b) the vadose-zone profile where, for modeling purposes, L is total depth, y is the distance into the system, and Δy is the change in distance. EPS, extracellular polymeric substances.

applicable differential equation that accounts for both diffusional substrate resupply and microbial reaction is:

$$D\frac{d^2S}{dy^2} = -\mu X Y \eta_i \qquad [7]$$

where D is effective diffusivity of limiting nutrient ($l^2\,t^{-1}$), y is distance along the axis of diffusion (L), and η_i is internal effectiveness factor, a dimensionless parameter that reduces microbial reaction to the rate constrained by diffusional rate limitation.

The boundary conditions for eqn [7] are: for $y = 0$, S is a maximum concentration occurring at the outside edge of the biofilm; and, for $y = L$, $dS/dy = 0$.

Equation [7] can be used to describe gas-phase substrate mass transfer and microbially mediated substrate reactions along a depth profile (**Figure 5b**), as long as the advection of gas through soil pores is relatively rapid and there are no rate-limiting sorption processes. Where mass flow through soil pores and/or sorption processes limit the rate of biogeochemical processes in the vadose zone, the

mathematical model includes explicit mass flow terms for the moving phase. Equations for describing mass flow of solution phase nutrients are typically based on the Darcy equation. Models of the latter kind will also include an applicable reaction term, e.g., eqn [4] or [6], if the purpose of the model is to describe the transport and fate of nutrients or pollutants.

Applications of Vadose-Zone Microbial Ecology

Pollutant Remediation

The vadose zone is frequently subject to pollution from either activities on the ground surface or below-ground leaking storage tanks. Depending on the physicochemical properties of the particular pollutant, the local environmental conditions, and the length of time following release, pollutants can be distributed through various regions of the vadose zone, as depicted in Figure 6. The vadose zone, either passively or actively, is frequently relied upon to attenuate pollutants that would otherwise migrate into groundwater. Active uses include leach fields for disposal of septic tank effluent and bioventing for the remediation of volatile organic pollutants. Passive uses rely upon the natural degradation capabilities of the vadose zone rather than an engineered system. Table 1 summarizes the various technologies that use the vadose zone to attenuate and remediate environmental pollution.

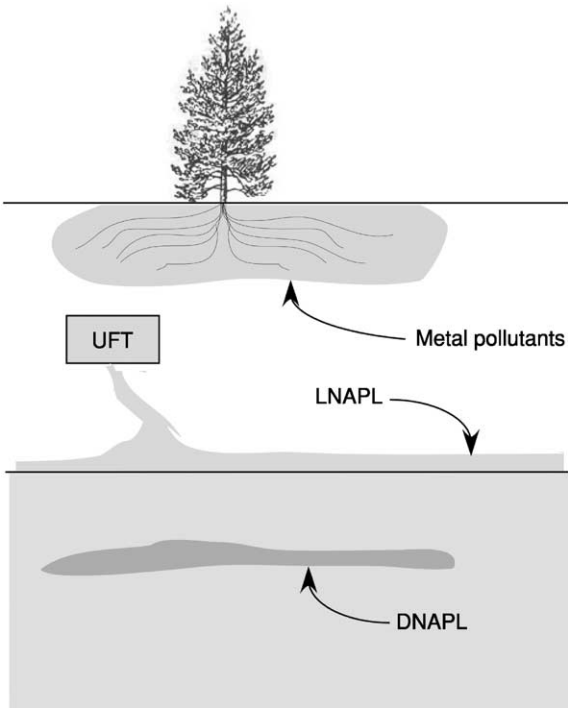


Figure 6 Profile of the polluted vadose zone depicting the orientation of a leaking underground fuel tank (UFT), light non-aqueous-phase liquid (LNAPL) and dense nonaqueous-phase liquid (DNAPL).

Pollution can alter vadose-zone microbial ecology. For example, sites used for waste disposal often have higher populations of culturable microbes and increased microbial activity relative to unimpacted sites. The effects of pollution on vadose-zone microbes and the effectiveness of microbial remediation depend on the characteristics of the specific pollutant and the nature of the microbial communities. Pollutant toxicity as well as bioavailability, i.e., the integrated mass transfer and biodegradation characteristics of the pollutant, are important determinants of the rate of pollutant bioattenuation in the vadose zone. Toxicity can be quantified in the laboratory using bioassays, but pollutants often occur in mixtures which have different toxicity characteristics to individual chemicals. Bioavailability largely depends on the physicochemical characteristics (e.g., solubility, partitioning characteristics, volatility) of the pollutant *in situ*.

The Vadose Zone as an Analog for Other Extreme Habitats

From the near surface to the extremely deep subsurface (several kilometers below the surface), the varying regions of subterranean Earth may physically and chemically mimic the harsh environments that exist on other planets. Although mostly beyond the scope of this chapter, a significant body of new knowledge has been gathered in the last two decades to confirm that microbes of varying taxonomies exist in the deep subsurface, living in high-temperature, high-pressure, and anoxic conditions. Much of the work that has generated the current knowledge of microbial ecology in the very deep subsurface has been funded by the US Department of Energy Subsurface Science Program. While much of that work has been focused on the microbiology of saturated zones, some research has been conducted on vadose-zone samples. Many questions in deep subsurface microbial ecology have been partly answered for specific sites, yet generalizable answers are still needed. How and when did microbes first inhabit deep vadose environments? How do they survive in environments that appear decoupled from aboveground primary productivity? What limits their proliferation: is it the lack of carbon or other nutrients, desiccation, unfavorable pH, and/or lack of terminal electron acceptors? Can subsurface microbiology reveal clues about the origins of life on Earth? Continued interest in these questions still fuels

Table 1 Selected waste treatment technologies and applications requiring vadose-zone microbes

Technology	*Application*	*Description*	*Comments*
Leach field	Septic tank effluent disposal	Organic and nutrient removal; effluent percolates from near-surface distribution pipes through underlying subsoil	Used frequently for rural domestic wastewater treatment; depending on percolation rates, nitrate and microbial pollutants may enter the groundwater
Phytoremediation	*In situ* bioremediation of metal and organic pollutants	Selected plants are cultivated in contaminated soil; plants sequester metals and enhance organic pollutant degradation in the rhizosphere	Low cost and high aesthetic value as compared to mechanical systems. Metal-enriched plant material requires special disposal; efficacy depends on the plant, pollutant, rooting depth, and extent
Bioventing	*In situ* bioremediation of volatile organic carbon pollutants	Air is piped belowground to enhance soil aeration, pollutant mobilization, and aerobic biodegradation	Used frequently for the remediation of refined petroleum that has leaked from underground storage tanks
Natural bioattenuation	Organic (volatile and semivolatile) and inorganic pollutant remediation	Natural physical, chemical, and biological processes distribute and possibly destroy waste; monitoring programs track performance	Used frequently, either intentionally or not; rate of remediation and pollutant fates are uncertain
Bioswale	Storm water	Storm water runoff collects in a constructed depression aboveground and percolates belowground, where dissolved and particulate pollutants are filtered	Used frequently in urban settings; depending on percolation characteristics, pollutants may enter groundwater directly. Sorption and biotransformation capacity for metals is uncertain
Land application	Disposal of treated wastewater	Effluent is applied by irrigation or spraying on to the soil surface and allowed to infiltrate for the removal of organic compounds and nutrients	Used for some domestic and livestock wastewater disposal; less effective in cold or wet climates. Pathogen distribution a potential problem when irrigating food crops

research in subsurface microbial ecology. Given that most of the biomass on Earth resides in the subsurface, it is imperative that we reach a better understanding of microbial life in the subsurface and the role of subsurface microbes in the Earth's biogeochemistry.

See also: **Microbial Processes:** Environmental Factors; Community Analysis; Kinetics; **Mineral–Organic–Microbial Interactions**; **Thermal Properties and Processes**

Further Reading

Fyfe WS (1996) The biosphere is going deep. *Science* 273: 448.

Holden PA (2001) Biofilms in unsaturated environments. In: Doyle R (ed.) *Methods of Enzymology* 337: 125–143

Kieft TL and Brockman FJ (2001) Vadose zone microbiology. In: Fredrickson JK and Fletcher M (eds) *Subsurface Microbiology and Biogeochemistry*, pp. 141–169. New York: Wiley-Liss.

Konopka A and Turco R (1991) Biodegradation of organic compounds in the vadose zone and aquifer sediments. *Applied and Environmental Microbiology* 57: 2260–2268.

Madsen EL (1998) Epistemology of environmental microbiology. *Environmental Science & Technology* 32: 429–439.

Papendick RI and Campbell GS (1981) Theory and measurement of water potential. In: Parr JF, Gardner WR, and Elliot LF (eds) *Water Potential Relations in Soil Microbiology*, pp. 1–22. SSSA Special Publication No. 9. Madison, WI: Soil Science Society of America.

Richter DD and Markewitz D (1995) How deep is soil? *Bioscience* 45: 600–609.

Stephens DB (1996) *Vadose Zone Hydrology*. Boca Raton, FL: CRC Lewis Publishers.

Wan J, Wilson JL, and Kieft TL (1994) Influence of the gas–water interface on transport of microorganisms through unsaturated porous media. *Applied and Environmental Microbiology* 60: 509–516.

Whitman WB, Coleman DC, and Wiebe WJ (1998) Prokaryotes: the unseen majority. *Proceedings of the National Academy of Sciences* 95: 6578–6583.

Viruses *See* Bacteriophage

VOLCANIC SOILS

G Uehara, University of Hawaii, Honolulu, HI, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Molten magma extruded as liquid lava or ejected as cinders or ash from deep within the Earth adds to the variety of parent materials and soils we find on the landscape. Unlike their plutonic counterparts that cool and solidify slowly underground and are subjected to weathering only after uplift and erosion, volcanic rocks cool and solidify immediately upon exposure to the atmosphere. The faster cooling rate results in formation of glassy and finer rock textures in volcanic rocks so that plutonic and volcanic rocks with identical chemical composition can lead to formation of different soils. This is especially true for finely divided pyroclastic ejecta that cool even before they land. The porosity and texture of such cindery or ashy materials differ so greatly from those of other parent materials that glassy volcanic ash develops into soils that fall in the exclusive Andisol soil order.

In addition to rock texture and porosity, volcanic materials also vary in chemical and mineral composition. In fact it is the chemical composition of the magma that determines, to a large extent, the texture and porosity of the extruded material. Highly glassy and porous pyroclastic rocks are more common in high-silica magma that form acidic rocks than the less viscous, basic magma that erupt less explosively and produce basalitc lavas.

The higher viscosity of silicic magma results in the formation of steep-sided, cone-shaped volcanoes exemplified by Mount Fuji in Japan, whereas the basic, low-viscosity magma form broader, shield volcanoes exemplified by Mauna Loa in the island of Hawaii. Both types of volcanoes erupt pyroclastic materials and lava, but the silicic magma have more frequent explosive eruptions and produce more ash than the shield volcanoes. The 1980 eruption of Mount St. Helens in the USA illustrates the explosive nature of intermediate and silicic eruptions. Silicic eruptions often force nearby residents to evacuate the area, whereas sightseers flock to view molten lava pouring out from vents and fissures along rift zones of shield volcanoes.

The range of silica content of volcanic rocks is shown in Table 1. As silica decreases, the content of iron, magnesium, calcium, and sodium increases. These differences in chemical composition influence the rocks' susceptibility to weathering. In general, the basic rocks weather more rapidly owing to the ease with which sodium-, calcium-, magnesium-, and iron-bearing silicate minerals undergo hydrolysis and dissolution. The rate of weathering is even faster if the erupted material is ash.

Table 1 Chemical composition of volcanic rocks

	Alkali rhyolite	*Dacite*	*Andesite*	*Tholeiitic olivine basalt*	*Nephelinite*
SiO_2	74.57	63.58	54.2	47.9	39.07
TiO_2	0.17	0.64	1.31	1.65	3.86
Al_2O_3	12.58	16.67	17.17	11.84	12.82
Fe_2O_3	1.30	2.24	3.48	2.32	8.75
FeO	1.02	3.00	5.49	9.8	6.39
MnO	0.05	0.11	0.15	0.15	0.26
MgO	0.11	2.12	4.36	14.07	6.14
CaO	0.61	5.53	7.92	9.29	14.20
Na_2O	4.13	3.98	3.67	1.60	4.09
K_2O	4.73	1.40	1.11	0.54	2.07

Source: Nockolds SR (1954) Average chemical compositions of some igneous rocks. *Bulletin of the Geological Society of America* 65: 1007–1032.

Rock Weathering

As a rule, the weathering rate of volcanic rocks is influenced by its chemical composition, glass content, and particle-size distribution. Volcanic ash, for example, weathers very rapidly owing to its high glass content and high specific surface associated with smaller particle size. Volcanic glass, however, varies in chemical composition. Felsic glass is high in silicon, potassium, and sodium and is more resistant to weathering than mafic glass high in iron and magnesium.

Artifacts such as arrowheads made of obsidian show little sign of weathering even after centuries of burial, primarily because they are felsic in composition and have low surface-area-to-mass ratios. It is a different matter with ash particles, called shards, which geologists and soil scientists often use as stratigraphic markers and evidence of ancient volcanic eruptions. Shards have high surface-area-to-mass ratios so that volcanic ash particles still identifiable as shards after deposition and buried for thousands of years are more likely to be felsic, as mafic shards would long have been weathered to clay minerals.

Soil Formation

Any student of soil science who has walked over a lava flow or freshly deposited volcanic ash can sense

that soil formation on such materials would be rapid. Owing to the high porosity and large pores of most lavas and ash deposits, very little runoff occurs, and plants are able to establish themselves with relative ease. Nitrogen-fixing lichens are the first to colonize lava flows, and, under warm and humid climates, dense forests can cover lava fields in less than a century. In humid areas, the soils that first form are Folists of the Histosol order. Folists are well-drained Histosols that consist primarily of an organic surface horizon resting on rocks or fragmental materials. If the underlying rock is A'a lava the soil is often a Typic Udifolist, but if the rock is Pahoehoe lava the soil is commonly a Lithic Udifolist. This distinction between a Typic and Lithic Udifolist is crucial from an agricultural standpoint. A Typic Udifolist with its clinkery A'a substrate is ideal for use as orchards, but its Lithic counterpart, with an unbroken and smooth rock substrate, is not. Figure 1 shows papaya production on a Typic Udifolist.

In the case of volcanic ash, plants readily send roots deep on to the loose ash deposit, and rapid buildup of organic matter occurs throughout the rooting depth. The first recognizable soil would most likely be a Vitrand. As the name implies, the soil is an Andisol consisting mainly of unweathered glass fragments.

Folists and Vitrands do not remain unchanged for very long. While soils change little in a human lifespan, one can, given the right circumstances, understand how volcanic soils change with time. Such circumstances exist in the Hawaiian Islands which formed from the movement of the Pacific plate over a stationary hotspot. Islands that form from outpouring of lavas from the hotspot are displaced in a northeasterly direction, providing a chain of islands that is now the Hawaiian islands. The age of the rocks in the island chain ranges from 0.4 million years on the island of Hawaii to 5 million years on the island of Kauai. Ten of the 12 soil orders of Soil Taxonomy occur in the island chain. Unlike Folists and Vitrands that can form in less than a century, and are common on the youngest island of Hawaii, Oxisols and Ulitisols that require prolonged weathering and leaching to develop do not exist there. On Kauai, the oldest island, the opposite is true, wherein Oxisols and Ultisols dominate and Folists and Vitrands are nowhere to be found. The greatest diversity of soils is found on the island of Oahu, midway between Hawaii and Kauai. Here soils found on the old islands dominate, but young, posterosional volcanic eruptions have rejuvenated part of the landscape. A soil order that does not occur on the young islands, but peaks in the intermediate islands and slightly diminishes in abundance on the oldest island, is the Vertisol. The Vertisols form in valley floors on the dry, leeward side of the older islands. On the leeward side, rainfall is lower and the bases that accumulate in the valley floor provide the right conditions for smectite formation. Vertisols do not occur on the young islands because the valleys have not yet formed. In the oldest islands, Vertisols in the wetter zones lose their bases and eventually become kaolinitic.

Figure 1 Papaya production on a young volcanic soil, which was originally a Udifolist.

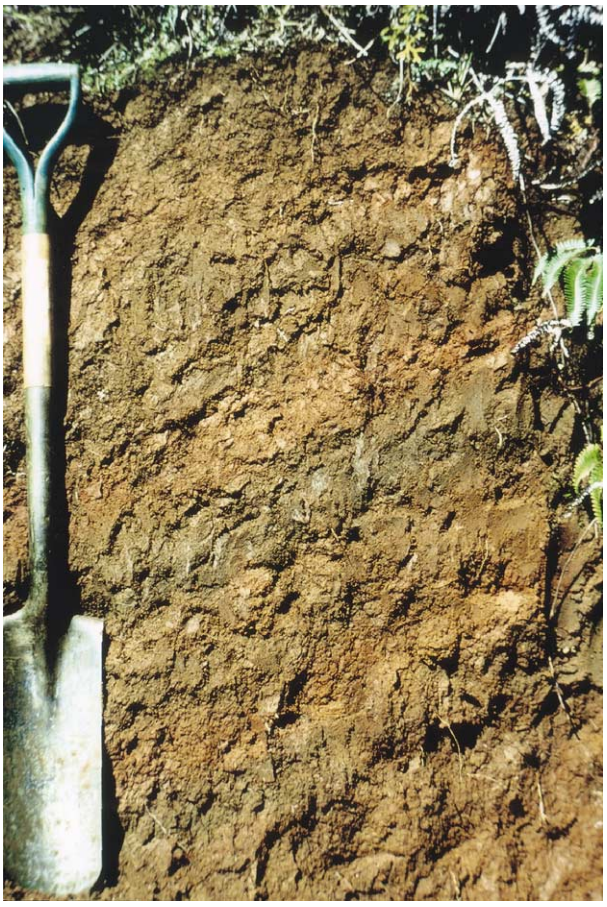


Figure 2 A highly weathered Andisol, an Acrudoxic Hydrudand.

The soils developed from volcanic ash age much faster. Vitrands quickly turn to Haplustands in the dryer zone and to Hydrudands in the wetter areas. The Hydrudands are some of the lightest soils in the world, with bulk densities less than $0.5\,cm^{-3}\,g$ and field water content by weight generally exceeding 100%. Owing to their porous nature, they are readily leached of bases and silica, leaving a residue high in hydrated iron and aluminum oxides. These highly weathered, prematurely old soils occur on the youngest islands (Figure 2).

Vertisols

A combination of basic parent material and low-leaching environment often serves as a necessary condition for Vertisol formation. Sediment deposited by streams and rivers cutting through basaltic rock often provides the first condition, and the semiarid tropics that circle the globe north and south of the equatorial humid tropics provide the second. The Vertisols of Sudan's Gezira region, situated between the White and Blue Nile, is such an example. There are no volcanoes in the Gezira, but the soils there owe their origin to the Blue Nile, which transports basaltic sediments from the highlands of Ethiopia to the Gezira. A similar situation exists in the experimental fields of the International Crop Research Institute for the Semi-Arid Tropics located near Hyderabad, India. The black Vertisols in the Hyderabad area sit side by side with red Alfisols to form a red-and-black soil complex. Here again, there are no signs of volcanoes nearby, for the Vertisols are products of sediment transported by water from basalts of the Deccan Flats.

In Table 2, the chemical composition and chemical properties of a Vertisol from Hawaii are shown. The Vertisols of the semiarid tropics are remarkably alike. Often referred to as tropical black earths, they have low organic carbon content and their dark color has been attributed to dark iron minerals and manganese coatings. Their chemical and physical properties are mainly determined by the high clay and smectite content.

Oxisols

A key feature of Oxisols is the low cation exchange capacity (CEC) of their clay. This condition is most likely to occur when the clay fraction is low in layered silicate clay and rich in iron oxides. Volcanic rocks low in silica and rich in ferromagnesium minerals weather rapidly to Oxisols under warm and humid conditions. While time and intense leaching can produce Oxisols from a variety of materials, volcanic materials, due to their porous nature, produce Oxisols with greater ease. The chemical composition and selected properties of an Oxisol formed from basic volcanic rock are shown in Table 3. The extreme weathering of the soil is indicated by its placement in the Acrudoxic subgroup and also by the rise in the $1\,mol\,l^{-1}$ KCl pH in the oxic horizons midway in the profile. In the oxic horizons the pH measured in $1\,mol\,l^{-1}$ KCl is higher than the pH measured in water, indicating that the net charge of the soil material is positive. Soils with these characteristics are infertile, but can be made productive with phosphorus fertilization, as well as liming with calcium silicate rather than calcium carbonate.

Ultisols

From a taxonomic standpoint, 'volcanic soil' is not a very useful term as a great variety of soils can have volcanic origins. Ultisols can form from volcanic rocks where rainfall is sufficiently high to leach bases, and conditions exist that favor activation and

Table 2 Chemical composition and properties of a Vertisol developed from basaltic alluvium (Lualualei series: fine, smectitic, isohyperthermic, Typic Gypsitorrerts)

												Chemical properties								
	Chemical composition (% of whole soil)													*Exchangeable bases (cmol kg^{-1})*				*CEC*	*pH (1:5)*	
Depth (cm)	*SiO_2*	*TiO_2*	*Al_2O_3*	*Fe_2O_3*	*MnO_2*	*MgO*	*CaO*	*Na_2O*	*K_2O*	*P_2O_5*	*LOI*	*C (%)*	*N (%)*	*Ca*	*Mg*	*Na*	*K*	*(cmol kg^{-1})*	*H_2O*	*1 mol l^{-1} KCl*
0–3	30.54	7.71	19.29	26.26	0.30	1.50	1.57	0.26	0.07	0.50	11.63	0.66	0.08	17.1	15.2	0.8	1.4	34.1	7.1	5.9
3–25	31.74	7.59	20.34	25.81	0.33	1.07	1.72	0.28	0.08	0.48	10.81	0.42	0.06	15.1	15.1	1.3	0.4	32.9	7.2	5.4
25–35	33.91	7.52	19.04	26.28	0.45	1.57	0.62	0.29	0.05	0.51	10.87	0.21	0.05	14.6	13.4	2.5	0.2	32.9	7.2	5.4
35–75	33.54	7.59	19.57	25.82	0.43	1.72	0.34	0.37	0.06	0.36	10.54	0.17	0.04	16.0	10.5	3.9	0.2	32.1	6.8	5.2
75–123	32.25	6.18	20.02	24.24	0.33	1.58	2.56	0.44	0.04	0.38	11.39	0.17	–	73.1	9.3	7.2	0.2	29.8	5.6	4.9
>123	31.52	6.57	20.39	22.37	0.38	1.50	4.39	0.49	0.07	0.29	11.56	0.16	–	73.8	10.7	8.4	0.3	30.6	5.8	5.0

LOI, loss on ignition; CEC, cation exchange capacity.

Source: USDA–SCS (1978) *Soil Survey Laboratory Data and Description for Some Soils of Hawaii*. Soil Survey Investigations Report No. 29. USDA Soil Conservation Service/Hawaii Agricultural Experimental Station/ Hawaiian Sugar Planters' Association. Washington, DC: US Government Printing Office.

Table 3 Chemical composition and properties of an Oxisol developed from nepheline basalt (Kapaa series: very fine, sesquic, isohyperthrmic, Anionic Acrudox)

												Chemical properties								
	Chemical composition (% of whole soil)													*Exchangeable bases (cmol kg^{-1})*				*CEC*	*pH (1:5)*	
Depth (cm)	*SiO_2*	*TiO_2*	*Al_2O_3*	*Fe_2O_3*	*MnO_2*	*MgO*	*CaO*	*Na_2O*	*K_2O*	*P_2O_5*	*LOI*	*C (%)*	*N (%)*	*Ca*	*Mg*	*Na*	*K*	*(cmol kg^{-1})*	*H_2O*	*1 mol l^{-1} KCl*
0–30	6.9	7.6	24.6	38.5	0.06	0.32	–	0.03	0.28	0.63	21.7	3.92	0.22	0.7	0.4	0.20	0.40	15.9	5.0	4.5
30–40	4.4	7.2	28.1	39.3	0.09	0.25	–	0.03	0.11	0.55	19.9	1.46	0.07	0.2	0.0	0.10	0.20	3.7	5.5	5.7
40–63	4.1	8.1	28.8	38.5	0.08	0.08	–	0.03	0.07	0.58	19.7	1.09	0.04	0.4	0.1	0.20	0.10	2.6	5.5	5.7
63–90	4.3	7.5	32.9	34.3	0.17	0.13	–	0.03	0.07	0.63	20.0	0.64	0.02	0.1	0.1	0.20	0.10	1.9	5.8	5.7
90–123	7.1	7.5	30.9	35.2	0.15	0.11	–	0.03	0.07	0.57	18.9	0.46	0.02	0.8	0.1	0.20	0.10	3.7	5.3	5.4
123–150	13.2	8.4	24.8	36.8	0.15	0.38	–	0.03	0.07	0.61	15.5	0.48	0.02	1.0	0.1	0.30	0.10	6.1	5.5	4.9

LOI, loss on ignition; CEC, cation exchange capacity.

Source: USDA–SCS (1978) *Soil Service Laboratory Data and Description for Some Soils of Hawaii*. Soil Survey Investigations Report No. 29. USDA Soil Conservation Service/Hawaii Agricultural Experimental Station/ Hawaiian Sugar Planters' Association. Washington, DC: US Government Printing Office.

illuviation of clay. It is not unusual to find two soil orders, in close proximity, forming from the same parent rock. Ultisols and Oxisols, for example, have developed synchronously on the same geomorphic surface on the volcanic island of Kauai in the Hawaiian Islands. The Ultisols, however, always occur on the knick points of the landscape, where soil creep activates and mobilizes clay along sheer planes in the subsoil. The chemical composition and properties of such an Ultisol are given in Table 4. The same shearing action leads to formation of Alfisols in drier locations, where leaching is less intense and basic cations are retained in the profile.

In climates where Ultisols dominate, water is plentiful but nutrients are in short supply, whereas, in semiarid regions where Alfisols predominate, the land is fertile but short of water.

Volcanic soils may have special meaning to people living in humid regions where soils that under normal circumstances would be Ultisols are maintained in an Alfisol-like condition by periodic additions of nutrient-rich ash from a neighboring volcano. The advantage of having an adequate supply of water and nutrients in contrast to having one or the other was, for our ancestors, the difference between being able to form creative, sedentary communities in contrast to simply surviving by hunting and gathering. Thus it may not be the soil itself but the presence of an active volcano nearby that makes volcanic soils what they are.

Andisols

If one had to choose a soil to represent the class of soils we call volcanic, it would be an Andisol. This is because all Andisols are volcanic soils, but the reverse is not true. Andisols have two characteristics which make them unique. They contain high amounts of short-range-order and noncrystalline materials, and have low bulk densities. Two short-range-order minerals commonly found in highly leached and weathered Andisols are allophane and imogolite. These minerals have high specific surfaces (approximately equal to $1000\,m^2g^{-1}$) and pH-dependent charge. The high specific surface of the inorganic colloids enables the soil to sequester large quantities of organic matter. The prefix 'Ando' on Andisol is the Japanese word for 'dark,' used to describe the low chroma of organic-rich volcanic ash soils of Japan. In Japan, where the ash is generally high in silica and low in iron, organic matter imparts a nearly black color to the soil. In Andisols developed from low-silica, high-iron and -magnesium ash, the iron oxide masks the black color of organic matter to produce a dark-brown soil (Figure 2).

The chemical composition and selected chemical properties of two Andisols, one developed in a perudic, and the other in an ustic moisture regime, are shown in Tables 5 and 6. Both soils are excellent agricultural soils, with the wetter soil requiring careful nutrient management and the drier soil requiring irrigation for profitable farming.

Human Interactions with Volcanic Soils

Our human ancestors may have avoided extinction and evolved by adapting to the diverse and ever-changing volcanic environment of the African Rift. The rich soil, constantly rejuvenated by sprinkling of volcanic ash, even today supports large herds of herbivores and their predators, and some of the highest nonurban human population densities in Africa.

When modern humans exited Africa through the Rift zone, then crossed Asia and the Bering Straits, they encountered the North, Central, and South American segment of the 'Ring of Fire' that circles the Pacific. Like the rich alluvial soils of the Nile, Tigris, Euphrates, and Indus rivers that enabled humans to shift their energies from hunting and gathering to creative endeavors, the fertile and well-watered volcanic soils of Central and South America provided the new settlers with the means to do the same.

Even today, the influence of volcanic soils on human society is evident. In Indonesia, for example, nearly half of its 200 million people live on the tiny volcanic island of Java. Efforts to move whole communities from Java to the larger and less densely populated neighboring islands have not been easy. Villagers accustomed to farming Andisols with characteristics similar to the one shown in Table 5 found it difficult to survive on Ultisols with features similar to those shown in Table 4.

Today science and technology enable farmers to understand and overcome soil constraints, but only a century ago soil fertility determined to a large degree outcomes of human interactions with the environment. Unfortunately this is still true in many developing countries, where science-based soil management practices have yet to be adopted. The volcanic soils of Africa, Asia, and Latin America that helped people in the past now need our help to perform and contribute to human well-being as they have in the past.

Table 4 Chemical composition and properties of an Ultisol developed from basalt (Haiku Series: very fine, isohyperthermic, Ustic Palehumult)

	Chemical composition (% of whole soil)											Chemical properties								
														Exchangeable bases ($cmol\,kg^{-1}$)				CEC	pH (1:5)	
Depth (cm)	SiO_2	TiO_2	Al_2O_3	Fe_2O_3	MnO_2	MgO	CaO	Na_2O	K_2O	P_2O_5	LOI	C (%)	N (%)	Ca	Mg	Na	K	($cmol\,kg^{-1}$)	H_2O	1 $mol\,l^{-1}$ KCl
0–18	16.8	11.6	7.6	46.3	0.15	0.97	–	0.06	1.15	0.55	15.9	3.08	0.26	–	0.6	0.1	0.3	15.9	5.1	4.1
18–33	16.8	11.2	11.1	43.0	0.12	0.77	–	0.07	1.23	0.51	15.5	2.79	0.23	–	0.6	0.1	0.2	14.3	5.0	4.0
33–45	14.1	7.9	15.2	42.7	0.07	0.66	–	–	0.99	0.57	17.6	1.97	0.16	0.2	0.8	0.2	0.1	15.5	4.9	4.1
45–70	11.9	6.5	23.2	36.9	0.09	0.78	–	0.02	0.65	0.52	19.1	1.78	0.13	0.6	0.7	0.4	0.1	12.2	5.2	4.4
70–98	18.9	5.2	27.4	28.4	0.12	0.98	–	–	0.31	0.39	18.1	1.08	0.08	1.0	0.8	1.0	0.1	12.7	5.1	4.0
98–155	17.3	5.7	30.1	26.4	0.12	1.62	–	–	0.11	0.42	17.6	0.91	0.06	0.6	0.2	0.8	0.1	12.0	5.0	4.0
155–175	17.0	5.2	32.9	24.7	0.11	1.54	–	–	–	0.32	18.0	0.74	0.04	0.4	0.3	>0.8	0.1	12.4	4.9	4.0

LOI, loss on ignition; CEC, cation exchange capacity.

Source: USDA–SCS (1978) *Soil Survey Laboratory Data and Description for some Soils of Hawaii.* Soil Survey Investigations Report No. 29. USDA Soil Conservation Services/Hawaii Agricultural Experimental Station/ Hawaiian Sugar Planters' Association. Washington, DC: US Government Printing Office.

Table 5 Chemical composition and properties of a Haplotorrand developed from basaltic ash (Waikaloa series: medial, amorphic, isothermic, Typic Haplotorrands)

	Chemical composition (% of whole soil)											Chemical properties								
														Exchangeable bases ($cmol\,kg^{-1}$)				CEC	pH (1:5)	
Depth (cm)	SiO_2	TiO_2	Al_2O_3	Fe_2O_3	MnO_2	MgO	CaO	Na_2O	K_2O	P_2O_5	LOI	C (%)	N (%)	Ca	Mg	Na	K	($cmol\,kg^{-1}$)	H_2O	1 $mol\,l^{-1}$ KCl
0–13	32.3	3.7	21.1	14.5	0.44	3.30	3.44	1.76	0.80	0.63	18.3	7.53	0.32	27.9	7.8	0.40	4.80	81.0	6.6	5.7
13–25	32.9	3.5	23.9	15.0	0.42	2.86	3.55	2.24	0.87	0.61	14.7	3.69	0.16	28.0	6.4	0.40	5.60	78.0	7.1	6.2
25–50	36.1	3.2	24.6	15.4	0.43	2.03	2.45	2.26	0.88	0.44	11.7	1.88	0.11	32.6	7.1	0.80	4.80	88.0	7.3	6.2
50–63	38.6	3.4	24.0	15.1	0.47	2.07	3.44	2.25	0.82	0.38	10.0	1.02	0.09	34.6	9.8	1.10	3.00	97.0	7.5	6.3
63–78	40.4	4.3	22.5	14.8	0.43	2.04	3.17	2.41	0.92	0.36	9.2	0.77	–	33.3	12.8	1.60	0.60	93.0	7.6	6.4
78–98	44.8	3.2	20.9	13.9	0.40	2.08	2.11	2.42	1.42	0.26	8.6	0.31	–	33.0	15.9	2.20	0.30	>100	7.8	6.5
98–125	51.9	2.5	20.4	6.7	0.29	1.00	2.61	4.20	2.70	0.09	8.1	0.15	–	34.0	12.4	5.40	0.50	>100	8.1	6.8
125–163	55.7	1.6	21.1	4.8	0.31	0.77	1.68	4.19	2.66	0.09	7.6	0.11	–	33.3	14.4	6.10	1.30	>100	8.2	6.9

LOI, loss on ignition; CEC, cation exchange capacity.

Source: USDA–SCS (1978) *Soil Survey Laboratory Data and Description for some Soils of Hawaii.* Soil Survey Investigations Report No. 29. USDA Soil Conservation Service/Hawaii Agricultural Experimental Station/Hawaiian Sugar Planters' Association. Washington, DC: US Government Printing Office.

Table 6 Chemical composition and properties of a Hydrudand developed from basaltic ash (Hilo series: medial over hydrous, ferrihydritic, isohyperthermic Acrudoxic Hydrudands)

	Chemical composition (% of whole soil)											*Chemical properties*								
														Exchangeable bases ($cmol\,kg^{-1}$)				*CEC*	*pH (1:5)*	
Depth (cm)	*SiO_2*	*TiO_2*	*Al_2O_3*	*Fe_2O_3*	*MnO_2*	*MgO*	*CaO*	*Na_2O*	*K_2O*	*P_2O_5*	*LOI*	*C (%)*	*N (%)*	*Ca*	*Mg*	*Na*	*K*	*($cmol\,kg^{-1}$)*	*H_2O*	*$1\,mol\,l^{-1}$ KCl*
0–40	12.90	5.24	24.32	27.56	0.15	–	–	–	–	–	27.30	5.30	0.41	2.0	1.8	0.1	0.1	67.6	5.8	5.6
40–53	8.78	5.58	34.40	26.26	0.25	–	–	–	–	–	25.82	3.06	0.22	2.2	0.6	–	0.1	68.4	6.1	6.2
53–65	8.58	5.13	35.92	26.10	0.34	–	–	–	–	–	24.90	2.61	0.17	2.1	0.3	–	–	60.4	6.3	6.5
65–80	8.54	5.30	35.04	26.62	0.37	–	–	–	–	–	25.50	2.67	0.19	1.2	0.6	–	–	62.9	6.3	6.5
80–123	10.82	5.38	33.78	26.38	0.36	–	–	–	–	–	24.46	2.80	0.20	1.8	0.4	–	–	60.5	6.3	6.4
123–140	10.73	5.60	32.56	25.97	0.25	–	–	–	–	–	24.13	2.57	0.16	1.8	0.5	–	–	65.5	6.4	6.4
140–168	10.68	5.68	30.04	28.95	0.25	–	–	–	–	–	24.43	2.24	0.15	2.5	0.5	–	–	71.0	6.3	6.4

LOI, loss on ignition; CEC, cation exchange capacity.

Source: USDA–SCS (1978) *Soil Survey Laboratory Data and Description for some Soils of Hawaii*. Soil Survey Investigations Report No. 29. USDA Soil Conservation Service/Hawaii Agricultural Experimental Station/ Hawaiian Sugar Planters' Association. Washington, DC: US Government Printing Office.

List of Technical Nomenclature

Acrodoxic Hydrudands — These are highly weathered soils derived from volcanic ash that occur in high-rainfall areas. The subsoil material contains more that 100% water on a weight basis and dries irreversibly when exposed to the atmosphere

Udifolists — These are well-drained organic soils that occur mainly in the high-rainfall areas of the Hawaiian Islands on forested lava flows

Further Reading

Beinroth FH, Uehara G, and Ikawa H (1974) Geomorphic relationships of Oxisols and Ultisols on Kauai, Hawaii. *Soil Science Society of America Proceedings* 38: 128–131.

Bleeker P and Parfitt RL (1974) Volcanic ash and its clay mineralogy at Cape Hoskins, New Britian, Papua New Guinea. *Geoderma* 11: 123–135.

Christiansen RL and Peterson DW (1981) Chronology of the 1980 eruption activity. In: Lipman PW and Mullineaux DR (eds) *The 1980 Eruptions of Mount St. Helens*, pp. 17–30. Geological Survey Professional Paper 1250. Washington, DC: US Government Printing Office.

Clague DA, Dalrymple GB, Wright TL *et al.* (1989) The Hawaiian/Emperor chain. In: Winterer EL, Hussong DM, and Decker RW (eds) *The Geology of North America*, vol. N. *The Eastern Pacific Ocean and Hawaii*, pp. 187–287. Geological Society of America.

Dethier DP, Pevear DR, and Frank D (1981) Alteration of new volcanic deposits. In: Lipman PW and Mullineaux DR (eds) *The 1980 Eruptions of Mount St. Helens*, pp. 649–665. Geological Survey Professional Paper 1250. Washington, DC: US Government Printing Office.

Jones RC, Babcock CJ, and Knowlton WB (2000) Estimation of the total amorphous contact of Hawaii soils by the Rietveld method. *Soil Science Society of America Journal* 64: 1100–1108.

Lowe DJ (1986) Controls on the rate of weathering and clay mineral genesis in airfall tephras: a review and New Zealand case study. In: Coleman SM and Dethier DP (eds) *Rates of Chemical Weathering of Rocks and Minerals*, pp. 265–330. New York: Academic Press.

Nockolds SR (1954) Average chemical compositions of some igneous rocks. *Bulletin of the Geological Society of America* 65: 1007–1032.

Parfitt RL (1975) Clay minerals in recent volcanic ash soils from Papua New Guinea. *Bulletin of the Royal Society of New Zealand* 13: 241–245.

Plucknett DL (1971) The use of soluble silicates in Hawaiian agriculture. *University of Queensland Papers* 1: 203–223.

Raymundo ME (1965) *The Properties of the Black Earths of Hawaii*. PhD Dissertation. Honolulu, HI: Department of Agronomy and Soil Science, University of Hawaii.

Saigusa M, Shoji S, and Kato T (1978) Origin and nature of halloysite in Ando soils from Towada tephra, Japan. *Geoderma* 20: 115–129.

Shoji S, Fujiwara Y, Yamada I, and Saigusa M (1982) Chemistry and clay mineralogy of Ando soils, Brown forest soils, and Podzolic soils formed from recent Towada ashes, northeastern Japan. *Soil Science* 133: 69–86.

Stevens KF and Vucetich CG (1984) Pedogenic weathering of Upper Quaternary tephras in New Zealand. Isovolumetric geochemical evidence of cation movement. *Chemical Geology* 47: 285–302.

Uehara G and Gillman GD (1980) Change characteristics of soils with variable and permanent change minerals. I. Theory. *Soil Science Society of America Journal* 44: 250–252.

USDA–SCS (1978) *Soil Survey Laboratory Data and Description for Some Soils of Hawaii*. Soil Survey Investigations Report No. 29. USDA Soil Conservation Service/Hawaii Agricultural Experimental Station/ Hawaiian Sugar Planters' Association. Washington, DC: US Government Printing Office.

USDA–SCS (1999) *Soil Taxonomy: A Basic System of Soil Classification for Making and Interpreting Soil Surveys*, 2nd edn. Washington, DC: Natural Resource Conservation Service.

Wada K (1985) The distinctive properties of Andosols. In: Stewart BA (ed.) *Advances in Soil Science*, vol. 2, pp. 173–229. New York: Springer-Verlag.

W

WAKSMAN, SELMAN A.

H B Woodruff, Soil Microbiology Associates, Watchung, NJ, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Selman A. Waksman, professor and research specialist at Rutgers University, the State University of New Jersey, USA, was a highly creative soil scientist. From the time of preparation of his senior college thesis in 1915 until his death in 1973 he specialized in the study of the microbes of the soil. During that 58-year period, he published 447 scientific reports. They covered the full range of the early development of soil microbiology, extending from the descriptive phases of the microorganisms of the soil to details of their physiology, from investigations of their influence on soil fertility to studies of their interactions together within soils. His research on the microbial associations led to a demonstration that soil microorganisms produce antagonistic agents. He proposed the name 'antibiotic' for such products. Certain of the antibiotics derived from soil microbes, for example streptomycin, discovered in his laboratory, have proven useful in the treatment of infections of humans, animals, and plants.

Selman Waksman was a prolific writer, publishing papers in a wide range of scientific journals, in several languages. He was author or co-author of 28 books. His *Principles of Soil Microbiology*, an 897-page volume, the first edition published in 1927, for years became the standard textbook of his field. He guided 78 students to graduate degrees. He was an inspiring lecturer. His presentations were filled with stories of his relationships with leaders of his field and were avidly followed by his students and by general audiences. He was beloved by his students, who spoke of him familiarly as 'The Old Maestro.' Waksman's name is generally included with Winogradsky, Omeliansky, Beijerinck, and the Americans Lipman and Thom in lists of the pioneers of soil microbiology.

Illustrative of the breadth of Selman Waksman's interests and the scope of his accomplishments is the quantitative record of his publications, written in association with his students. Fifty-eight were descriptive reports of the microbes of the soil, 53 of the enzymology and biochemistry of soil microbes, 11 on sulfur oxidation, 35 on peat, composts, and humus, 18 on microbes of the sea, 20 on taxonomy and methodology, eight on organic acids and fermentation products of molds, 18 on general physiology, and 31 were biographies and discussions on the philosophy of science. Finally, 195 were on antibiotics, including the announcement of 17 new antibiotics discovered in his laboratory.

Waksman was greatly honored, based on his productive career. By the time of his 80th birthday, he had received 22 honorary degrees, 12 from universities of Europe, Asia, and South America. He had been granted 63 prizes, awards, and medals. He was elected to the prestigious National Academy of Sciences, served as President of the American Society for Microbiology and was awarded the 1952 Nobel Prize in Physiology or Medicine.

What were the background, family, experiences, and education that led to such a creative career?

Selman Abraham Waksman was born 8 July, 1888 according to the old Russian calendar (now 22 July) in Novaia-Priluki, a small village in the Russian Ukraine. The villages and small towns of western Russia and Ukraine were populated by recently liberated serfs, who subsisted by small-scale farming and livestock husbandry, and by Jewish artisans and tradesmen who engaged in crafts, shop-keeping, and the processing and marketing of farm and forest products. These activities influenced the young child, and provided a motivation for his later decision to specialize in unraveling the chemistry of living things.

The inhabitants of Novaia-Priluki were poor. There was a rigid separation in culture and approaches to life between two groups, the Russian and the ethnic Jewish populations. The Jews stressed education, but a specialized religious education, directed by the synagogue Rabbinate, devoted almost exclusively to the Bible and the Talmud. The adult men spent much of their free time in study, discussion, and

arguments in their religious halls. Yiddish was their universal language, only a minority learning the intricacies of Hebrew and developing the ability to write in Russian. Waksman's mother was more accomplished than most. She vowed that her son Selman would become even more so, therefore she strove constantly to see that objective fulfilled.

No formal state schools existed in Novaia-Priluki. The sole opportunity for a broad education was private tutors, who either taught small groups or, if financial circumstances permitted, individual students. Waksman was taught privately. He learned the rudiments of language and mathematics and limited geography. He had an exceptional ability to memorize almost instantly anything read, thus progressed more rapidly than many of his student associates. Somewhat chagrined by the advantages offered him, compared to his friends, he, with some associates, organized a volunteer school, in which they passed on to economically deprived young folks the results of their studies. It was a most fortunate experience. It not only honed the lecturing skills Waksman exemplified later as a college professor, but provided the means whereby he could remove himself from the drudgery of mere subsistence.

Once he had sampled the full extent of educational opportunities available in his home town, Waksman moved to a larger town, Zhitomir, capital of a nearby prefecture. It had a Gymnasium for advanced studies, so accomplished teachers were available who would accept students for private study. From Zhitomir he progressed to an even larger city, Odessa, where highly qualified professors were located. In these locations at a distance from home, his prior experience as a volunteer instructor became significant. He offered himself as an instructor in the daytime, thus earning funds to cover living costs and tutors' fees, taking his own lessons during evening hours.

After years of study, at 22 years of age, Waksman felt qualified to take the 2-week extensive written and oral exams whose passage was required to obtain permission to enter a university. He did so, and passed.

Now that the essential qualification had been earned, entry into a university became possible. For Jews in Russia, however, that was very difficult to accomplish because of a quota system. Waksman realized that, graduating in Russia as a specialist in his chosen area, he would face few opportunities for employment. Therefore, he gave serious consideration to emigrating to the USA. Close relatives had done so and their letters contained descriptions of the opportunities for advancement that existed in America. The turning point came with the death of his mother. His father quickly remarried. Feeling that he no longer had a home to return to and with little opportunity available to advance in his objective for life work, he emigrated.

His port of entry to the USA was Philadelphia. Relatives, the Kornblatts, had advanced from tenant farmers to ownership of a farm specializing in poultry. They lived nearby in Metuchen, New Jersey. He joined them, working on the farm to earn his board. For the first time he became directly responsible for agricultural practices, obtaining best yields of vegetables and enhancing egg production.

The experience aiding the Kornblatts on their farm strengthened Waksman's intent to study the application of chemistry to biology. He applied to Columbia University in New York City to enter a premedical program, was accepted, but had no means to cover the costs involved. He visited nearby Rutgers College, at that time a private school that had associated with it federally supported Schools of Engineering and of Agriculture, supported financially through the Federal Morrill Land Grant Act, and the site of an agricultural experiment station supported by the state of New Jersey. There he met Jacob G. Lipman, professor of soil chemistry and microbiology, an excellent experimentalist and skilled leader, who was slated to become Director of the Agricultural Experiment Station. In Lipman, Waksman found his mentor. In soil microbiology he discovered the topic he had been striving toward all his life, providing him the opportunity to study the relationship of chemistry to living systems. He applied for a scholarship at Rutgers College, received one and became an undergraduate student, the sole person in his class majoring in soil microbiology.

Waksman was highly successful in his studies. Despite his limited knowledge of the English language, he received high grades. As he entered his senior year, he had already fulfilled most of the requirements for graduation. Therefore, a senior thesis became his prime objective for the year. His life as an experimentalist in the field of soil microbiology had begun.

The prior investigations in soil microbiology had been directed primarily toward individual organisms, the nitrogen-fixing rhizobia and azotobacter, or soil organisms that caused plant diseases. Waksman chose to apply a holistic approach to the topic. He enumerated the types of organisms existing in soils, defined the changes in numbers and types that occurred at different soil depths, among differing soil types, at different temperatures and moisture levels. From trenches he had dug on the college farm, using good aseptic techniques, he isolated and classified hundreds of organisms. At the end of his senior year, he tabulated his results. They were presented orally by his professor, Jacob Lipman, at a Society of American Bacteriologists meeting. An abstract of the talk was

published in the first issue of the newly established *Journal of Bacteriology*.

Highly appreciative of Waksman's ability to perform science independently, Lipman encouraged him to remain at Rutgers as a graduate student, working towards an MS degree in soil microbiology. He provided the financial wherewithal to make it possible. Waksman had become fascinated with the small, hard, slowly growing microbes that appeared almost universally in his agar platings of soils at high dilutions, often by count a dominant proportion of the soil population. They were known to others, but were very little studied. They had been generally ignored because of the smallness of their colonies and slowness of their growth. These microorganisms were the actinomycetes, the cells of which are typical of bacteria, approximately the same size in diameter, but in morphology filamentous, resembling the molds. Waksman decided to concentrate on them as an MS thesis objective and with an associate he isolated many, studying each in detail. When the isolates were typical of past descriptions, he applied the established species names, but often he found cultures that differed distinctly from previously known actinomycetes. He created names for these, often related to his personal experiences: *lipmanii* and *halstedii*, his favorite teachers; *rutgersensis* and *californicus*, his university and choice for his PhD education; *bobili*, the nickname for his wife.

Upon completion of his MS program, Waksman realized the need to enhance his knowledge of biochemistry, a special branch of chemistry, and to do so he accepted a student appointment to the Department of Biochemistry at the University of California in Berkeley. As subjects for his investigations there, he took with him a selection of the cultures isolated at Rutgers and with them concentrated on defining the enzymes they produced.

Upon receiving his PhD degree at Berkeley, Waksman was offered a position at Rutgers as a college instructor in soil microbiology, with simultaneous appointment as researcher in the Agricultural Experiment Station, at a small salary. He returned to New Brunswick and designed a graduate-level lecture program for soil microbiology. He continued his research studies on the actinomycetes and extended them to include the fungi and protozoa of the soil. For a few years he also worked part-time at a commercial concern to supplement his salary.

Assertive in his approaches, he came into conflict with recognized specialists in the field of soil microbiology. H.J. Conn of Cornell University was convinced that the fungi observed in soil platings were of little significance, that they were derived from wind-blown fungal spores. E.J. Russell in England, a member of the prestigious Rothamsted Experiment Station, had proposed that soil protozoa are controlling members of soil populations because they consume soil microorganisms as a source of nutrition. Waksman objected to both claims, did research to prove the correctness of his ideas, and, through his publications on soil microorganisms he became well-known in the field. He, with the aid of his students, greatly expanded the understanding of microbes of the soil, especially the actinomycetes. New students were attracted to his laboratory, as well as many non-degree-seeking visitors who desired to work several weeks or months conducting studies in association with him.

By 1924 Waksman had completed studies defining the nature of the common members of the soil population and he faced the question of an objective for his future program. To search for the answer, he chose to tour the well-known laboratories of Europe, to spend hours in conversation with specialists in his field and allied subjects. For 5 months he did so, visiting a variety of persons. Mostly, he was received with enthusiasm by those who appreciated his scientific record. Sometimes he had to negotiate an entry where unknown, but always he entered into detailed discussions, sometimes somewhat heated if he disagreed with accepted dogma. He had firmly fixed ideas about many chemical and biochemical processes, a few not in tune with current thinking, and he avidly promoted them.

Waksman came home thoroughly refreshed, convinced that his holistic approach to the study of soil microbes was correct. He had chosen the new objective of defining what the grand mixture of organisms present in soils was accomplishing. What products did they produce in the soil? How did they affect soil fertility? He became convinced that a significant factor in defining fertility was humus, the resistant organic matter resulting from the microbial attack on plant residues, composts, other organic substances that come into contact with soils. Using a system of analysis, dubbed proximate analysis, he followed the changes in organic matter in the soil, then followed the slower attack on humus itself by other microbes, causing its gradual breakdown. He became convinced that the resistant organic complexes in soils have fertility-defining significance.

He extended his studies to follow details of the microbial processes that occur during the composting of waste agricultural products. He also investigated the special geographic situations in which humus is broken down less rapidly than it is formed, resulting in the formation of peat bogs. Here, likewise, he stated his opinions assertively, which eventually brought him into conflict with members of the

chemistry section of his department. In contrast to his belief that humus is a significant substance in soil fertility, some chemists at Rutgers, led by the excellent experimentalist Jacob Joffe, differed. Joffe had previously been Waksman's associate in research on sulfur-oxidizing organisms and was co-discoverer with him of *Thiobacillus thiooxidans*. He and fellow chemists considered soils to be the equivalent of ion exchange substances, binding essential plant nutrients as they entered the soil, gradually releasing them to support plant growth, with the soil organic fraction being relatively insignificant for fertility. In many aspects, the arguments advanced by members of both sections of the department had validity, but their inability to come to an agreement led to a permanent subdivision of the Soils Department into two separate departments, Soil Chemistry and Soil Microbiology. Waksman was appointed chairman of the latter.

In his new capacity, Waksman had one professional associate, Robert Starkey, one graduate student position, and one technician funded permanently by the University and its associated Experiment Station. He greatly expanded his research program by receiving financial support from private funding agencies and occasionally by government grants. His work was also supplemented by accepting commercial support for projects with practical objectives that had relation to his ongoing programs. One that had special significance was support offered by Merck & Co., Inc., at that time a commercial concern in New Jersey especially interested in producing bulk chemicals for pharmaceutical and industrial uses. Merck directors felt the need to expand into organic acid manufacture and provided a fellowship to Waksman to investigate organic acid production by soil microbes. Initially, the effort was directed at fumaric acid, a product useful in the dry-cleaning industry. It was soon extended to citric acid, which was then being produced for the beverage industry by culturing *Aspergillus niger* in pure culture in large trays. Waksman and his students isolated from soils new highly efficient organisms that produced the desired acids. Because of close contacts with A.J. Kluyver of the Technische Hoogschool in Delft (Waksman's associate Robert Starkey had recently spent a sabbatical year in the Netherlands with him), Waksman became familiar with the value of cultivating molds aerobically for citric acid production by submerged culture on rotary shaking machines. With adequate oxygen available at all times, mold growth was far more rapid and it was much more dense in the shaken cultures than that obtained by the prior stationary culture techniques. This was an advanced method of culture for microbes, not only valuable in organic acid production but highly significant when Waksman extended his program utilizing submerged culture procedures into the antibiotic field.

Supplementing his broad attack on understanding the activities of microbes within the soil, Waksman became involved with many related activities. Much effort was spent in developing a classification system for the filamentous bacteria, commonly termed the actinomycetes. In 1943, he, with A.T. Henrici, proposed a new generic name *Streptomyces* for the frequently found spore-forming members of the order Actinomycetales. The new genus name became official and was included in Bergey's *Manual of Determinative Bacteriology.*

Waksman was consulted concerning the role of bacteria in initiating the fouling of ship bottoms with barnacles and other marine life. Because of the special wartime needs that developed soon thereafter, the project rapidly expanded and Waksman started summer residences at the Oceanographic Institution in Woods Hole, Massachusetts, supervising students there. For several years he applied the basic procedures of soil microbiology to investigation of the microbes of the sea, until the demands of the program required full-time supervisory effort, whereupon he resigned. Eighteen publications resulted from this scientific side excursion.

Waksman spoke frequently of threefold objectives for research in soil microbiology: first, defining the nature of the organisms present in soils; second, defining the microbiological processes that occur in soils that influence soil fertility; and, finally, elucidation of the interactions, both growth-promoting and growth-inhibiting, that occur among the multitudes of organisms that exist and interact together in soils.

Waksman's approach to the latter objective was based on observations he had made throughout his career. As early as 1923, Starkey and Waksman, in a paper describing results of their study of the microbiological changes that occur during recovery from partial sterilizations of soils, had stated: "Certain actinomycetes produce substances toxic to bacteria – around an actinomycete colony upon a plate a zone is found free from bacterial growth." They had seen such inhibitions as well as growth enhancements repeatedly when they plated soils. In 1938, Waksman decided it was appropriate to initiate a formal study of the interaction of microbes upon one another in soils. Three companion publications resulted. The first was a historical review of microbial interactions. The second, prepared with a student, covered antagonistic actions among microbes grown in the laboratory in artificial substrates. The third, with another student, covered interactions during the decomposition of plant residues. Waksman was fully prepared, therefore, to appreciate an experiment

reported the following year by Rene Dubos. Dubos, as a former Waksman student, had utilized the technique of soil enrichment at Rutgers to isolate cellulose-degrading organisms from the soil. Waksman had recommended him to Oswald Avery of the Rockefeller Institute for a project to find an enzyme capable of hydrolyzing the resistant mucoid capsule of the pneumococcus. Dubos had not only been successful in that endeavor, using the soil-enrichment procedure learned under Waksman, but had extended his studies, enriching soil with living cocci and from that enriched soil isolating a bacillus that killed cocci, including the pneumococci. The antibacterial agent responsible for the killing, which he named tyrothricin, was described in a 1939 publication. Waksman immediately encouraged his cooperators at Merck to aid Dubos by growing the tyrothricin producer in large quantities.

Waksman also saw the potential of the method as a means of discovering other important antibacterial agents. He decided to organize a project aimed at their discovery as products of actinomycetes. He assigned his university-supported graduate student to full-time study of microbial associations. Quickly, a new extremely active inhibitory substance was obtained from an actinomycete. It was crystallized in association with a Merck chemist. It was named actinomycin (Figure 1). Additional graduate students were enlisted to the task, supplemented by short-time visitors who arrived at the laboratory. The scientific publication describing actinomycin had resulted in intense newspaper publicity, attracting an even greater number of visitors than usual. Thus was organized an expanded program on antibacterial substances derived from soil microbes. When requested by an editor of an abstract journal to coin a name under which reports of such substances could be grouped, Waksman proposed the term 'antibiotic' and it became accepted worldwide.

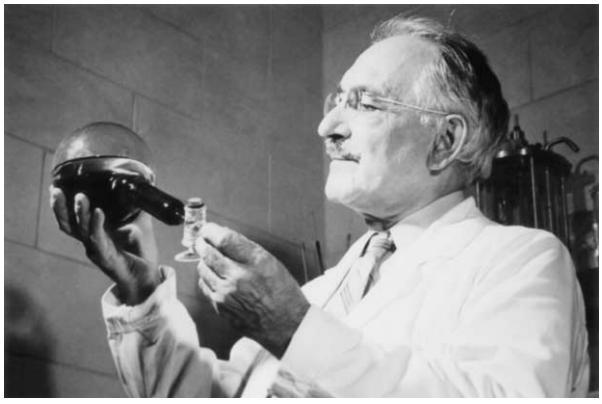


Figure 1 S.A. Waksman purifying actinomycin, his first antibiotic. (From the public files of Rutgers University, photographer unknown.)

The antibiotic discovered in Waksman's laboratory that has had greatest commercial significance is streptomycin, a broad-spectrum inhibitor, that investigators at the Mayo Clinic showed to be active against experimental tuberculosis infections in guinea pigs. Clinical trials of streptomycin soon followed. No effective therapeutic agent existed for human tuberculosis, therefore streptomycin quickly became the dominant approach for treatment of the disease, resulting in many striking cures, in spite of frequently expressed side-effects. Commercial production of streptomycin was initiated in America, initially by Merck & Co., Inc., but with conversion of the streptomycin product patent to a nonexclusive basis at Waksman's and Rutgers University's request, production was undertaken by several additional companies in the USA as well as abroad.

The commercial success of steptomycin led to an intensive search for improved antibiotic producers among the soil population, especially among actinomycetes, both at Rutgers and elsewhere. Many useful agents were found and actinomycete antibiotics have achieved great commercial value.

As the result of Waksman's prominence in the field of antibiotics and the success stories concerning streptomycin, his presence was sought constantly for advice and for the presentation of lectures, especially after he was awarded the Nobel Prize for Physiology or Medicine in 1952 for his studies on the microbes of the soil leading to the discovery of streptomycin. Students flocked to his laboratory, far more than budgets could support or for whom space was available. Waksman recommended, and the University assented, that the royalties earned from the sale of streptomycin, as well as other discoveries from his laboratory, be applied for the construction and support of a specialized Institute of Microbiology located on the Rutgers University campus. Its objective was not to be limited to research on microbes of the soil but to be applied to research on microbes broadly. Through the financial support of the royalties, the Institute of Microbiology was built and became fully staffed. Selman Waksman served as its first Director. In the Institute, important fundamental discoveries have been made and basic investigative work on microorganisms is continuing.

Supplementing this expanded activity, a research and teaching program in agricultural microbiology has been continued in Cook College, which is the successor name for the Rutgers University Agricultural College. Present in that college is a museum devoted to the development of the science of soil microbiology. It celebrates the organization in 1901 of a specialized Department of Soil Microbiology, the first such university department organized in the USA. Special

emphasis in the museum is given to Selman A. Waksman's creative leadership and to the honors accorded, including award of a Nobel Prize in 1952.

To honor Selman Waksman, in 1973 the name of the Institute for Microbiology at Rutgers University was changed to the Waksman Institute for Microbiology. For Selman A. Waksman, a dedicated research scientist for more than 50 years, and this author's mentor during the period when the first Rutgers antibiotics were discovered, the honor provided by the revised Institute name assures permanent recognition for his creative leadership in the field of soil microbiology.

See also: **Bacteria:** Soil; **Microbial Processes:** Environmental Factors; Community Analysis; Kinetics

Further Reading

Waksman SA (1927) *Principles of Soil Microbiology*. Baltimore: Williams and Wilkins.

Waksman SA (1938) *Humus. Origin, Chemical Composition, and Importance in Nature*, 2nd edn. Baltimore: Williams and Wilkins.

Waksman SA (1949) *Streptomycin – Its Nature and Practical Application*. Baltimore: Williams and Wilkins.

Waksman SA (1954) *My Life with the Microbes*. New York: Simon and Schuster.

Waksman SA (1959) *The Actinomycetes*. Vol. I. *Nature, Occurrence and Activities*, Baltimore: Williams and Wilkins.

Waksman SA (1961) *The Actinomycetes*. Vol. 2. *Classification, Identification and Descriptions of Genera and Species*, Baltimore: Williams and Wilkins.

Waksman SA and Curtis RE (1916) The actinomycetes of the soil. *Soil Science* 1: 99–134.

Waksman SA and Lechevalier HA (1963) *The Actinomycetes*. Vol. 3. *The Antibiotics of Actinomycetes*, Baltimore: Williams and Wilkins.

Waksman SA and Starkey RL (1931) *The Soil and the Microbe*. New York: John Wiley.

Woodruff HB (1968) *Scientific Contributions of Selman A. Waksman*. New Brunswick, NJ: Rutgers University Press.

WASTE DISPOSAL ON LAND

Contents
Liquid
Municipal

Liquid

C P Gerba, University of Arizona, Tucson, AZ, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

The land application of human wastes has been practiced for centuries. The benefit of the nutrients provided in these wastes has been recognized as a major benefit in crop production. However, because of the presence of pathogenic organisms in such wastes, the use of untreated wastes is no longer practiced in developed countries. The main objectives of land application of wastewater today are further effluent treatment, groundwater recharge, and provision of nutrients for agricultural crops. Land treatment is defined as the controlled application of wastewater onto the land surface to achieve a planned degree of treatment through natural physical, chemical, and biological processes within the plant–soil–water matrix. During land treatment of wastewater effluents, biological and chemical pollutants are removed by physical (settling, filtration), chemical (adsorption, precipitation), and biological (e.g., plant uptake, microbial transformation) processes.

Land-treatment systems are capable of reducing the levels of pathogenic microorganisms, biochemical oxygen demand (BOD), suspended solids, nutrients (nitrogen and phosphorus), toxic metals, and trace organics. Suspended solids are removed by filtration and sedimentation. Soluble organics are removed by microbial action in the soil. Nitrogen is removed by sedimentation-filtration (e.g., particle-associated organic nitrogen), adsorption to soil, volatilization (e.g., NH_4), uptake by crops, and biological denitrification. Phosphorus is removed by adsorption to soil particles, chemical precipitation, and uptake by vegetation.

There are three major land-treatment processes: slow rate, rapid infiltration, and overland flow. A comparison of the design features of these land-treatment

Table 1 General characteristics of the three methods used for land application of sewage effluent

	Application method		
Factor	*Low-rate irrigation*	*Overland flow*	*High-rate infiltration*
Main objectives	Reuse of nutrients and water Wastewater treatment	Wastewater treatment	Wastewater treatment Groundwater recharge
Soil permeability	Moderate (sandy to clay soils)	Slow (clay soils)	Rapid (sandy soils)
Need for vegetation	Required	Required	Optional
Loading rate	1.3–10 cm week^{-1}	5–14 cm week^{-1}	>50 cm
Application technique	Spray, surface	Usually spray	Surface flooding
Land required for flow of 10^6 l day^{-1}	8–66 hectares	5–16 ha	0.25–7 ha
Required depth to groundwater	Approx. 2 m	Undetermined	5 m or more
BOD and suspended solid removal	90–99%	90–99%	90–99%
N removal	85–90%	70–90%	0–80%
P removal	80–90%	50–60%	75–90%

BOD, biochemical oxygen demand.

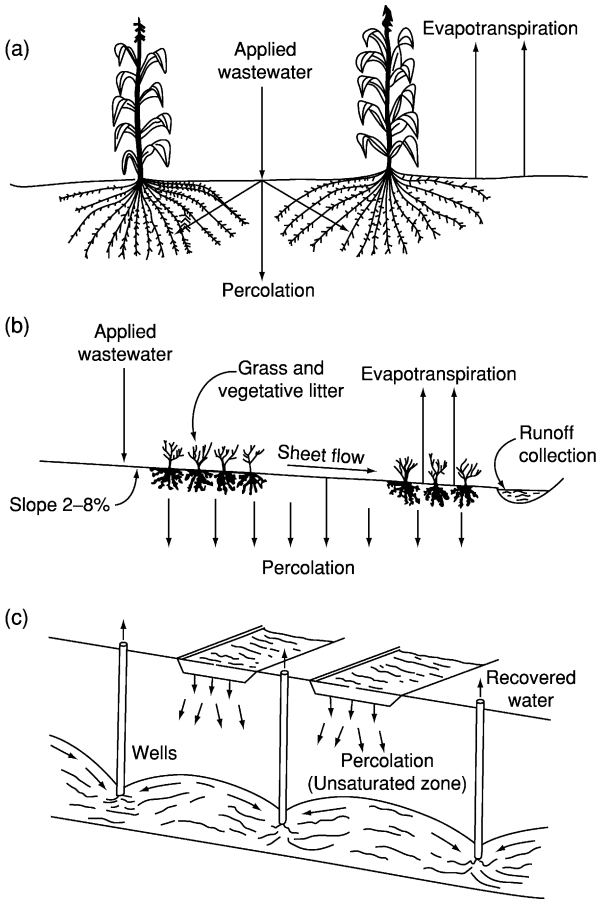


Figure 1 (a) Slow-rate irrigation system; (b) overland-flow system; (c) rapid-infiltration system. (Reproduced from the US Environmental Protection Agency (1981) *Technology Transfer Process Design Manual for Land Treatment of Municipal Wastewater*. EPA/1-81-013. Cincinnati, OH: USEPA.)

processes is shown in Table 1 and Figure 1. Expected removals of BOD, nitrogen, and phosphorus are also shown in Table 1.

In recent years, rapid-infiltration and slow-rate systems have been studied most. Rapid-infiltration systems are seen as important in the recharge of drinking-water aquifers, and slow-rate systems as a means of crop production in the arid regions of the world. Because of the potential impact of contaminants on underground drinking-water supplies, the fate of trace organics and pathogens have received increasing attention where rapid infiltration is used to supplement groundwater supplies. Because of the transport of food crops around the world today, the fate of pathogens in slow-rate systems has also received more attention.

Types of Land-Treatment Processes

Slow-Rate Process

Slow-rate irrigation systems are the most frequently used land-treatment system. There are approximately 1200 systems in the USA. Slow-rate land treatment is the application of wastewater to a vegetated land surface, with the applied wastewater being treated as it flows through the plant–soil matrix. Some of the flow percolates to the groundwater and some is used by the vegetation. Off-site surface runoff of the applied wastewater is generally avoided in the design. Nitrogen is often the limiting factor in these systems because of strict standards for levels of nitrate in drinking water. In arid lands, salts at acceptable levels for crop production may be limiting. Usually sandy loams or clay loams are the preferred soil type.

In slow-rate systems, the soil can be as shallow as 0.3 m for grass crops, but 1.5 m is preferred for complete wastewater treatment. Retention of contaminants is a function of residence time of the wastewater in the soil and the degree of contact with the soil colloids.

BOD in most slow-rate treatment systems is applied at rates of less than 11 kg ha^{-1} day^{-1}, which is

an order of magnitude lower than the capacity of the soil. High-strength food-processing wastewater systems are loaded as high as $9500\,kg\,ha^{-1}\,day^{-1}$. The BOD reductions in most systems exceed 98%. Filtration through the soil results in removal of 99% or more of total suspended solids.

Nitrogen removal is achieved using a nitrogen balance that matches the expected removal plus a percolate nitrate nitrogen of less than $10\,mg\,l^{-1}$, which is the drinking-water standard in the USA. Slow-rate systems use forage crops to remove much of the applied nitrogen. To achieve the nitrogen uptake rates expected for the crop, excess nitrogen must be applied so that the crop can compete effectively with soil microorganisms for the available nitrogen. Biological nitrogen removal occurs by nitrification–denitrification. The loss due to denitrification depends on the BOD-to-N ratio and the soil temperature, pH, and moisture. Intermittent application, which is characteristic of slow-rate systems, serves to enhance nitrification followed by denitrification. Ammonia volatilization losses of 10% can be expected if the soil pH is greater than 7.8 and the cation exchange capacity is low (low absorption of ammonium by the soil). For soils with less than 2% organic matter, nitrogen storage in the soil can be a significant loss for the first 3–4 years of operation of the system. Eventually, equilibrium is reached and the net storage of nitrogen stops.

Phosphorus removal has been found to be generally around 98–99%. Removal is by formation of insoluble phosphates with metals at the soil surface.

The concern with pathogens in slow-rate systems is related to their potential for reaching the groundwater, contamination of adjacent surface waters, and contamination of the crop. The low rate of wastewater application usually limits the potential for groundwater contamination, and contamination of surface waters can be controlled by proper management and site selection. However, enteric viruses have been detected in groundwater beneath slow-rate systems when high rates of wastewater are applied to areas with high water tables. Food-crop contamination with pathogens easily occurs during furrow irrigation, but can be limited by drip irrigation and covering the soil surface with plastic sheeting.

Rapid-Infiltration Process

In rapid infiltration, most of the applied wastewater percolates through the soil and the treated effluent percolates through to the groundwater, where it may be collected in recovery wells for drinking water or irrigation. The wastewater is applied intermittently at high loading rates ($6–125\,m\,year^{-1}$) onto a permeable soil (e.g., sands or loamy sands). Application is usually in basins and vegetation is usually not planted. Recovery wells may be used to reclaim the treated water for irrigation or as a drinking-water source. The treatment potential of rapid-infiltration systems is lower than in slow-rate systems. Removal of nitrogen is generally low but may be enhanced by encouraging denitrification. Denitrification requires adequate carbon (as found in primary wastewater effluents) and low oxygen levels, necessitating flooding periods as long as 9 days, followed by drying periods of approximately 2 weeks. Climate and season will affect the intervals of flooding and drying due to their influence on microbial activity.

Typical BOD loadings for rapid-infiltration systems using municipal wastewater range from 27 to $175\,kg\,ha^{-1}\,day^{-1}$ with removals ranging from 74 to 96%. BOD loadings of industrial systems range from 112 to $676\,kg\,ha^{-1}\,day^{-1}$. Suspended-solids loadings of $112–224\,kg\,ha^{-1}\,day^{-1}$ or more require more frequent disking or scarifying of the infiltration basin surface to avoid plugging of the soil.

Nitrification–denitrification is the principal mechanism of ammonia and nitrogen removal from wastewater in rapid-infiltration systems. Ammonia adsorption also plays an important role in retaining ammonia in the soil long enough for biological conversion. Nitrification and denitrification are affected by temperature and the amount of organic carbon available. It has been found that nitrification rates of up to $67\,kg\,ha^{-1}\,day^{-1}$ can be achieved under favorable moisture and temperature conditions. Nitrogen removal is a function of detention time, BOD-to-N ratio, and anoxic conditions. Detention time is related to hydraulic loading rates through the soil profile. For effective nitrogen removal (80% or more), the loading rate should not exceed 3–4 cm. The BOD-to-N ratio needs to be 3:1 or more to ensure adequate carbon to drive the denitrification reaction. Secondary wastewater effluent will have a BOD-to-N ratio of approximately 1:1, while primary effluent usually has a BOD-to-N ratio of 3:1. To overcome the low BOD-to-N ratio in secondary effluent, a longer period of application (7–9 days) is necessary. Typical removal of total nitrogen ranges from 38 to 93%.

Phosphorus removal is accomplished by adsorption and chemical precipitation. The adsorption occurs quickly and the chemical precipitation occurs more slowly; removal typically ranges from 40 to 97%.

Removal of bacteria, helminths, and protozoan parasites is largely accomplished by filtration. Almost all are removed within a meter of the soil surface under unsaturated conditions. Virus removal is dependent upon adsorption to the soil surface; viruses have the potential to travel long distances under the

right conditions (50 m or more). The depth of the unsaturated or vadose zone is important in the degree of removal of viruses before the wastewater reaches the aquifer. Expected removals of virus should be 99% or more.

Overland-Flow Process

In overland-flow treatment, wastewater is applied at the upper reaches of grass-covered slopes of 2–10% grade and allowed to flow over the vegetated surface to runoff collection ditches. Typical slopes are 2–4% in grade and 36–45 m long. The overland-flow process is best suited to sites having relatively impermeable soils. The most suitable soils are clay or clay-loamy soils with a permeability equal to or less than $0.5\,cm\,h^{-1}$. About 50 overland-flow systems have been built in the USA. The wastewater is treated by a thin biofilm and plants as it flows down the slope. Removal of nitrogen is due to nitrification followed by denitrification, and uptake by the grass or other plants. Phosphorus removal is via adsorption and precipitation.

In municipal systems, the BOD loading rate typically ranges from 12.3 to $4\,kg\,ha^{-1}$. Biological oxidation accounts for the 90–95% removal of BOD normally found in overland-flow systems. A typical BOD concentration in treated runoff water is approximately $10\,mg\,l^{-1}$. Overland flow is effective in removing most suspended solids, with effluent total suspended solids between 10 and $15\,mg\,l^{-1}$.

The removal of nitrogen depends on nitrification–denitrification and the crop uptake of nitrogen. Denitrification can account for 60–90% of the nitrogen removed, with denitrification rates of $160\,kg\,ha^{-1}$. Phosphorus removal in overland-flow systems is limited to approximately 40–50% because of the lack of soil–wastewater contact.

Overland flow is not very effective in removing microorganisms. Fecal coliforms can be reduced by approximately 90% when raw or primary effluent is applied; however, minimal removal occurs when secondary effluents are applied. Enteric virus removals of up to 85% have been observed with overland flow.

Treatment Mechanisms

In the land application of wastewater, the soil is used as a treatment medium. Some substances pass through the soil and into the groundwater, some are utilized by growing plants, while others are retained almost indefinitely within the soil. Proper design of land-application facilities must relate the fate of pollutants to the properties of soil with which they may interact and minimize the fraction of contaminants passing through to groundwater. There are several separate-unit processes that can be adopted for removal of suspended solids, total dissolved solids, biochemical oxygen demand, nitrogen, phosphorus, and toxic substances. Waste-treatment mechanisms that occur in soils can conveniently be categorized as physical, chemical, and biological. Within each category, various processes act to remove or alter specific pollutants.

Filtration

As wastewater moves into and through soil pores, suspended solids are removed by mechanical filtration. The actual depth at which removal occurs varies with the size of suspended particles, soil texture, and the rate of application. The larger the hydraulic loading rate and the coarser the soil, the greater the distance required. However, when loading rates are such that much of the applied water is held in the soil, additional removal can occur as suspended materials settle or adhere to the surfaces of the soil particles. Some organic particles and some organisms such as worms and protozoa are large enough so that removal by simple blockage occurs. Both bacteria and viruses, however, are small enough that they can move through the soil pores. While filtration in the soil matrix limits the movement of most bacteria to a meter or less, viruses are too small to be affected, and filtration is not believed to play a significant role in their removal. In practice an organic mat or 'Schmutsdecke' forms at the soil surface, enhancing the removal of particulates. Suspended matter can evidently clog soil pores and thereby severely reduce the infiltration rate. In rapid-infiltration operations, flooding and drying cycles help restore infiltration rates. During flooding the infiltration rate is reduced over time. Drying of the Scmutszdecke results in its cracking and removal during the next infiltration event.

Adsorption and Precipitation

Chemical reactions among dissolved ions or compounds and interactions with the soil solid-phase alter the mobility of waste pollutants. Some dissolved constituents are retained within the profile indefinitely, while the movement of others is only temporarily restricted or unaffected. Two processes, adsorption and precipitation, account for most of the retention. Adsorption refers to the net effect of interactions between dissolved and particulate (i.e., viruses) matter of the soil and humus particles. Precipitation is the formation of an insoluble form of a substance that was originally in solution. Ion exchange is one type

of adsorption process. Anion exchange is of very minor importance in land-application systems. Phosphorus is the only anion appreciably retained in soils, though the primary mechanism is not anion exchange. The predominant ion exchange reactions in most soils in the USA involve cations and are related to the cation exchange capacity. In wastewater-treatment systems, the retention of dissolved cations depends largely on their concentration in solution entering the soil. Large concentrations overcome a higher degree of attraction of other cations toward the soil matrix. Many of the heavy metal cations such as zinc, copper, nickel, cadmium, mercury, lead, and chromium are present in wastewater at concentrations too low to be appreciably affected by cation exchange reactions. These cations seem to be incorporated into the soil solid-phase in a nonexchangeable form.

Chemical Reactions

Phosphorus and nitrogen are at least partially controlled by chemical mechanisms in the soil. The most important chemical mechanism for nitrogen removal is the reaction of positively charged ammonium ions with the soil cation-exchange complex.

Volatilization

Volatilization refers to the evaporation of chemical vapor from soil or water and its subsequent loss to the atmosphere. Many organic compounds are volatile in water, as are some nitrogen compounds (e.g., ammonia, nitrous oxide) generated from biological transformations. In addition, some inorganic chemicals (e.g., selenium compounds) may be rendered volatile through biological reactions. For a given chemical, the extent of volatilization is very dependent on the soil and atmospheric conditions. In general, volatilization is greatly reduced in soil compared with water, because the soil solid-phase retains the chemical mass, thereby reducing its vapor pressure. In addition, the soil can offer substantial resistance to the transport of the chemical from the soil profile to the surface, particularly if the soil is wet and little upward flow of water is occurring. In rapid-infiltration processes where the wastewater is ponded over the surface for prolonged periods of time, the primary route of volatilization loss is from the surface of the standing water. During the drainage cycle, when the soil becomes unsaturated, volatile constituents in solution near the surface can evaporate and escape to the atmosphere.

Ammonia present in wastewater is very volatile and vaporizes from the anhydrous form immediately upon exposure to air. Nitrite may be transformed anaerobically by biological denitrification to several gas species (primarily N_2O and N_2) when soil-water content is high and a source of organic carbon is present.

Biological Mechanisms

Soil microorganisms alter the waste constituents through organic matter decomposition, inorganic transformations, and nutrient assimilation. These processes are largely restricted to the upper meter of soil. The ability of soil microorganisms to decompose organic matter is a function of their population complexity. The diversity of microorganisms enhances the capability of soil to degrade a wide variety of organic substances; some prey on pathogenic bacteria and reduce the survival of viruses. The rate at which organic matter decomposition occurs and the exact nature of the intermediate and end products depend in part on the composition of the added organic matter in the wastewater. Soil factors, however, exert considerable control as well. The presence or absence of oxygen, more than any other single factor, determines the rate and end products. The oxygen status of the soil is a function of soil porosity, and properties that favor rapid transmission of water also favor oxygen movement unless the soil is completely saturated.

Decomposition proceeds most rapidly under aerobic conditions. Under aerobic conditions, up to 60% of the organic carbon may be respired by organisms as carbon dioxide in the initial stage of decomposition. Much of the rest is incorporated into microbial cells, and some of this is subsequently respired when the population declines. Through organic matter decomposition, nitrogen, phosphorus, sulfur, and a host of other trace elements are converted from organic to inorganic forms, i.e., mineralization. Many of the elements that are mineralized during organic matter decomposition are then subject to inorganic transformation in the soil and are assimilated by plants in overland-flow and irrigation-treatment systems. Vegetation thus functions as a sink where waste nutrients, most notably nitrogen and phosphorus, can be effectively immobilized. When harvested, vegetation plays an integral part in the renovation or 'treatment' of the applied wastewater.

Fate of Specific Contaminants

Metals and Trace Elements

Factors that affect the retention of trace elements by soils include soil texture, pH, soil organic matter, and contents of amorphous oxides of Fe, Al, and Mn. Most studies have a high capacity to attenuate the concentration of copper and lead. The retention of

other trace elements studied is best correlated with the clay content of free iron oxides of the soil. Capacities for attenuating the cationic elements (Cu, Pb, Be, Zn, Cd, Ni, and Hg) in soils tends to increase with their clay and iron oxide contents. For the anionic trace elements (SeO_3, VO_3, AsO_4, CrO_4), retention in the soil also increases with increasing clay and iron oxide content of the soils. In general, the solubility of the cationic trace-element species increases as the pH of the soil decreases. By contrast, the solubilities of anionic trace-element species in the soil tend to increase as the pH of the soil increases.

Metals and trace elements may be removed by a number of mechanisms during land application of wastewater. Those associated with suspended matter are retained at or near the soil surface. In rapid-infiltration systems, they accumulate in the colloid material in the surface-clogging layer, which eventually must be removed to restore infiltration. Smaller suspended particulates that can move through soil pores without becoming trapped are also attenuated by sorption to mineral surfaces in the soil matrix. For dissolved elements, ion exchange, precipitation, surface adsorption, and complexing with organic compounds are important. The principal mechanisms that immobilize dissolved trace elements in the wastewater within the soil appear to be absorption and precipitation. Most soils appear to have a high capacity to remove trace elements during land treatment of wastewater. The amounts of trace elements removed by crops are small compared with the amounts applied to the soils though application of wastewater (Table 2).

Because soil sorption may be finite for most trace elements, there may be long-term accumulation of metals in the soil. Groundwater concentrations of silver, barium, cadmium, cobalt, and chromium below the rapid-infiltration site at Hollister, California, have been unaffected by the additions of wastewater to the overlaying soil. However, manganese, nickel, iron, zinc, lead, and copper are above background levels. Soil samples taken at the Whittier Narrows infiltration facility after more than 20 years of operation have shown elevated levels of cadmium, chromium, copper, nickel, lead, and zinc in the top 60 cm, but not below that depth, suggesting that the soil has the capacity to remove metals for many more years of operation before groundwater is affected.

Organic Compounds

Organic compounds vary greatly in their mobility, volatility, and persistence in soil. In land-application systems, volatile compounds volatilize prior to application, and only soluble organic compounds enter the soil. The fate of the soluble compound depends on the degree to which it is chemically or biologically transformed during its passage through the system. Organic compounds degrade by hydrolysis, photodecomposition, or redox reactions. Microbial conversion occurs chiefly at or near the soil surface, where bacterial populations and organic carbon levels are high. Low organic carbon levels limit microbial action in the deeper regions of the vadose zone during treatment in rapid-infiltration systems. The travel time of an organic compound may be roughly estimated by its retardation factor in a given soil. The overall action of the chemical and microbiological processes transforming an organic compound moving through the soil may be crudely expressed as the half-life or degradation rate constant. The length of the half-life compared with the travel time can be used as an index of the potential for the compound to survive its contact and passage through the soil.

The dissolved organic carbon (DOC) in domestic, secondary-treated wastewater during rapid infiltration is reduced by 50% after passage of a few meters through the soil. DOC is primarily removed by microbial action during passage through the soil, with more than half being removed during passage through the top 8 cm of soil. With long residence

Table 2 Expected trace element removal by vegetation from wastewater-irrigated soils

Element	*Typical concentration in wastewater ($g\,l^{-1}$)*	*Annual input ($g\,ha^{-1}$)*	*Typical concentration in vegetation ($g\,g^{-1}$)*	*Annual removal ($g\,ha^{-1}$)*	*Removal (%)*
As	<0.005	<60	1	5	8.3
Cd	0.005	60	0.5	2.5	4.2
Cr	0.025	300	0.5	2.5	0.8
Cu	0.10	1200	15	75	6.3
Hg	0.0009	11	0.02	0.1	0.9
Pb	0.05	600	2	10	1.7
Zn	0.15	1800	50	250	13.9

At an application rate of 1.2 m $year^{-1}$. Assuming annual dry matter yield of 5 t ha^{-1} e.g., potatoes.
Adapted from Page AL and Chang AC (1984). Fate of wastewater constituents in soil and groundwater: trace elements. pp. 31-1 through 13-16. In: Pettygrove GS and Asano T (eds) *Irrigation with Reclaimed Muncipal Wastewater. A Guidance Manual.* Report Number 84-1. Sacramento, CA: California State Water Resources Control Board.

times in the soil, aquifer DOC is reduced to 1 mg l^{-1} after 12–24 months.

Disinfection By-products

In most cases, wastewater is disinfected before land application with chlorine, chloramine, ozone, or ultraviolet light. In the case of chemical oxidizers, this results in the formation of disinfection by-products which may potentially be harmful to human health. Some of these compounds are listed in Table 3 along with their probable fate during land application of wastewater. In addition to known disinfection by-products, residual halogen is present when chlorine is used. This material may then appear as adsorbable organic halogen (AOX). Sorption processes do not play a significant role in their removal. Under anoxic conditions removal is probably based on co-metabolism and is related to the DOC that is needed as a substrate. Ozone disinfection of wastewater improves the biodegradability of refractory organic compounds and co-substrate concentration for AOX co-metabolism in soil columns. AOX removal in 1-m soil columns to which secondary effluent is applied averages 30%.

Endocrine Disruptors

Many endocrine-disrupting compounds are present in wastewater at trace quantities (i.e., estrogens, pesticides, polychlorinated biphenyls, phthalates, pharmaceuticals). Estrogens are probably the largest contributors to the endocrine-disruption activity in domestic wastewater, since daily urination adds estrogen to household wastewater. Humans are known to excrete between 10 000 and 100 000 ng l^{-1} of 17beta-estradiol per day. Synthetic estrogens such as 17alpha-ethinyl estradiol have been detected in wastewater and in surface waters affected by effluent discharge. Measurements of endocrine-disrupting compounds displaced by a fluorescent estrogen assay or cell proliferation assay have shown the reduction of endocrine disruptors during rapid infiltration. By cell proliferation assay, the estradiol equivalent value of 71 ng l^{-1} was found to be reduced to 13.7 ng l^{-1} after passage of 63 m through the vadose zone. The equivalent 17beta-estradiol activity measured by a binding assay was reduced by 97% after travel through the vadose zone to reach the aquifer. Since the study site had been in operation for more than 10 years, it seems unlikely that endocrine disruptors would ever break through the vadose zone.

Table 3 Fate of disinfection by-products (DBP) during rapid infiltration

DBP	*Disinfectant*	*Estimated fate during recharge*
Chloroform	Chlorine	Volatilization
	Chloramines	Sorption
Bromodichloromethane	Chlorine	Volatilization
	Ozone	Sorption
Trichloroacetic acid	Chlorine	Mineral sorption
MX	Chlorine	Sorption
		Degradation
Bromate	Ozone	Reaction with soil

MX, 3-chloro-4(*d*-chloromethyl-5-hydroxyl-2(*H*)-furanone.
Adapted from NRC (1994) *Groundwater Recharge Using Waters of Impaired Quality*. Washington, DC: US Government Printing Office.

Pharmaceuticals

There has been increasing concern about the fate of micropollutants originating from pharmaceuticals and active ingredients in personal-care products (e.g., soaps, deodorants) that are introduced into domestic wastewater. There is evidence that substances of pharmaceutical origin are not completely eliminated during wastewater treatment or biodegraded in the environment. At two rapid-infiltration sites in Arizona, neither ibuprofen nor naproxen or any other acidic drug were detected in groundwater wells downgradient of the infiltration basins (Table 4). These results indicate a high potential for degradation of anti-inflammatory and lipid-regulator drugs. However, antiepileptic, carbamazepine, and primidone drugs have been detected in all of the wells, suggesting that they are able to persist in soil-treatment systems.

Microorganisms

From the standpoint of acute diseases, microbial contaminants are of greatest concern during the land application of wastewater. Microorganisms are responsible for more than 90% of all waterborne outbreaks reported each year in the USA. Hundreds of different types of pathogenic microorganisms (i.e., bacteria, viruses, protozoa, helminths) are excreted in fecal material of infected hosts and these can find their way into municipal wastewater. The number of types of pathogenic microorganisms present in wastewater varies by location and over time at a given location. A variety of factors influence pathogen content of wastewater, including the incidence of the disease in the population, the season of the year, the economic status of the population, and water-use patterns. Diseases caused by waterborne organisms range from mild gastroenteritis to severe and life-threatening illness such as hepatitis, cholera, typhoid, paralysis, meningitis, heart disease, etc. New enteric pathogens are discovered almost yearly and the significance of new ones changes over time. *Cyclospora*

Table 4 Fate of pharmaceuticals after rapid infiltration

Use/origin	*Compound*	*Secondary-treated wastewater (mg l^{-1})*	*Detection in groundwater (ng l^{-1})*
Lipid regulator	Gemfibrozil	1235	Not detected
Antiepileptic	Carbamazepine	Not detected	455
	Primidone	110	115
Analgesic/anti-inflammatory	Ibuprofen	3380	Not detected
	Naproxen	6280	20
	Fenoprofen	35	Not detected
	Propyphenazone	20	15

Adapted from Drewes JE, Heberer T, and Reddersen K (2002) Fate of pharmaceuticals during indirect potable reuse. *Water Science and Technology* 46: 73–80.

and TT hepatitis virus are examples of new pathogens that have been recognized only recently, and the significance of caliciviruses (Norovirus) as a major cause of food-borne illness has only recently been recognized.

During land application of wastewater, pathogens are removed from the system by a combination of die-off (inactivation in the case of viruses) and physical filtration or adsorption by the soil or plant material. Pathogen removal by overland flow systems has been little-studied, but it is not expected to result in large reductions of enteric bacteria or viruses. Irrigation systems and rapid-infiltration systems have been studied more because of the potential for crop and aquifer contamination.

Transmission of food-borne illness by enteric pathogens due to irrigation with untreated wastewater has been well-established for more than 100 years. For this reason, irrigation with untreated wastewater for food, crops is usually forbidden. Wastewater to be used for food crop production should meet the same standards as drinking water and be intensely monitored. Thus almost all irrigation with wastewater is for the production of nonfood crops or fruit crops. However, recent outbreaks of disease associated with apples and raspberries caused by the parasites *Cryptosporidium* and *Cyclospora* suggest that only wastewater treated to potable standards should be used in any food crop production.

Factors controlling the survival of pathogens are shown in Table 5. Table 6 shows the range of reported survival times of selected enteric pathogens in the environment. Temperature is probably the most important factor controlling persistence of pathogens on plant material and soil. The die-off of pathogens increases as the temperature increases; however, under proper conditions (high humidity), some growth of bacterial pathogens may occur on the plants in the field. Rotavirus and enteroviruses appear to be inactivated very rapidly on irrigated grass under the conditions found in the arid southwestern USA. Poliovirus type one is inactivated at a rate of $0.06\,\log_{10}\,h^{-1}$ and rotavirus SA-11 at $0.04\,\log_{10}\,h^{-1}$ during the winter (4–10°C) in Arizona (USA). The rates of die-off during the summer are $0.37\,\log_{10}\,h^{-1}$ and $0.2\,\log_{10}\,h^{-1}$. It would appear that 8–10 h is needed during the summer and 16–24 h during the winter before inactivation of these viruses. *Salmonella* and other enteric bacteria can survive for several weeks on grass if sufficient organic matter and moisture is available. Helminth eggs such as *Ascaris* are

Table 5 Factors influencing enteric pathogen survival in soils

Factor	*Comments*
Temperature	Lower temperatures promote increased survival
Moisture content	Survival is decreased in drying soils. Increases may result in bacterial growth
pH	Survival is decreased pH extremes
Soil type	Survival influenced by chemical, textural, and mineralogical properties such as pH, exchangeable ion content, and water-retention capacity
Organic matter	Survival enhanced and growth of bacteria possible
Soil microflora	Survival is decreased in nonsterile soil
Salt species and concentration	Cationic type and ionic strength may influence survival
Soil surfaces	Survival may be enhanced by adsorption to soil particles
Saturation	Air–water interface under unsaturated conditions may enhance viral inactivation

Table 6 Survival of pathogens in soils

Organism	*Survival time (days)*
Coliforms	38
Fecal streptococci	26–77
Salmonellae	15–90
Salmonella typhi	1–120
Entamoeba histolytica cysts	6–8
Enteroviruses	8–175
Ascaris ova	Up to 2 years

believed to survive for 30–60 days, although they may survive many months in the soil itself.

Temperature also significantly influences the survival of enteric pathogens in soil. Poliovirus persists for 8 days in saturated sand and sandy loam soils at 4°C, whereas more than 175 days are required at 8°C. Survival of bacteria in soils is also favored by cold temperatures. However, freeze–thawing increases mortality of enteric bacteria and protozoa. *Cryptosporidium* may persist in soils for several weeks, although its survival time is reduced at temperatures above 35°C and if multiple freeze–thaw cycles in the soil take place. Extreme acidic or alkaline conditions (pH <6.0 or >8.0) tend to adversely affect the survival of both enteric bacteria and viruses in soil. Greater moisture also enhances survival of most enteric organisms in soil as does greater concentrations of organic matter in the wastewater.

Transport of microorganisms through the soil is influenced by many of the same factors that affect survival of microorganisms (Table 7). Because of their large size, bacteria, protozoan parasites, and helminths are usually removed within a meter or less of the soil surface. Because of their small size (20–200 nm), viruses have the potential to travel the greatest distances through soil. Removal largely occurs by adsorption to the soil surface and is influenced by the surface properties of the virus and soil. Electrostatic and hydrophobic interactions control the degree of adsorption. Transport is favored in sandy soils with a low clay content. Adsorption largely occurs near the soil surface because of the presence of iron oxides to which the negatively charged viruses readily adsorb. Transport is more limited under unsaturated conditions and at pH levels greater than 8.0. Organic matter in wastewater reduces adsorption because of competition for adsorption sites on the soil surface.

Table 7 Factors influencing microbial transport

Factor	*Comments*
Size of microorganism	Transport of bacteria and parasites is limited because of size in most agricultural soils
Adsorption	Major factor in limiting virus transport through soils. Also plays a role in bacterial transport
pH	Lower pH enhances adsorption
Salt species and concentration	Increased cation valency and ion concentration increase retention; low ionic strength promotes desorption and transport
Organic matter	Some organics (e.g., humic and fulvic acids) interfere with adsorption and cause desorption
Metal oxides	Virus adsorption to iron oxides occurs readily in soils
Hydraulic conditions and moisture content	Increase flow rates and saturated flow decrease adsorption

See also: **Groundwater and Aquifers**; **Pollutants:** Persistent Organic (POPs); **Pollution:** Groundwater; **Waste Disposal on Land:** Municipal

Further Reading

Amy G, Wilson LG, Conroy A *et al.* (1993) Fate of chlorination by-products and nitrogen species during effluent recharge and soil aquifer treatment (SAT). *Water Environment Research* 65: 726–734.

Asano T (1985) *Artificial Recharge of Groundwater.* Boston, MA: Butterworth.

Asano T (1998) *Wastewater Reclamation and Reuse*. Lancaster, PA: Technomic.

Bitton G (1999) *Wastewater Microbiology.* New York: Wiley-Liss.

Crites R and Tchobanoglous G (1998) *Small and Decentralized Wastewater Management Systems*. New York: McGraw-Hill.

Drewes JE and Jekel M (1998) Behavior of DOC and AOX using advanced treated wastewater for groundwater recharge. *Water Research* 32: 3125–3133.

Drewes JE, Heberer T, and Reddersen K (2002) Fate of pharmaceuticals during indirect potable reuse. *Water Science and Technology* 46: 73–80.

Fox P, Narayanaswamy K, Genz A, and Drewes DE (2001) Water quality transformation during soil aquifer treatment at the Mesa Northwest Water Reclamation Plant, USA. *Water Science and Technology* 43: 343–350.

Hurst CJ, Gerba CP, and Cech I (1980) Effects of environmental variables and soil characteristics on virus survival in soil. *Applied and Environmental Microbiology* 40: 1067–1079.

Page AL and Chang AC (1984) Fate of wastewater constituents in soil and groundwater: trace elements. pp. 31-1 through 13-16. In: Pettygrove GS and Asano T (eds) *Irrigation with Reclaimed Muncipal Wastewater. A Guidance Manual*. Report Number 84-1. Sacramento, CA: California State Water Resources Control Board.

Quanrud DM, Arnold RG, Wilson LG *et al.* (1996) Fate of organics during column studies of soil aquifer treatment. *Journal of Environmental Engineering-ASCE* 122: 314–321.

Reed SC and Crites RW (1984) *Handbook of Land Treatment Systems for Industrial and Municipal Wastes*. Park Ridge, NJ: Noyes Publications.

Reed SC, Crites RW, and Middlebrooks EJ (1995) *Natural Systems for Waste Management and Treatment*, 2nd edn. New York: McGraw-Hill.

US Environmental Protection Agency (1981) *Technology Transfer Process Design Manual for Land Treatment of Municipal Wastewater.* EPA/1-81-013. Cincinnati, OH: EPA.

US National Research Council (1994) *Ground Water Recharge Using Waters of Impaired Quality*. Washington, DC: US Government Printing Office.

Yates MV and Gerba CP (1998) Microbial considerations in wastewater reclamation and reuse. In: Asano T (ed.) *Wastewater Reclamation and Reuse*, pp. 437–488. Lancaster, PA: Technomic.

Municipal

D A C Manning, University of Newcastle, Newcastle upon Tyne, UK

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Waste materials are produced by all societies. The nature of waste and the quantities requiring management vary according to both geographic and historical factors, largely because the definition of waste is governed by both political and economic considerations. Given that at the present time it is required that waste is managed to minimize possible pollution of water resources (drinking water and fisheries and/or recreation) and the atmosphere, disposal to land is the dominant procedure used in North America and Europe for the disposal of the residues left after recovery of potentially useful materials and energy. In other parts of the world, disposal to land is widespread, depending on local political, economic, and environmental factors.

The artificial process of disposal to land by landfill, land-raising, or land-spreading produces accumulations that may be as much as 30 m thick of organic matter derived from plants and animals, mineral matter (including glass), metals, and plastics. Once deposited, these materials are subjected to natural processes of decomposition, and in good practice the site is restored to a new use, through appropriate capping and revegetation. Disposal to land thus generates an artificial 'soil' profile that can affect root systems and that influences pore water and gas compositions in the unsaturated and saturated zones. To understand the consequences of land disposal on soil systems, in the broadest sense, it is necessary to understand the chemical processes that take place within landfill and how they change with time.

Municipal solid waste (MSW) is waste generated by urban populations and collected for disposal by municipal authorities. It is typically composed of putrescible waste from food production, distribution, and consumption, together with discarded consumer goods, packaging, and papers. It differs from industrial wastes that are specific to particular factories or processes and so have a focused impact requiring attention on a case-by-case basis. It differs from mine waste and demolition wastes, which are 'inert' through their lack of putrescible material (although important because of, for example, reactions involving sulfides or sulfates).

The amounts of MSW generated by the UK, the USA, and Japan are summarized in Table 1. This shows variation in the relative proportions of key waste types. Japan has little land available for landfill disposal, whereas land is abundant in the USA. Like all European countries, the UK has a very tightly regulated waste management system, coupled with shortage of land close to urban centers where there is considerable popular opposition to waste incineration.

MSW Composition and Degradation

The composition of municipal solid waste (Figure 1) is dominated by paper (33%) and putrescible waste (20%). Putrescible waste includes raw and cooked

Table 1 Amounts of waste produced annually in the USA, UK, and Japan, and relative proportions of methods of disposal (1995)

	USA	*UK*	*Japan*
Household waste (t year^{-1})	210	20	50
Landfill (%)	63	90	20
Incinerated (%)	16	5	75
Recycled/reused (%)	17	5	5

Based on data in Williams PT (1998) *Waste Disposal and Treatment*. Chichester, UK: John Wiley & Sons, Ltd.

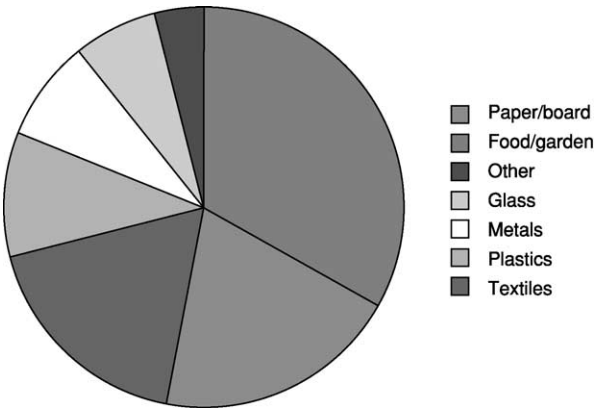


Figure 1 Approximate composition (percentage weight) of municipal solid waste (UK, 1992). (Data from Williams PT (1998) *Waste Disposal and Treatment*. Chichester, UK: John Wiley & Sons, Ltd.)

food waste and waste from domestic gardens or yards. It is composed of fats, carbohydrates and proteins, all of which decompose readily through microbial-mediated putrefaction. Paper waste is dominated by cellulose, with a significant mineral component (up to 25%) through the use of kaolin and calcite as paper-coating and -filling agents. Compared with putrescible waste, paper is relatively inert within a landfill and may not degrade for many years.

Within a landfill, waste is deposited ideally under dry conditions and is rapidly compacted and buried beneath a daily cover (often a local clay or an inert waste stream). Anaerobic conditions become established rapidly, and microbial degradation is dominated by communities of anaerobic bacteria. Although moisture ingress is minimized as far as possible, the inherent moisture content of the waste and some input from rainfall is sufficient to produce a liquid, 'landfill leachate,' that drains to the bottom of the waste pile. At the same time, microbial reactions within the leachate and on surfaces within the waste produce methane gas, which rises through the waste pile. Landfill management revolves around the control of both leachate and gas, to prevent harm that might arise from their escape, and to recover energy from landfill gas.

Typically, landfill gas is composed of approximately 66% methane and 34% carbon dioxide, with varying proportions of other gas species (Table 2). This 'end-member' composition is diluted with nitrogen from the atmosphere (oxygen having been consumed prior to the onset of anaerobic conditions). Reported landfill gas compositions may contain oxygen due to ingress of air to the sample prior to analysis. High levels of hydrogen sulfide are unusual, but may arise in circumstances where a site has received gypsum waste (especially plaster-based building products) or other sulfate-bearing industrial waste.

Landfill leachate composition varies according to the age of a landfill and the types of waste that have been deposited within it. Typical compositions are given in Table 3. The dominant cationic species

Table 2 Typical composition of landfill gas

Component	*Typical value (vol %)*	*Observed maximum (vol %)*
Methane	63.8	88.0
Carbon dioxide	33.6	89.3
Oxygen	0.16	20.9
Nitrogen	2.4	87.0
Hydrogen	0.05	21.1
Carbon monoxide	0.01	0.09
Hydrogen sulfide	0.00002	35.0
Others	Less than 0.02	NA

Based on data in UK Department of the Environment (1989) Waste Management Paper 27. *Landfill Gas*. London, UK: The Stationery Office.

Table 3 Typical contents of major constituents of landfill leachate (derived from Department of the Environment, 1995)

		Acetogenic leachate		*Methanogenic leachate*	
Component	*Units*	*Minimum*	*Maximum*	*Minimum*	*Maximum*
pH	pH units	5.12	7.8	6.8	8.2
Conductivity	$\mu S\,cm^{-1}$	5800	52 000	5990	19 300
COD[a]	$mg\,l^{-1}$	2740	152 000	622	8000
BOD_5[b]	$mg\,l^{-1}$	2000	68 000	97	1770
TOC[c]	$mg\,l^{-1}$	1010	29 000	184	2270
VFA[d] (as C)	$mg\,l^{-1}$	963	22 414	<5	146
Alkalinity	$mg\,l^{-1}$ $CaCO_3$	2720	15 870	3000	9130
Chloride	$mg\,l^{-1}$	659	4670	570	4710
Nitrate	$mg\,l^{-1}$	<0.9	79.7	0.9	9.3
Sulfate	$mg\,l^{-1}$	<5	1560	<5	322
Phosphate	$mg\,l^{-1}$	0.6	22.6	0.3	18.4
Ammonium	$mg\,l^{-1}$	249	4641	364	2623
Sodium	$mg\,l^{-1}$	474	2400	474	3650
Potassium	$mg\,l^{-1}$	350	3100	100	1580
Magnesium	$mg\,l^{-1}$	25	820	40	1580
Calcium	$mg\,l^{-1}$	270	6240	23	501
Iron	$mg\,l^{-1}$	48.3	2300	1.6	160
Manganese	$mg\,l^{-1}$	1.40	164.0	0.04	3.59

[a]Chemical oxygen demand.
[b]Biochemical oxygen demand.
[c]Total organic carbon.
[d]Total volatile fatty acids, as C.
Adapted from UK Department of the Environment (1995) Waste Management Paper 26B. *Landfill Design, Construction and Operational Practice*. London, UK: The Stationery Office.

are ammonium, Na, K, Mg, Ca, and Fe. The dominant anionic species include chloride, bicarbonate, and organic acid anions. Leachates differ from surface waters and from most natural shallow groundwaters in having high ammonium and organic acid anion contents, sometimes as the dominant cation or anion, respectively.

Variation in landfill gas and leachate composition with time is shown in Figure 2. Atmospheric oxygen is consumed rapidly, and anaerobic conditions lead initially to the formation of hydrogen (during acetogenesis, which is the stage when ethanoate (acetate) is a dominant product of decomposition) and then methane (methanogenesis; which is the stage when methane is the dominant product of decomposition). Gas production peaks and then declines over periods of months to years (depending on local conditions). Over a similar period, leachate composition shows a major change as acetogenic conditions are replaced by methanogenesis. As organic acid anions are consumed quantitatively by microbial communities that produce methane, inorganic solutes also reduce to very low levels. Ammonium is not affected greatly by this change, remaining at 100- to 1000-mg l^{-1} levels.

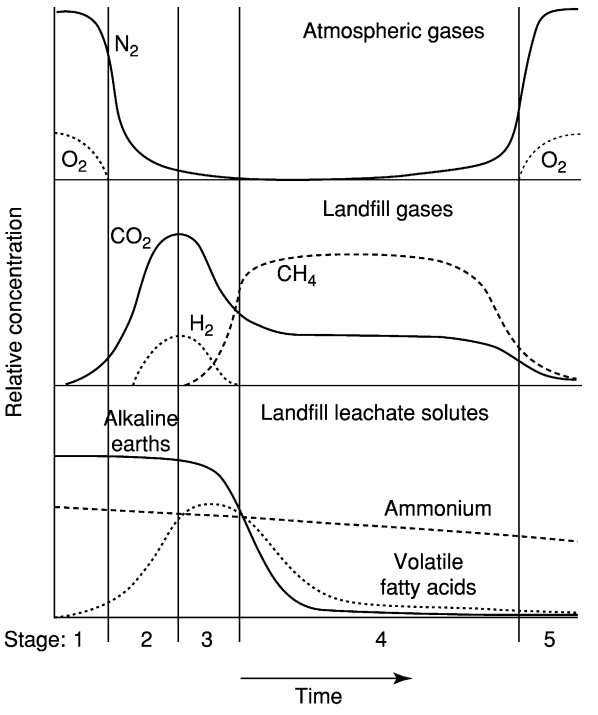


Figure 2 Changes with time in (a) atmospheric gases, (b) landfill gas, and (c) leachate composition (expressed relative to chloride). Stages 1 and 5 are dominated by aerobic conditions. Stages 2, 3, and 4 are anaerobic acidogenesis, acetogenesis, and methanogenesis, respectively. (Adapted from Rees JF (1980) The fate of organic compounds in the landfill disposal of organic matter. *Journal of Chemical Technology and Biotechnology* 30: 161–175; Robinson HD (1995) *A Review of the Composition of Leachates from Domestic Wastes in Landfill Sites*. Report CWM/072/95. Department of the Environment. London, UK: The Stationery Office.)

Microbial Reactions and Waste Degradation

The decomposition of the constituents of the putrescible fraction (proteins, lipids, and carbohydrates) is summarized in Figure 3. Initially, relative molecular masses are reduced through aerobic hydrolysis and anaerobic fermentation reactions that reduce chain length largely without destroying the characteristic functional groups, and that produce intermediate products that are readily water-soluble. Thus proteins degrade to give amino acids, carbohydrates yield simple sugars, and lipids form glycerol and long-chain fatty acids.

During the acetogenic stage, acetate is produced as an end-member of the degradation of amino acids (which also yield the ammonium ion) and long-chain fatty acids. The microbial communities responsible for the degradation of the fatty acids typically include two components in an obligate syntrophic relationship – one population oxidizes the volatile fatty acid anion, yielding hydrogen, and another reduces hydrogen or an inorganic species (sulfate, nitrate, or iron, all acting as electron donors). Examples of reactions known to occur for specific bacterial communities (e.g., *Syntrophobacter wolinii*

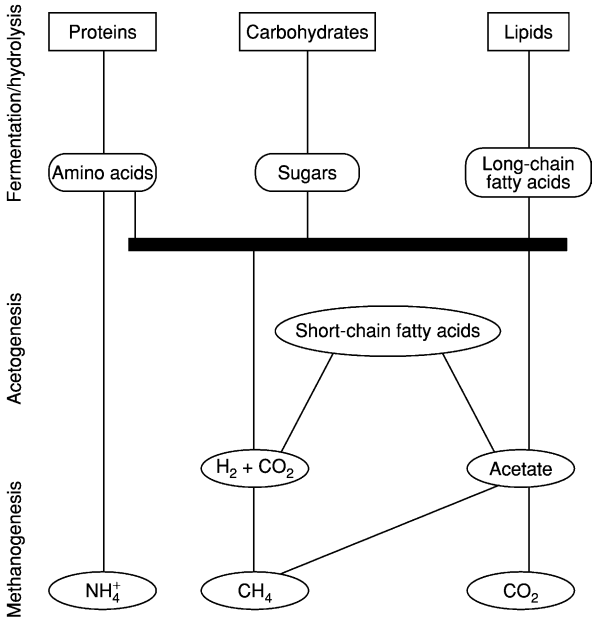


Figure 3 Schematic to summarize the decomposition of putrescible matter. (Adapted from Rees JF (1980) The fate of organic compounds in the landfill disposal of organic matter. *Journal of Chemical Technology and Biotechnology* 30: 161–175.)

Table 4 Examples of microbially mediated acetogenic reactions within landfill

	Microorganism	*Reaction*	ΔG_r^o *(kJ)*
Propionate-acetate	*Syntrophobacter wolinii*	$CH_3CH_2COO^- + 3H_2O = CH_3COO^- + H^+ + 3H_2 + HCO_3^-$	+76.1
Butanate-acetate	*Syntrophomonas wolfei*	$CH_3CH_2CH_2COO^- + 2H_2O = 2CH_3COO^- + H^+ + 2H_2$	+48.1
Sulfate reduction[a] *with oxidation of:*			
Hydrogen	*Desulfovibrio vulgaris*	$4H_2 + SO_4^{2-} + H^+ = 4H_2O + HS^-$	−152.2
Acetate	*Desulfobacter postgatei*	$CH_3COO^- + SO_4^{2-} = 2HCO_3^- + HS^-$	−47.6
Propionate	*Desulfobulbus* sp. or *S. wolinii* + *D. vulgaris*	$4CH_3CH_2COO^- + 3SO_4^{2-} = 4CH_3COO^- + 4HCO_3^- + 3HS^- + H^+$	−150.6
Methane	Species unknown in 1988; still controversial in 2003	$CH_4 + SO_4^{2-} = HCO_3^- + HS^- + H_2O$	−16.6

[a]Sulfate reducers can reduce nitrate to ammonium.

+ *Desulfovibrio vulgaris*) are summarized in Table 4. The 1:1 relationship for acetate and propionate that is observed in landfill leachates (Figure 4) appears to be characteristic of anaerobic systems of this type, as it is also observed for sedimentary waters (oil-field waters).

During the methanogenic stage, methane is produced in two ways once sulfate, nitrate, and iron in solution have been exhausted. First, decomposition of acetate in solution is facilitated by acetoclastic bacterial communities (e.g., *Methanosarcina barkeri*, *Methanosarcina mazei*, *Methanothrix soehngenii*). Secondly, methanogenic bacteria produce methane directly from hydrogen and carbon dioxide (or bicarbonate):

- General reaction (all methanogens except acetoclasts; syntrophic):

$$4H_2 + HCO_3^- + H^+ = CH_4 + 3H_2O$$
$$\Delta G_r^o = -135.6\,\text{kJ};$$

- Acetoclastic methanogenesis (*Methanosarcina barkeri*, *Methanosarcina mazei*, *Methanothrix soehngenii*):

$$CH_3COO^- + H_2O = CH_4 + HCO_3^-$$
$$\Delta G_r^o = -31.0\,\text{kJ}.$$

In all landfill leachates, the nature of the microbial community will be complex and variable in time as well as spatially. The production of methane is rarely homogeneously distributed throughout a landfill site, partly as a consequence of leachate-management procedures that mix leachates of different ages in separate parts of a landfill site.

Mineralogical Reactions Within Waste

The changes in the inorganic solute constituents of landfill leachates that are observed relate directly to mineralogical reactions that take place within the waste. These reactions can be predicted through use of the technique of geochemical modeling of landfill leachate compositions, in which a chemical analysis is recalculated to give thermodynamic concentrations of dissolved species, and then this information is used to calculate saturation indices (SI) for selected minerals. Those mineral species that have $SI > 0$ are predicted to precipitate, and those with $SI < 0$ are predicted to dissolve. Evidence that predicted mineral precipitation reactions take place can be obtained by determining the mineralogical composition of the suspended solids within leachate, or of scale that forms within leachate drainage and pumping systems.

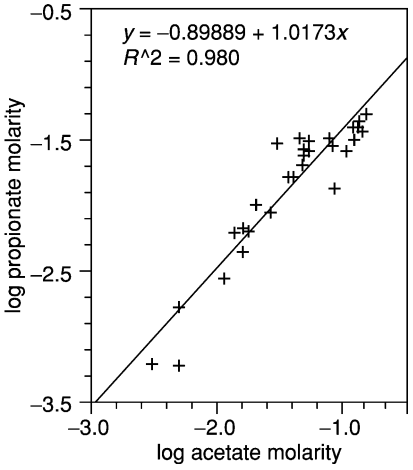


Figure 4 Relationship between acetate and propionate within landfill leachates. (Data from Manning DAC (1997) Acetate and propionate in landfill leachates: implications for the recognition of microbiological influences on the composition of waters in sedimentary systems. *Geology* 25: 279–281.)

Typically, landfill leachates have saturation indices for calcite and carbonate minerals that are greater than zero and so precipitation of calcite is predicted. Figure 5 shows the change in calcite SI with time, showing a general decrease with age. Also, calcite

saturation is observed over a wide range of pH values (Figure 6). Calcite precipitation is evident from the common observation that calcite forms scale within landfill leachate-management systems, and from the presence of calcite in a number of morphologies within leachate-suspended solids.

Silicate minerals typically show saturation indices greater than zero and so are predicted to precipitate. However, the ubiquitous presence of quartz and clays within waste- and landfill-containment systems means that newly precipitated silicate minerals cannot be recognized with confidence. Additionally, the very slow reaction kinetics of silicate mineral precipitation suggest that these minerals will not form within landfill systems under typical temperature conditions.

The recalculation of leachate compositions as activities allows mineralogical controls on leachate chemistry to be elucidated. Cation exchange reactions involving leachate and clay minerals clearly explain the distribution of K and ammonium in solution (Figure 7). Silica activities are tightly constrained to values consistent with a control by equilibrium between illitic and smectitic clays (Figure 8). These patterns are seen irrespective of the type of containment system used to hold the waste and instead relate

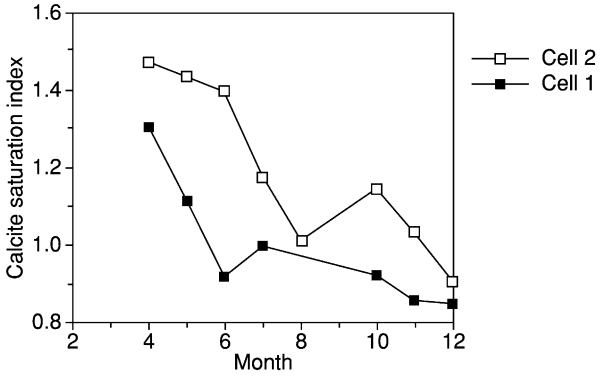


Figure 5 Variation in calcite saturation index calculated from leachate compositions for two landfill cells over time.

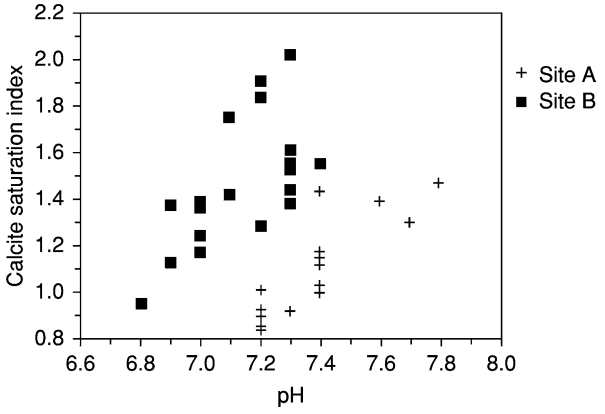


Figure 6 Variation in calcite saturation index calculated for leachate compositions with varying pH values.

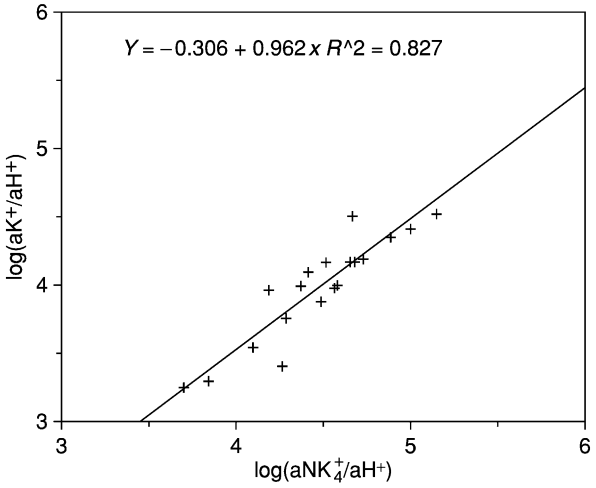


Figure 7 Ion activity diagram showing ammonium-to-hydrogen ion activity ratios $\log(aNH_4^+/aH^+)$, versus potassium-to-hydrogen ion activity ratios ($\log(aK^+/aH^+)$) for leachate data reported by Owen and Manning. A slope of 1 is consistent with mineral-controlled ammonium–potassium cation exchange. The term 'activity' refers to the thermodynamic concentration of the chemical species. (Adapted from Owen JA and Manning DAC (1997) Silica in landfill leachates: implications for clay mineral stabilities. *Applied Geochemistry* 12: 267–280.)

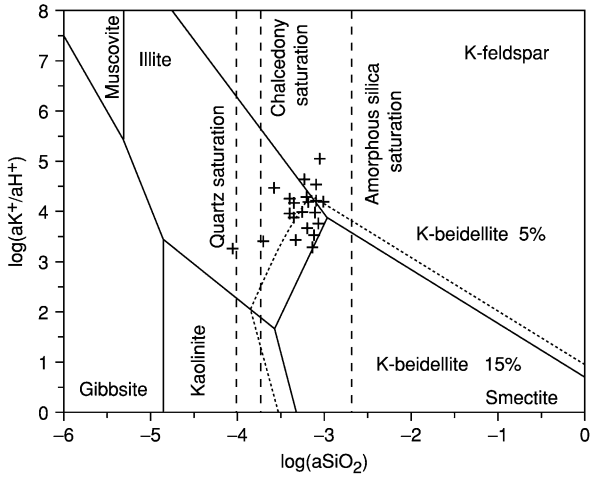


Figure 8 Ion activity diagram showing mineral stability fields calculated for soil minerals (including two smectite compositions: K-beidellite end-member, 5% (dashed) and 15% (solid)) as a function of potassium-to-hydrogen ion activity ratio ($\log(aK^+/aH^+)$) versus dissolved silica activity ($\log(aSiO_2)$), with activity ratios for landfill leachates. (Adapted from Owen JA and Manning DAC (1997) Silica in landfill leachates: implications for clay mineral stabilities. *Applied Geochemistry* 12: 267–280.)

to the nature of the minerals that are present within municipal waste as constituents of consumer goods. Sorbents such as animal litters are based on smectite–illite mixtures; clays within paper are kaolinite–illite mixtures.

Summary

Landfill disposal of municipal waste involves the artificial creation of an anaerobic environment in which degradation of putrescible matter and mineral reactions produce solutions and gases whose composition is controlled by natural processes analogous to those that take place in soils. Degradation products include methane and carbon dioxide, and calcium carbonate mineral precipitate. Clay minerals within waste control solution composition through cation exchange, and through the illite-smectite reaction.

List of Technical Nomenclature

–	pH (pH units)
ΔG_r^o	Gibbs free energy of reaction (kilojoules)
$\mu S\,cm^{-1}$	Conductivity (micro-Siemens per centimeter)
a	Activity
BOD_5	biochemical oxygen demand (5 days' incubation)
COD	chemical oxygen demand
kJ	Energy (kilojoules)
m	Molarity (moles per liter)
$mg\,l^{-1}$	Concentration (milligrams per litre)
TOC	total organic carbon
VFA	volatile fatty acids

See also: **Groundwater and Aquifers**; **Manure Management**; **Pollutants:** Persistent Organic (POPs); **Pollution:** Groundwater; **Waste Disposal on Land:** Liquid

Further Reading

Chistensen TH, Kjeldsen P, Bjerg PL *et al.* (2001) Biogeochemistry of landfill leachate plumes. *Applied Geochemistry* 16: 659–718.

Dearlove J (1998) *An Introduction to Landfill*. London, UK: Arnold.

Maliva RG, Missimer TM, Leo KC *et al.* (2000) Unusual calcite stromatolites and pisoids from a landfill leachate collection system. *Geology* 28: 931–934.

Manning DAC (1997) Acetate and propionate in landfill leachates: implications for the recognition of microbiological influences on the composition of waters in sedimentary systems. *Geology* 25: 279–281.

Manning DAC (2001) Calcite precipitation in landfills: an essential product of waste stabilization. *Mineralogical Magazine* 65: 603–610.

Owen JA and Manning DAC (1997) Silica in landfill leachates: implications for clay mineral stabilities. *Applied Geochemistry* 12: 267–280.

Rees JF (1980) The fate of organic compounds in the landfill disposal of organic matter. *Journal of Chemical Technology and Biotechnology* 30: 161–175.

Robinson HD (1995) *A Review of the Composition of Leachates from Domestic Wastes in Landfill Sites*. Report CWM/072/95. Department of the Environment. London, UK: The Stationery Office.

UK Department of the Environment (1989) Waste Management Paper 27. *Landfill Gas*. London, UK: The Stationery Office.

UK Department of the Environment (1995) Waste Management Paper 26B. *Landfill Design, Construction and Operational Practice*. London, UK: The Stationery Office.

Williams PT (1998) *Waste Disposal and Treatment*. Chichester, UK: John Wiley & Sons, Ltd.

Zehnder AJB (1988) *Biology of Anaerobic Microorganisms*. New York: John Wiley & Sons, Inc.

Water Availability *See* **Plant–Soil–Water Relations**

WATER CONTENT AND POTENTIAL, MEASUREMENT

G S Campbell and C S Campbell, Decagon Devices, Inc., Pullman, WA, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Soil-water content is a measure of the amount of water (volume or mass) contained in a unit volume or mass of soil. If the measure is the volume of water per unit volume of soil, the water content is called the 'volume wetness.' If it is the mass of water per unit mass of soil it is called the 'mass wetness.' Mass wetness in soils is typically defined as the mass of water per unit mass of dry soil. Volume and mass wetness are related by:

$$\theta = \frac{\rho_b}{\rho_w} w \qquad [1]$$

where θ (cubic meter per cubic meter) is the volume wetness, w (kilogram per kilogram) is the mass wetness, ρ_b (kilogram per cubic meter) is the bulk density of the soil (mass of dry soil per unit volume) and ρ_w (kilogram per cubic meter) is the density of water (10^3 kg m^{-3}).

Soil-water potential is a measure of the potential energy per unit mass, volume, or weight of soil water, compared with that of pure, free water. It is the work required, per unit quantity of water, to remove an infinitesimal quantity of water from the soil to a pool of pure, free water. The energy of soil water responds to the hydrostatic or pneumatic pressure on the water, to the concentration of solutes in the water, to the forces that adsorb the water to the surfaces of soil particles, and to the position of the water in the gravitational field. The components of the water potential arising from these interactions are, respectively: ψ_p, the pressure potential; ψ_o, the osmotic potential; ψ_m, the matric potential; and ψ_g, the gravitational potential. The pressure and gravitational components can be positive or negative. The osmotic and matric components can only be negative.

Both water content and water potential must be known to characterize the status of water in soil. Water content tells how much water is there, but gives no information about the availability of the water for plant uptake or microbial activity, and no information about the direction of movement of the water. Water tends to move from regions of high potential to regions of lower potential. A gradient in water potential is the driving force for water flow. A plant or microbe is therefore able to obtain water from the soil as long as it can maintain an internal water potential below that of the soil. Since there is a limit to the lowest potentials attainable by living organisms, organisms may cease to be able to obtain water from soil when the soil still contains a substantial quantity of water.

The gravitational and pressure potentials are important for determining rates and directions of water flow when the soil is saturated or near saturation. Both are proportional to the distance from a reference plane to the soil location in question. They are therefore measured with a ruler. Our focus here will be on the other components of the water potential, the osmotic and matric, which are more difficult to measure.

Measurement of Water Content

Gravimetric Methods

Mass wetness is typically measured by oven-drying a sample. The mass wetness is the mass loss divided by the dry mass of the sample. Samples are typically dried for approximately 24 h at 105°C. Aside from errors in weighing, the two most serious errors come from loss of mass of nonaqueous volatiles in the sample and uncertainty in the point at which the sample is really 'dry.' Increasing or decreasing the drying temperature changes the mass loss, and this change can be substantial in some soils. Variation in oven humidity can also lead to some uncertainty. Typical errors are on the order of 0.005 kg kg^{-1}. There are a number of faster drying methods which give quicker results at the expense of some accuracy. Drying of samples in a microwave oven can decrease the time required for a measurement to approximately 20 min, while still giving acceptable accuracy for many purposes.

Mass wetness is frequently used in laboratory studies, but has little relevance in the field, where one is mainly interested in the ability of the soil to store water for later transpiration and evaporation. Here volume wetness is the relevant measurement. The only direct method for measuring volume wetness is to find the mass wetness, through oven-drying, of a sample of known volume. From the known volume and the soil dry mass, the bulk density is computed. The volume wetness is then obtained using Eqn [1]. Since this is a tedious and time-consuming

measurement and is subject to large uncertainty owing to the large spatial variation found in the field (*See* **Spatial Patterns**), volume wetness is often inferred from measurements of other soil properties. These include dielectric properties, thermal properties, and interactions with nuclear radiation.

Dielectric Properties

The relative dielectric permittivity of water is approximately 80 at microwave frequencies, while the permittivity of soil minerals, organic matter, and ice ranges from 3 to 5. A measure of soil permittivity is therefore strongly influenced by the amount of water present in the soil. One of two methods is typically used to measure the permittivity of the soil: time-domain reflectrometry (TDR) and capacitance. For the TDR method, parallel conductors are placed in the soil forming a transmission line, and one measures the time required for an electromagnetic pulse to traverse the transmission line, reflect from the distal end, and return along the same line. The speed of propagation of an electromagnetic disturbance in the transmission line is determined by the dielectric permittivity of the surrounding medium. It is computed from:

$$\epsilon = \left(\frac{ct}{2L}\right)^2 \qquad [2]$$

where c is the speed of light ($3 \times 10^8\,\mathrm{m\,s^{-1}}$), t is the time to traverse the transmission line, and L is the length of the transmission line. (*See* **Time-Domain Reflectometry**.)

TDR measurements obviously require very fast response circuitry. The time resolution must be on the order of 100 ps. Equipment with this kind of time resolution is necessarily expensive. When slower circuitry is connected to the transmission line and a voltage pulse applied through a resistor of value R, the transmission line acts like a capacitor, with the soil as the dielectric. The time constant for charging the capacitor (time to charge to 63% of its final voltage) is RC, where C is the capacitance of the probe. The capacitance is related to geometric factors and is directly proportional to the dielectric permittivity of the medium surrounding the probe. The charging time is therefore directly related to the soil permittivity.

The dielectric constant of the soil can therefore be measured by either method. TDR, however, due to its higher-frequency excitation, is less influenced by electrical conductivity of the soil. Since ice has a dielectric permittivity near that of soil minerals, dielectric methods measure unfrozen water content in frozen soil.

A number of sensors are available commercially, both for TDR and capacitance measurement of soil moisture. An example of a capacitance sensor is shown in Figure 1. Sensors of this type have been used extensively for the past 20 years or so. Since they measure the properties of the dielectric within the electromagnetic field of the probe or transmission line, it is important that a representative sample of the soil lie within the field of the sensor. Unfortunately, the field lies close to the surface of the sensor, so any air gaps or compaction around the sensor may adversely affect the measurement.

Thermal Properties

Both thermal conductivity and heat capacity of soil vary with water content. (*See* **Thermal Properties and Processes**.) The relationship with heat capacity is an easy one to use for determining water content. The water content can be computed from the volumetric heat capacity of soil (*See* Eqn [13] in **Thermal Properties and Processes**):

$$\theta = \frac{C - C_s\rho_b/\rho_s}{C_w} \qquad [3]$$

Here C, C_s, and C_w (joules per cubic meter per kelvin) are the volumetric heat capacities of the bulk soil, the soil minerals, and water; and ρ_b and ρ_s are the bulk and particle densities. Since C_w is typically about twice as large as C_s, changes in water content have a fairly large effect on C. If the bulk density of the soil is known and C is measured, all values on the right-hand side of Eqn [3] are known, so water content can be computed. Typically C is measured using a dual-needle thermal properties sensor. Figure 2 shows a typical dual-needle sensor.

Nuclear Methods

Gamma rays and neutrons both interact strongly with soil water. The methods used most widely for water-content measurement involve attenuation

Figure 1 An example of a capacitance soil-water content sensor. The circuitry to measure the capacitance is in the overmolded head of the probe. The capacitance sensor is in the blade and consists of parallel conductors sandwiched between insulating layers.

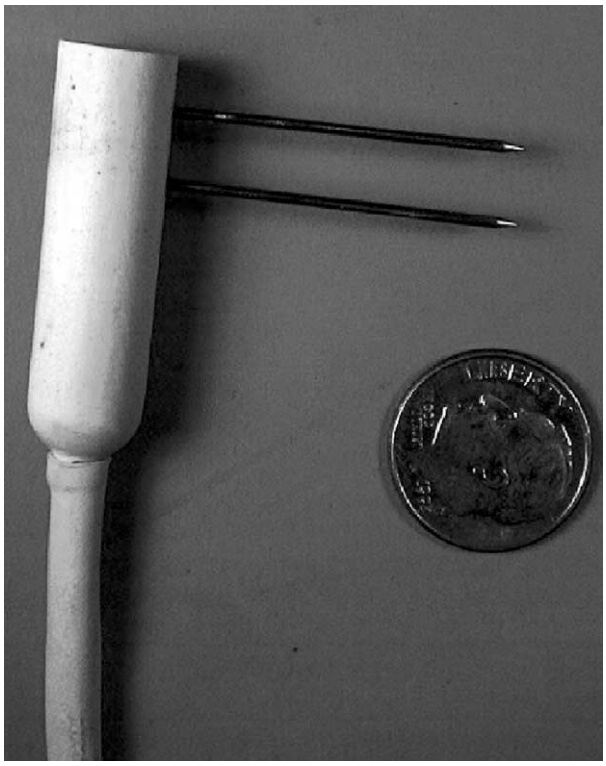


Figure 2 Dual-needle probe for measuring soil heat capacity. One needle contains a heater, the other contains a temperature sensor. The temperature rise in the temperature sensor resulting from a heat pulse to the heater is measured.

of transmitted gamma radiation and backscattering of neutrons, though neutron transmission and gamma backscatter have also been used. Gamma attenuation is typically a laboratory method. A soil column, 5–20 cm thick, is irridiated by a collimated beam of gamma radiation. Sources with appropriate energy levels are ^{137}Cs and ^{241}Am. The volume wetness is computed from:

$$\theta = \frac{\ln\left(\frac{I_d}{I_w}\right)}{\mu_w S} \qquad [4]$$

where I_d and I_w are the numbers of gamma photons passing through the dry and wet soil in unit time, respectively, μ_w is the mass attenuation coefficient for water, and S is the soil-column thickness. It is possible to determine the bulk density and the water content of a sample without measuring the gamma attenuation in the dry soil column by measuring the attenuation of two gamma beams having different energies.

Neutron scattering is a method for measuring volume wetness in the field. In a typical application, a neutron source such as ^{241}Am/Be is lowered into an access tube in the soil. The neutrons from the source initially have high energy, but, as they collide with low-atomic-mass nuclei (mainly hydrogen) in the soil, they slow and become 'thermal' neutrons, with energies typical of atoms at room temperature. These slow neutrons can be counted with a relatively simple detector. Since the main source of hydrogen nuclei in the soil is water, and since the number of scattered neutrons is proportional to the number of collisions with these hydrogen nuclei, there is a relationship between neutron count and water content:

$$\theta = a + b\frac{I}{I_{std}} \qquad [5]$$

where a and b are soil-specific empirical constants, I is the count in soil, and I_{std} is the count in a standard scattering medium. Neutron scattering is much less affected by air gaps around the sensor than are dielectric methods. The scattering volume for the neutrons varies with soil-water content from approximately 15 cm diameter in saturated soil to, e.g., 70 cm in dry soil.

Measurement of Water Potential

While the water potential concept is straightforward, no method exists for directly measuring the energy status of water in soil. All of the methods currently available equilibrate an external phase of some sort with the water in the soil and then measure the potential of the equilibrated phase. The sensing phase can be solid, liquid, or gas. Typical solid-phase sensors are heat-dissipation matric-potential sensors, gypsum blocks, dielectric matric-potential sensors, and filter paper. The tensiometer is a liquid-phase sensor. Thermocouple psychrometers and dew-point potentiometers are vapor-phase sensors.

Solid-Phase Sensors

For each of the solid-phase sensors, a porous matrix of ceramic, gypsum, or cellulose equilibrates with the soil matric potential. The water content of the matrix is then determined by weighing and drying (filter paper), measuring thermal conductivity (heat-dissipation sensor), electrical conductivity (gypsum block), or dielectric constant (dielectric sensor). From predetermined relationships between these measures of water content and matric potential, the measurements are converted to water potential.

Heat-Dissipation Matric-Potential Sensors

A heat-dissipation sensor is shown in Figure 3. It has a cylindrical, ceramic outer matrix that fully encases a stainless-steel needle. Inside the needle is an electrical resistor that runs the length of the probe

Figure 3 Heat-dissipation matric-potential sensor. The darker, upper part of the cylinder is the overmold, containing wires and connections. The lighter, lower part is the ceramic matrix, which equilibrates with the soil.

(typically 3 cm) and a temperature sensor, either a thermocouple or thermistor. To measure thermal conductivity, a known amount of current is passed through the resistor for a preset amount of time and data are collected on the temperature change (δT) over time (t). The thermal conductivity (λ_h) can be calculated from the slope (m) of the δT versus logarithm of time curve using:

$$\lambda_h \cong \frac{q}{4\pi m} \qquad [6]$$

where q is the power applied to the internal resistor (watts per meter). Because the ceramic matrix properties remain fixed, the thermal conductivity is a measure of the water content of the matrix, which, in turn, is related to the water potential of the soil. These sensors must be calibrated and, at midrange water potentials, have a strong temperature dependence. The heat-dissipation probe functions over an extremely wide range of water potentials (field capacity to air dryness).

Electrical Resistance Matric Potential Sensors

Electrical resistance sensors measure the electrical resistance between two electrodes packed in a granular (sand) or gypsum matrix to obtain water potential. As the sensor matrix comes into equilibrium with the surrounding soil, water fills the voids in the matrix. The increase in water causes the electrical resistance between the electrodes to decrease. The resistance of the sensor is related to the water content, and a value of water potential can be calculated using its predetermined relationship with water content. Electrical resistance sensors are the most inexpensive sensors available for measuring water potential, and this may be why they are also the most popular. They are easy to install and simple to automate for data-logging purposes; however, they are sensitive to changes in salinity and temperature. The largest pores in a gypsum matrix are still fairly small, so these sensors have very low sensitivity in wet soil.

Dielectric Matric-Potential Sensors

Dielectric matric potential sensors differ very little from the other solid-phase sensors discussed above. Instead of a heated needle or two electrodes, two or more wave guides are separated by a porous matrix. The wave guides measure the dielectric permittivity of the porous matrix. An empirical relationship between the matrix permittivity and matric potential is used to infer matric potential of the matrix, which is equal to the matric potential of the soil at equilibrium. The high accuracy of the dielectric sensor provides excellent resolution of water potential; in addition, careful selection of a matrix allows relatively quick equilibration in a wide range of soils. However, although dielectric sensors are not sensitive to temperature and salinity to the same degree as the electrical resistance sensor, they are not completely insensitive either.

Filter-Paper Method

The filter-paper method is a simple method for measuring water potential. A single disk of Whatman No. 42 filter paper is brought to equilibrium with a soil sample, which generally takes 2–3 days. The filter paper is then weighed and dried. The water content of the paper is used to calculate water potential using the relation:

$$\psi_m = -11w^{-3.68} \qquad [7]$$

where the relationship between the water content of the filter paper and its water potential in joules per kilogram is defined by the two constants, which are specific for the specified filter paper.

Tensiometer

The tensiometer equilibrates water inside a tube with the water in the soil matrix. The soil is separated from the water in the tube by a porous ceramic which is permeable to water and solutes, but not to soil colloids and air. When the water in the tube equilibrates with the soil water, the suction on

Table 1 Water-potential units

	$J\,kg^{-1}$	MPa	$m\ H_2O$	Relative humidity	Freezing point (°C)	pF	Pore diameter (μm)
	−1	−0.001	−0.1	0.999993	−0.0008	1.01	290.08
	−10	−0.01	−1.02	0.999926	−0.0076	2.01	29.01
FC	−33	−0.033	−3.37	0.999756	−0.0252	2.53	8.79
	−100	−0.1	−10.2	0.999261	−0.0764	3.01	2.9
	−1000	−1	−102.04	0.992638	−0.7635	4.01	0.29
PWP	−1500	−1.5	−153.06	0.988977	−1.1453	4.18	0.19
	−10 000	−10	−1020.41	0.928772	−7.6352	5.01	0.03
Air dry	−100 000	−100	−10 204	0.477632		6.01	
Oven dry	−1 000 000	−1000	−102 041	0.000618		7.01	

Relative humidity computed assuming a temperature of 20°C.
FC, a typical value for the drained upper limit of water potential in soil; PWP, a typical value for the lower limit of plant-available water in soil; air dry, typical value; varies with atmospheric humidity; oven dry, typical value; varies with oven humidity and oven temperature; pF, base 10 logarithm of the water potential in centimeters of water; pore diameter, bubble pressure or diameter of the largest water-filled pore at the indicated potential.

Figure 4 Dew-point potentiometer for measuring water potential of soil samples using vapor equilibration. The sample is placed in the cup, sealed inside the instrument, and the dew point and sample temperature are measured.

the water is equal to the soil matric potential. (*See* **Tensiometry**.)

Vapor-Pressure Methods

If soil water is allowed to come to temperature and vapor equilibrium with a vapor phase, the potential in the vapor phase becomes the same as that in the liquid phase. Since a vapor barrier is impermeable to solutes as well as soil colloids, the potential in the vapor phase is equal to the sum of the osmotic and matric potentials in the soil. The potential in the vapor phase can be determined by measuring its relative humidity. The Kelvin equation relates humidity and water potential:

$$\psi = \frac{RT}{M_w} \ln h \qquad [8]$$

where R is the gas constant, T is temperature (Kelvin) M_w is the molecular mass of water, and h is the relative humidity (a fraction, not a percentage). It is possible to measure h by measuring either the wet-bulb depression or the dew-point depression of the headspace air. For soil wet enough to sustain plant growth, the relative humidity is greater than approximately 0.99, so the psychrometer or dew-point meter must be extremely accurate and sensitive. A number of such devices are available which resolve changes in humidity of 0.0001 at these high humidities. Figure 4 shows a dew-point meter for water-potential measurement.

Water-Potential Units

Water potential has been expressed in a variety of ways and with a variety of units. As stated earlier, water potential is the potential energy per unit quantity of water, where the quantity is sometimes mass, sometimes volume, and sometimes weight. Potential energy per unit volume is dimensionally equivalent to pressure, so pressure units are often used; equivalent concentrations of ideal solute, relative humidity, and freezing point have also been used. Table 1 shows some of these units.

See also: **Hydrodynamics in Soils**; **Neutron Scattering**; **Spatial Patterns**; **Tensiometry**; **Thermal Properties and Processes**; **Time-Domain Reflectometry**; **Water Potential**

Further Reading

Campbell GS, Calissendorff C, and Williams J (1991) Probe for measuring soil specific heat using a heat pulse method. *Soil Science Society of America Journal* 55: 291–293.

Flint AL, Campbell GS, Ellett KM, and Calissendorff C (2002) Calibration and temperature correction of heat dissipation matric potential sensors. *Soil Science Society of America Journal* 66: 1439–1445.

Gardner WH (1986) Water content. In: Klute A (ed.) *Methods of Soil Analysis*, 2nd edn, part 1. *Physical and Mineralogical Methods*. Agronomy Monographs No. 9, pp. 493–544. Madison, WI: American Society of Agronomy.

WATER CYCLE

D K Cassel and B B Thapa, North Carolina State University, Raleigh, NC, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

The water cycle is an endless process of water circulation on the planet Earth. Water is an essential constituent of all plants and animals, most of which contain more than 60% water, and many contain more than 95%. The planet Earth is dominated by the hydrosphere, which contains all of the Earth's water. Approximately 1.39 billion km^3 of liquid water is stored in depressions, primarily oceans, which occupy over 70% of the Earth's surface. Water is unique in many ways: it has a high heat capacity, high heats of vaporization and fusion, and can exist in solid, liquid, and gaseous phases at temperatures commonly encountered on Earth. The ease with which water can move from one phase to another in response to additions or losses of heat or energy allows water to move or cycle from one storage field to another.

Overview

Approximately 97% of the Earth's water is salty and is stored in oceans (Table 1). The remaining 3% is fresh water and, if uncontaminated, is considered to be potable and drinkable by animals and utilized by nonsalt-tolerant plants. Fresh water is stored primarily as ice in glaciers and icecaps, but also in groundwater, lakes, streams, and rivers. Water retained as soil moisture is a small fraction of the total, but is extremely important for the production of food and fiber. The amount of water existing in the gaseous state is relatively constant and is estimated to be 12 900 km^3.

Table 1 Estimate of global water distribution

Source	*Volume × 1000 km³*	*Total water (%)*	*Fresh water (%)*
Oceans	1 338 000	96.5	–
Ice, glaciers, snow	24 064	1.74	68.7
Groundwater			
Fresh	10 530	0.76	30.1
Saline	12 870	0.94	–
Soil moisture	16.5	0.001	0.05
Ground ice and permafrost	300	0.022	0.86
Lakes			
Fresh	91.0	0.007	0.26
Saline	85.4	0.006	–
Atmosphere	12.9	0.001	0.04
Swamps	11.47	0.0008	0.03
Rivers	2.12	0.0002	0.006
Biological water	1.12	0.0001	0.003
Total fresh and saline	1 385 984	100.0	100.0

Reproduced with permission from Shilkomanov IA (1993) World fresh water resources. In: Gleick PH (ed.) *Water in Crisis: A Guide to the World's Fresh Water Resources*. New York: Oxford University Press.

The water cycle, commonly called the 'hydrologic cycle,' is a conceptual model that relates to the transport of water to and from the various storage fields or pools on the Earth. Movement of water from one storage field to another depends on the energy sources available for the processes of evaporation, condensation, precipitation or deposition, runoff, infiltration, melting, and groundwater flow. It is estimated that 380 000 km^3 of water are recycled each year. Human recognition of this water-recycling process has led to the utilization of renewable (recycling) flowing water to generate hydroelectricity in many parts of the world. A given water molecule could follow an infinite number of pathways as it progresses through the various processes in the water cycle.

At the global scale (macro scale), the water-cycle model appears to be a simple process of evaporation and precipitation and is typically viewed as transport of water from oceans and lakes to land masses and back to oceans and lakes. The water, once on land, is eventually returned to the oceans in runoff water or as precipitation fed by water evaporated from the soil or transpired by plants. When considered at the regional scale, human intervention in the water cycle often exists, and the water-cycle model becomes more complicated. For example, in arid regions, in addition to precipitation, water previously stored in the groundwater might be 'mined' for domestic purposes, or water falling in another region might be transported to the arid region for domestic or commercial uses or to satisfy the transpiration demand to grow crops. Conversely, water in humid regions might be impounded in dams and transported by canal or pipeline to another region or watersheds.

The water cycle becomes even more complex at the local or field scale, and we often view it as the water balance. The individual mechanisms of water transport become more important at smaller scales. For example, at the local scale, a high percentage of the water might be transferred to the land by irrigation canals rather than by precipitation. Likewise, lateral subsoil transport of water from sloping fields might be the major mode of transport rather than runoff. In

many agricultural fields, a conscious effort is made to reduce or at least control runoff to allow more water to infiltrate and to control soil erosion. The soil moisture component of the water cycle may appear to be insignificant at the global scale (Table 1), but at the field scale it is elevated to primary importance in growing plants.

The quality of waters transferred by runoff, groundwater flow, and precipitation is an important consideration. Pollutant-laden runoff has significant negative impacts on aquatic habitats and decreases the usefulness of water for further use. As the population increases, not only is more water needed, but human activities tend to impair water quality as it continues to travel in the water cycle. Examples of water-quality impairment are the contamination of groundwater by organic and inorganic chemicals, and the fouling of lakes, streams, and oceans by rubbish and human and animal wastes. Acid rain due to sulfur, a by-product of some industrial operations, has damaged plant and animal habitats. These impacts on water quality can lead to health hazards and drastic environmental and economic impacts at the regional and local scales. Fortunately, the evaporation process in the water cycle purifies polluted water, allowing it eventually to reappear as fresh precipitation to be reused to support plant and animal life.

Water Cycle Processes

The water cycle is a conceptual model to describe the transfer and temporary storage of water among various storage fields or reservoirs. These reservoirs include lakes, streams, oceans, groundwater, soil moisture, glaciers, snowfields, the atmosphere, and the biosphere. Energy balance and the water cycle are closely linked. Water is transferred from one storage field to another by processes of evaporation, transpiration, sublimation, condensation, precipitation, infiltration, runoff, and groundwater flow (Figure 1). An estimate of the approximate residence time in various water reservoirs is given in Table 2.

Evaporation converts liquid water to the gaseous phase. Evaporation is a distillation process that leaves all impurities and contaminants behind in the soil or water body from which it evaporates. Evaporation of water requires a large quantity of energy, most of which comes from the Sun. The heat required to vaporize water is $540\,\mathrm{MJ\,kg^{-1}}$ at 373 K (101 kPa) and increases to $2.26\,\mathrm{MJ\,kg^{-1}}$ at 273 K. Evaporation occurs when the vapor pressure of the atmosphere is less than the vapor pressure of liquid water at the evaporating surface. The evaporation rate is controlled by the energy available, temperature, humidity, wind speed, and salinity of the water. On a global

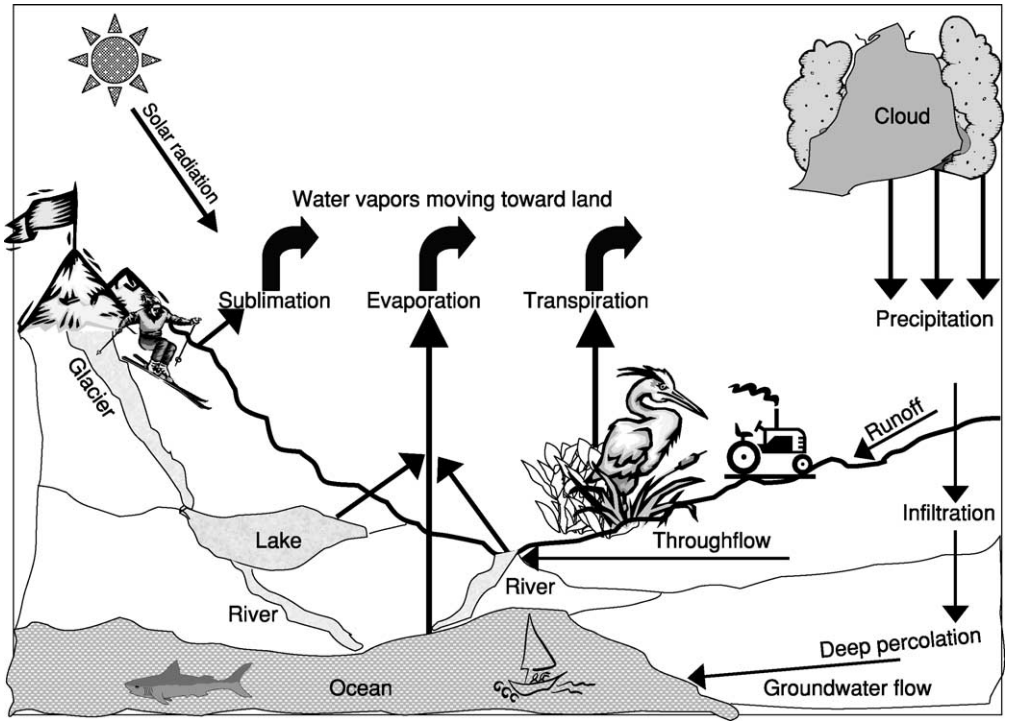


Figure 1 The water cycle.

Table 2 Approximate residence time of water in various reservoirs

Reservoir	*Residence time (years)*
Glaciers	40
Seasonal snow	0.4
Soil moisture	0.2
Groundwater, shallow	200
Groundwater, deep	10 000
Lakes	100
Rivers	0.04

Adapted from Wetzel RG (1983) *Limnology*, 2nd edn. New York: Saunders College Publishing.

scale, evaporation of water is most intense in the subtropical oceans, where solar radiation is relatively constant throughout the year and provides a constant energy source to vaporize water.

Sublimation is conversion of solid-phase water (ice) directly to the vapor phase without passing through the liquid phase. This process requires approximately 2.85 MJ kg^{-1}.

Transpiration is the conversion of water from liquid to gas as it passes through plant stomata, small openings on the undersides of leaves of vascular plants. The stomata of some plants have the ability to open and close. Transpiration is a passive process, being controlled by several factors. When soil water content is high (absolute value of soil water pressure is small), the transpiration rate of plants with extensive root systems tends to be controlled by the solar energy available. Wind tends to increase transpiration rate. In general, 99% of the water entering plant roots passes through the plant and is lost to transpiration.

The processes of evaporation and transpiration on a given area of land often occur simultaneously, and it is difficult to differentiate between the evaporative loss of water from the land surface and the transpiration loss by plants. In such cases the combined loss of water by these two processes is referred to as evapotranspiration. Generally, potential evapotranspiration exceeds precipitation from landmasses during the summer at middle and high latitudes.

Condensation is the process whereby water is converted from the gaseous phase to the liquid phase. This process releases a vast amount of heat. Condensation occurs when the atmospheric temperature cools below the point at which the relative humidity approaches 100%. When this condition is reached, condensation nuclei such as dust, smoke particles, or salts provide sites for water vapor to condense to form water droplets. These condensed water droplets form clouds or fog. The water droplets in the clouds continue to increase in size when the right conditions are encountered and eventually coalesce into ice crystals or a liquid. Condensation nuclei such as salts, e.g., NaCl, are hygroscopic, thus allowing water to condense at a relative humidity of less than 100%. As clouds form, air currents move them, thus dispersing the water vapor around the Earth. When clouds eventually encounter environmental conditions where they can no longer hold the moisture, they release some or part of it as precipitation. Another result of the condensation process is the formation of dew during summer and autumn nights, when the air temperature cools to the point where the relative humidity at the plant or land surfaces rises to 100%.

Precipitation is any aqueous deposit, whether in liquid or solid form, that develops in a water-vapor-saturated atmosphere and falls to land or water surfaces. In general, precipitation over continents tends to exceed evaporation, whereas over oceans, evaporation tends to exceed precipitation. Most precipitation forms from natural clouds, but most clouds exist without yielding precipitation. On the other hand, small precipitation events derived from water vapor emitted from smoke stacks are not uncommon. Precipitable water is the amount of water potentially available in the atmosphere for precipitation, usually measured in a vertical column that extends from the Earth's surface to the upper edge of the troposphere. In many clouds, water droplets and ice crystals are too small to overcome natural updrafts found in the atmosphere. Hailstones are formed by the continual thickening of ice crystals until they become heavy enough to overcome the updraft.

The distribution of rainfall reaching the soil surface can be modified by vegetation covering all or part of the soil surface. Deforestation in some regions of the world has changed the rainfall pattern. The amount of rainfall intercepted by vegetative cover depends on leaf density and the branching structure of the plants. Some of the water is evaporated directly from the vegetative surfaces, while some of it might drip from leaves or move by 'stemflow' down the leaves and stems to the soil surface.

The intensity, duration, and amount of rainfall events control to some degree the pathway water takes once it arrives at the soil surface. Rainfall intensity varies during a storm event. If the intensity is high, or if the duration is long, the likelihood is great that some of the water reaching the soil surface will run off. The energy associated with high-intensity rainfall increases the opportunity for soil erosion, a process that degrades productive soil as well as the quality of the runoff water.

Irrigation is the application of water to the soil, usually for the purpose of growing food or fiber. This very important process in the water cycle is human-imposed, but has tremendous impact on other processes in the water cycle. Approximately

40% of the world's food comes from irrigated land. One common source of irrigation is water diverted from large, constructed reservoirs. These reservoirs are often hundreds of kilometers away from the land being irrigated. Other sources of irrigation water are ponds, streams and rivers, and groundwater. Irrigation accounts for over 65% of the fresh water use, but, because of inefficiencies at various points in the irrigation process, less than half of that water reaches the plant roots. The remaining half is lost to evaporation during transport of water to the land to be irrigated, deep percolation, and runoff. Runoff water leaving the lower boundary or an irrigated field (irrigation return flow) often has a high salt content, thus reducing water quality of streams into which it drains. Irrigation water is applied to the soil by flooding, spraying, sprinkling, furrowing, and trickle methods. Groundwater resources continue to be depleted in some regions of the world due to pumping more water than is being recharged.

Infiltration is the process of water entry into the soil, generally by downward flow through the soil surface. Liquid water falling on the soil surface is temporarily stored on the surface, runs off, or infiltrates. The infiltrability, that is, the rate at which water enters the soil when water at atmospheric pressure is available at the soil surface, decreases with time. The infiltration rate of unsaturated soils is controlled by a number of factors, including the condition and cover of the soil surface, soil texture, pore-size distribution and porosity, soil water pressure or soil suction, soil structure, the rate at which water arrives at the soil surface, antecedent soil water content, and the permeability of underlying soil horizons. If a soil is saturated, water will not infiltrate because there is no pore space for it to enter. Coarse-textured soils usually have high infiltration rates unless they have high antecedent water contents. Fine-textured soils tend to have low infiltration rates, increasing the chances for water to run off. For some soils, water infiltrates along preferred pathways due to the presence of cracks and fissures, soil 'pipes,' and large pores created by animal or plant roots.

Runoff is excess water that does not infiltrate into the soil. When liquid water as precipitation or irrigation arrives at the soil surface at a rate that exceeds the soil's infiltration rate, the excess water is temporarily stored in small depressions on the soil surface. Once these depressions on sloping land are filled, additional water arriving at the soil surface that does not infiltrate is transported across the soil surface as runoff. Gravity provides the energy to transport runoff, or 'overland flow,' from higher to lower elevations. Overland flow occurs as sheet flow on the soil surface. Runoff water concentrates in rivulets or rills, waterways, and gullies, eventually emptying into streams and rivers, and finally into lakes and oceans. This runoff water provides energy that can transport detached soil materials in the processes of sheet, rill, and gully erosion. Flooding can occur if too much water runs off in a short time period. In some arid regions, land is set aside and managed to increase the amount of runoff water that is usually collected in reservoirs at lower elevations for various consumptive uses.

Drainage When soils become saturated or nearly saturated, the possibility that some of the water will continue to drain or percolate deeper into the soil or into the underlying water table increases. This process recharges ground water supplies and also provides 'base flow' water for streams and rivers. When irrigation is practiced, it is imperative that some water drains below the root zone to remove salts that are added to the soil in the irrigation water. Open ditches or buried drains carry the drainage water to a discharge outlet, where it is returned to a river or stream. Mobile chemicals present in the soil move in the percolating and drainage waters. Transport of chemicals has led to the contamination of groundwater, streams, rivers, and reservoirs throughout the world and is a continuing problem.

Environmental and Economic Implications of Water-Cycle Processes

Tremendous environmental impacts are associated with the processes in the water cycle. At the global scale, the geography of the entire world along with its many beautiful geologic features has been shaped primarily by the various components of the water cycle, without human intervention, acting over extended periods of time. For millions of years the water-cycle processes, driven and interacting with the energy-distribution patterns on the Earth's surface, have shaped Earth's surface. Geologic erosion, deposition of eroded and transported solids, the formation of rivers, lakes, and streams by runoff water, and the formation of wetlands in low areas are results of water-cycle processes. All of these processes result from a combination of ordinary and episodic events associated with the water cycle.

At the regional scale, one can only imagine the flooding that occurred as huge amounts of runoff water loosened and scoured minerals from higher elevations, transporting them to lower elevations to settle and eventually develop into sedimentary and metamorphic rocks. Phosphates, carbonates, and placer gold are but a few of the numerous minerals and chemicals that are mined that have been concentrated in certain regions throughout the world by

water-cycle processes. Similarly, very fertile alluvial flood plains, such as the Indo-Gangetic Plain in Asia, have developed.

Glaciation, a process that occurred in some regions of the Earth in the past and presently is ongoing in some regions, transports vast amounts of ice that scour and shape the surface features of the Earth. As glaciers melt, rivers, lakes, prairie potholes (closed depressions), wetlands, and numerous other features are formed. At the local scale, for example, the magnitudes of the individual processes of the water cycle for a prairie pothole in eastern North Dakota differ from those for land just a few hundred meters away. Surface runoff from land in the small, depressional watershed surrounding the pothole contributes to the water stored in the pothole, where it will be evaporated or percolate downward to the groundwater. On the other hand, runoff water from the nearby land outside the depressional watershed contributes its runoff to streams. This seemingly small difference creates great diversity in ecology within a distance of 100 m or less.

Extreme variation in the magnitude of one or more components of the water cycle, for example, precipitation, causes catastrophic events such as flooding or drought, although these two terms are defined in terms of norms as perceived by humans today. Yet flooding as a catastrophic event has occurred for billions of years and has been a key process in the formation of sedimentary rocks.

On a shorter time scale, vagaries in various processes of the water cycle have caused disasters for humans. In fact, humans have altered the water cycle at various scales for millennia, and currently the process of global warming, which is still a highly debated topic, is intimately linked with the water cycle. Impacts of global warming that have been predicted using models are shifts in temperature, rainfall distribution, and food production. These shifts are predicted to have greater impacts in some regions of the world than in others. Acid precipitation is another man-induced problem that affects the quality of water and is changing the ecology of large regions. Nitrogen and sulfur oxides released into the air as a result of industrial processes and fuel combustion are further oxidized and converted into nitric and sulfuric acids. These acids are washed from the air by precipitation. Acid precipitation refers to wet forms of acid pollution contained in rain, sleet, snow, fog, and vapor transport. Still another problem affecting the quality of water is the input of chemicals and heat that is discharged into stream waters. Excess nitrogen and phosphorus can lead to excessive algal growth. Siltation of lakes and rivers and accompanying fish kills are common occurrences.

Water in lakes and streams has a large impact on quality of life and the economy. Water for transportation led to the development of ancient and modern civilization. Water for industrial and domestic use is diverted from lakes, streams, and groundwater supplies. After the liquid water is used, much of it continues its travel in the water cycle as 'wastewater' discharged into streams and lakes. This water is reused further downstream for additional domestic and industrial applications. In some locations energy in the form of hydroelectricity is 'harvested' from the flowing water. Without this source of energy, many areas of the world could not have been developed. The use and reuse of water continue until the water eventually evaporates or reaches the ocean and then evaporates. The development of cities, industries, agriculture, and recreational parks are dependent on water as it passes along the various stages in the water cycle.

Land-Management Impacts

Human interventions in the 'natural' water cycle at many scales have occurred as a result of implemented land- and water-management (or mismanagement) practices. Several examples of the effects of management practices involving the water cycle at different scales, either directly or indirectly, follow. Intervention of the water cycle at the regional scale began in ancient times with the construction of irrigation canals that diverted water from one region to another, thus increasing the amounts of water applied to irrigated areas. The anticipated response of increasing food production by applying greater amounts of water to the lands was obtained. However, an unanticipated regional-scale response was that, in time, the soils became saline and were eventually abandoned for crop production. Failure to provide soil drainage led to the gradual increase in salt content of irrigated soils. Eventually, even salt-tolerant crops failed to grow. This undesirable salinization process was a direct effect of inadequate land and water management. Unfortunately, today we are still losing arable land due to salinization caused by inadequate drainage of irrigation water.

Another type of salinization problem, caused by modification of the water cycle as a result of changes in land management at the field or landscape scale, arose in the subarid Northern Great Plains (USA) due to the adoption of 'summer fallow.' Dryland small grain production was, and still is, common in this area even though precipitation in a typical year is often insufficient to produce a grain crop. The practice of summer fallow was developed to accumulate water in the soil profile, thereby using the precipitation occurring over a 2-year period to produce one

grain crop every 2 years. Typically, the soil was kept bare the year before planting the grain crop (year 1) so that water falling as precipitation during that year would infiltrate, thus increasing the amount of stored soil moisture. The grain crop planted in year 2 would derive benefit from the stored moisture in year 1 in addition to precipitation occurring in year 2. Use of this practice over decades gave rise to 'saline seeps.' In essence, summer fallow was inefficient in conserving water for the following crop. Much of the water evaporated from the soil surface in year 1. However, some of the water in higher-rainfall years percolated through the soil. This percolating water carried salts leached from underlying strata having high salt contents. When the drainage water encountered an aquatard, it moved obliquely downslope and eventually seeped from the hillside, causing excessively wet, salty nonarable soils at lower elevations.

As the human population increases, more and more land is being altered and subjected to different land uses. The land use changes have impacts on the processes of the water cycle, the scale of impact being dictated by the extent and nature of land management. Compaction of soils by farm and construction machinery, as well as human and animal footprints, decreases infiltration with a resultant increase in runoff. The eroded soil material from farmland and construction sites may include adsorbed hazardous chemicals. The construction of sedimentation basins to collect runoff water to allow solids removed in runoff water to settle out is a management practice that has arisen to help control the environmental degradation.

The construction of dams and ponds occurs at many scales and, regardless of scale, they cause some alteration of the magnitudes of the processes in the water cycle. The water that fills ponds and lakes impounded by dams arises from runoff waters, groundwater, seeps, and springs. The impounded water in farm ponds is often used for irrigation. At a regional scale, for example, some of the large dams constructed in California impound water that is transported hundreds of miles and used for domestic consumption, industry, and irrigation. In some regions of the world, so much water is diverted for irrigation and other human uses that rivers run dry for parts of the year. Rivers affected in this manner include the Yellow River in China, the Ganges in South Asia, and the Colorado in the United States. In addition, in many regions the groundwater supplies are being mined for irrigation and other uses at rates exceeding their recharge rates, resulting in dropping water tables.

Urbanization has drastically altered the water cycle at various scales. Precipitation falling on roofs of houses and buildings, on streets and highways, and paved parking lots near industries, schools, and shopping centers does not infiltrate the soil, but increases the runoff component of the water cycle. This stormwater runoff carries suspended solids and numerous chemicals, such as spilled petroleum products, and nitrogen and phosphorus fertilizers that degrade water quality and cause eutrophication. In addition, it is common for the increase in amount of runoff water to initiate or exacerbate downstream flooding. Following Hurricane Floyd in 1999, the state of North Carolina began a program to develop new flood plain maps for five river basins which incorporates changes in flood plain elevation caused by two decades of development.

Summary

Human life cannot exist without water: life originated in water, and civilization began near water; water is the shelter for many marine creatures; water fulfills human recreational needs. The importance of water in the world today cannot be overstated. In addition to being the major constituent of human life forms, water is a required input to produce food and clothing, manufacture products, transport people and products around the world, and produce a significant percentage of the world's energy.

Water molecules are in continuous motion and keep moving from one storage field to another. Liquid water in rivers moves downslope to the ocean, where solar radiation converts it to the gaseous phase, which in turn transports the water vapor back to the landmasses, where it condenses to form precipitation. The precipitation is divided into water that infiltrates the soil and runoff water that carries with it suspended solids, chemicals, pesticides, and organic and inorganic wastes. This endless cycle is called the water cycle or hydrologic cycle.

The total amount of water on Earth remains constant, but it is separated into different processes or pathways depending on environmental factors and human interventions. Thus processes such as evaporation, condensation, precipitation, infiltration, and runoff and subsurface water movement induce the transfer of the three phases of water. The evaporation processes can be viewed as a distillation process that purifies water, leaving impurities in the ocean, whereas runoff and percolation are water-polluting processes of the water cycle. The runoff process affects the quality of surface waters. Fish kills as a result of algal growth and eutrophication are primarily due to runoff generated in response to inappropriate land management, deforestation, and misuse of heavy machinery, chemical fertilizers, and pesticides.

Percolated water can leach undesirable chemical compounds below the rooting zone of plants and into drainage water or through the vadose zone into the groundwater. Even though the ocean serves as a huge sink for salts and pollutants, the contamination of streams, rivers, lakes, and groundwater resources poses a great threat to plant and animal health, the environment, the economy, and civilization.

Further Reading

Berner EK and Berner RA (1996) *Global Environment: Water, Air, and Geochemical Cycles*. Upper Saddle River, NJ: Prentice-Hall.

Browning KA and Gurney RJ (1999) *Global Energy and Water Cycles*. Cambridge, UK: Cambridge University Press.

Carpenter SR, Fisher SG, Grimm NB, and Kitchell JF (1992) Global change and freshwater ecosystems. *Annual Review of Ecology and Systematics* 23: 119–139.

Gleick PH (1993) An introduction to global fresh water issues. In: Gleick PH (ed.) *Water in Crisis: A Guide to the World's Fresh Water Resources*, pp. 3–12. New York: Oxford University Press.

Guymon GL (1994) *Unsaturated Zone Hydrology*. Englewood Cliffs, NJ: Prentice-Hall.

Hillel D (1998) *Environmental Soil Physics*. San Diego, CA: Academic Press.

Howells G (1995) *Acid Rain and Acid Waters*, 2nd edn. New York: Ellis Harwood.

Postel S (1999) *Pillar of Sand: Can the Miracle Last?* Saddle Brook, NJ: WW Norton.

Vorosmarty CJ, Green P, Salisbury J, and Lammers RB (2000) Global water resources: vulnerability from climate change and population growth. *Science* 289: 284–288.

Wetzel RG (1983) *Limnology*, 2nd edn. New York: Saunders College.

Water Erosion *See* **Erosion:** Water-Induced

WATER HARVESTING

D Hillel, Columbia University, New York, NY, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Definition

The term 'water harvesting' generally refers to the collection of rainstorm-generated runoff from a particular area (a catchment) in order to provide water for human, animal, or crop use. The water thus collected can either be utilized immediately, as for irrigation, or be stored in aboveground ponds or in subsurface reservoirs, such as cisterns or shallow aquifers, for subsequent utilization. As such, water harvesting is an ancient practice that has enabled some societies to subsist in semiarid and arid areas where other sources of fresh water (e.g., rivers, lakes, or aquifers) are scant or unavailable.

Surface Runoff

Whenever the rate at which rainwater is applied to the soil surface exceeds the rate of infiltration into the soil, the excess tends to accumulate over the surface. Where the surface is not perfectly flat and smooth, the excess water collects in depressions, forming puddles. The total volume of water thus held, per unit area, is called 'surface-storage capacity.' It depends on the geometric irregularities (roughness) of the surface as well as on the overall slope of the land (Figure 1).

Only when the surface storage is filled and the puddles begin to overflow can actual runoff begin. The term 'surface runoff' thus represents the portion of the water supply to the surface that is neither absorbed by the soil nor accumulates on its surface, but that runs downslope.

Surface runoff typically begins as sheet flow but, as it accelerates and gains in erosive power, it eventually scours the soil surface to create channels. There exists

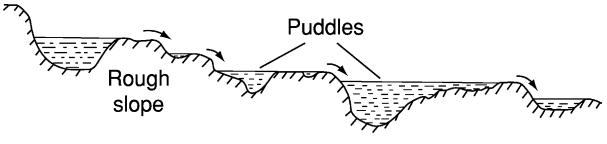


Figure 1 Effect of surface roughness and slope on surface storage of rainfall excess. Reprinted from *Environmental Soil Physics*, Hillel D (ed.). Copyright (1998), with permission from Elsevier.

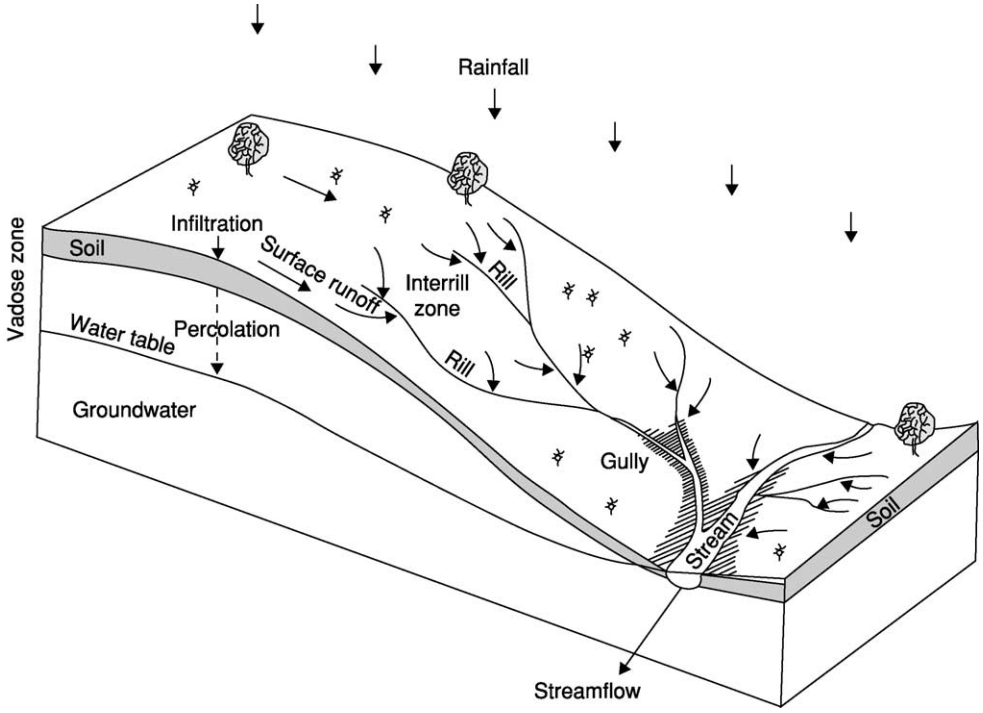


Figure 2 A sloping area exhibiting overland flow as well as runoff through rills and gullies. Reprinted from *Environmental Soil Physics*, Hillel D (ed.). Copyright (1998), with permission from Elsevier.

a wide spectrum of channel geometries and flow patterns. On the one extreme is the thin, sheet-like runoff called 'overland flow.' It is likely to be the primary form of surface runoff from small areas or fields having little topographic relief. The next distinctive form of flow takes place in small channels called rills. The latter gather the overland flow in a continuous fashion along their length to form the lowest order of stream flow. As these small streamlets merge with one another, they form higher-order channels, called gullies, which collect concentrated tributaries as well as distributed (lateral) overland flows (Figure 2).

Runoff Control and Utilization

Uncontrolled runoff is never desirable, as it is likely to cause soil erosion on slopes as well as flooding and silting in bottomlands. In humid regions, where rainfall may be excessive, measures may be needed to ensure the safe routing and conveyance of the runoff. Such measures, called 'surface drainage,' include shaping the land and treating it so as to direct the runoff via protected (grassed or even concrete-lined) channels. In semiarid regions, by way of contrast, natural rainfall is barely sufficient for crops, hence farmers typically strive – by such means as terracing and mulching – to cause as much of the rainfall as possible to infiltrate the soil, and thereby to minimize runoff.

The situation is fundamentally different in arid (as distinct from semiarid) regions. In many arid regions, large tracts of land are basically unsuitable for conventional rainfed farming, owing to the paucity and instability of rainfall, the nature of the soil (too shallow, stony, or saline to permit cultivation), or the rough topography. From the point of view of farmers (though not of ecologists, who are concerned about an area's natural biota rather than about crop production per se), rain falling on such lands is almost totally lost – being insufficient either to recharge groundwater or to support an economically viable crop.

Most of the meager rainwater generally infiltrates the soil to a shallow depth only, and it is quickly returned to the atmosphere – either by direct evaporation from the soil or by transpiration from native vegetation. Occasional intense rainstorms may none the less generate runoff and cause sudden flash floods. Although runoff from any particular rainstorm may be high, under natural conditions the total seasonal runoff seldom exceeds 10% of the annual precipitation.

The possibility of controlling and even increasing the amount of surface runoff obtainable from such lands can be of great importance. Particularly where no other dependable water source is available, the runoff thus obtained may constitute the major supply of an inhabited area.

Ancient Methods

The art of inducing, collecting, and utilizing runoff has been practiced by desert-dwelling communities since antiquity, in such disparate regions as southwestern Asia, northern Africa, and southwestern North America. Remnants of extensive water-harvesting systems are found in these regions.

Highly noteworthy are the works of the ancient Nabateans, who inhabited the Negev Desert of southern Israel some 2000 years ago. They began as caravan drivers, conveying aromatic, spicy, and medicinal plant-products, as well as other precious objects of trade, from distant sources across the desert to the major population centers along the shores of the Mediterranean Sea. The Nabateans established stations along their trading routes, and these grew in time into permanent settlements and even into cities. To sustain their population in the desert environment, the Nabateans built extensive terraces and channels, and hewed out numerous cisterns. Remains of their cities (most notable among them being the city of Petra, in southern Jordan) can be visited today.

The first imperative of desert settlement was the provision of potable water for humans and livestock. This was done by means of cisterns, which are artificially constructed reservoirs filled with surface flows during infrequent rains. The early cisterns were undoubtedly leaky and inefficient. Building efficient cisterns became possible only with the advent of watertight plaster, made of burned and slaked lime. Also crucial was the ability to recognize suitable rock formations, such as soft marly chalk, which could be hewed out readily and was not as fissured and leaky as the hard limestone that is also prevalent in the same region. The Nabateans were also skilled at choosing appropriate sites for their cisterns and at ensuring that they could be filled with water annually.

Where cisterns could be located alongside natural streams, they were filled directly by flash floods. However, most cisterns in the Negev were built along the lower reaches of hillsides and depended on the direct collection of runoff from the higher slopes. Many hundreds of such cisterns were built in the Negev, and they are clearly discernible landmarks even today. A typical one resembles a giant necklace, with the glistening white pile of excavated rock appearing to hang like a pendant from the two collection channels that ring the hill and curve down its sides from opposite directions (Figure 3). To parched travelers through the desert, to whom these cisterns would beckon from afar, no sight could be more gladdening.

The application of water-harvesting techniques to enable farming in desert areas (that is to say, collecting runoff from sloping areas and using it to irrigate plots or tracts of flat land, where crops could be grown) has been called 'runoff farming.' In ancient times, stone dikes were constructed across tributary streambeds, thus forming a series of terraced fields. These fields were watered by flows arriving from the upper watershed of the stream, as well as from the adjacent hillslopes. The slopes were often divided into sections, the runoff from each of which was led by means of constructed channels to specific fields (Figure 4). The water retained and infiltrated in each field was then utilized by annual crops (e.g., wheat or barley) or by perennial crops (e.g., grapevines, olive trees, or other fruit trees).

Water Spreading

Another form of water harvesting, called 'water spreading,' consists of diverting flash floods from intermittent streams (known as *wadis* in the Middle East, *arroyos* in the American Southwest, and *dongas* in parts of Africa) on to adjacent tracts of land. It is a simple form of flood irrigation, controlled by dikes, check dams, or channels designed to direct and spread the expected flow. It is generally used to irrigate pastures and rangelands, but in places also to sustain groves of trees in arid areas.

Water spreading is typically practiced on small watersheds of a few square kilometers. In larger watersheds, the floods may occasionally be too violent or torrential to be controlled by simple diversion structures. In such cases, a more complex system may be required: the construction of dams designed not to retain the floods but merely to detain and regulate them, so as to provide farm units located downstream from the dam with controlled flows. Such dams, called 'detention dams,' are built across a stream in order temporarily to impound the flood. A large-diameter open pipe is laid through the dam to permit downstream flow at a predetermined rate, as through the drain of a bathtub. Thus, a flash flood that would normally last just a few hours is made to flow through the pipe and on to a series of fields for perhaps several days. The field dikes (made of compacted earth or stone) can then be built economically and safely to withstand floods of a known maximal intensity, so farming operations can be planned accordingly (Figure 5).

Runoff Inducement

The builders of runoff utilization systems in the Negev in ancient times had to contend with the paucity of natural runoff, and there is evidence that they actually strove to augment it. They developed techniques for inducing a greater portion of the rainfall to trickle downslope as runoff.

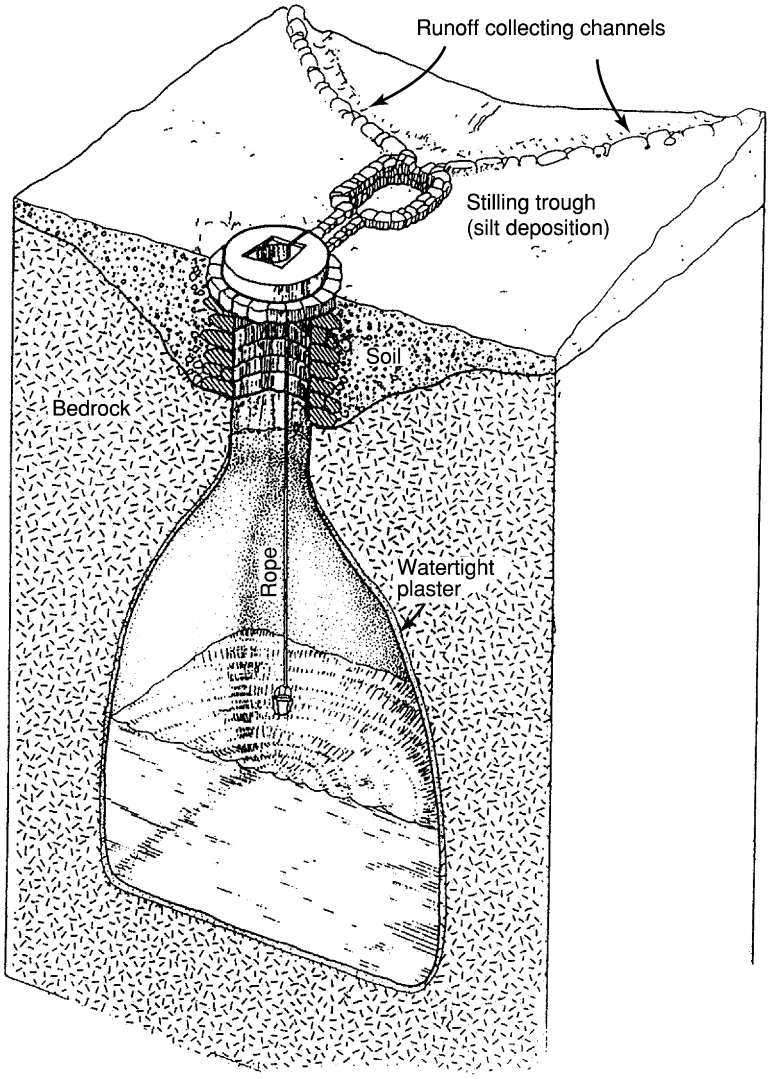


Figure 3 A typical hillside cistern excavated in the bedrock and filled by runoff from a sloping catchment. Reproduced from Hillel D (1982) *Negev: Land, Water, and Life in a Desert Environment.* New York: Praeger Publishers. © D Hillel.

Noticing that the soil had a natural tendency to crust, which was obstructed, however, by the desert's natural cover of loose gravel (commonly known as 'desert pavement'), the ancient inhabitants of the Negev region in southern Israel raked the stones off the surface in order to expose the finer soil material and to induce the formation of a surface seal. Even so, the ancient runoff farmers needed a water-contributing area approximately 20 times larger than the area to which water was directed for crop production. Similar techniques were used to collect runoff in cisterns for subsequent use as drinking water for humans and domestic animals (goats, sheep, and camels).

The importance of runoff inducement is greater than the mere increase in runoff yield that it may produce. The practice can also lower the 'runoff threshold' of a rainstorm, i.e., the minimal rainstorm (in terms of intensity, duration, and total amount of rain) needed to initiate runoff. This decrease in the threshold correspondingly increases the probability of obtaining runoff a sufficient number of times during the season to provide for the needs of domestic human use, as well as for agricultural or industrial purposes. This is especially important in view of the fact that most of the storms in an arid region result in light rains only, and that only a few storms (typically no more than two or three per season) are of sufficient intensity and quantity to yield runoff under natural conditions.

Still another climatic feature of arid regions that adds to the importance of runoff inducement is the interannual variability of rainfall. Most years provide less than the average (or mean) rainfall, and only a few anomalously rainy years skew the mean. (In statistical terms, mode and median rainfall tend to be less than the long-term mean rainfall.) In such regions, droughts are relatively frequent and may be very severe. In some years, there may be practically no

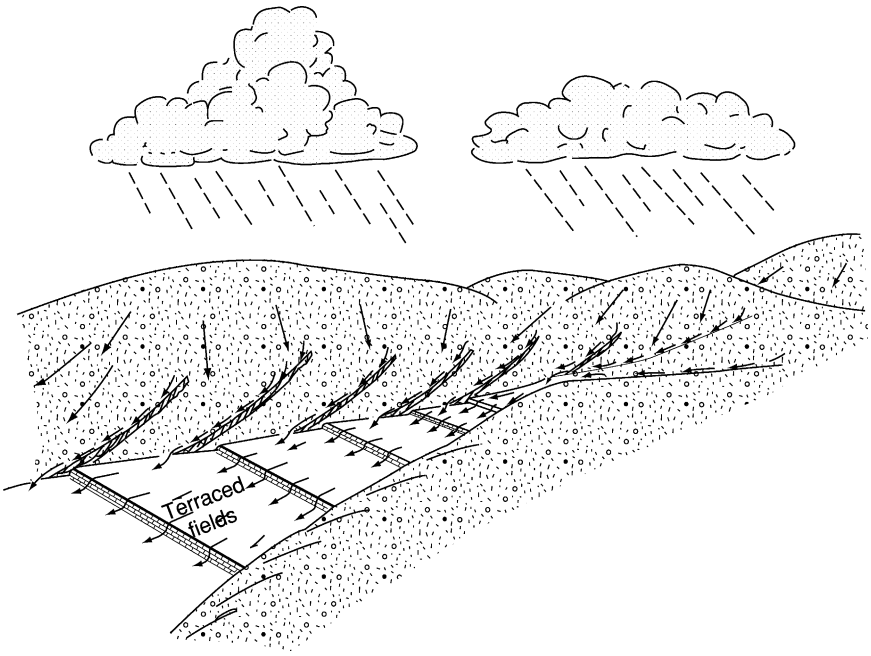


Figure 4 Ancient runoff farming in the Negev: water trickling off barren slopes was directed to terraced fields in the *wadi* bed. Reprinted from *Environmental Soil Physics*, Hillel D (ed.). Copyright (1998), with permission from Elsevier.

intense rainstorms and therefore no substantial natural runoff. Human enterprises dependent on the collection and utilization of runoff must therefore ensure, first of all, that there be a minimally sufficient water supply even in years of drought. Methods of runoff inducement can greatly contribute to this end.

Modern Methods

Modern technology holds the promise of more efficient runoff inducement than was possible in ancient times. The simplest, though generally most expensive, way to induce runoff is to cover the surface with an impervious apron of such materials as plastic, rubber, or aluminum sheeting, or by asphalt or concrete paving. A possibly more economical approach is to cause the soil itself to shed, rather than absorb, the rain. Runoff from natural surfaces can be increased several-fold by means of mechanical treatments (stone clearing, smoothing, and compaction), as well as by a variety of chemical treatments to encrust and stabilize the surface so as to prevent erosion of the runoff-contributing areas. Accordingly, the soil surface can be made water-repellent and relatively impermeable by treating it with sprayable clay-dispersants, sealers, and hydrophobic agents. The following series of treatments can be applied:

1. Eradication of vegetation and removal of surface stones, to reduce interception of rain and obstruction of overland flow, and to permit the formation of a continuous surface crust;
2. Smoothing of land surface, to obliterate surface depressions and prevent the retention of water in puddles;
3. Compaction of the soil top-layer to reduce its permeability. This can be done by means of a roller at optimal soil moisture content;
4. Dispersion of soil colloids to cause crusting, by means of sprayable solutions of sodic salts. This treatment pertains to soils that contain sufficient clay to be dispersible, but not so much as to exhibit marked shrinkage and cracking;
5. Impregnation of the surface with a sealing and binding substance such as an emulsion of asphalt that can form a water-repellent and stable coating.

With such methods, it is possible not only to increase the total yield per unit area of the watershed, but also to decrease the threshold of rain needed to form runoff and thus to increase the frequency of runoff supply and contribute to the efficiency and economic feasibility of agricultural and engineering systems designed to utilize runoff. In a desert with a seasonal rainfall of 250 mm, for instance, yields as high as 200 000 m^3 of water may be obtainable per square kilometer of treated area per season.

Several systems have been tried with respect to the size and arrangement of the contributing area in relation to water-receiving areas or reservoirs. A small watershed may be treated in its entirety so as to provide the maximal amount of water at the outflow

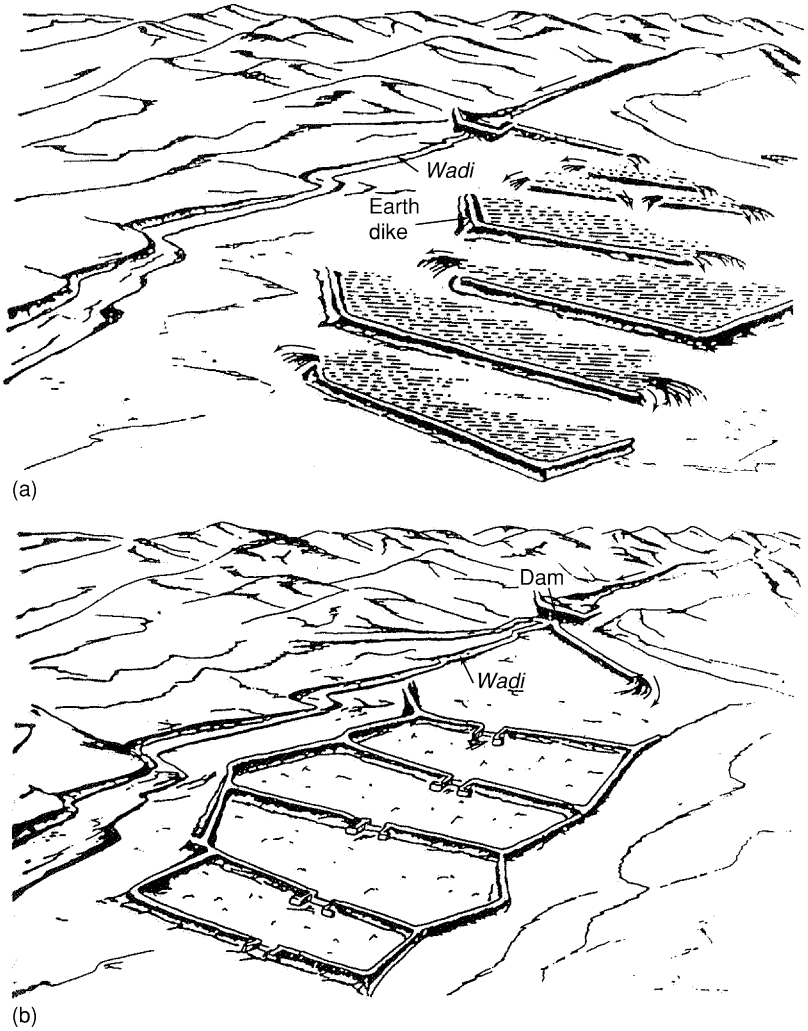


Figure 5 Two general methods of water spreading by diversion from an intermittent stream (a *wadi*) on to adjacent land: (a) diversion of an uncontrolled flow over unleveled land by means of a zigzag series of earthen dikes; and (b) diversion of a controlled flow from a detention dam to a series of level basins with concrete- or stone-lined spillways. Reproduced from Hillel D (1982) *Negev: Land, Water, and Life in a Desert Environment*. New York: Praeger Publishers. © D Hillel.

of the basin for conveyance to a pond or a series of irrigated fields. Another system of runoff farming may consist of strips treated to shed runoff, alternating with basins or areas treated to receive and absorb the runoff (Figure 6). A third possibility is to form microwatersheds, wherein each contributing area serves a single tree or row of plants.

In the ancient runoff-farming systems found in the Negev of Israel, the measured ratio of runoff-contributing areas to the runoff-receiving areas is generally between 30:1 and 20:1. In a region with a mean annual rainfall of 100 mm and a mean runoff yield of 10%, a 'runoff-to-runon' area ratio of 10:1 would double the effective water supply to an agricultural plot, a 20:1 area ratio would triple it, and a 30:1 ratio would quadruple it. The latter ratio would provide 400 mm (100 mm of rain plus 300 mm of runon) to a plot of arable land that would otherwise receive only 100 mm. (These data are hypothetical.) If using methods of runoff inducement can increase the mean runoff yield to approximately 50%, then the runoff-to-runon area ratio may be reduced to about

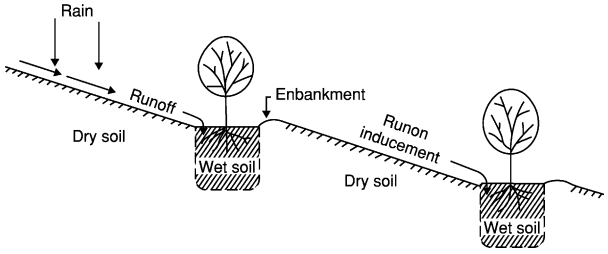


Figure 6 Schematic of a modern runoff-farming system. Reprinted from *Environmental Soil Physics*, Hillel D (ed.). Copyright (1998), with permission from Elsevier.

6:1 or even less. Moreover, the number of months or seasons without a minimally sufficient water supply (owing to the paucity of rain during a drought) would diminish and the entire system could thereby operate with a greater probability of success.

See also: **Infiltration**; **Overland Flow**

Further Reading

Bruins HM, Evenari M, and Nessler U (1986) Rainwater-harvesting agriculture for food production in arid zones. *Applied Geography* 6: 13–33.

Critchey W and Siegert K (1991) *Water Harvesting Manual*. Rome: Food and Agriculture Organization.

Dutt GR, Hutchinson CF, and Anaya Garduno M (eds) (1981) *Rainfall Collection for Agriculture in Arid and Semiarid Regions*. Farnham Royal, UK: Commonwealth Agricultural Bureaux.

Evenari M, Shanan L, and Tadmor N (1971) *The Negev: The Challenge of a Desert*. Cambridge, MA: Harvard University Press.

Hillel D (1982) *Negev: Land, Water, and Life in a Desert Environment*. New York: Praeger Publishers.

Hillel D (1994) *Rivers of Eden: The Struggle for Water and the Quest for Peace in the Middle East*. New York: Oxford University Press.

Water Management *See* **Crop Water Requirements**

WATER POTENTIAL

D Or, University of Connecticut, Storrs, CN, USA
M Tuller, University of Idaho, Moscow, ID, USA
J M Wraith, Montana State University, Bozeman, MT, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Water status in soils is characterized by both the amount of water present and its energy state. Soil water is subjected to forces of variable origin and intensity, thereby acquiring different quantities and forms of energy. The two primary forms of energy of interest here are kinetic and potential. Kinetic energy is acquired by virtue of motion and is proportional to velocity squared. However, because the movement of water in soils is relatively slow (usually less than 0.1 $m\,h^{-1}$) its kinetic energy is negligible. Potential energy, which is defined by the position of soil water within a soil body and by internal conditions, is largely responsible for determining soil water status under isothermal conditions.

Like all other matter, soil water tends to move from where the potential energy is higher to where it is lower, in pursuit of equilibrium with its surroundings. The magnitude of the driving force behind such spontaneous motion is a difference in potential energy across a distance between two points of interest. At a macroscopic scale, we can define potential energy relative to a reference state. The standard state for soil water is defined as pure and free water (no solutes and no external forces other than gravity) at a reference pressure, temperature, and elevation, and is arbitrarily given the value of zero.

The 'Total' Soil Water Potential and its Components

Soil water is subject to several force fields, the combined effects of which result in a deviation in potential energy relative to the reference state, called the 'total soil water potential' (ψ_T) defined as: "The amount of work that an infinitesimal unit quantity of water at equilibrium is capable of doing when it moves (isothermally and reversibly) to a pool of water at similar standard (reference) state, i.e., similar pressure, elevation, temperature and chemical composition." It should be emphasized that there are alternative definitions of soil water potential using concepts of chemical potential or specific free energy of the chemical species water (which is different from the soil solution termed 'soil water'). Recognizing that these fundamental concepts are subject to ongoing debate, presented here are simple and widely accepted definitions which are applicable at macroscopic scales and which yield an appropriate framework for practical applications.

The primary forces acting on soil water held within a rigid soil matrix under isothermal conditions can be conveniently grouped as: (1) matric forces

resulting from interactions of the solid phase with the liquid and gaseous phases; (2) osmotic forces owing to differences in chemical composition of soil solution; and (3) body forces induced by gravitational and other (e.g., centrifugal) inertial force fields.

The thermodynamic approach, whereby potential energy rather than forces are used, is particularly useful for equilibrium and flow considerations. Equilibrium would require the vector sum of these different forces acting on a body of water in different directions to be zero; this is an extremely difficult criterion to deal with in soils. On the other hand, potential energy, mathematically defined as the negative integral of the force over the path taken by an infinitesimal amount of water when it moves from a reference location to the point under consideration, is a scalar (not a vector) quantity. Subsequently, we can express the total potential as the algebraic sum of the component potentials corresponding to the different fields acting on soil water as:

$$\psi_T = \psi_m + \psi_s + \psi_p + \psi_z \qquad [1]$$

where the component potentials ψ_i are discussed below.

ψ_m is the matric potential resulting from the combined effects of capillarity and adsorptive forces within the soil matrix. The primary mechanisms for these effects include: (1) capillarity caused by liquid–gas interfaces forming and interacting within the irregular soil-pore geometry; (2) adhesion of water molecules to solid surfaces due to short-range London–van der Waals forces and extension of these effects by cohesion through hydrogen bonds formed in the liquid; and (3) ion hydration and water participating in diffuse double layers (particularly near clay surfaces). There is some disagreement regarding the practical definition of this component of the total potential. Some consider all contributions other than gravity and solute interactions (at a reference atmospheric pressure). Others use a device known as a tensiometer to measure and provide a practical definition of the matric potential in a soil volume of interest in contact with a tensiometer's porous cup. The value of ψ_m ranges from zero when the soil is saturated to increasingly negative values as the soil becomes drier (note that $\psi_m = 0$ mm is greater than $\psi_m = -1000$ mm; in analogy, a temperature of 0°C is greater than −10°C). (*See* **Capillarity.**)

Applied theories for flow and transport in unsaturated porous media, particularly at low water content, commonly lump capillary and adsorptive forces without distinguishing individual contributions to the matric potential. Advanced frameworks simultaneously consider individual contributions of capillary and adsorptive forces for calculation of liquid–vapor interfacial configurations in angular pore spaces. The liquid–vapor interface is considered as a surface of constant, partial specific Gibbs free-energy (or matric potential) made up of an adsorptive component (A) and a capillary component (C):

$$\psi_m = A(h) + C(\kappa) \qquad [2]$$

with κ as the mean curvature of the liquid–vapor interface, and h as the distance from the solid to the liquid–vapor interface, taken normal to the solid surface (thickness of the adsorbed film). The capillary component C is given by the classic Young–Laplace equation:

$$C(\kappa) = \frac{-2 \cdot \sigma \cdot \kappa}{\rho} \qquad [3]$$

where κ is positive for an interface concave outward from the liquid, σ is the surface tension at the interface, and ρ is the density of the liquid. (*See* **Capillarity.**)

The adsorptive component in eqn [2] is attributed to two types of surface forces. The first kind includes long-range (>500 Å) electrostatic forces (e.g., diffuse double layer, DDL), and short-range (<100 Å) van der Waals and hydration forces, responsible for molecular interactions and structural changes in water molecules near the solid surface. The second kind is comprised of long-range forces due to the overlapping of two interfacial regions (e.g., mutual attraction between two clay platelets across a slit-shaped pore space). The combined effect of interfacial interactions results in a difference in chemical potentials between the liquid in the adsorbed film and the bulk liquid phase. This difference in chemical potentials may be expressed as an equivalent interfacial force per unit area of the interface, termed the 'disjoining pressure' (Π). The disjoining pressure is a function of liquid film thickness (h), and it can also be viewed as the difference between a normal component of film pressure, P_N (in equilibrium with the gaseous phase $P_N = P_G$), and the pressure in the bulk liquid phase, P_L:

$$\Pi(h) = P_N(h) - P_L = P_G - P_L \qquad [4]$$

The disjoining pressure is related to more conventional thermodynamic quantities such as Gibbs free energy. Gibbs free energy (G) per unit area of the interface may be defined on the basis of $\Pi(h)$ isotherms for constant pressure P_L, temperature T, and chemical (μ) and electric potentials of the liquid–gaseous and the liquid–solid interfaces as:

$$G(h) = -\int_{\infty}^{h} \Pi(h)\mathrm{d}h \qquad [5]$$

The value of $G(h)$ is equal to the work of thinning the film in a reversible isobaric–isothermal process from ∞ to a finite thickness h, with $\Pi(h) = -(\partial G/\partial h)_{T,\, P_L,\, \mu,\, \psi_g,\, \psi_s}$. The use of $\Pi(h)$ as the basic thermodynamic property is not a mere change of notation, but $\Pi(h)$ has advantages in cases where Gibbs thermodynamic theory is difficult to define, such as when interfacial zones overlap to the extent that the film does not retain the intensive properties of the bulk phase. The use of the disjoining pressure is advantageous from an experimental point of view because of the relative ease in accounting for different contributions (e.g., electrostatic effects).

The disjoining pressure is a sum of several components, similar to the concept of total soil water potential discussed above. The primary components of $\Pi(h)$ in porous media are: molecular, $\Pi_m(h)$; electrostatic, $\Pi_e(h)$; structural, $\Pi_s(h)$; and adsorptive $\Pi_a(h)$:

$$\Pi(h) = \Pi_m(h) + \Pi_e(h) + \Pi_s(h) + \Pi_a(h) \qquad [6]$$

$\Pi_m(h)$

The molecular component originates from van der Waals interaction between macro-objects (e.g., parallel clay plates). Various expressions, with $\Pi_m(h)$ often proportional to h^{-3}, were derived by Paunov *et al.* and Iwamatsu and Horii.

$\Pi_e(h)$

The electrostatic component of the disjoining pressure is calculated from the solution of the Poisson–Boltzmann equation for the DDL with appropriate boundary conditions. Approximate solutions are adequate for many applications, often with $\Pi_e(h) \propto h^{-2}$.

$\Pi_s(h)$

Some controversy exists regarding the origin of the structural component; some attribute it to changes in the structure (density) of water adjacent to solid surfaces and deformation of hydrated shells, while others attribute this force to the presence of a layer with a lower dielectric constant near the surface. Regardless of its exact origin, this component is responsible for the so-called hydration repulsion which stabilizes dispersion and prevents coagulation of some colloidal particles, even at high electrolyte concentrations: $\Pi_s(h) \propto h^{-1}$.

$\Pi_a(h)$

The adsorptive component of the disjoining pressure results from nonuniform concentrations in the water film due to unequal interaction energies of solute and solvent with interfaces in nonionic solutions. This is different to the nonuniform distribution of charged ions. This component of the disjoining pressure is likely to become very important for interactions between nonpolar molecules (e.g., NAPLs) which give rise to repulsive forces in the liquid film.

The form of the disjoining pressure isotherm $\Pi(h)$ is determined by the nature of surface forces. While the molecular component $\Pi_m(h)$ is always present, the influence of other components depends on surface properties, liquid polarity and its composition, and adsorption of dissolved components. The range of the electrostatic forces in dilute solutions of a 1:1 electrolyte (10^{-6}–10^{-7} mol l^{-1}) is in the range of 0.3–1.0 μm. Consequently, thick films of water and aqueous electrolyte solutions ($h > 500$ Å) are stable mainly through the $\Pi_e(h)$ component of disjoining pressure. The magnitude and contribution of $\Pi_e(h)$ primarily depend on the charges of the film and substrate surfaces. Dispersion forces become appreciable in the range $h < 500$ Å, and their influence is enhanced by large differences between the permittivity of the liquid and the solid. The forces of structural repulsion may come to play when the film thickness is less than 100 Å.

ψ_s is the solute or osmotic potential determined by the presence of solutes in soil water, which lower its potential energy and its vapor pressure. The effects of ψ_s are important when: (1) there are appreciable amounts of solutes in the soil; and (2) in the presence of a selectively permeable membrane or a diffusion barrier which transmits water more readily than salts. The effects of ψ_s are otherwise generally negligible when only liquid water flow is considered and no diffusion barrier exists. The two most important diffusion barriers in the soil are: (1) soil–plant root interfaces (cell membranes are selectively permeable); and (2) air–water interfaces; thus when water evaporates, salts are left behind. In dilute solutions the solute potential, also called the osmotic pressure, is proportional to the concentration and temperature according to:

$$\psi_s = -RT\,C_s \qquad [7]$$

where ψ_s is in kilopascals, R is the universal gas constant (8.314×10^{-3} kPa m^3/(mol K)), T is absolute temperature (Kelvin), and C_s is solute concentration (moles per cubic meter). A useful approximation which may be used to estimate ψ_s in kilopascals from the electrical conductivity of the soil solution at saturation (EC_s) in deciSiemens per meter (dS m^{-1}) is:

$$\psi_s \approx -36\,EC_s \qquad [8]$$

ψ_p is the pressure potential, defined as the hydrostatic pressure exerted by unsupported water that

saturates the soil and overlays a point of interest. Using units of energy per unit weight provides a simple and practical definition of ψ_p as the vertical distance from the point of interest to the free water surface (unconfined water table elevation). The convention used here is that ψ_p is always positive below a water table, or zero if the point of interest is at or above the water table. In this sense, nonzero magnitudes of ψ_p and ψ_m are mutually exclusive: either ψ_p is positive and ψ_m is zero (saturated conditions), or ψ_m is negative and ψ_p is zero (unsaturated conditions), or $\psi_p = \psi_m = 0$ at the free water table elevation. Note that some prefer to combine the pressure and matric components into a single term, which assumes positive values under saturated conditions and negative values under unsaturated conditions. Based on operational and explanatory considerations, we prefer to adopt the more commonly used separate components protocol.

ψ_z is the gravitational potential, which is determined solely by the elevation of a point relative to some arbitrary reference point, and is equal to the work needed to raise a body against the Earth's gravitational pull from a reference level to its present position. When expressed as energy per unit weight, the gravitational potential is simply the vertical distance from a reference level to the point of interest. The numerical value of ψ_z itself is thus not important (it is defined with respect to an arbitrary reference level) – what is important is the difference (or gradient) in ψ_z between any two points of interest. This value is invariant of the reference level location.

Soil water is at equilibrium when the net force on an infinitestimal body of water equals zero everywhere, or when the total potential is constant in the system. Though the last statement is a logical consequence of the definitions above, it is not strictly true. Constant total potential is a necessary but not a sufficient condition, and, for thermodynamic equilibrium to prevail, three conditions must be met simultaneously: (1) thermal equilibrium or uniform temperature; (2) mechanical equilibrium, meaning no net convection-producing force; and (3) chemical equilibrium, meaning no net diffusional transport or chemical reaction. In most practical applications, however, the macroscopic definition of the total potential and equilibrium conditions based on it are completely adequate.

The difference in chemical and mechanical potentials between soil water and pure water at the same temperature is known as the soil water potential (ψ_w):

$$\psi_w = \psi_m + \psi_s + \psi_p \qquad [9]$$

Note that the gravitational component (ψ_z) is absent in this definition. Soil water potential is thus the result of inherent properties of soil water itself, and of its physical and chemical interactions with its surroundings; whereas the total potential includes the effects of gravity (an 'external' and ubiquitous force field).

Total soil water potential and its components may be expressed in several ways depending on the definition of a 'unit quantity of water.' Potential may be expressed as (1) energy per unit of mass; (2) energy per unit of volume; or (3) energy per unit of weight. A summary of the resulting dimensions, common symbols, and units are presented in Table 1.

Only μ has actual units of potential; ψ has units of pressure, and h of head of water. However, the above terminology (i.e., potential energy expressions rather than units of potential per se) is widely used in a generic sense in the soil and plant sciences. The various expressions of soil water energy status are equivalent, with:

$$\mu = \frac{\psi}{\rho_w} = gh \qquad [10]$$

where ρ_w is density of water ($1000\,kg\,m^{-3}$ at 20°C) and g is gravitational acceleration ($9.81\,m\,s^{-2}$).

Measurement of Potential Components

Water Potential

A psychrometer (Figure 1) is commonly used for measurement of total water potential (ψ_w) in soils. The potential of the soil solution is in thermodynamic equilibrium with its ambient water vapor. Taking the vapor pressure above pure water at reference state ($\psi_w = 0$) as e_o, the vapor pressure (e) over a salt

Table 1 Units, dimensions, and common symbols for potential energy of soil water

Units	*Symbol*	*Name*	*Dimensions*	*SI units*	*CGS units*
Energy/mass	μ	Chemical potential	$L^2 t^{-2}$	$J\,kg^{-1}$	$erg\,g^{-1}$
Energy/volume	ψ	Soil water potential, suction, or tension	$ML^{-1}t^{-2}$	$N\,m^{-2}$(Pa)	$dyn\,cm^{-2}$
Energy/weight	h	Pressure head	L	m	cm

L, length; *M*, mass; *t*, time.

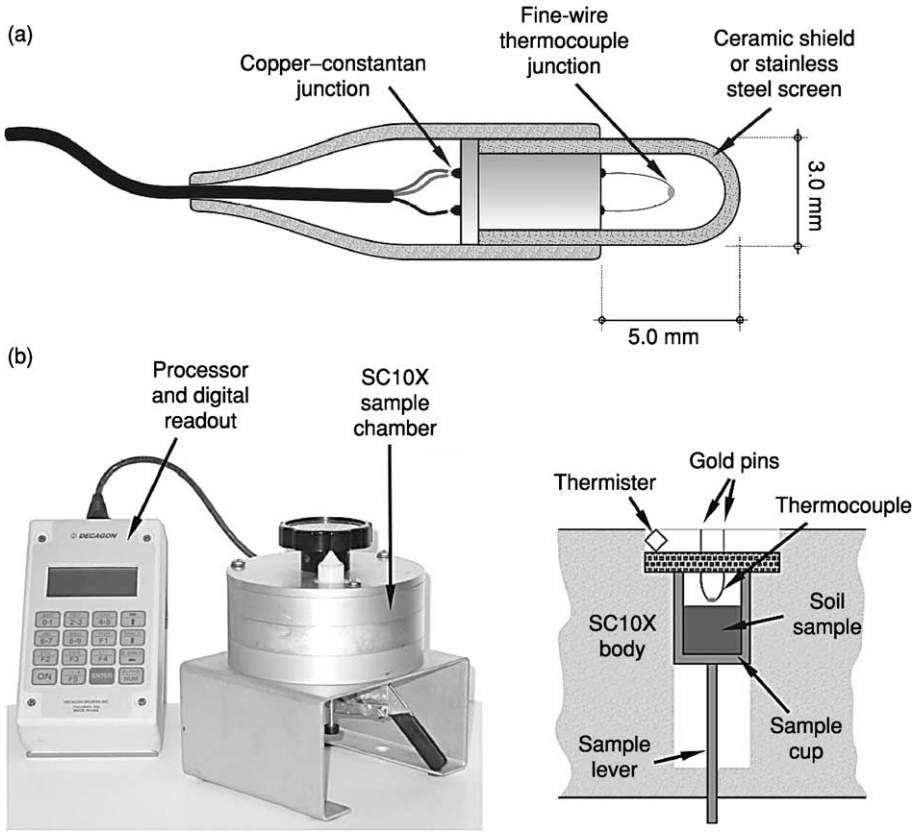


Figure 1 (a) A field psychrometer with porous ceramic shield (Wescor Inc., Logan, Utah, USA); and (b) SC10X sample chamber for psychrometric laboratory measurements of soil water potential (Decagon Devices Inc., Pullman, Washington, USA).

solution or soil water held in soil pores by matric forces is depressed relative to the reference state, i.e., $e < e_o$. A convenient measure obtained by the psychrometer is the relative vapor pressure of the ambient soil atmosphere, which is related to the water potential (ψ_w) of soil water through the well-known Kelvin equation:

$$\mathrm{RH} = \frac{e}{e_o} = \exp\left[\frac{M_w \psi_w}{\rho_w R T}\right] \qquad [11]$$

where e is water vapor pressure (kilopascals), e_o is saturated vapor pressure at the same temperature, M_w is the molecular weight of water ($0.018\,\mathrm{kg\,mol^{-1}}$), R is the ideal gas constant ($8.31\,\mathrm{J\,K^{-1}\,mol^{-1}}$ or $0.008314\,\mathrm{kPa\,m^3\,mol^{-1}\,K^{-1}}$), T is absolute temperature (Kelvin), and ρ_w is the density of water ($1000\,\mathrm{kg\,m^{-3}}$ at 20°C). Rearranging and taking a log-transformation of eqn [11] yields an expression for water potential ψ_w:

$$\psi_w = \frac{R T \rho_w}{M_w} \ln\left(\frac{e}{e_o}\right) \qquad [12]$$

The water potential in drier soils is lower such that fewer water molecules 'escape' into the ambient atmosphere, resulting in lower relative humidity (lower relative vapor pressure). Concentrated soil solutions having lower osmotic potentials have similar effect on reducing vapor pressure, as more water molecules are associated with hydrated salt molecules and are less free to 'escape' the liquid state. The inability to distinguish between matric and osmotic effects limits psychrometric measurements to soil water potential only. In some cases where the osmotic potential is negligible, psychrometric measurements are used to infer the matric potential.

Pressure Potential

Piezometers are commonly applied to measure ψ_p. A piezometer (Figure 2) is a tube that is placed in the soil to depths below the water table and that extends to the soil surface and is open to the atmosphere. The bottom of the piezometer is perforated to allow soil water under positive hydrostatic pressure to enter the tube. Water enters the tube and rises to a height equal to that of the unconfined water table. The elevation of the free water table is measured relative to the soil surface using a steel tape with bell sounder, or other electro-optic devices that indicate water table depth. The value of pressure potential

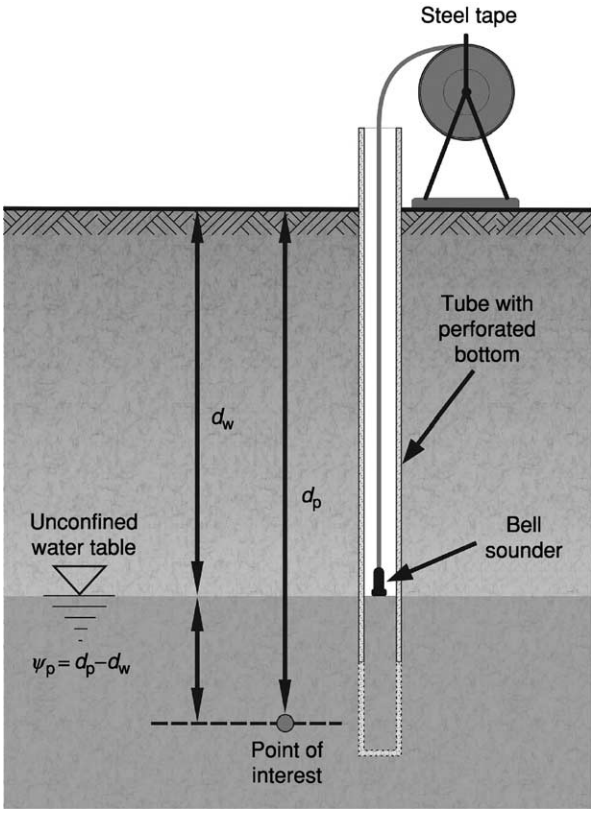


Figure 2 The concept of piezometer measurements.

expressed as energy per weight is simply the vertical distance from a point of interest to the surface of the free water table. Pressure potentials above the water table surface are always zero (nonzero pressure and matric potentials are mutually exclusive).

Matric Potential

Tensiometers or heat-dissipation sensors are commonly applied to measure soil matric potential. A tensiometer consists of a porous cup, usually made of ceramic and having very fine pores, connected to a vacuum gauge through a water-filled tube (Figure 3). The porous cup is placed in intimate contact with the bulk soil at the depth of measurement. When the matric potential of the soil is lower (more negative) than inside the tensiometer, water moves from the tensiometer along a potential energy gradient to the soil through the saturated porous cup, thereby creating suction sensed by the gauge. Water flow into the soil continues until equilibrium is reached and the suction inside the tensiometer equals the soil matric potential. When the soil is wetted, flow may occur in the reverse direction, i.e., soil water enters the tensiometer until a new equilibrium is attained. The tensiometer equation is:

$$\psi_m = \psi_{gauge} + (z_{gauge} - z_{cup}) \quad [13]$$

with ψ_{gauge} the reading at the vacuum gauge location and z indicating depth. The vertical distance from the gauge plane to the cup, expressed as a negative quantity, must be added to the matric potential measured by the gauge (ψ_{gauge}) in order to obtain the matric potential at the depth of the cup. This accounts for the positive head exerted by the overlying tensiometer water column at the depth of the ceramic cup. Note that using the difference in vertical elevation is appropriate only when potentials are expressed per unit of weight. Electronic sensors called pressure transducers often replace mechanical vacuum gauges. The transducers convert mechanical pressure into an electric signal which can be more easily and more precisely measured. In practice, pressure transducers can provide more accurate readings than other gauges, and in combination with data-logging equipment are able to supply continuous measurements of matric potential.

The tensiometer range is limited to suctions (absolute values of the matric potential) of less than 100 kPa, i.e., 1 bar, 10 m head of water, or ~1 atmosphere. Therefore other means are needed for matric potential measurement under drier conditions.

Heat-dissipation sensors may be applied for a matric potential range from −10 to −1000 kPa. The rate of heat dissipation in a porous medium is dependent on the medium's specific heat capacity, thermal conductivity, and density. The heat capacity and thermal conductivity of a porous matrix are affected by its water content. Heat dissipation sensors contain heating elements in line or point source configurations embedded in a rigid porous matrix with fixed pore space. The measurement is based on application of a heat pulse by applying a constant current through the heating element for specified time period, and analysis of the temperature response measured by a thermocouple placed at a fixed distance from the heating source. With the heat dissipation sensor buried in the soil, changes in soil water matric potential result in a gradient between the soil and the porous ceramic matrix, inducing water flow between the two materials until a new equilibrium is established. The water flow changes the water content of the ceramic matrix which, in turn, changes the thermal conductivity and heat capacity of the sensor, and hence the measured temperature response to the applied heat pulse.

As already mentioned above, for cases where the osmotic potential is negligible, psychrometric measurements can be used to infer the matric potential. A typical range for psychrometers is −800 to −10 000 kPa.

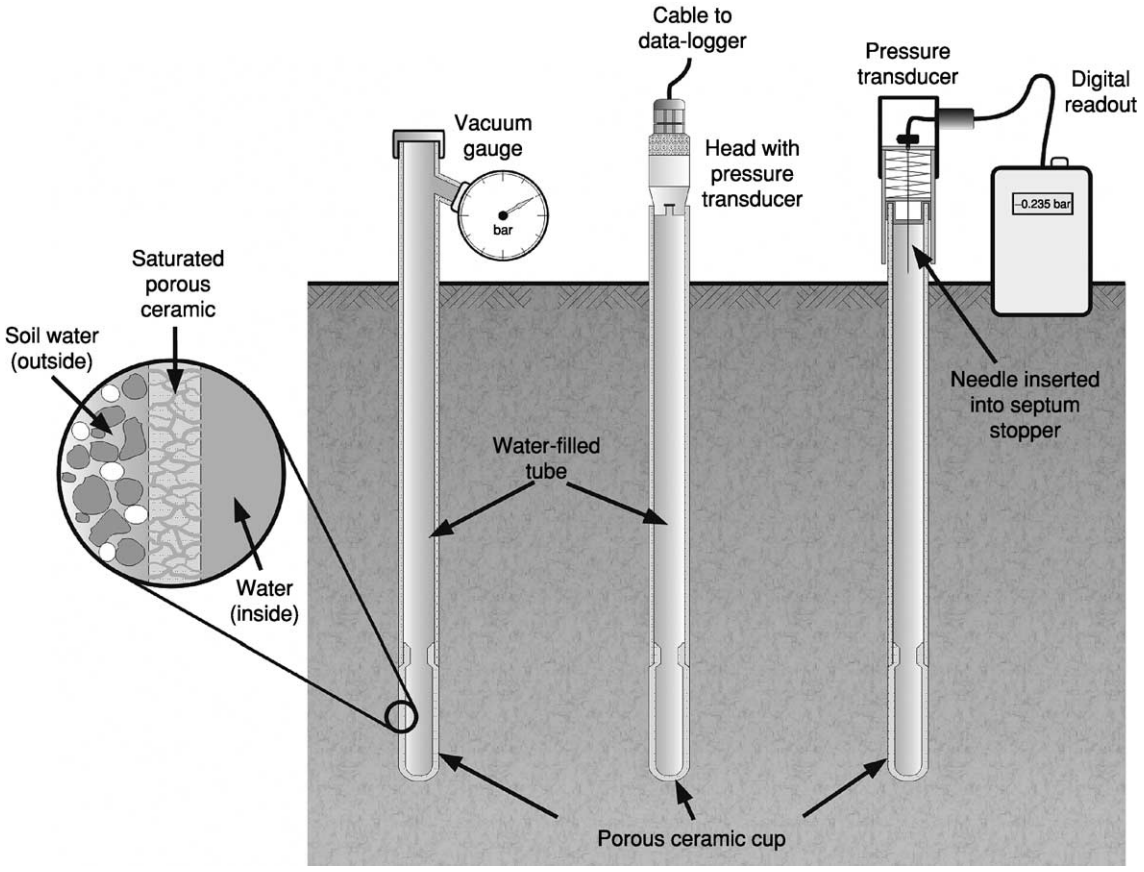


Figure 3 Tensiometers for matric potential measurement using vacuum gauges and electronic pressure transducers.

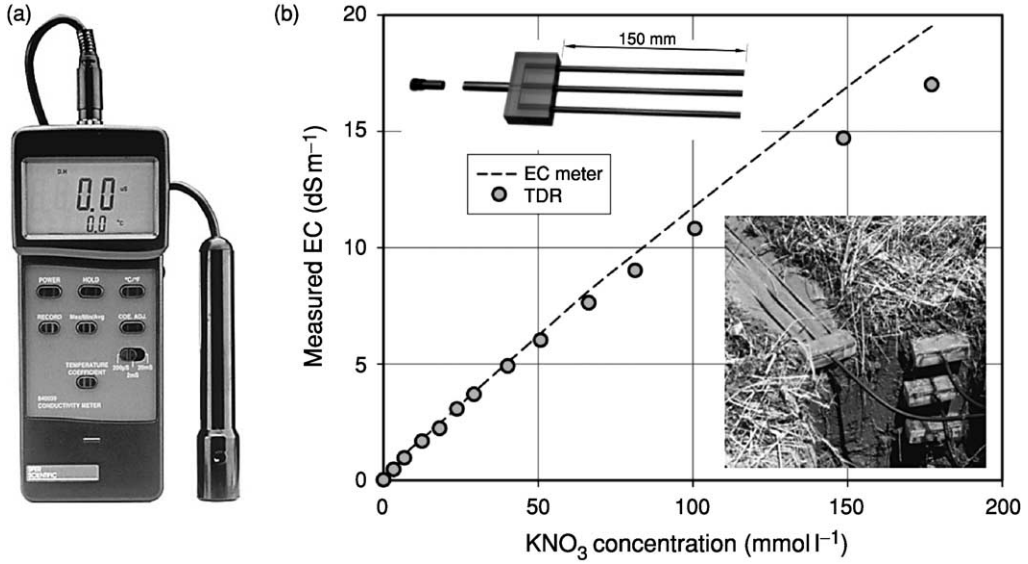


Figure 4 (a) Handheld electrical conductivity (EC) meter; (b) time-domain reflectometry (TDR) probes and solution EC versus concentration measured with TDR and EC meter. (Reproduced from Mmolawa KB and Or D (2000) Root zone solute dynamics under drip irrigation: a review. *Plant and Soil* 222: 163–190.)

Osmotic Potential

Soil water solutions contain varied quantities and compositions of dissolved salts. The relationships between the salt concentration and ψ_s, and the possibility for estimating ψ_s from the electrical conductivity (EC) of the soil solution, were discussed above. Conventional measurement of soil solution EC involves solution extraction from saturated soil samples and measuring the EC using an electrical conductivity meter (**Figure 4a**).

Electrical conductivity meters rely on Ohm's law:

$$E = I \cdot R \qquad [14]$$

with E the electromotive force (volts), I the current flow (amperes), and R the resistance (ohms). For constant voltage, the current flowing through the solution is inversely proportional to the electrical resistance, or directly proportional to the electrical conductance. The solution EC is thus determined from known voltage and electrode geometry and measurement of the electric current. More recently, a variety of *in situ* methods such as time-domain reflectometry (TDR) have been used to deduce soil bulk EC from electromagnetic signal attenuation, hence enabling simultaneous measurements of water content and soil EC in the same undisturbed soil volume. Concurrent knowledge of θ and EC can be used to infer the soil solution EC, and hence to estimate ψ_s.

See also: **Capillarity**

Further Reading

Adamson AW (1990) *Physical Chemistry of Surfaces*, 5th edn. New York: John Wiley.

Bolt GH (1976) Soil physics terminology. *International Society Soil Science Bulletin* 49: 16–22.

Bristow KL, Campbell GS, and Calissendroff K (1993) Test of a heat-pulse probe for measuring changes in soil water content. *Soil Science Society of America Journal* 57: 930–934.

Corey AT and Klute A (1985) Application of the potential concept to soil water equilibrium and transport. *Soil Science Society of America Journal* 49: 3–11.

Dalton FN, Herkelrath WN, Rawlins DS, and Rhoades JD (1984) Time-domain reflectometry: simultaneous measurement of soil water content and electrical conductivity with a single probe. *Science* 224: 989–990.

Day PR, Bolt GH, and Anderson DM (1967) Nature of soil water. In: Hagan RM, Haise HR, and Edminster TW (eds) *Irrigation of Agricultural Lands*, pp. 193–208. Madison, WI: American Society of Agronomy.

Derjaguin BV, Churaev NV, and Muller VM (1987) *Surface Forces*. New York: Plenum.

Edlefsen NE and Anderson ABC (1943) Thermodynamics of soil moisture. *Hilgardia* 15: 31–298.

Hanks RJ (1992) *Applied Soil Physics*, 2nd edn. New York: Springer-Verlag.

Hendrickx JMH, Wraith JM, Corwin DL, and Kachanoski RG (2002) Solute content and concentration. Dane JH and Topp GC (eds) *Methods of Soil Analysis*, part 4, *Physical Methods*, pp. 1253–1322. Madison, WI: American Society of Agronomy.

Iwamatsu MI and Horii K (1996) Capillary condensation and adhesion of two wetter surfaces. *Journal of Colloid Interface Science* 182: 400–406.

Iwata S, Tabuchi T, and Warkentin BP (1988) *Soil Water Interactions*. New York: Marcel Dekker.

Kutilek M and Nielsen DR (1994) *Soil Hydrology*. Cremlingen-Destedt, Germany: Catena-Verlag.

Mitlin VS and Sharma MM (1993) A local gradient theory for structural forces in thin fluid films. *Journal of Colloid Interface Science* 157: 447–464.

Mmolawa KB and Or D (2000) Root zone solute dynamics under drip irrigation: a review. *Plant and Soil* 222: 163–190.

Nitao JJ and Bear J (1996) Potentials and their role in transport in porous media. *Water Resources Research* 32(2): 225–250.

Novy RA, Toledo PG, Davis HT, and Scriven LE (1989) Capillary dispersion in porous media at low wetting phase saturations. *Chemical Engineering Science* 44(9): 1785–1797.

Paunov VN, Dimova RI, Kralchevsky PA, Broze G, and Mehreteab A (1996) The hydration repulsion between charged surfaces as an interplay of volume exclusion and dielectric saturation effects. *Journal of Colloid Interface Science* 182: 239–248.

Phene CJ, Hoffman GJ, and Rawlins SL (1971) Measuring soil matric potential *in situ* by sensing heat dissipation within a porous body. 1. Theory and sensor construction. *Soil Science Society of America Proceedings* 35: 27–33.

Rhoades JD and Oster JD (1986) Solute content. Klute A (ed.) *Methods of Soil Analysis*, part 1, *Physical and Mineralogical Methods*, 2nd edn, pp. 985–1006. Madison, WI: American Society of Agronomy.

Philip JR (1977) Unitary approach to capillary condensation and adsorption. *Journal of Chemical Physics* 66(11): 5069–5075.

Tuller M, Or D, and Dudley LM (1999) Adsorption and capillary condensation in porous media: liquid retention and interfacial configurations in angular pores. *Water Resources Research* 35(7): 1949–1964.

Water Requirements *See* **Crop Water Requirements**

WATER RETENTION AND CHARACTERISTIC CURVE

M Tuller, University of Idaho, Moscow, ID, USA
D Or, University of Connecticut, Storrs, CT, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

A soil-water characteristic (SWC) curve describes the amount of water retained in a soil (expressed as mass or volume water content, θ_m or θ_v) under equilibrium at a given matric potential. An SWC is an important hydraulic property, related to size and connectedness of pore spaces, hence strongly affected by soil texture and structure, and by other constituents, including organic matter. Modeling water distribution and flow in partially saturated soils requires knowledge of the SWC, therefore plays a critical role in water management and in prediction of solute and contaminant transport in the environment. Typically a SWC is highly nonlinear and is relatively difficult to obtain accurately. Because the matric potential extends over several orders of magnitude for the range of water contents commonly encountered in practical applications, the matric potential is often plotted on a logarithmic scale. Figure 1 depicts representative SWC curves for soils of different textures, demonstrating the effects of porosity (saturated water content) and the varied slopes of the relationships resulting from variable pore-size distributions.

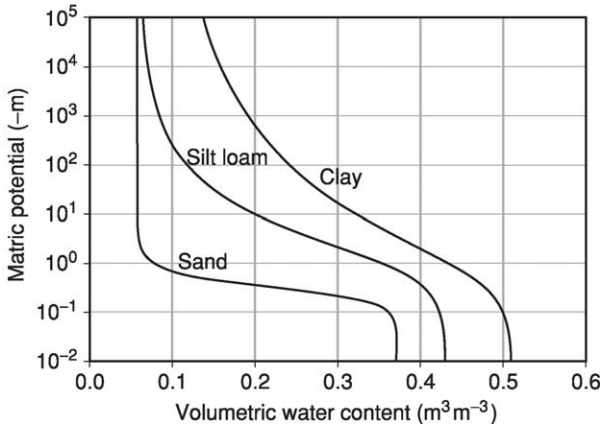


Figure 1 Typical soil-water characteristic curves for soils of different texture.

The Matric Potential

The matric potential is attributed to capillary and adsorptive forces acting between liquid, gaseous, and solid phases. Capillarity results from the surface tension of water and its contact angle with the solid particles. Under partially saturated conditions (i.e., in the presence of the nonwetting air phase), curved liquid–vapor interfaces (menisci) form within the porous soil system. Menisci radii of curvature (R) are a function of capillary pressure (P_c) and are calculated according to the Young–Laplace equation:

$$P_0 - P_c = \Delta P = \frac{2\sigma}{R} \qquad [1]$$

where P_0 is the atmospheric pressure (conventionally referenced as zero), P_c is the pressure of the soil water, and σ is the surface tension of the liquid–vapor interface. If soil pores were behaving as a bundle of capillary tubes, capillarity would be sufficient to describe the relationships between matric potential and soil pore radii. However, in addition to capillarity, the soil also exhibits adsorption, which forms hydration envelopes over the particle surfaces. The presence of water in films is most important in clayey soils with large surface area and is influenced by the electric double layer and the exchangeable cations present. In sandy soils, adsorption is relatively insignificant and the capillary effect predominates. In general, however, matric potential results from the combined effect of capillarity and surface adsorption. The capillary 'wedges' are at a state of internal equilibrium with adsorption 'films,' and the two cannot be considered separately. (*See* **Water Potential; Capillarity.**)

The Bundle of Cylindrical Capillaries Model

Early conceptual models for the SWC and liquid distribution in partially saturated porous media are based on the 'bundle of cylindrical capillaries' (BCC)

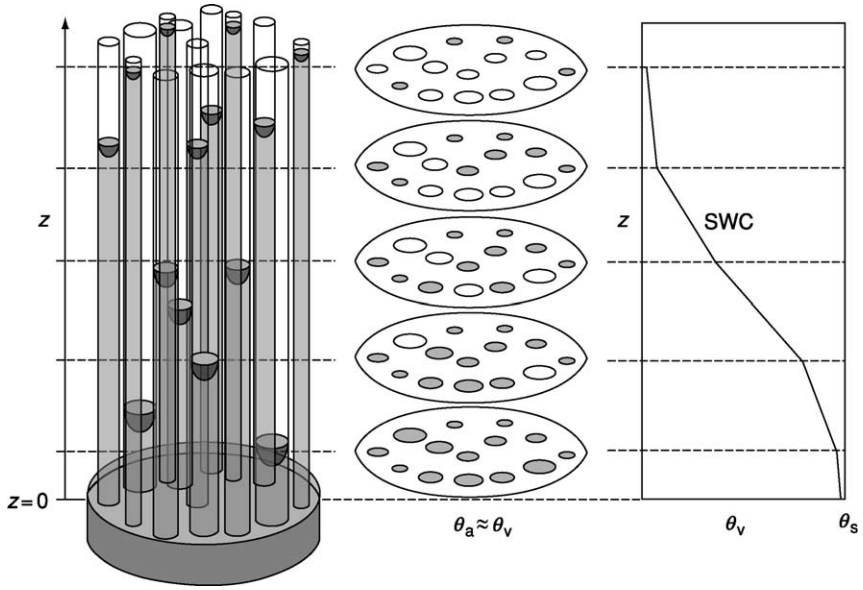


Figure 2 Relationship between the pore space, as represented by the bundle of cylindrical capillaries model, and the soil-water characteristic (SWC). Greater capillary rise occurs in smaller pores, which have smaller radii of meniscus curvature. (Note that z indicates the elevation above free water, θ_a stands for pore volume, θ_v for volumetric water content, and θ_s for volumetric water content at saturation.)

representation of pore-space geometry. The BCC representation postulates that at a given matric potential a portion of interconnected cylindrical pores is completely liquid-filled, whereas larger pores are completely empty (Figure 2). This convenient idealization of soil pore space enables a linkage between the soil pore size distribution and the SWC based on the capillary-rise equation. (*See* **Capillarity**.) However, such representation imposes serious limitations on the practical application of the BCC model to natural porous media as discussed in the following section, including the unrealistic underlying geometry that precludes dual water–air occupancy within the same pores and lack of consideration of adsorbed water films.

Liquid Retention and Pore Shape

Liquid retention in the porous soil matrix is highly dependent on the shape and angularity of individual pores. Inspection of thin sections or soil micrographs (Figure 3) reveals that natural pore spaces do not resemble cylindrical capillaries, as often assumed for an idealized representation. Because natural porous media are formed by aggregation of primary particles and mineral surfaces, the resulting pore space is more realistically described by angular or slit-shaped pore cross-sections than by cylindrical capillaries. In addition to a more realistic representation of natural pore spaces, angular pores offer other advantages over cylindrical tubes in terms of liquid behavior. When angular pores are drained, a fraction of the wetting phase remains in the pore corners (Figure 3c). This aspect of 'dual occupancy' of the invaded portion of the tube, not possible in cylindrical tubes, more realistically represents liquid configurations and mechanisms for maintaining hydraulic continuity in porous media. Liquid-filled corners and crevices play an important role in displacement rates of oil and in other transport processes in partially saturated porous media.

The relationships between liquid retention and pore angularity are known. The water-filled cross-sectional area A_{wt} for all regular and irregular triangles and for regular, higher-order polygons is given by the following expression:

$$A_{wt} = r^2 \cdot F(\gamma) \qquad [2]$$

where r is the radius of curvature of the liquid–vapor interface that is dependent on chemical potential (μ) or capillary pressure (p_c) according to the Young–Laplace equation:

$$\mu = \frac{\sigma}{r\rho} \qquad \text{or} \qquad p_c = \frac{\sigma}{r} \qquad [3]$$

with σ as the surface tension of the liquid, ρ the density of the liquid, and $F(\gamma)$ a shape or angularity factor dependent on angularity of the pore cross-section only. Note that in this discussion we consider

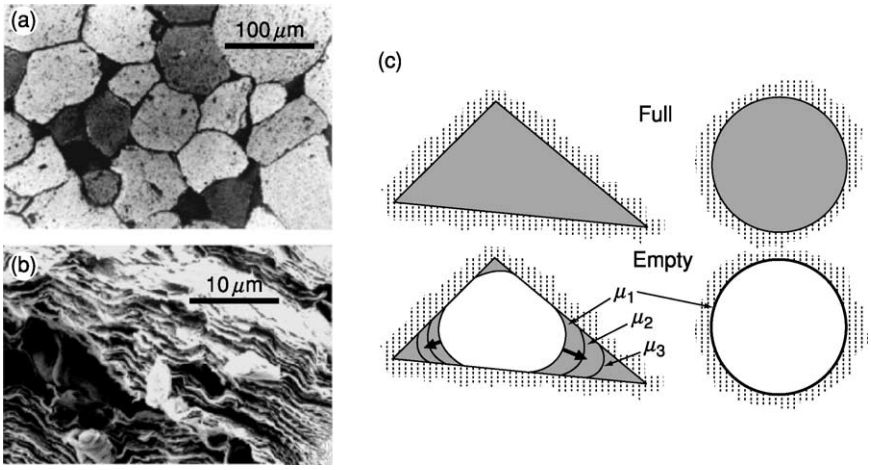


Figure 3 Angular pore space in natural porous media: (a) thin section of Devonian sandstone. (Reproduced with permission of the AAPG, whose permission is required for further use, from Scholle PA (1979) *A Color Illustrated Guide to Constituents, Textures, Cements, and Porosities of Sandstones and Associated Rocks*; photograph by RF Sippel, AAPG © 1979. AAPG Mem. 28); (b) scanning electron micrograph (SEM) of calcium-saturated montmorillonite; (c) liquid retention in triangular and cylindrical pores. μ, chemical potential.

capillary behavior only, ignoring adsorbed liquid films that cover the flat pore surfaces.

In contrast to the piston-like filling or emptying of circular capillaries, angular pores undergo different filling stages and spontaneous displacement in the transition from dry to wet or vice versa. At relatively dry conditions (low chemical potentials), thin liquid films are adsorbed on all flat surfaces of the angular pore, and liquid accumulates in corners due to capillary forces. A further increase in chemical potential leads to further increase in film thickness and to an increase in the radius of capillary interface curvature in the corners, until the capillary corner menisci contact to form an inscribed circle. At this critical potential, liquid fills up the central pore spontaneously (snap-off). The radius of interface curvature at this critical point is equal to the radius of an inscribed circle in the pore cross-section:

$$r_{\text{imb}} = \frac{2 \cdot A}{P} = \frac{P}{4[F(\gamma) + \pi]} \qquad [4]$$

where r_{imb} is the radius of spontaneous imbibition, A is pore cross-sectional area, and P is the perimeter of the pore cross-section. For drainage, at a certain potential a liquid–vapor meniscus invades the pore and liquid is displaced from its center, leaving liquid in the corners. The radius of curvature at onset of drainage, r_{d}, is expressed in terms of perimeter (P) and the angularity factor $F(\gamma)$:

$$r_{\text{d}} = \frac{P}{2 \cdot [(F(\gamma) + \pi) + \sqrt{\pi \cdot (F(\gamma) + \pi)}]} \qquad [5]$$

An example of the effects of pore shape (and angularity) on imbibition and drainage processes is depicted in Figure 4, showing marked differences between cylindrical and angular pores having equal pore cross-sectional areas. Liquid displacement in cylindrical tubes during drainage is piston-like, leaving no liquid in the cross-section after the drainage threshold. Angular pores, on the other hand, show that liquid is displaced from the central region first (at a radius of curvature given by eqn [5], leaving some liquid in the corners. Subsequent decrease in chemical potential results in commensurate decreasing amounts of liquid in the corners ('lower' chemical potential or capillary pressure indicates more negative values, much like the terminology for the subzero temperature scale). The threshold chemical potential for different pore shapes with the same cross-sectional area increases (becomes less negative) with increasing angularity factor $F(\gamma)$. The same holds for the amount of liquid held in the corners.

The conditions during imbibition are slightly different. Liquid–vapor interfaces in corners of angular pores expand with increasing chemical potential to the point of snap-off (eqn [4]), where the pore completely fills up with liquid. The threshold chemical potential for snap-off increases with angularity $F(\gamma)$, and the amount of liquid held in the corners at a given chemical potential is directly proportional to angularity. Highly angular pore shapes such as triangular pores retain more liquid at the same potential than squares or hexagons (Figure 4). In the extreme case of cylindrical tubes, no liquid is held prior to spontaneous filling by snap-off (the 'empty–full' behavior).

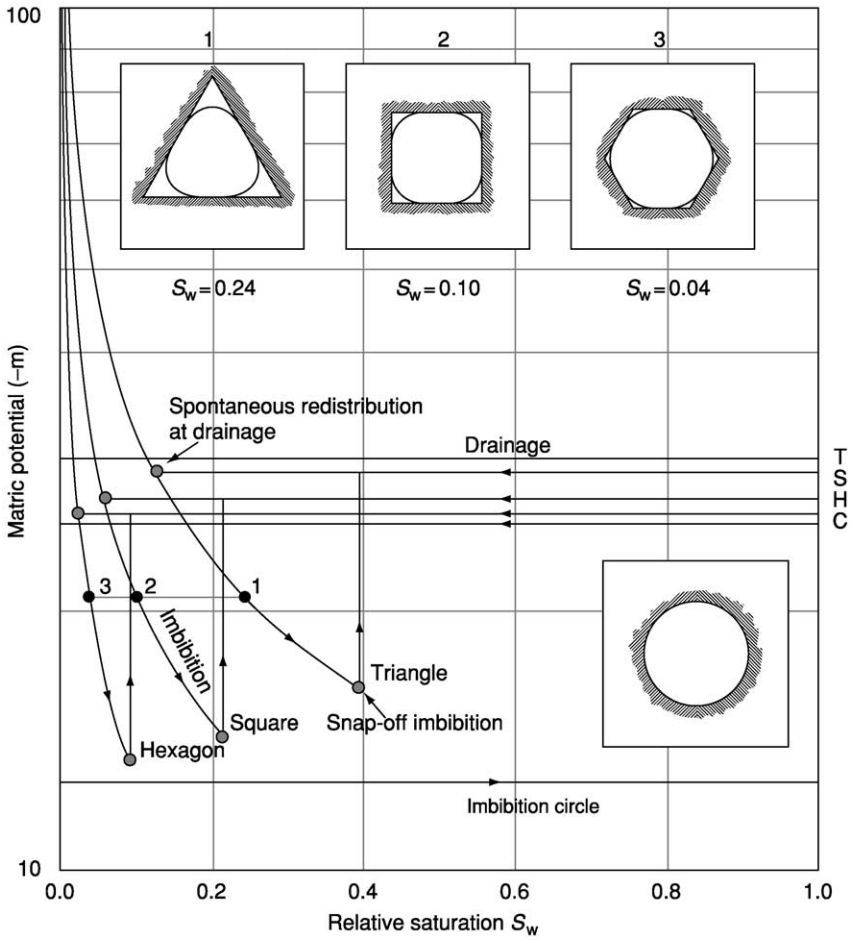


Figure 4 Capillary phenomena (imbibition and drainage) in pores with different cross-section shapes but identical pore cross-sectional area. T, triangle; S, square; H, hexagon; C, circle.

Modeling SWC

Measuring the SWC is laborious and time-consuming. Measured θ–ψ_m pairs are often fragmentary, and usually constitute relatively few measurements over the wetness range of interest. For modeling and analysis purposes, and for characterization and comparison between different soils and scenarios, it is therefore beneficial to represent the SWC in a mathematically continuous form. Several approaches, ranging from empirical parametric expressions to physically based models with parameters derived from measurable medium properties, to pore network or lattice Boltzmann simulations can be employed to represent a continuous SWC.

Empirical SWC Models

Key requirements for all parametric SWC expressions are parsimony (as few parameters as possible), to simplify parameter estimation, and accurate description of SWC behavior at the limits (wet and dry ends) while closely fitting the nonlinear shape of θ–ψ_m measurements.

An effective and commonly used parametric model for relating water content or effective saturation (Θ) to the matric potential was proposed by van Genuchten and is denoted as VG:

$$\Theta = \frac{\theta - \theta_r}{\theta_s - \theta_r} = \left[\frac{1}{1 + (\alpha\,\psi_m)^n}\right]^m \qquad [6]$$

θ_r and θ_s are the residual and saturated water contents, respectively, ψ_m is matric potential, and α, n, and m are parameters directly dependent on the shape of the $\theta(\psi_m)$ curve. A common simplification is to assume that $m = 1 - 1/n$. Thus the parameters required for estimation of the model are θ_r, θ_s, α, and n. θ_s is sometimes known and is easy to measure, leaving only the three unknown parameters θ_r, α, and n to be estimated from the experimental data in many cases. Note that θ_r is sometimes taken as θ at -1.5 MPa, $\theta_{air\ dry}$, or a similar meaningful

value, though it is often advantageous to use it as a fitting parameter.

Another well-established parametric model is denoted by BC (Brooks–Corey):

$$\Theta = \frac{\theta - \theta_r}{\theta_s - \theta_r} = \left(\frac{\psi_b}{\psi_m}\right)^{\lambda} \qquad \psi_m > \psi_b$$
$$\Theta = 1 \qquad \psi_m \leq \psi_b \qquad [7]$$

where ψ_b is a parameter related to the soil matric potential at air entry (b represents 'bubbling pressure'), and λ is related to the soil pore-size distribution. Matric potentials are expressed as positive quantities (i.e., in absolute values) in both VG and BC parametric expressions. Campbell uses the same power law function as BC to express degree of saturation ($S = \theta/\theta_s$) as a function of the air entry potential ψ_b and a factor b that can be related to soil texture:

$$\frac{\theta}{\theta_s} = \left(\frac{\psi_b}{\psi_m}\right)^{\frac{1}{b}} \qquad [8]$$

ψ_b and b are derived as functions of the geometric mean diameter and geometric standard deviation of the particle-size distribution. Equation [8] is often used in relation to the fractal idealization of the soil porous system as discussed below.

Estimation of VG or BC parameters from experimental data requires sufficient data points to characterize the shape of the SWC, and a program to perform nonlinear regression. Many computer spreadsheet software packages provide relatively simple and effective mechanisms to perform nonlinear regression.

Figure 5 depicts fitted parametric VG and BC models to silt loam $\theta(\psi_m)$ data. The resulting best-fit parameters for the VG model are: $\alpha = 0.417\,m^{-1}$; $n = 1.75$; $\theta_s = 0.513\,m^3\,m^{-3}$; and $\theta_r = 0.05\,m^3\,m^{-3}$ (with $r^2 = 0.99$). For the BC model, the best-fit parameters are: $\lambda = 0.54$; $\psi_b = 1.48\,m$; $\theta_s = 0.513\,m^3\,m^{-3}$; and $\theta_r = 0.03\,m^3\,m^{-3}$ (with $r^2 = 0.98$). Note that the most striking difference between the VG and the BC models is the discontinuity at $\psi = \psi_b$ in the BC model.

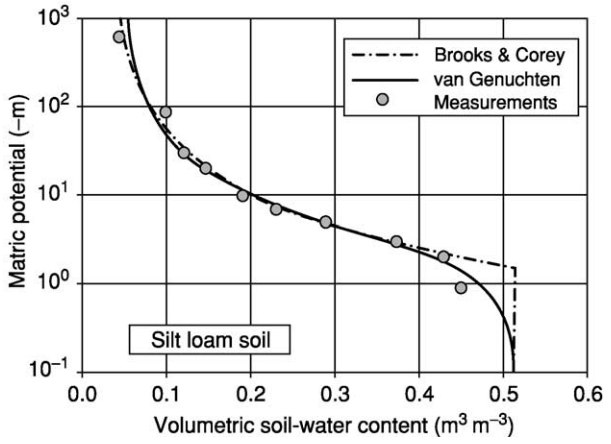


Figure 5 van Genuchten and Brooks and Corey parametric models fitted to measured data for silt loam soil.

Representative measured SWC information based on the VG and BC parametric models is available from a variety of sources. The UNSODA computer database compiled by the US Salinity Laboratory contains an exhaustive collection of soil-water retention and unsaturated hydraulic conductivity information for soils of different textures from around the world. While the authors have attempted to provide some indices concerning quality or reliability of the compiled data, users are advised (as always) to use their own experience and discretion in adapting others' data to their own applications. Regression studies provide a wealth of information on the BC parameter values for many soils. These include estimation of the hydraulic parameters based on other, often more easily available soil properties. These estimates may be sufficiently accurate for some applications and could be used to obtain first-order approximations. Table 1 contains listed values for the VG parameters α and n, and the residual and saturated water contents for various soil textural classes compiled from the UNSODA database. Note that substantial variation in SWC relationships for given soil textural classes is to be expected.

The models introduced so far can be categorized as empirical curve-fitting functions with free model parameters related to the specific shape of the employed

Table 1 Typical van Genuchten model parameters (α, n) including residual (θ_r) and saturated (θ_s) water contents compiled from the UNSODA database

Textural class	*N*	θ_r ($cm^3\,cm^{-3}$)	θ_s ($cm^3\,cm^{-3}$)	α (cm^{-1})	n
Sand	126	0.058	0.37	0.035	3.19
Loamy sand	51	0.074	0.39	0.035	2.39
Sandy loam	78	0.067	0.37	0.021	1.61
Loam	61	0.083	0.46	0.025	1.31
Silt	3	0.123	0.48	0.006	1.53
Silt loam	101	0.061	0.43	0.012	1.39
Sandy clay loam	37	0.086	0.40	0.033	1.49
Clay loam	23	0.129	0.47	0.030	1.37
Silty clay loam	20	0.098	0.55	0.027	1.41
Silty clay	12	0.163	0.47	0.023	1.39
Clay	25	0.102	0.51	0.021	1.20

N, the number of soils or samples of a given textural class from which the mean values are compiled.

Reproduced from Leij FJ, Alves WJ, van Genuchten MT, and Williams JR (1996) *The UNSODA Unsaturated Hydraulic Database*. EPA/600/ R-96/095. Cincinnati, OH: US Environmental Protection Agency.

mathematical function, rather than to physical properties of the porous medium. Relationships between the SWC and the pore-size distribution have been established, which can be described as a statistical log-normal distribution function. Brutsaert, and more recently Kosugi, established relationships between the SWC and the pore-size distribution, which can be described as a statistical log-normal distribution function. They idealized the soil porous system as a BCC with log-normal distributed radii, and apply the capillary-rise equation to establish a relationship between the matric potential and effective medium saturation. (*See* **Capillarity.**)

The applicability of the latter models is limited to coarse-textured soils and to conditions where capillary forces are the major contributor to the matric potential. They tend to fail in finer-textured soils with high specific surface area (e.g., clay soils) where adsorptive surface forces dominate, especially under drier conditions. (*See* **Water Potential.**)

Fractal Representation of the Soil Pore Space and the SWC

Fractals are hierarchical, often highly complex, spatial, or temporal systems that are generated with iterative algorithms obeying simple scaling rules. Patterns within such systems repeat themselves over a defined range of scales (self-similarity). This enables the reproduction of statistical properties of a particular pattern at other length or time scales. Fractal geometry can be applied to describe quantitatively irregularity and shape of natural objects by estimating their fractal dimension. Several theoretical models have been proposed to derive the SWC from the fractal representations of the soil porous system. There are two general approaches, based on either surface or mass fractals. Surface fractal models assume that water is only present in the form of adsorbed liquid films on pore surfaces, whereas mass fractal models assume that only capillary water obeying the capillary-rise equation is present within the fractal system. As with the BCC approach, fractal models for the SWC are based on derivation of the pore-size distribution from the fractal structure under consideration, neglecting pore connectivity and topology issues. Crawford presented the following relationship between the mass-fractal dimension (D_m) and the degree of saturation (S):

$$S = \left(\frac{\psi_m}{\psi_b}\right)^{(D_m - d_e)} \quad [9]$$

where ψ_m is the matric potential under consideration, ψ_b is the matric potential at the air entry point, and d_e is the embedding dimension. The embedding dimension equals 2 in two-dimensional systems and 3 in three-dimensional space. Note that the mass-fractal dimension D_m is always less than d_e. Due to the identical functional form, eqn [9] is commonly equated with Campbell's version of the B&C SWC function (eqn [8]) to derive Campbell's b-factor from the fractal dimension ($b = -1/(D_m - d_e)$).

Fractal approaches are limited due to the assumption that fractal scaling is valid over an infinite range of matric potentials. In reality natural porous media have lower and upper scaling limits related to their minimum and maximum pore sizes. Perfect proposed a mass-based model for the SWC that accounts for the finite range of the matric potential. Systems that are fractal over a finite range of scales are called prefractal.

Physically Based Models for the SWC

In contrast to the BCC and fractal approaches, angular pore-space models accommodate both capillary and adsorptive phenomena on internal surfaces. The basic unit pore element (**Figure 6**) is comprised of an angular central pore attached to slit-shaped spaces and is defined by the dimensionless slit-width and slit-length scaling parameters α and β, and the central pore length L. Variation of the cell parameters allows accommodation of a wide range of soil textural and structural classes. Sandy soils, for example, may be represented by a relatively large central pore length, and small values for β (i.e., small specific surface area). Fine-textured soils (e.g., clays) on the other

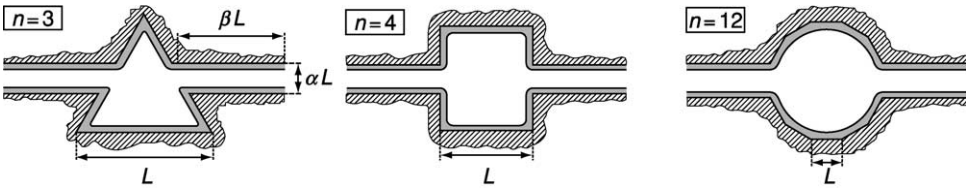


Figure 6 Basic unit elements of the angular pore space model of Tuller *et al.* (Reproduced from Tuller M, Or D, and Dudley LM (1999) Adsorption and capillary condensation in porous media-liquid retention and interfacial configurations in angular pores. *Water Resources Research* 35(7): 1949–1964.)

hand can be represented by relatively small L and large β to account for the high internal surface areas often observed in such soils.

Despite the higher complexity of the angular pore-space model, it has distinct advantages over the more commonly applied BCC approach. First, angular pores more realistically represent natural porous medium pore space and allow dual occupancy of wetting and nonwetting phases (see the section Liquid Retention and Pore Shape, above). The potential of accommodating adsorptive surface forces leads to a more accurate derivation of the SWC for porous media with high specific surface areas as well as under dry conditions.

A modified form of the augmented Young–Laplace equation that considers capillary and adsorptive contributions to the matric potential is employed to calculate liquid–vapor interfaces within a cross-section of the angular pore model. In addition, dynamic liquid displacement mechanisms (see the section Liquid Retention and Pore Shape, above) derive SWC functions for imbibition or drainage at the pore scale. Figure 7 depicts a typical transition of a unit element from dry to wet (imbibition) that includes spontaneous liquid displacement (snap-off) in slits (Figure 7b) and in the central pore (Figure 7d).

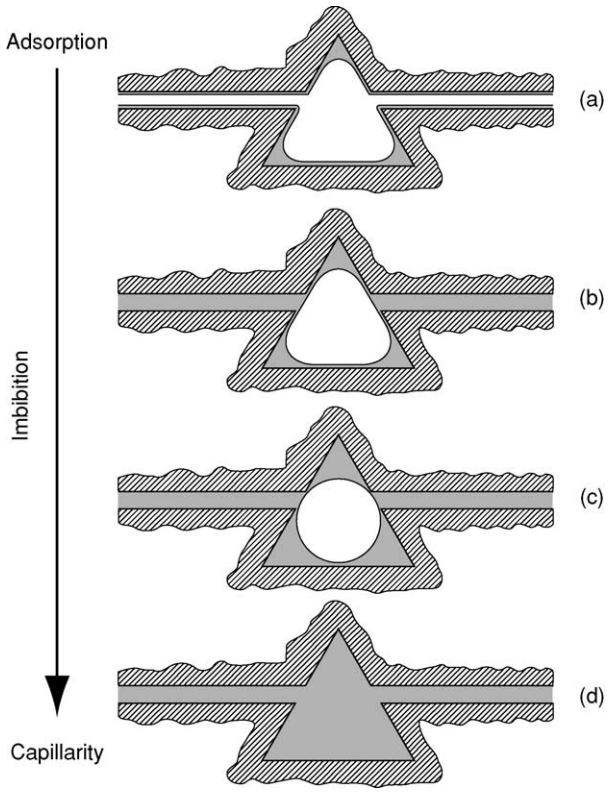


Figure 7 Unit element filling stages during imbibition. (a) Liquid films adsorbed on pore and slit walls and liquid held in corners due to capillary forces at low matric potentials; (b) spontaneous slit fill-up (capillary condensation); (c) pore snap-off; (d) full unit element.

A statistical upscaling scheme using a gamma density function for the central pore length L can be used to represent liquid behavior at the sample scale. The gamma distribution is given preference because it resembles the commonly observed positive skewness of soil pore-size distributions, and at the same time facilitates the derivation of analytical solutions for the SWC.

The upscaling procedure leads to a physically based expression for the SWC, where the degree of saturation (S) defined as a function of matric potential is expressed as the sum of functions corresponding to various pore-filling mechanisms and stages, as depicted in Figure 7. Figure 8a depicts a sample application for the Millville silt loam soil. Because the model calculates the configuration of liquid–vapor interfaces within pores and films, adsorptive and capillary contributions to the SWC can be separated (Figure 8a). Such separation is not possible with any of the empirical or semiempirical approaches presented above. Another important feature of this model is the ability to predict liquid–vapor interfacial area as shown in Figure 8b, comparing model predictions for sand with known surface area of 0.01–0.05 $m^2\,g^{-1}$. The magnitude and changes in liquid vapor interfacial area play an important role in multiphase flow processes, bioremediation of contaminated soils, gas-exchange phenomena, and virus and colloid transport.

Lattice Boltzmann Approach

Lattice Boltzmann models (LBM) are descendants of lattice gas cellular automata, which follow the motions of individual particles and were first presented as a viable means of solving Navier–Stokes equations. In the most simplistic sense, LBMs work with distributions of particles at each lattice point rather than with individual particles. LBMs simulate interactions of hypothetical particles confined to a regular lattice, which greatly enhances solvability of the Boltzmann equation. This has advantages for certain types of simulations in that averaging is not required to obtain smooth velocity fields.

The lattice Boltzmann method has emerged as a powerful tool for simulation of multiphase fluid systems, including water and water vapor. The LBM incorporates complex details of pore shape that characterize realistic porous media, fluid factors, and solid–fluid interactions in a physically sound way, leading to realistic simulation of interfaces between different fluids or between a liquid and its vapor and adsorption of wetting films. Interfaces arise, deform,

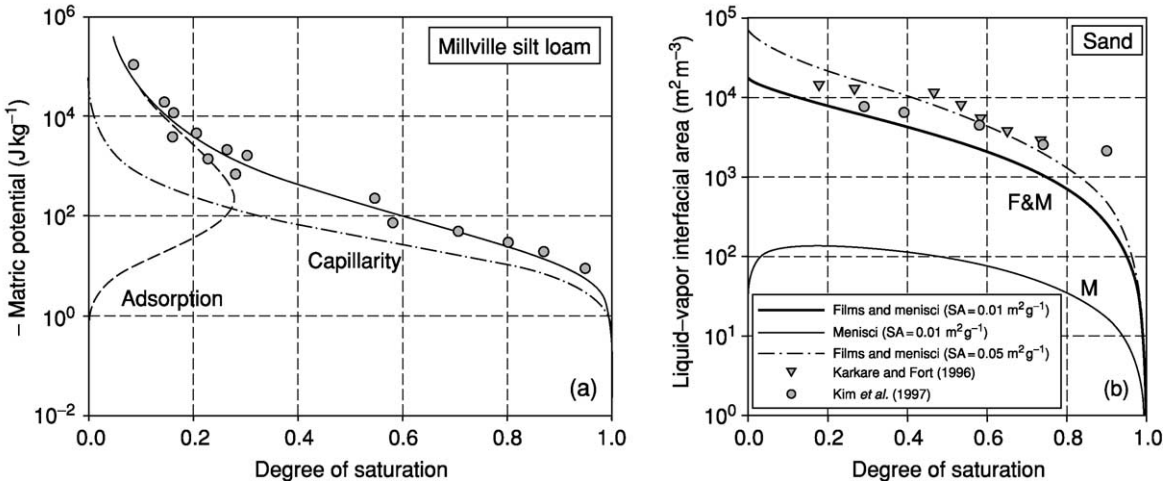


Figure 8 (a) Application of the angular pore-space model for Millville silt loam soil. Note the capillary and absorptive contributions to the soil-water characteristic (SWC); (b) comparison of model-calculated capillary (menisci) and adsorptive (films) contributions to liquid–vapor interfacial area as a function of saturation in an artificial sand mixture and interfacial area measurement results obtained with interfacial tracers. SA, surface area; F&M, films and menisci, Karkare & Fort (1996), Karkare MV and Fort J (1996) Determination of the air–water interfacial area in wet ''unsaturated'' porous media. *Langmuir* 12: 2041–2044; Kim *et al.* (1997), Kim H, Rao PSC, and Annable MD (1997) Determination of effective air–water interfacial area in partially saturated porous media using surfactant adsorption. *Water Resources Research* 33: 2705–2711. (Reproduced from Or D and Tuller M (1999) Liquid retention and interfacial area in variably saturated porous media: upscaling from single-pore to sample-scale model. *Water Resources Research* 35(12): 3591–3606.)

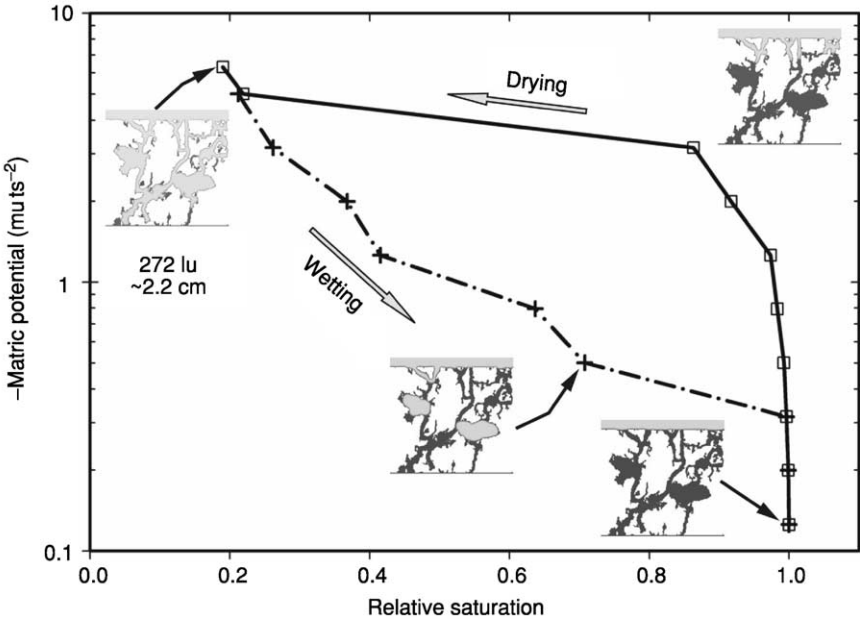


Figure 9 Lattice Boltzmann model simulation of liquid distribution in a complex pore system and computation of soil-water characteristic. (Reproduced from Sukop M and Or D (2003) Lattice Boltzmann method for modeling liquid–vapor interface configurations in porous media. *Water Resources Research* 40(1): W01509.)

and migrate in virtually any pore geometry rather naturally without the need for complex interface tracking algorithms.

An example of LBM application to compute water retention in complex porous media is shown as insets in Figure 9, which are based on two-dimensional (2-D) imagery of a real soil. Vapor and liquid boundaries of equal pressure were applied to the top and bottom of the domain, respectively, in steps corresponding to equal increments of log(−matric

potential). The matric potential is expressed in mass units per time step squared (equivalent to energy per unit volume in a 3-D system). Each potential step was terminated when the relative change in fluid mass in the domain was negligibly small. The simulated liquid behavior and SWC curve show appreciable differences in fluid configurations during wetting and drying that resulted from hysteresis.

Pore Network Models and the SWC

Network models were first developed by Fatt, based on the idea that pore space may be represented as an interconnected network of capillary tubes whose radii represent the dimensions of the pores within a porous medium. For a given matric potential, liquid–vapor interfacial configurations within the network can be determined exactly, based on pore-scale capillary and dynamic displacement considerations. The macroscopic SWC is then determined based on geometric volume averaging of the spatially distributed liquid within the network. A primary advantage of pore network models is explicit consideration of pore connectivity and topology in a simplified and mathematically tractable framework. Some of the limitations involve oversimplification of pore-scale physics (e.g., neglect of adsorptive pore-scale processes), incomplete understanding of interface migration and routing, inadequate technologies for inference of network parameters from real samples, and significant computational burden for detailed 3-D networks even at a core scale (greater than 100 mm).

Hysteresis of the SWC

Water content and the potential energy of soil water are not uniquely related, because the amount of water present at a given matric potential is dependent on the pore-size distribution and the properties of air–water–solid interfaces. An SWC relationship may be obtained by: (1) taking an initially saturated sample and applying suction or pressure to desaturate it (desorption), or (2) gradually wetting an initially dry soil (sorption). These two pathways produce curves that in most cases are not identical; the water content in the 'drying' curve is higher for a given matric potential than that in the 'wetting' branch (Figure 10a). This is called 'hysteresis,' defined as "the phenomenon exhibited by a system in which the reaction of the system to changes is dependent upon its past reactions to change."

The hysteresis in SWC can be related to several phenomena, including: (1) the 'ink bottle' effect, resulting from nonuniformity in shape and sizes of interconnected pores; drainage is governed by the smaller pore radius r, whereas wetting is dependent on the large radius R (Figure 10c); (2) different liquid–solid contact angles for advancing and receding water menisci (Figure 10b); (3) entrapped air in

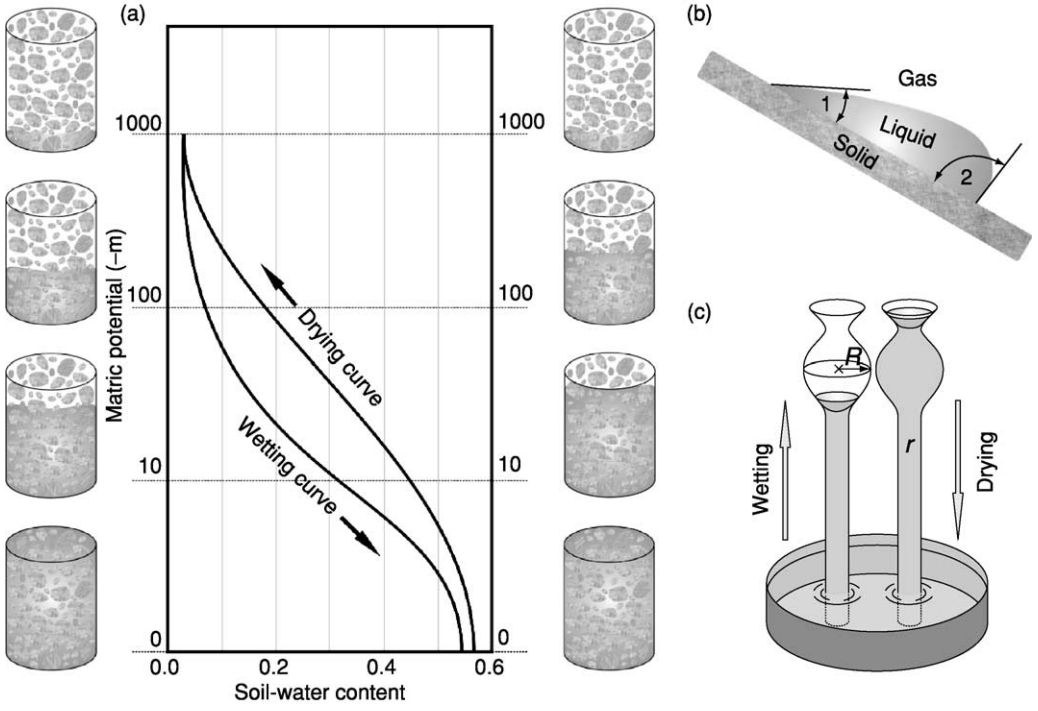


Figure 10 (a) Hysteresis of the soil-water characteristic. (b) The contact angle effect and (c) the ink bottle effect as potential mechanisms for hysteresis.

a newly wetted soil (e.g., pore doublet); and (4) swelling and shrinking of the soil under wetting and drying. The role of individual factors remains unclear and is subject to ongoing research.

Part of the hysteresis phenomena may be attributed to measurement artifacts, for example due to differences between tension- and pressure-induced desaturation. A potentially important aspect of desorption methods under tension is the possibility of liquid displacement (drainage) even in the absence of a continuous gaseous phase due to cavitation initiated by encapsulated gas bubbles or the liquid's own vapor pressure. Surface heterogeneity and impurities in soil and rock water are conducive to lowering the cavitation threshold.

Measurement of SWC Relationships

A variety of methods may be used to obtain requisite θ and ψ_m values to estimate the SWC. Potential experimental problems include: the limited functional range of the tensiometer, which is often used for *in situ* measurements; inaccurate θ measurements in some cases; the difficulty in obtaining undisturbed samples for laboratory determinations; and a slow rate of equilibrium under low matric potential (i.e., dry soils).

In situ methods are preferred in defining SWCs, as is measuring over a wide range of ψ_m and θ values. An effective method to obtain simultaneous measurements of ψ_m and θ_v is by installing time-domain

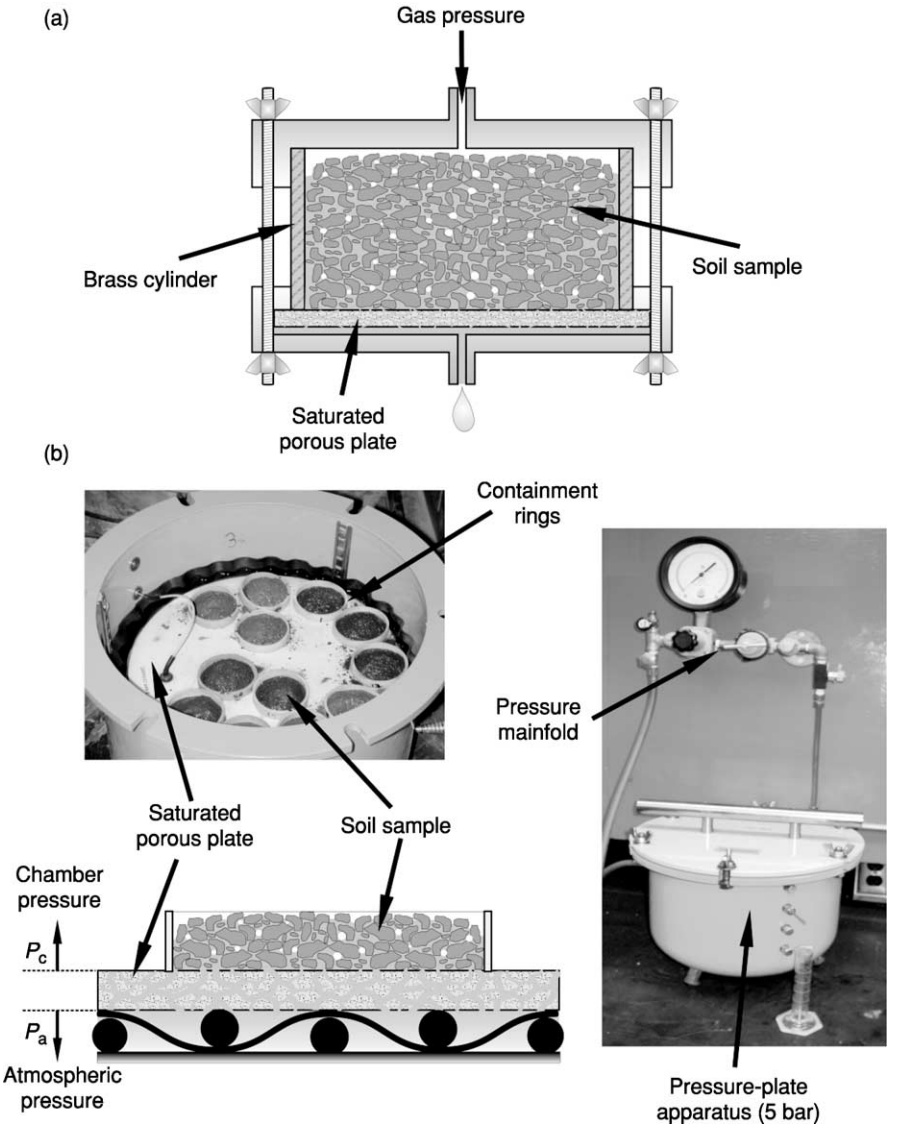


Figure 11 (a) Tempe pressure and flow cell; and (b) pressure-plate apparatus used to desaturate soil samples to specified matric potentials.

reflectometry (TDR) probes in close proximity to transducer tensiometers, with the attributes monitored during variable soil wetness. Large changes in ψ_m and θ_v are expected under highly evaporative conditions near the soil surface, or in the presence of active plant roots.

Pressure-Plate Apparatus and Pressure-Flow Cells

The pressure-plate apparatus comprises a pressure chamber enclosing a water-saturated porous plate, which allows water but not air to flow through its pores (Figure 11b). The porous plate is at atmospheric pressure at the bottom, while the top surface is at the applied pressure of the chamber. Soil samples, usually sieved to less than 2 mm, are placed in retaining rings in contact with the porous plate and allowed to saturate by immersion in water. The porous plate with saturated soil samples is then placed in the chamber and a known N_2 or air gas pressure is applied to force water out of the soil through the plate. Water flows out of the soil until equilibrium between the force exerted by the air pressure and the force by which soil water is being held by the soil (ψ_m) is attained.

Soil water retention at the wet end (less than 1 bar) is strongly influenced by the soil structure and its natural pore-size distribution. Hence, 'undisturbed' intact soil samples (cores) are preferred over repacked samples for this portion of the SWC. The pressure-flow cell (also known as Tempe cell) can hold intact soil samples encased in metal rings (Figure 11a). The operation of the Tempe cell follows that of the pressure plate, except its pressure range is usually limited to 0–1 bar or 0.1 MPa. The porous ceramic material used in pressure plates and flow cells must be completely water-saturated prior to use. Following equilibrium between soil matric potential and the applied air pressure, the soil samples are removed from the apparatus, weighed wet, then oven-dried to determine the mass water content gravimetrically. These may be converted to volume water contents through knowledge of the sample bulk densities. The water content of repacked soils at a given matric potential should not be used to infer θ of intact soils at the same ψ_m, due to modified pore sizes and pore geometry.

Multiple pressure steps may be applied to the same soil samples when using Tempe cells. The cells may be sequentially disconnected from the pressure source, weighed to determine the change in water content from the previous step, then reconnected and the next pressure step applied. Outflow of water from the cells may be collected to calculate or confirm changes in sample water contents.

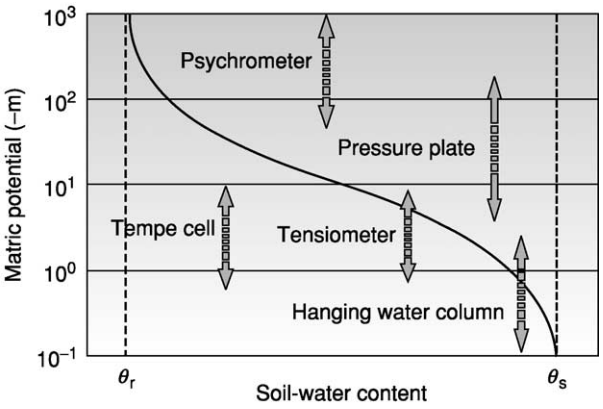


Figure 12 Some common soil-water characteristic measurement methods and their corresponding matric potential ranges.

Paired Sensors for Field SWC Measurement

In spite of the importance of measuring SWCs *in situ*, few suitable methods are available for field or *in situ* application. Paired sensors may be used to measure θ and ψ_m in the same or closely adjacent soil volumes, over a range of soil wetness. Examples include use of neutron moisture-meter access tubes or TDR waveguides, together with tensiometers. Application of paired sensors is often constrained by differences in soil volumes sampled by the respective sensors, different time constants required (e.g., many sensors obtain instantaneous measurements, while many ψ_m methods require equilibrium), and the fact that few matric-potential techniques function over the entire range of interest for wetness. This commonly provides limited overlap in soil water retention measured using different techniques, and measurements obtained using different methods may not be consistent in the ranges of overlap.

Figure 12 presents some common methods to measure or infer soil matric potential, including their respective ranges of application. Many of the available techniques have a limited range of overlap or do not overlap at all, and few of the methods are amenable to field measurements.

See also: **Capillarity**; **Hydrodynamics in Soils**; **Hysteresis**; **Porosity and Pore-Size Distribution**; **Water Content and Potential, Measurement**; **Water Potential**

Further Reading

Ahl C and Niemeyer J (1989) The fractal dimension of the pore volume inside soils. *Zeitschrift für Pflanzenernährung und Bodenkunde* 152: 457–458.

Berkowitz B and Ewing RP (1998) Percolation theory and network modeling applications in soil physics. *Surveys in Geophysics* 19: 23–72.

Bird NRA, Bartoli F, and Dexter AR (1996) Water retention models for fractal soil structures. *European Journal of Soil Science* 47: 1–6.

Brooks RH and Corey AT (1964) *Hydraulic Properties of Porous Media*. Hydrology Papers, No. 3. Fort Collins, CO: Colorado State University Press.

Celia MA, Reeves PC, and Ferrand LA (1995) Recent advances in pore scale models for multiphase flow in porous media. US National Report of the International Union of Geodesy and Geophysics 1991–1994. *Reviews of Geophysics* 33: 1049–1057.

Chen S and Doolen GD (1998) Lattice Boltzmann method for fluid flows. *Annual Review of Fluid Mechanics* 30: 329–364.

Crawford JW (1994) The relationship between structure and the hydraulic conductivity of soil. *European Journal of Soil Science* 45: 493–502.

Dullien FAL (1992) *Porous Media: Fluid Transport and Pore Structure*, 2nd edn. San Diego, CA: Academic Press.

Giménez D, Perfect E, Rawls WJ, and Pachepsky YA (1997) Fractal models for predicting soil hydraulic properties: a review. *Engineering Geology* 48: 161–183.

Haines WB (1930) Studies in the physical properties of soil. V. The hysteresis effect in capillary properties, and the modes of moisture distribution associated therewith. *Journal of Agricultural Science* 20: 97–116.

Kutilek M and Nielsen DR (1994) *Soil Hydrology*. Reiskirchen, Germany: Catena-Verlag.

Leij FJ, Alves WJ, van Genuchten MT, and Williams JR (1996) *The UNSODA Unsaturated Hydraulic Database*. EPA/600/R-96/095. Cincinnati, OH: US Environmental Protection Agency.

Li Y and Wardlaw NC (1986) Mechanisms of nonwetting phase trapping during imbibition at slow rates. *Journal of Colloid Interface Science* 109: 473–486.

Mandelbrot BB (1975) *Les Objets Fractals: Forme, Hasard et Dimension*. Paris, France: Flammarion.

Mandelbrot BB (1982) *The Fractal Geometry of Nature*. San Francisco, CA: Freeman.

Mualem Y (1976) A new model for predicting the hydraulic conductivity of unsaturated porous media. *Water Resources Research* 12(3): 513–522.

Mualem Y (1984) A modified dependent domain theory of hysteresis. *Soil Science* 137: 283–291.

Or D and Tuller M (1999) Liquid retention and interfacial area in variably saturated porous media: upscaling from single-pore to sample-scale model. *Water Resources Research* 35(12): 3591–3606.

Or D and Tuller M (2002) Cavitation during desaturation of porous media under tension. *Water Resources Research* 38(5): 19.

Or D, Groeneveld DP, Loague K, and Rubin Y (1991) *Evaluation of Single and Multi-parameter Methods for Estimating Soil-Water Characteristic Curves*. Geotechnical Engineering Report No. UCB/GT/91-07. Berkeley, CA: University of California Press.

Pachepsky YA, Shcherbakov RA, and Korsunskaya LP (1995) Scaling of soil water retention using a fractal model. *Soil Science* 159: 99–104.

Perfect E (1999) Estimating soil mass fractal dimensions from water retention curves. *Geoderma* 88: 221–231.

Rawls WJ and Brakensiek DL (1989) Estimation of soil water retention and hydraulic properties. In: Morel-Seytoux HJ (ed.) *Unsaturated Flow in Hydraulic Modeling Theory and Practice*, pp. 275–300. NATO ASI Series. Series C: Mathematical and physical sciences, No. 275.

Sahimi M (1995) *Flow and Transport in Porous Media and Fractured Rock*. Weinheim, Germany: VCH-Verlag.

Scholle PA (1979) *A Color Illustrated Guide to Constituents, Textures, Cements, and Porosities of Sandstones and Associated Rocks*. American Association of Petroleum Geologists Memorandum 28. Julsa, OK: American Association of Petroleum Geologists.

Tuller M and Or D (2001) Hydraulic conductivity of variably saturated porous media – laminar film and corner flow in angular pore space. *Water Resources Research* 37(5): 1257–1276.

Tuller M, Or D, and Dudley LM (1999) Adsorption and capillary condensation in porous media-liquid retention and interfacial configurations in angular pores. *Water Resources Research* 35(7): 1949–1964.

Tyler SWI and Wheatcraft SW (1990) Fractal processes in soil water retention. *Water Resources Research* 26: 1047–1054.

van Genuchten MT (1980) A closed-form equation for predicting the hydraulic conductivity of unsaturated soils. *Soil Science Society of America Journal* 44: 892–898.

Water Table *See* **Groundwater and Aquifers**

WATER, PROPERTIES

D Hillel, Columbia University, New York, NY, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Our planet is the planet of life primarily because it is blessed with the precise ranges of temperature and pressure that make possible the existence in a liquid state of a singular substance called water. So ubiquitous is water on our globe, covering nearly three-quarters of its surface, that the entire planet really should be called 'Water' rather than 'Earth.' However, as Coleridge's Ancient Mariner complained, most of the water everywhere is unfit to drink. Less than 1% of the water on earth is 'fresh' (i.e., non-saline) water, and that amount is unevenly distributed. Humid regions are endowed with an abundance of it, even with a surfeit, so that often the problem is how to dispose of excess water. Arid and semiarid regions, on the other hand, are afflicted with a chronic shortage.

Life as we know it began in an aquatic medium and water is still the principal constituent of living organisms. It is, literally, the essence of life. As Vladimir Vernadsky wrote a century ago: "Life is animated water." Though we appear to be solid, we are really liquid bodies, similar to gelatin, which also seems solid but is in fact largely water, made consistent by the presence of organic material. The analogous material in our bodies is protoplasm. The water content of a newborn infant is nearly 90% water by mass, and even in adults it is over 65%. Actively growing herbaceous plants typically contain over 90% water. Far from being a bland, inert liquid, water is a highly reactive substance, a solvent and a transporter of numerous substances.

The importance of water was recognized early in history, yet little was known about its real nature. In the Middle Ages, people believed that fresh water emanated magically from the bowels of the earth. They could not imagine that all the water flowing in innumerable springs and mighty rivers (such as the Nile, which appeared to the ancient Egyptians to come out of the driest desert!) could possibly result from so seemingly feeble a source as rain and snow. The first to conjecture this was Leonardo da Vinci, but only in the latter part of the seventeenth century did the English astronomer Edmond Halley and, separately, the Frenchman Claude Perrault prove the principle by calculation and measurement. Water was long thought to be a single element, until early in the eighteenth century, when it was found to consist of hydrogen and oxygen in combination.

Notwithstanding its ubiquity, water remains something of an enigma, possessing unusual and anomalous attributes. Perhaps the first anomaly is that, being a compound of two gases and having relatively low molecular weight, water is a liquid and not a gas at normal temperatures. (Its sister compound, hydrogen sulfide, H_2S, has a boiling-point temperature of $-60.7°C$.) Compared with other common liquids, water has unusually high melting and boiling points, heats of fusion and vaporization, specific heat, dielectric constant, viscosity, and surface tension.

Molecular Structure

One cubic meter of liquid water at 20°C contains about 3.4×10^{28} (34 billion billion billion) molecules, the diameter of which is about 3×10^{-10} meter ($3\times10^{-4}\,\mu$m, or about 3 Angstrom units). The chemical formula of water is H_2O, which signifies that each molecule consists of two atoms of hydrogen and one of oxygen. There are three isotopes of hydrogen (^{1}H, ^{2}H, ^{3}H) and three of oxygen (^{16}O, ^{17}O, ^{18}O), which can form 18 combinations. However, all isotopes but ^{1}H and ^{16}O are quite rare.

The hydrogen atom consists of a positively charged proton and a negatively charged electron. The oxygen atom consists of a nucleus having a positive charge of eight protons, surrounded by eight electrons, of which six are in the outer shell. Since the outer electron shell of hydrogen lacks one electron and that of oxygen lacks two electrons, one atom of oxygen can combine with two atoms of hydrogen in an electron-sharing molecule.

The strong intermolecular forces in liquid water are caused by the electrical polarity of the water molecule, which in turn is a consequence of the arrangement of electrons in its oxygen and hydrogen atoms (Figure 1). The oxygen atom shares a pair of electrons with each of the two hydrogen atoms, through overlap of the 1s orbitals of the hydrogen atoms with two hybridized sp^3 orbitals of the oxygen atom. The H–O–H bond in water is not linear but bent

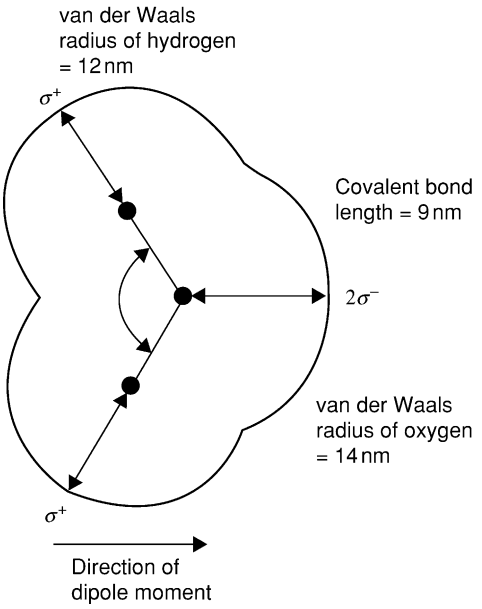


Figure 1 Model of a water molecule. The curved lines represent the borders at which van der Waals attractions are counterbalanced by repulsive forces. Reprinted from *Environmental Soil Physics*, Hillel D (ed.). Copyright (1998), with permission from Elsevier.

at an angle of 104.5°. That angle deviates slightly from a perfectly tetrahedral arrangement of the oxygen atom's four possible sp^3 orbitals, which would have an angle of 109.5°. The mean H-O interatomic distance is 9.65×10^{-5} μm. The arrangement of electrons in the molecule gives it electrical asymmetry. The electronegative oxygen atom tends to attract the single electrons of the hydrogen atoms, leaving the hydrogen nuclei bare. Hence, each of the two hydrogen atoms has a local partial positive charge. The oxygen atom, in turn, has a partial negative charge, located in the zone of the unshared orbitals. Thus, though the water molecule has no net charge, it forms an electrical dipole.

Hydrogen Bonding

Every hydrogen proton, while attached primarily to a particular molecule, is also attracted to the oxygen of the neighboring molecule, with which it forms a secondary link known as a hydrogen bond. Though the intermolecular link resulting from dipole attraction is not as strong as the primary link of the hydrogen to the oxygen of its own molecule, water can be regarded as a polymer of hydrogen-bonded molecules. This structure is most complete in ice crystals, in which each molecule is linked to four neighbors via four hydrogen bonds, thus forming a hexagonal lattice with a rather open structure (Figure 2). When the ice melts, this rigid structure collapses partially, so additional molecules can

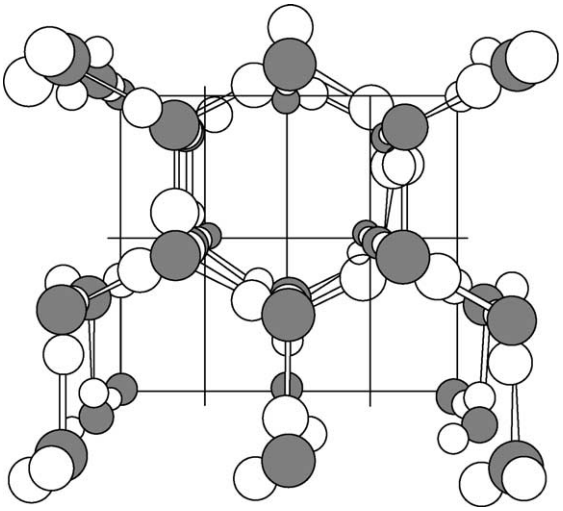


Figure 2 Schematic structure of an ice crystal. The oxygen atoms are shown in gray and the hydrogen atoms in white. The pegs linking adjacent molecules represent hydrogen bonds. Reprinted from *Environmental Soil Physics*, Hillel D (ed.). Copyright (1998), with permission from Elsevier.

enter the intermolecular spaces and each molecule can have more than four near neighbors. For this reason, liquid water can be more dense than ice at the same temperature, and thus lakes and ponds develop a surface ice sheet in winter rather than freeze solid from bottom to top as they would if ice were denser than liquid water.

States of Water

In the vapor or gaseous state, water molecules are largely independent of one another and occur mostly as monomers signified as $(H_2O)_1$. Occasionally, colliding molecules may fuse to form dimers $(H_2O)_2$ or even trimers, $(H_2O)_3$, but such combinations are rare. However, in the solid state a rigidly structured lattice forms with a tetrahedral configuration (Figure 2) that can be schematically depicted as sheets of puckered hexagonal rings (Figure 3). As many as nine alternative ice forms can occur when water freezes, depending on prevailing temperature and pressure conditions. Figure 3 pertains to ice 1, the familiar form, which occurs and is stable at ordinary atmospheric pressure.

The orderly structure of ice does not totally disappear in the liquid state. The polarity and hydrogen bonds continue to bind water molecules together, though the structural forms that develop in the liquid state are much more flexible and transient than in the rigidly structured solid state. Hydrogen bonds in liquid water form an extensive three-dimensional network, the detailed features of which appear to be short-lived. According to the 'flickering cluster' model discovered by Frank and Wen and modified

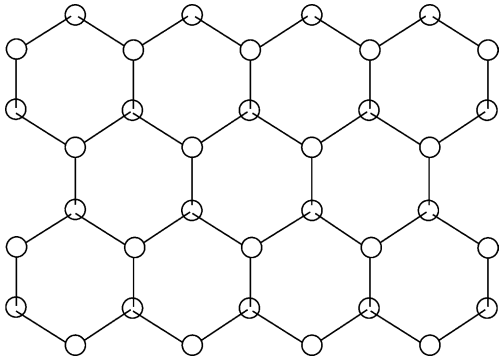


Figure 3 The crystalline structure of ice. Reprinted from *Environmental Soil Physics*, Hillel D (ed.). Copyright (1998), with permission from Elsevier.

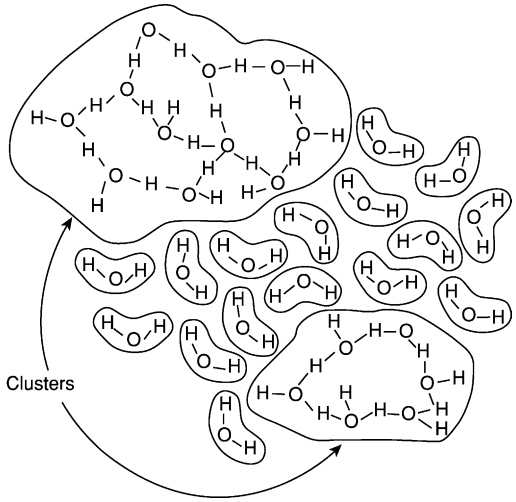


Figure 4 Schematic illustration of 'flickering clusters,' showing polymeric associations and monomeric molecules in liquid water. Reprinted from *Environmental Soil Physics*, Hillel D (ed.). Copyright (1998), with permission from Elsevier.

by Erland, the molecules of liquid water associate and dissociate repeatedly in transitory or flickering polymer groups, designated $(H_2O)_n$, having a quasi-crystalline internal structure. These microcrystals, as it were, form and melt so rapidly on the scale of picoseconds and randomly that, on a macroscopic scale, water appears to behave as a homogeneous liquid (Figure 4).

In transition from solid to liquid, and from liquid to gas, hydrogen bonds must be broken (while in freezing and condensation they are re-established). Hence relatively high temperatures and energies are required to achieve these transitions. To thaw 1 kg of ice, 3.35×10^5 J (80 cal g^{-1}) must be supplied. Conversely, the same energy (the latent heat of fusion) is released in freezing. At the boiling point (100°C at atmospheric pressure), water passes from the liquid to the gaseous state and in so doing it absorbs 2.26×10^6 J kg^{-1} (540 cal g^{-1}). This amount of heat is known as the latent heat of vaporization. Water can be vaporized at temperatures below 100°C, but such vaporization requires greater heat or lower atmospheric pressure. At 30°C, the latent heat is about 2.43×10^6 J kg^{-1} (580 cal g^{-1}). Sublimation is the direct transition from the solid state to vapor, and the heat absorbed by it is equal to the sum of the latent heats of fusion and of vaporization.

Ionization and pH

Because of its small mass and the tightness with which its single electron is bound to the oxygen atom, the nucleus of the hydrogen atom in the water molecule exhibits a finite tendency to dissociate from the oxygen with which it is covalently associated and to 'jump' to the adjacent water molecule, to which it is hydrogen-bonded. Such an event produces two ions: the hydronium ion (H_3O^+) and the hydroxyl ion (OH^-). The reaction described is reversible, and should be written as:

$$2H_2O \rightleftharpoons (H_3O)^+ + OH^-$$

However, by convention it is written simply as:

$$H_2O \rightleftharpoons H^+ + OH^-$$

and one speaks of 'hydrogen ions' rather than of 'hydronium ions.'

Although the self-ionization of water is small, its consequences are extremely important. The ionization is reversible, and it tends to an equilibrium state in which the rate of dissociation into ions equals the rate of ion reassociation to form molecules once again. For such a system in equilibrium (at which the concentration of each of the species H_2O, H^+, and OH^- remains constant), the law of mass action applies; i.e., the ratio of concentrations of the products and the reactants must be constant. Using brackets to denote concentration, we can write this in the following way:

$$K_{equil} = [H^+][OH^-]/[H_2O] \qquad [1]$$

Since the number of water molecules undergoing dissociation at any given time is very small relative to the total number of water molecules present, $[H_2O]$ can be considered constant. Assuming this concentration to be 55.5 mol l^{-1} (the number of grams per liter divided by the gram molecular weight: 1000/18 = 55.5 mol l^{-1}), we can simplify the equilibrium constant expression as follows:

$$55.5 \times K_{equil} = [H^+][OH^-], \text{or } K_w = [H^+][OH^-] \qquad [2]$$

in which K_w is a composite constant called the ion product of water. In fact, the concentrations of H^+ and OH^- ions in pure water at 25°C are 10^{-7} mol l^{-1}, an extremely small value when compared to the overall concentrations of (largely undissociated) water, namely 55.5 mol l^{-1}. Thus, K_w at 25°C is 10^{-14}. If the hydroxyl ion concentration $[OH^-]$ is changed, the hydrogen ion concentration $[H^+]$ changes automatically to maintain the constancy of the product, and vice versa. An excess concentration of hydrogen ions over the concentration of hydroxyl ions imparts to the aqueous medium the property of acidity, whereas a predominance of hydroxyl ions produces the opposite property of alkalinity or basicity. A condition in which the concentrations of H^+ and OH^- are equal is called neutrality.

The ion product of water, K_w, is the basis for the pH scale, a measure of the concentration of H^+ (and of OH^- as well) in any aqueous solution in the range of concentration between 1.0 mol l^{-1} H^+ and 1.0 mol l^{-1} OH^-. The pH is defined as

$$\mathrm{pH} = \log_{10} 1/[H^+] = -\log_{10}[H^+] \qquad [3]$$

As already stated, in a precisely neutral solution at 25°C,

$$[H^+] = [OH^-] = 1.0 \times 10^{-7} \mathrm{mol\,l^{-1}}$$

The pH of such a solution is

$$\mathrm{pH} = \log_{10}[1/(1 \times 10^{-7})] = 7.0$$

The pH value of 7.0 for a neutral solution is thus not arbitrary, but derives from the value of the ion product of water at 25°C. The pH scale is logarithmic, not arithmetic. If two solutions differ in pH by one unit, then one solution has 10 times the hydrogen ion concentration of the other. Thus, a pH of 6 implies a hydrogen ion concentration of 10^{-6} and a hydroxyl ion concentration of 10^{-8}. A pH of 5 indicates $[H^+] = 10^{-5}$ (acidity 10 times that of the above) and $[OH^+] = 10^{-9}$.

Solvent Properties of Water

Water dissolves or disperses many substances because of its polar nature. Hence it has been called the universal solvent. All chemical substances have finite solubilities in water, but these solubilities range widely. Many crystalline salts and other ionic compounds readily dissolve in water but are nearly insoluble in nonpolar liquids such as chloroform or benzene. Since the crystal lattice of salts, such as sodium chloride, is held together by very strong electrostatic attractions between alternating positive and negative ions, considerable energy is required to pull these ions away from one another. However, water dissolves sodium chloride because the strong electrostatic attraction between water dipoles and the Na^+ and Cl^- ions, forming stable hydrated Na^+ and Cl^- ions, exceeds the attraction of these ions to each other. In the case of Na^+, hydration is represented by the process:

$$Na^+ + (n+4)H_2O - Na(H_2O)_4(H_2O)_n^+$$

illustrated in Figure 5. In addition to hydration, there is also the hydrolysis of metal species, a reaction in which the metal ion displaces one of the protons (hydrogen) of water to form basic hydroxides.

Ion solvation is also aided by the tendency of the solvent to oppose the electrostatic attraction between ions of opposite charges. This is characterized by the dielectric constant D, which is defined by the relationship:

$$F = \kappa e_1 e_2 / \mathrm{D} r^2 \qquad [4]$$

Here F is the attractive force between two ions of opposite charge, κ is Coulomb's constant, e_1 and e_2 are charges on the ions, and r is the distance between them. Water has an extremely high dielectric constant, as shown in Table 1. For instance, the attractive force between Na^+ and Cl^- ions at a given distance in water is less than one-third that in ethanol and only

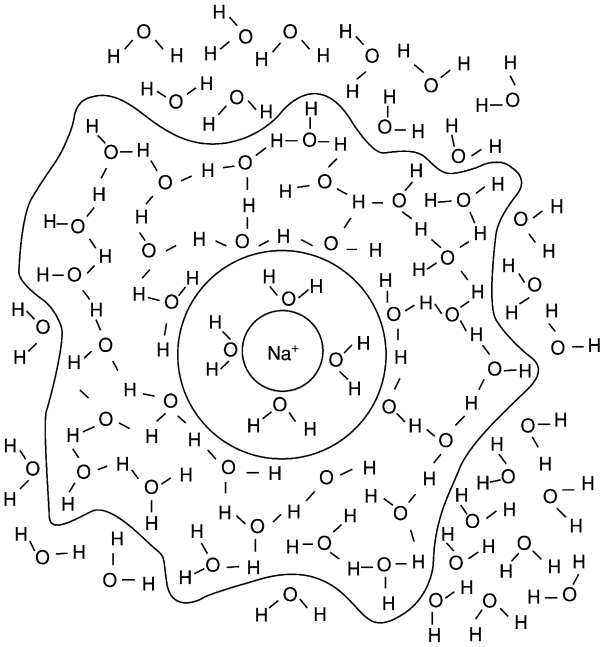


Figure 5 A model of the hydration 'atmosphere' of a sodium ion. An inner shell of more or less rigidly structured water is surrounded by a cluster of looser but still structure-enhanced water, the whole floating in a sea of 'free' water. Reprinted from *Environmental Soil Physics*, Hillel D (ed.). Copyright (1998), with permission from Elsevier.

Table 1 Dielectric constants of some liquids (20°C)

Water	80	Acetone	21.4
Methanol	33	Benzene	2.3
Ethanol	24	Hexane	1.9

Reprinted from *Environmental Soil Physics*, Hillel D (ed.). Copyright (1998), with permission from Elsevier.

one-40th that in benzene. This fact greatly facilitates hydration of ions and dissolution of the crystal lattices of salts in water.

The effect of a solute on the solvent is manifest in a set of properties, namely, the colligative properties of solutions, which depend on the number of solute particles per unit volume of solvent. Solutes produce such characteristic effects in the solvent as depression of the freezing point and of the vapor pressure, and elevation of the boiling point. They also endow a solution with the property of osmotic pressure. Theoretically, 1 mol of an ideal solute dissolved in 1 kg of water at a pressure of 760 mm of mercury (0.1 MPa, or 1 bar) depresses the freezing point by 1.86°C and elevates the boiling point by 0.543°C. Such a solution also yields an osmotic pressure of 2.24 MPa (22.4 atm) in an appropriate apparatus. However, aqueous solutions usually deviate considerably from ideal behavior, and the deviations are greater the higher the concentrations. The quantitative relationships given above are exact only at infinite dilution.

Osmotic Pressure

Owing to the constant thermal motion of all molecules in a fluid (above a temperature of absolute zero), solute species spread throughout the solution in a spontaneous tendency toward a state of equal concentration throughout. This migration of solutes in response to spatial differences in concentration is called diffusion.

If a physical barrier is interposed between two regions, across the path of diffusion, and if that barrier is permeable to molecules of the solvent but not to those of the solute, the former will diffuse through the barrier in a process called osmosis (from the Greek $\omega\sigma\mu o\sigma$, meaning 'push'). As in the case of unhindered diffusion, this process tends toward a state of uniform concentration even across the barrier. Barriers permeable to one substance in a solution but not to another are called selective or semipermeable membranes. Membranes surrounding cells in living organisms, for example, exhibit selective permeability to water while restricting the diffusion of solutes between the cells' interior and their exterior environment. Water molecules cross the membrane in both directions, but the net flow of water is from the more dilute solution to the more concentrated.

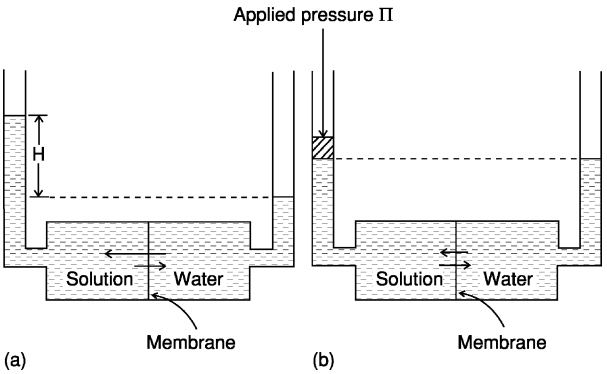


Figure 6 Osmosis and osmotic pressure: (a) in osmosis, the flow of water molecules through the membrane into the solutions is at first greater than the reverse flow from the solution into the water compartment. The hydrostatic pressure due to the column of expanded solution increases the rate of water flow from the solution to the water compartment until, at equilibrium, the opposite flows are equal. (b) The osmotic pressure of the solution is equal to the hydrostatic pressure Π, which must be applied to the solution to equalize the rate of flow to and from the solution and produce a net flow of zero. Reprinted from *Environmental Soil Physics*, Hillel D (ed.). Copyright (1998), with permission from Elsevier.

Figure 6 is a schematic representation of a pure solvent separated from a solution by a semipermeable membrane. Solvent will pass through the membrane and enter the solution compartment, driving the solution level up the left-hand tube until the hydrostatic pressure of the column of dilute solution on the left is sufficient to counter the diffusion pressure of the solvent molecules drawn into the solution through the membrane. The hydrostatic pressure at equilibrium, when solvent molecules are crossing the membranes in both directions at equal rates, is the osmotic pressure of the solution.

In dilute solutions, the osmotic pressure is proportional to the concentration of the solution and to its temperature according to the following equation:

$$\Pi = MRT \qquad [5]$$

Here Π is the osmotic pressure in atmospheres (to be multiplied by 0.101 to obtain megapascal units), M the total molar concentration of the solute (whether molecules or dissociated ions), T the temperature in degrees Kelvin, and R the gas constant (0.08205 liter atm deg^{-1} mole). The osmotic pressure increase with temperature is associated with the corresponding increase of the molecular diffusivity (self-diffusion coefficient) of water, D_w. According to the Einstein–Stokes equation,

$$D_w = kT/6\pi\eta$$

where $k = R/N$, the Boltzmann constant (1.38×10^{-23} JK^{-1}); r is the rotation radius of the molecule ($\sim 1.5 \times 10^{-4}$) and η is the viscosity.

Solubility of Gases

The concentration of dissolved gases in water in equilibrium with a gaseous phase generally increases with pressure and decreases with temperature. According to Henry's law, the mass concentration of a dissolved gas c_m is proportional to the partial pressure of the gas p_i:

$$c_m = s_o p_i / p_o \quad [6]$$

where s_o is the solubility coefficient of the particular gas in water and p_o is the total pressure of the atmosphere. The volume concentration is similarly proportional:

$$c_v = s_v p_i / p_o \quad [7]$$

where s_v is the solubility coefficient expressed in terms of volume ratios (i.e., c_v is the volume of dissolved gas relative to the volume of water). The values of s_c and s_v are determined experimentally. If the gas does not react chemically with the liquid, these properties should remain constant over a range of pressures, especially at low partial pressures of the dissolved gases. Solubility is, however, strongly influenced by temperature. Table 2 gives the s_v values of several atmospheric gases at various temperatures.

The solubilities of various gases (particularly oxygen) in varying conditions strongly influence such vital soil processes as oxidation and reduction, and respiration by roots and microorganisms.

Adsorption of Water on Solid Surfaces

Adsorption is an interfacial phenomenon resulting from the differential forces of attraction or repulsion occurring among molecules or ions of different phases at their exposed surfaces. As a result of cohesive and adhesive forces coming into play, the zones of contact among phases may exhibit a concentration or a density of material different from that inside the phases themselves. As different phases come in contact, various types of adsorption can occur: adsorption of gases on solids, of gases on liquid surfaces, and of liquids on solids.

Table 2 Solubility coefficients of gases in water

Temperature (°C)	*Nitrogen (N_2)*	*Oxygen (O_2)*	*Carbon dioxide (CO_2)*	*Air (without CO_2)*
0	0.0235	0.0489	1.713	0.0292
10	0.0186	0.0380	1.194	0.0228
20	0.0154	0.0310	0.878	0.0187
30	0.0134	0.0261	0.665	0.0156
40	0.0118	0.0231	0.530	—

Reprinted from *Environmental Soil Physics*, Hillel D (ed.). Copyright (1998), with permission from Elsevier.

The interfacial forces of attraction or repulsion may themselves be of different types, including electrostatic or ionic (Coulombic) forces, intermolecular forces such as van der Waals and London forces, and short-range repulsive (Born) forces. The adsorption of water upon solid surfaces is generally of an electrostatic nature. The polar water molecules attach to the charged faces of the solids and to the ions adsorbed on them. This adsorption of water is the mechanism causing the strong retention of water by clay at high suctions.

The interaction of the charges of the solid with the polar water molecules may impart to the adsorbed water a distinct structure in which the water dipoles assume an orientation dictated by the charge sites on the solids. This adsorbed 'phase' may have mechanical properties of strength and viscosity that differ from those of ordinary liquid water at the same temperature. The adsorption of water on clay surfaces is an exothermic process, resulting in the liberation of an amount of heat known as the heat of wetting.

Vapor Pressure

According to the kinetic theory, molecules in a liquid are in constant motion, which is an expression of their thermal energy. These molecules collide frequently, and occasionally one or another at the surface absorbs sufficient momentum to leap out of the liquid and into the atmosphere above it. Such a molecule, by virtue of its kinetic energy, thus changes from the liquid to the gaseous phase. This kinetic energy is then lost in overcoming the potential energy of intermolecular attraction while escaping from the liquid. At the same time, some of the randomly moving molecules in the gaseous phase may strike the surface of the liquid and be absorbed in it.

The relative rates of these two directions of movement depend upon the concentration of vapor in the atmosphere relative to its concentration at a state of equilibrium (i.e., when the movement in both directions is equal). An atmosphere that is at equilibrium with a body of pure water at standard atmospheric pressure is considered to be saturated with water vapor, and the partial pressure of the vapor in such an atmosphere is called the saturation (or equilibrium) vapor pressure. The vapor pressure at equilibrium with any body of water depends on the physical condition of the water (pressure and temperature) and its chemical condition (solutes) but is independent of the absolute or relative quantity of liquid or gas in the system.

The saturation vapor pressure rises with temperature. As the kinetic energy of the molecules in the liquid increases, so does the evaporation rate.

Consequently, a higher concentration of vapor in the atmosphere is required for the rate of return to the liquid to match the rate of escape from it. A liquid arrives at its boiling point when the vapor pressure becomes equal to the atmospheric pressure. If the temperature range is not too wide, the dependence of saturation vapor pressure on temperature is expressible by the equation (Table 3):

$$\ln p_o = a - b/T \qquad [8]$$

where $\ln p_o$ is the logarithm to the base e of the saturation vapor pressure p_o, T is the absolute temperature, and a and b are constants.

As mentioned, the vapor pressure also depends on the hydrostatic pressure of the liquid water. At equilibrium with drops of water (which have a hydrostatic pressure greater than atmospheric), the vapor pressure is greater than in a state of equilibrium with free water (which has a flat interface with the atmosphere). On the other hand, in equilibrium with adsorbed or capillary water under a hydrostatic pressure smaller than atmospheric, the vapor pressure is smaller than in equilibrium with free water. The curvature of drops is considered to be positive, as these drops are convex toward the atmosphere, whereas the curvature of capillary water menisci is considered negative, as they are concave toward the atmosphere.

For water in capillaries, in which the air–water interface is concave, the Kelvin equation applies:

$$-(\mu_1 - \mu_1^o) = RT \ln (p_1^o/p_1) = 2\gamma v_i \cos \alpha / r_c$$

in which $(\mu_1 - \mu_1^o)$ is the change in potential of the water due to the curvature of the air–water interface, γ is the surface tension of water, α is the contact angle, v_1 is the partial molar volume of water, and r_c is the radius of the capillary.

Water present in the soil invariably contains solutes, mainly electrolytic salts, in variable concentrations. Thus, soil water should properly be called the soil solution. The composition and concentration of the soil solution affect soil behavior. While in humid regions the soil solution may have a concentration of but a few parts per million, in arid regions the concentration may become as high as several percent. The ions commonly present are H^+, Ca^{2+}, Mg^{2+}, Na^+, K^+, NH_4^+, OH^-, Cl^-, HCO_3^-, NO_3^-, SO_4^-, and CO_3^-. The vapor pressure of electrolytic solutions is less than that of pure water. The equation is:

$$v_1 \Pi_o = RT \ln (p_1^o/p_1) = \mu_1 - \mu_1^o$$

wherein Π_o is the osmotic pressure of a nonvolatile solute, μ_1^o and p_1^o are the chemical potential and vapor pressure of the liquid in its pure state, and μ_1 and p_1 are the same for the solution. Thus the soil solution has a lower vapor pressure than pure water, even when the soil is saturated. In unsaturated soil the capillary and adsorptive effects further lower the potential and hence also the vapor pressure.

Surface Tension

Surface tension is a phenomenon occurring typically, but not exclusively, at the interface of a liquid and a gas. The liquid behaves as if it were covered by an elastic membrane in a constant state of tension that tends to cause the surface to contract. To be sure, no such membrane exists, yet the analogy is a useful one if not taken too literally. If we draw an arbitrary line of length L on a liquid surface, there will be a force F

Table 3 Physical properties of water vapor

	Saturation vapor pressure (torr)		*Vapor density in saturated air (kg m^{-3})*		
Temperature (°C)	*Over liquid*	*Over ice*	*Over liquid ($\times 10^{-3}$)*	*Over ice ($\times 10^{-3}$)*	*Diffusion coefficient (m^2 s^{-1}) ($\times 10^{-4}$)*
−10	2.15	1.95	2.36	2.14	0.211
−5	3.16	3.01	3.41	3.25	—
0	4.58	4.58	4.85	4.85	0.226
5	6.53	—	6.80	—	
10	9.20	—	9.40	—	0.241
15	12.78	—	12.85	—	
20	17.52	—	17.30	—	0.257
25	23.75	—	23.05	—	
30	31.82	—	30.38	—	0.273
35	42.20	—	39.63	—	
40	55.30	—	51.1	—	0.289
45	71.90	—	65.6	—	
50	92.50	—	83.2	—	

Reprinted from *Environmental Soil Physics*, Hillel D (ed.). Copyright (1998), with permission from Elsevier.

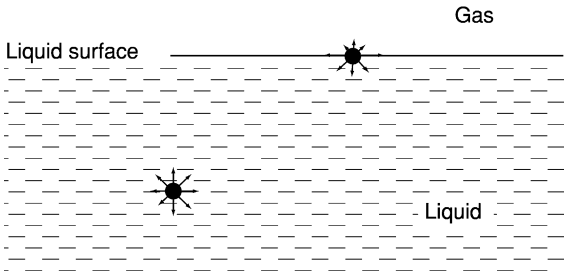


Figure 7 Cohesive forces acting on a molecule inside the liquid and at its surface.

pulling the surface to the right of the line and an equal force pulling the surface leftwards. The ratio F/L is the surface tension and its dimensions are those of force per unit length. The same phenomenon can also be described in terms of energy. Increasing the surface area of a liquid requires work, which remains stored as potential energy in the enlarged surface, just as energy can be stored in a stretched spring. That potential energy can perform work if the enlarged surface is allowed to contract again. Energy per unit area has the same dimensions as force per unit length.

An explanation for occurrence of surface tension is given in Figure 7. A molecule inside the liquid is attracted in all directions equally by the cohesive forces of neighboring molecules, while a molecule at the surface of the liquid is attracted into the relatively dense liquid phase by a net force greater than that attracting it toward the rarified gaseous phase. This unbalanced force draws the surface molecules inward into the liquid and results in the tendency for the surface to contract. This is why drops of a liquid in air as well as bubbles of air in a liquid assume the shape of a sphere, which is a body of minimal surface exposure relative to its volume.

Different liquids exhibit different surface tension values, as in the following list:

Water, 7.27×10^{-2} N m^{-1}(72.7 dyn cm^{-1}) at 20°C;
Ethyl ether, 1.7×10^{-2} N m^{-1} (17 dyn cm^{-1});
Ethyl alcohol, 2.2×10^{-2} N m^{-1}(22 dyn cm^{-1});
Benzene, 2.9×10^{-2} N m^{-1}(29 dyn cm^{-1});
Mercury, 0.43 N m^{-1} (430 dyn cm^{-1}).

Curvature of Water Surfaces and Hydrostatic Pressure

Wherever an interface between fluids (say, between water and air) is not planar but curved, the resolution of forces due to surface tension creates a pressure differential across that interface. For a spherical interface (as in the case of a bubble of air immersed in a body of water), the pressure difference is proportional

BOX 1 A Farewell to Teardrops

For centuries, conventional wisdom held that larger stones fall faster than small ones, that the Earth is flat yet the sun revolves around it, and that the sun and moon are of equal size. We would like to believe that in our time all baseless notions have been replaced by sound science. But have they?

Consider the shape of a raindrop. The conventional standard is a teardrop, rounded at the bottom and tapering to a point at the top. So prevalent is that image that it is printed in textbooks and used as a logo by irrigation companies and even by conferences sponsored by the United Nations. Alas, the vertically elongated, top-pointed raindrop is a physical impossibility.

A drop forming at the tip of a spout assumes that shape only at the very instant of detachment. Because of surface tension, any drop suspended in air 'balls up' spontaneously into a sphere. In a cloud, spherical droplets tend to grow by condensation or coalescence until reaching a critical weight, at which they begin to fall.

Air bypassing a falling drop acquires greater velocity around the curved 'waist' of the drop than near its top or bottom. By Bernoulli's law ($p + \rho v^2/2 = \text{constant}$), the pressure of the air alongside the drop is lowered relative to that of the air above and below the drop. Consequently the drop compresses vertically and comes to resemble an ellipsoid. Such is indeed the shape of small drops (< 2 mm).

In the case of larger drops, the laminar streams of air flowing past the drop may not converge smoothly above the top, but may leave a turbulent wake there. In such a wake, the pressure is lower than it would be in an ideally laminar flow regime. The reduced air pressure at the top causes the drop to bulge a bit there, thus acquiring the appetizing shape of a miniature hamburger bun (described by the appropriately named McDonald as early as 1954).

If a drop were to continue accelerating, it might eventually spread out to form a pancake. But Stokes' law intervenes, decreeing that the drop's acceleration be countered by the increasing viscous resistance of the air. When air resistance equals the gravitational force, acceleration ceases and the drop continues to fall at a constant 'terminal' velocity and with a more or less constant shape. It finally slaps the ground with its flattened face going 'plop'.

Some of us may think it unfair that such an exquisitely sculpted natural body should have no more glorious a fate than to splatter down on some dry bit of earth. But that is where – having at once lost its distinctive shape as it enters the labyrinthine interstices of the soil – our drop might well give life to a thirsty plant, perhaps even to a sunflower. Anyhow, we bid farewell to the sad countenance of the teardrop, a singularly inappropriate symbol for happy rain.

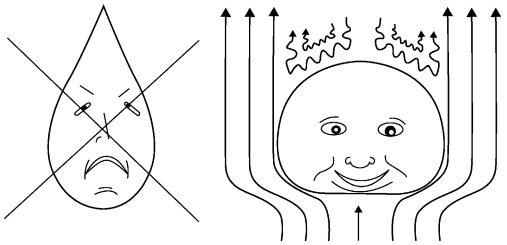


Figure B1 Conventional and real raindrops. Reprinted from *Environmental Soil Physics*, Hillel D (ed.). Copyright (1998), with permission from Elsevier.

to the surface tension and inversely proportional to the curvature:

$$\Delta P = 2\gamma/R \qquad [9]$$

Thus, the smaller the bubble is, the greater is its pressure.

If the bubble is not spherical, then instead of eqn [9] we obtain:

$$\Delta P = \gamma(1/R_1 + 1/R_2) \qquad [10]$$

where R_1 and R_2 are the principal radii of curvature for any given point on the interface. Eqn [10] reduces to [9] whenever $R_1 = R_2$.

Contact Angle of Water on Solid Surfaces

If we place a drop of liquid upon a dry solid surface, the liquid will usually displace the gas that covered the surface of the solid and it will spread over that surface to a certain extent. Where its spreading ceases and the edge of the drop comes to rest, it will form a typical angle with the surface of the solid. This angle, termed contact angle, is illustrated in **Figure 8**.

We now consider what factors determine the magnitude of the angle α. We can expect that angle to be acute if the adhesive affinity between the solid and liquid is strong relative to the cohesive forces inside the liquid itself and to the affinity between the gas and the solid. We can then say that the liquid 'wets' the solid. A contact angle of zero would mean the complete flattening of the drop and the perfect wetting of the solid surface by the liquid. On the other hand, a contact angle of 180° would imply a complete nonwetting or rejection of the liquid by the gas-covered solid. In that case the drop would retain its spherical shape without spreading over the surface at all (assuming no gravity effect). Surfaces on which water exhibits an obtuse contact angle are called water-repellent, or hydrophobic (Greek: 'water-hating').

The contact angle of a given liquid on a given solid is generally characteristic of its interaction under given physical conditions. This angle, however, may be different in the case of a liquid that is advancing over the solid surface than in the case of the same liquid receding over the surface. This phenomenon, where it occurs, is called contact angle hysteresis. The wetting angle of pure water upon clean and smooth mineral surfaces is generally zero, but where the surface is rough or coated with adsorbed surfactants of a hydrophobic nature, the contact angle, and especially the wetting angle, can be considerably greater than zero. This is illustrated in **Figure 9**.

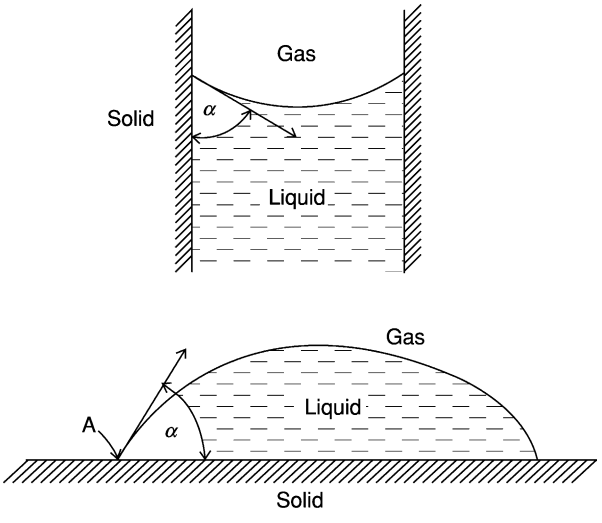


Figure 8 The contact angle of a meniscus in a capillary tube and a drop resting on the surface of a solid. Reprinted from *Environmental Soil Physics*, Hillel D (ed.). Copyright (1998), with permission from Elsevier.

The Phenomenon of Capillarity

A capillary tube dipped in a body of free water will form a meniscus as the result of the contact angle of water with the walls of the tube. The curvature of this meniscus will be greater (i.e., the radius of curvature smaller) the narrower the tube. The occurrence of curvature causes a pressure difference to develop across the liquid–gas interface. A liquid with an acute contact angle (e.g., water on glass) will form a concave meniscus, and therefore the liquid pressure under the meniscus (P_1) will be smaller than the atmospheric pressure (**Figure 10**). Hence, the water inside the tube will be driven up the tube from its initial location (shown as a dashed curve in **Figure 10**) by the greater pressure of the free water (i.e., water at atmospheric pressure, under a horizontal air–water interface) outside the tube at the same level. The

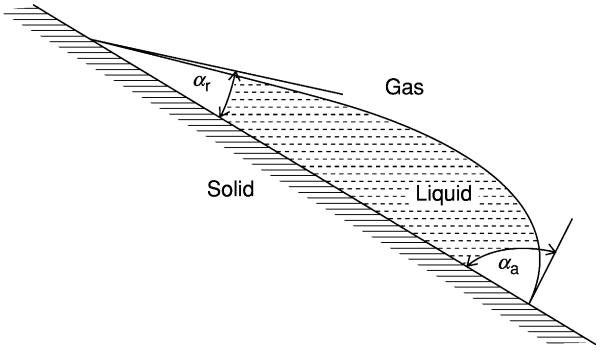


Figure 9 Hypothetical representation of a drop resting on an inclined surface. The contact angle α_a, at the advancing edge of the drop, is shown to be greater than the corresponding angle α_r at the receding edge. Reprinted from *Environmental Soil Physics*, Hillel D (ed.). Copyright (1998), with permission from Elsevier.

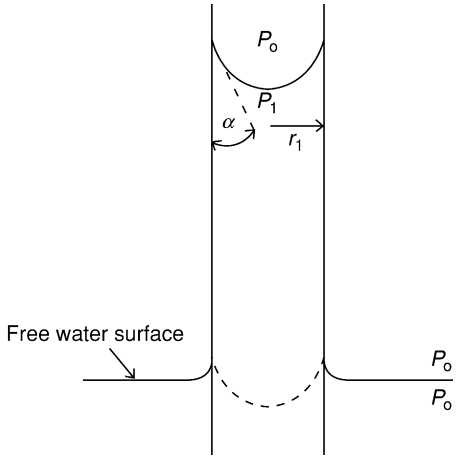


Figure 10 Capillary rise. Reprinted from *Environmental Soil Physics*, Hillel D (ed.). Copyright (1998), with permission from Elsevier.

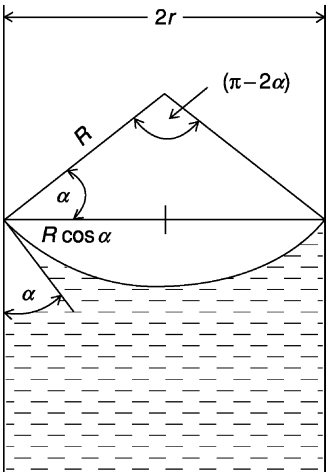


Figure 11 The geometric relationship of the radius of curvature R to the radius of the capillary r and the contact angle α. Reprinted from *Environmental Soil Physics*, Hillel D (ed.). Copyright (1998), with permission from Elsevier.

upward movement will stop when the pressure difference between the water inside the tube and the water under the flat surface outside the tube is countered by the hydrostatic pressure exerted by the water column in the capillary tube.

If the capillary tube is cylindrical and if the contact angle of the liquid on the walls of the tube is zero, the meniscus will be a hemisphere (and in a two-dimensional drawing can be represented as a semicircle) with its radius of curvature equal to the radius of the capillary tube. If, on the other hand, the liquid contacts the tube at an angle greater than zero but smaller than 90°, then the diameter of the tube ($2r$) is the length of a cord cutting a section of a circle with an angle of $\pi - 2\alpha$, as shown in Figure 11. Thus,

$$R = r/\cos\alpha \qquad [12]$$

where R is the radius of curvature of the meniscus, r the radius of the capillary, and α the contact angle.

The pressure difference ΔP between the capillary water (under the meniscus) and the atmosphere, therefore, is:

$$\Delta P = (2\gamma \cos\alpha)/r \qquad [13]$$

Recalling that hydrostatic pressure is proportional to the depth d below the free water surface (i.e., $P = \rho g d$, where ρ is liquid density and g is gravitational acceleration) we can infer that hydrostatic tension (negative pressure) in a capillary tube is proportional to the height h above the free water surface. Hence the height of capillary rise is:

$$h_c = (2\gamma \cos\alpha)/g(\rho_1 - \rho_o)r \qquad [14]$$

where ρ_g is the density of the gas (which is generally neglected), ρ_1 the density of the liquid, g the acceleration of gravity, r the capillary radius, α the contact angle, and γ the surface tension between the liquid and the air.

When the liquid surface is concave, the center of curvature lies outside the liquid and the curvature, by convention, is regarded as negative. Thus, for a concave meniscus such as that of water in a clean glass capillary, ΔP is negative with reference to the atmosphere, indicating a capillary pressure deficit, or subpressure, called tension. On the other hand, in a capillary tube that forms a convex meniscus (such as that of mercury in glass, or of water in an only or otherwise water-repellent tube), ΔP is positive and capillary depression, rather than capillary rise, will result.

Density and Compressibility

The open packing of water molecules in ice and liquid water accounts for their relatively low densities. Unlike most substances, water exhibits a point of maximum density (at 4°C), below which the substance expands due to the formation of a hexagonal lattice structure, and above which the expansion is due to the increasing thermal motion of the molecules. The coefficient of thermal expansion of water is rather low, and in the normal temperature range of, say, 4–50°C, the density diminishes only slightly, from 1000 to 988 kg m^{-3}.

The compressibility of water, C_w, can be defined as the relative change in density with change in pressure:

$$C_w = (1/\rho_w)(d\rho_w/dP) \qquad [15]$$

At 20°C and at atmospheric pressure, the compressibility of pure water is about 4.6×10^{-10} m^2 N^{-1}. In

Table 4 Physical properties of liquid water

Temperature (°C)	Density(kg m^{-3}) ($\times 10^3$)	Specific heat (J kg^{-1}deg) ($\times 10^3$)	Latent heat (vaporization) (J kg^{-1}) ($\times 10^6$)	Surface tension (kg s^{-2}) ($\times 10^{-2}$)	Thermal conductivity (J m^{-6} s deg)	Dynamic viscosity (kg m^{-1} s) ($\times 10^{-2}$)	Kinematic viscosity (m^2 s^{-1}) ($\times 10^{-6}$)
−10	0.99794	4.27	2.53	—	—	—	—
−5	0.99918	4.23	2.51	7.64	—	—	—
0	0.99987	4.22	2.50	7.56	0.561	0.1787	1.79
4	1.00000	4.21	2.49	7.5	0.570	0.1567	1.57
5	0.99999	4.207	2.49	7.48	0.574	0.1519	1.52
10	0.99973	4.194	2.48	7.42	0.587	0.1307	1.31
15	0.99913	4.19	2.47	7.34	0.595	0.1139	1.14
20	0.99823	4.186	2.46	7.27	0.603	0.1002	1.007
25	0.99708	4.18	2.44	7.19	0.612	0.0890	0.897
30	0.99568	4.18	2.43	7.11	0.620	0.0798	0.804
35	0.99406	4.18	2.42	7.03	0.629	0.0719	0.733
40	0.99225	4.18	2.41	6.95	0.633	0.0633	0.661
45	0.99024	4.18	2.40	6.87	0.641	0.0596	0.609
50	0.98807	4.186	2.38	6.79	0.645	0.0547	0.556

Reprinted from *Environmental Soil Physics*, Hillel D (ed.). Copyright (1998), with permission from Elsevier.

the normal situations encountered on the surface of the Earth, water can usually be considered incompressible. The compression of water cannot be ignored, however, in the case of deep confined aquifers, which may be subject to a pressure of, say, 10 MPa or more.

Dynamic and Kinematic Viscosity

When a fluid is moved in shear (i.e., when adjacent layers of fluid are made to slide over each other), the force required is proportional to the velocity of shear. The proportionality factor is called the viscosity. As such, it is the property of the fluid to resist the rate of shearing and this can be visualized as an internal friction. The coefficient of viscosity η is defined as the force per unit area necessary to maintain a velocity difference of 1 m s^{-1} between two parallel layers of fluid which are 1 m apart. The viscosity equation is:

$$\tau = F_s/A = \eta \, \mathrm{du/dx} \qquad [16]$$

wherein τ is the shearing stress, consisting of a force F_s acting on an area A; η (dimensions: mass/(length $\times$ time)) is the coefficient of dynamic viscosity; and du/dx is the velocity gradient perpendicular to the stressed area A.

The ratio of the dynamic viscosity of a fluid to its density is called the kinematic viscosity, designated ν. It expresses the shearing-rate resistance of a fluid independently of the density. Thus, while the dynamic viscosity of water exceeds that of air by a factor of about 50 (at room temperature), its kinematic viscosity is actually lower.

Fluids of lower viscosity flow more readily and are said to possess greater fluidity (which is the reciprocal of viscosity). As shown in Table 4, the viscosity of water diminishes by over 2% per 1°C rise in temperature, and thus decreases by more than half as the temperature increases from 5 to 35°C. The viscosity is also affected by the type and concentration of solutes present.

See also: **Capillarity**; **Hydrodynamics in Soils**; **Water Cycle**

Further Reading

Acheson DJ (1990) *Elementary Fluid Dynamics*. Oxford, UK: Clarendon.

Gleick PH (ed.) (1993) *Water in Crisis: A Guide to the World's Freshwater Resources*. New York: Oxford University Press.

Hillel D (1998) *Environmental Soil Physics*. Boston: Academic Press.

Kandel R (2003) *Water from Heaven*. New York: Columbia University Press.

Kramer PJ and Boyer JS (1995) *Water Relations of Plants and Soils*. San Diego: Academic Press.

WATER-REPELLENT SOILS

J Letey, University of California–Riverside, Riverside, CA, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Dry soil that does not spontaneously imbibe water, but rather has water remain on the surface, is referred to as a water-repellent or hydrophobic soil. Soils which spontaneously imbibe water are considered to be wettable but they can have different degrees of wettability.

The most common observations made on water-repellent soils are: (1) the water resists penetration and 'rolls' off the soil, and (2) very wet soil can be found next to very dry soil even though both receive equal water from rain or irrigation.

Water repellency has been reported in most continents of the world for varying land uses and climatic conditions and has been the topic of two international conferences. The first conference was held at Riverside, California, in 1968, and the second was convened 30 years later at Wageningen, the Netherlands.

The theory and description of soil-water processes described in soil physics textbooks are valid for wettable soils. That theory does not apply to water-repellent soils. Techniques that quantify the degree of soil water repellency are important for characterizing and mitigating such soil conditions.

Characterizing the Degree of Water Repellency

The extent of water repellency of a flat, smooth surface is commonly determined by measuring the water–solid contact angle when a drop of water is placed on the surface. Water will form a 'ball' on a very repellent surface. Soils do not have surfaces that conveniently allow the geometric measurement of a contact angle. Thus, an alternative to geometric measurement is required. Soils have pores and occasionally have been represented as being composed of a bundle of capillary tubes. The capillary tube is a vast oversimplification of the complex geometric arrangements of soil pores. Nevertheless, helpful insights can be achieved by assuming the capillary tube model for soils.

If the tip of a capillary tube is brought in contact with water, the water will be drawn into the tube. The height that the water will rise is given by eqn [1].

$$h = 2\gamma_w \cos\theta / r\rho g \qquad [1]$$

where h is the height of rise, γ_w the water–air surface tension, θ the water–solid contact angle, r the capillary radius, ρ the water density, and g the gravitational constant. According to eqn [1], water will not spontaneously enter the soil if cos θ is zero or a negative number ($\theta \geq 90°$).

A soil is commonly classified as being water-repellent if a drop of water placed on the soil does not spontaneously enter the soil. By this convention, a water-repellent soil is one in which the water–solid contact angle is equal to or greater than 90°. The water-drop penetration time (WDPT) is an index commonly used to specify the degree of water repellency. This procedure involves placing a drop of water on the soil and measuring the time for it to penetrate. Since water only penetrates the soil if $\theta < 90°$, WDPT is the measure of the time required for θ to change from its original value, which was more than 90°, to a value approaching 90°. Therefore, it is a measure of the stability of the repellency when the soil is brought in contact with water, and not necessarily an index of θ. The fact that the degree of repellency is not static but can change with time after contact with water complicates the analysis of temporal effects of water repellency on water flow.

Liquid surface tension that wets the soil material with a 90° contact angle is one index of water repellency. A series of aqueous ethanol solutions producing a range of surface tensions is prepared for this measurement. Drops of these solutions are placed on the soil. The higher-surface-tension solutions set on the surface and the lower-surface-tension solutions will spontaneously penetrate the soil. The 90° surface tension (γ_{nd}) is the surface tension of the solution where there is transition from sitting on the surface to penetration. Sometimes this procedure is followed but the molarity of the aqueous ethanol solution rather than the surface tension is reported as the index, referred to as the molarity of ethanol (MED) test. Others have reported the results of this test in terms of the volumetric ethanol percentage when penetration is initiated. The γ_{nd} has at times been referred to as the critical surface tension.

The solid–air surface tension, γ_s, is a fundamental physical–chemical property of a solid that affects its wetting properties. Therefore, characterizing the magnitude of water repellency by measuring the solid–air surface tension is valuable. Theoretical relationships between various surface tensions and contact angle have been combined to obtain the following two equations:

$$\gamma_s = \gamma_{nd}/4 \quad [2]$$

$$\cos\theta = [(\gamma_{nd}/\gamma_w)^{1/2} - 1] \quad [3]$$

Therefore the measurement of γ_{nd} can be used to calculate the value of γ_s and θ. The values of γ_s and θ determined by measuring γ_{nd} represent initial values before they have had an opportunity to change after contact with water. Both WDPT and γ_{nd} can be measured quickly and easily in either the laboratory or field. Measurement of γ_{nd} provides an index of the initial extent of water repellency, and WDPT provides an index of the stability of repellency after contact with water.

Water-Entry Pressure Head

When $\theta > 90°$, pressure must be applied to force the water into the soil. The water-entry pressure head, h_p, into a capillary tube is:

$$h_p = -2\gamma_w \cos\theta / r\rho g \quad [4]$$

For $\theta < 90°$, h_p is negative and water is drawn spontaneously into the tube. For $\theta > 90°$, h_p is positive and the water head equal to or greater than h_p must exist for water entry. Note that h_p depends on both the water-repellent index (θ) and the pore size (r). Measurement of h_p requires an apparatus that allows the depth of water on a soil column to increase gradually and some means of determining at what point the water penetrates the soil column.

Infiltration Rate

One of the most important effects of soil water repellency on the hydrologic cycle is its effect on infiltration rate. Infiltration will not occur until water has been in contact with the soil surface for a time equal to or greater than the WDPT unless the depth of ponding (h_o) is greater than the water-entry pressure (h_p). Even after conditions allowing water penetration, the infiltration and hydraulic property characteristics of water-repellent soils differ from wettable soils. Wettable soils typically have a high initial infiltration rate which decreases and then becomes relatively constant with increased time of water applied to the surface. In contrast, the infiltration rate into a water-repellent soil is slow during the initial phase of infiltration and increases with time. For soils with a finite WDPT, the degree of repellency changes with time after contact with water, so the increase in infiltration rate with time might be attributed to the changes in the degree of repellency.

In order to isolate and clearly identify the effects of water repellency that are not time-dependent on infiltration behavior, studies have been conducted on soils that have an infinite WDPT (water never penetrates the soil). Because natural soils do not typically have infinite WDPT values, soil materials artificially treated with octadecylamine provide a stable water repellency with an infinite WDPT. Different concentrations of octadecylamine create sands with different degrees of water repellency. Studies have been conducted using such treated sands.

Whereas the infiltration rate into wettable soils is only slightly affected by the depth of ponded water, infiltration into the water-repellent materials is drastically affected by the depth of ponded water (**Figure 1**). The h_p value for this material is 8.4 cm. As expected, no water infiltrates unless h_o is greater than h_p. At low values of h_o, the infiltration rate increases with time. However, for the higher values of h_o, infiltration rate decreases with increasing time, which is typical of wettable soils. Intermediate values of h_o ($h_o/h_p \approx 2.6$) produce a nearly constant infiltration rate as a function of time.

The hydraulic conductivity of the treated sand can be measured by ponding water on the surface of a column and measuring the steady-state water flow through the column. Increasing the depth of water ponding on the surface increases the hydraulic head gradient and therefore should induce an increased water flow. However, the hydraulic conductivity should be unaffected by the depth of water ponding, and this result is typically observed for wettable soils. In contrast, the hydraulic conductivity of the treated sands increases with depth of ponding until a critical depth of ponding is reached, after which the hydraulic conductivity becomes constant (**Figure 2**). The hydraulic conductivity in the water-repellent sand attains a value of the untreated sand when h_o is high enough. The ratio of h_o/h_p that results in maximum K equivalent to the untreated K is approximately 3.1 for each treatment.

Researchers have found that the average water content in the water-repellent sand decreases as the value of h_o decreases. Thus the decreasing K is caused by a decrease in water content, which is consistent with the well-recognized transport phenomenon that hydraulic conductivity is a function of soil water content. The water-entry pressure increases as the pore size decreases (eqn [4]); therefore the increase in water content could be the result of smaller pores being filled with water as h_o increases. Alternatively, finger flow through water-repellent sand could have caused only a portion of the sand to be wet. Whether the decrease in average water content is uniformly distributed in the sand or the result of finger flows

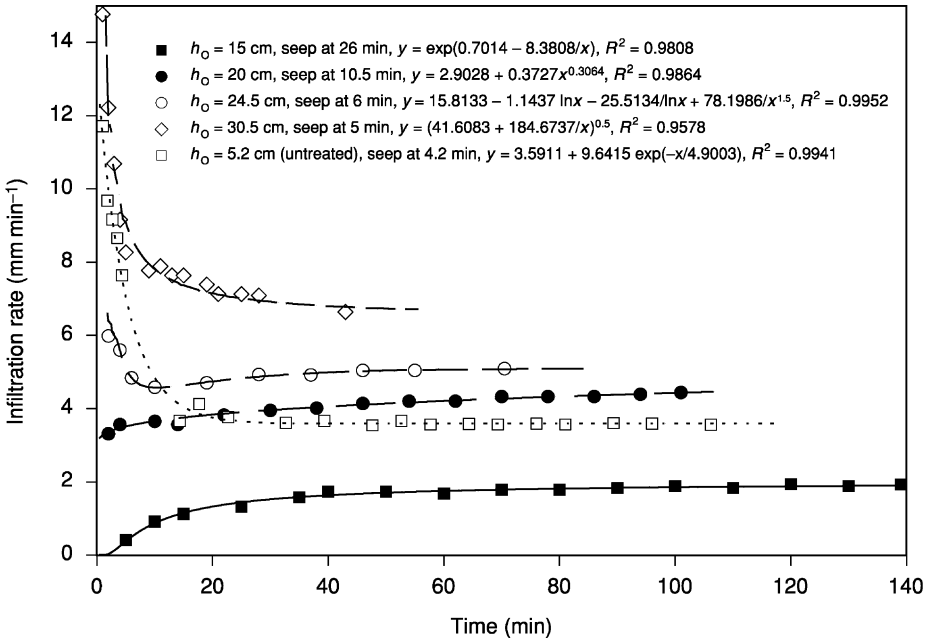


Figure 1 The infiltration rate as a function of time for sand ($h_p = 8.4$ cm) for different values of ponded-water depth (h_o). The equations are for the best-fit curve to the data points. (Reproduced with permission from Feng GL, Letey J, and Wu L (2001) Water ponding depths affect temporal infiltration rates in a water-repellent sand. *Soil Science Society of America Journal* 65: 315–320.)

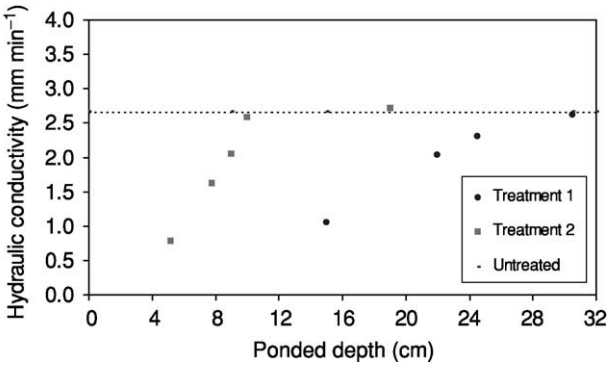


Figure 2 The relationship between hydraulic conductivity and ponded-water depth for sands receiving two treatments to make them water-repellent. The dotted line is the hydraulic conductivity of the untreated sand, which is independent of ponded depth. (Reproduced with permission from Feng GL, Letey J, and Wu L (2001) Water ponding depths affect temporal infiltration rates in a water-repellent sand. *Soil Science Society of America Journal* 65: 315–320.)

where only a fraction of the sand is wet has not been determined conclusively.

The h_o/h_p value at which there is a transition from infiltration rate increasing with time to decreasing with time (Figure 1) is close to the h_o/h_p value that produces a hydraulic conductivity close to that of untreated sand.

The effect on infiltration rate of placing a layer of wettable, untreated sand over the water-repellent sand has been studied. Infiltration rate is rapid into the overlying wettable sand, but stops when reaching the repellent layer until the hydraulic head at the interface exceeds the value of h_p. Thereafter, infiltration into the repellent layer is initiated. The consequence of having the untreated layer is approximately equivalent to increasing the ponded depth. In other words, ponding water 1 cm above a 5-cm layer of wettable sand is equivalent to ponding 6 cm of water on a soil that is water-repellent at the surface.

The hydraulic heads immediately behind the wetting front differ between wettable and water-repellent soils. In a wettable soil, the hydraulic head is negative immediately behind the wetting front and increases as

it moves upward from the wetting front. Conversely, the hydraulic head immediately behind the wetting front is positive in a water-repellent soil and decreases away from the wetting front. These are conditions that contribute to unstable or preferential flow in the soil.

The term 'preferential flow' has been used to describe nonuniform water flow or wetting-front advance through the soil. However, the term 'preferential flow' can have different connotations. Ambiguity arises because of different dimensional scales, ranging from pore size to several centimeters. Indeed, wet and dry zones in a water-repellent field may have dimensions in meters where the term 'preferential flow' might not be appropriate.

Instability of wetting-front advance has been observed in the laboratory on water-repellent soil materials. Instability of a dynamic wetting front is defined as the unconstrained growth of randomly occurring, small perturbations of the wetting front. The effect is to have a wetting-front advance rapidly in random zones, creating a 'finger' wetting pattern.

The depth of ponded water h_o has a profound effect on the extent of finger flow in a water-repellent soil. If the water-repellent layer is at a depth L below the soil surface, the effect of L is comparable to the effect of h_o on finger formation. Water penetrates the wettable surface layer but is constrained by the water-repellent sublayer, which allows pressure to build up in the water to a value comparable to the value that would exist without the wettable layer. One study has found that no water penetrates the water-repellent layer for values of $(h_o + L)/h_p < 1$; unstable flow develops for the values between 1 and 1.5; and a stable wetting front develops for values greater than 1.5.

The basic concepts of flux through a water-repellent layer as affected by the total head may explain the finger formation. Instability of water flow occurs if the flow through the system is less than the saturated hydraulic conductivity. The fact that finger flow and reduced hydraulic conductivity (**Figure 2**) are both associated with low values of h_o is consistent with the theory that instability of water flow occurs when the flow through the system is less than the saturated hydraulic conductivity.

Temporal effects of infiltration rate and preferential flow are influenced by the WDPT value. Infiltration will occur when a WDPT has been reached even without a ponding depth. However, one study has found that the wetting front is unconditionally unstable when h_o is less than h_p, resulting in finger flow. Conversely, the flow has been observed to be stable when h_o is greater than h_p if WDPT is not too large. Finger flow does not occur for a soil that has a WDPT of 1 min. When the WDPT is 10 min, fingers form; however, the fingers broaden and converge after continued flow and an almost uniform wetting front eventually develops.

Uneven wetting patterns, whereby zones of very dry soil can exist next to zones of very wet soils, are commonly observed in the field on water-repellent soils. Location of wet and dry zones usually cannot be determined by visual observations of the landscape. The phenomenon, however, can be triggered by any factor that initiates infiltration. Once infiltration occurs, the infiltration rate increases with time, drawing water away from adjacent zones where infiltration has not been initiated. The 'trigger' may be associated with special distribution of WDPT values, whereby the water starts to infiltrate in those zones with a lower WDPT than an adjacent zone with a higher WDPT. Also the value of h_p is associated not only with differences in repellency, but also with pore-size distribution. Possibly, slight differences in texture or packing that lead to different pore-size distributions could contribute to the variable wetting. Also slight differences in topography that allow slightly more depth of water ponding on one zone than the other could initiate the uneven infiltration.

Wildfires have been found either to create or enhance the severity of water repellency on watersheds. Increases in streamflow, quickflow volumes, and sediment in bedload yields have been observed on fire-induced water-repellent soils. Fire-induced water repellency produces localized runoff and sediment movement on hillsides, but does not appreciably affect the performance of a whole watershed. Because of the effects of cracks, animals burrows, root channels, etc., it is difficult to characterize the effects of water repellency at the watershed scale.

Mitigating Water Repellency

Surfactants can be used to mitigate the effects of soil water repellency. Adding a surfactant to water lowers the surface tension of the water. Reduction of the liquid surface tension also reduces the liquid–solid contact angle, which makes the soil behave as if it were wettable. However, if a soil is wettable, reducing the surface tension of water may actually decrease the infiltration rate. Surfactant molecules are adsorbed by the soil as the solution moves downward. As the surfactant is adsorbed from the solution, surface tension of the solution increases and its ability to wet diminishes. Therefore, surfactants are most effective in treating soil water repellency when it is associated with the surface layer of the soil.

Adsorption of the surfactant molecule converts a water-repellent soil to a wettable soil upon drying.

This phenomenon has been referred to as the 'rewet properties of a surfactant.'

Surfactants have been extensively used to overcome water repellency in turf. Water repellency is a common phenomenon in highly managed golf course soils. Water repellency has been managed through the application of surfactants to areas of turf exhibiting visual symptoms commonly referred to as localized dry spots.

Although there are theoretical relationships between liquid surface tension and soil wetting that should allow the computation of the optimal concentration of surfactant to use for treatment, the adsorption of surfactant by the soil (which alters the liquid surface tension) makes a theoretical basis for describing surfactant treatment virtually impossible. Therefore, developing the most effective surfactant treatment is most commonly done by an empirical approach of applying various treatments and observing the effects.

The application and incorporation of clay into the surface 10-cm layer of a water-repellent sand ameliorates water repellency. The clay, which is hydrophilic, converts the surface layer from a water-repellent to a wettable surface. A wettable layer overlying a water-repellent layer acts as if the depth of water ponding is increased equal to the depth of the wettable layer. Therefore if the water entry pressure is less than 10 cm, providing a 10-cm layer of wettable material by incorporating clay would allow the water to penetrate the underlying water-repellent soil.

One field study has compared soil-wetting patterns on two adjacent plots: one with a water-repellent top layer, and one treated with clay to remove the water repellency in the top 30-cm layer. Dye placed on the soil surface to identify wet and dry zones revealed a mosaic surface wetting of water-repellent soil and a uniform surface wetting of wettable soil. Although water repellency existed in the treated plot below the 30-cm depth, it was uniformly wet. Thus stable flow occurs if the depth to the water-repellent layer is sufficiently deep.

Summary

Naturally occurring water-repellent soils have been reported in many areas of the world. Wildfires either create or enhance the water repellency of watersheds. Processes within the hydrologic cycle such as infiltration and erosion are drastically affected by soil water repellency. Basic water flow principles that have been developed for wettable soils must be modified to be applicable to water-repellent soils. Except under unique conditions where treatment with surfactants is practical, large-scale mitigation of water-repellency is not feasible.

List of Technical Nomenclature

γ_{nd}	Surface tension of solution that wets soil at 90° contact angle
γ_s	Surface tension of solid
γ_w	Surface tension of water
θ	Contact angle (degrees)
h	Height of capillary rise (cm)
h_o	Depth of ponded water (cm)
h_p	Water entry pressure head (cm)
K	Hydraulic conductivity
WDPT	Water-drop penetration time

See also: **Capillarity**; **Infiltration**

Further Reading

Bauters TWJ, DiCarlo DA, Steenhuis TS, and Parlange J-Y (1998) Preferential flow in water-repellent sands. *Soil Science Society of America Journal* 62: 1185–1190.

Cann MA (2000) Clay spreading on water repellent sands in the south east of South Australia – promoting sustainable agriculture. *Journal of Hydrology* 231–232: 333–341.

Carrillo MLK, Letey J, and Yates SR (1999) Measurement of initial soil–water contact angle of water repellent soils. *Soil Science Society of America Journal* 63: 433–436.

Carrillo MLK, Letey J, and Yates SR (2000) Unstable water flow in a layered soil. I. Effects of a stable water-repellent layer. *Soil Science Society of America Journal* 674: 450–455.

DeBano LF (2000) Water repellency in soils: a historical overview. *Journal of Hydrology* 231–232: 4–32.

DeBano LF and Dekker LW (2000) Water repellency bibliography. *Journal of Hydrology* 231–232: 409–432.

Dekker LW and Ritsema CJ (1994) How water moves in a water repellent sandy soil. I. Potential and actual water repellency. *Water Resources Research* 30: 2507–2517.

Dekker LW and Ritsema CJ (2000) Wetting patterns and moisture variability in water repellent Dutch soils. *Journal of Hydrology* 231–232: 148–164.

Feng, GL, Letey J, and Wu L (2001) Water ponding depths affect temporal infiltration rates in a water-repellent sand. *Soil Science Society of America Journal* 65: 315–320.

Hendrickx JMH, Dekker LW, and Boersma OH (1993) Unstable wetting fronts in water repellent field soils. *Journal of Environmental Quality* 22: 109–118.

Kostka SJ (2000) Amelioration of water repellency in highly managed soils and the enhancement of turfgrass

performance through the systematic application of surfactants. *Journal of Hydrology* 231–232: 359–368.
Letey J (2001) Causes and consequences of fire-induced water repellency. *Hydrological Processes* 15: 2867–2875.
Ritsema CJ and Dekker LW (2000) Preferential flow in water repellent sandy soils: principles and modeling implications. *Journal of Hydrology* 231–232: 308–319.
Robichaud PR and Waldrop TA (1994) A comparison of surface runoff and sediment yields from low- and high-severity site preparation burns. *Water Resources Bulletin* 30: 27–34.
Scott DF and van Wyk DB (1990) The effects of wildfire on soil wettability and hydrological behavior of an afforested catchment. *Journal of Hydrology* 121: 239–256.
Shakesby RA, Doerr SH, and Walsh RPD (2000) The erosional impact of soil hydrophobicity: current problems and future research directions. *Journal of Hydrology* 231–232: 178–191.
Watson CL and Letey J (1970) Indices for characterizing soil-water repellency based on contact angle–surface tension relationships. *Soil Science Society of America Proceedings* 34: 841–844.

WATERSHED MANAGEMENT

M D Tomer, USDA Agricultural Research Service, Ames, IA, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Watershed management may be defined as a set of resource-management practices that are planned and implemented to provide a sufficient source of quality water to sustain human society and natural ecosystems. The practice of watershed management is interdisciplinary, because it recognizes linkages between land and water resources, and because it seeks to balance the needs of society with the capacities of natural resources to meet them. Land-use practices have impacts on hydrologic processes and water quality. Identifying ways to manage these impacts across a mosaic of soils, terrain, and land-use settings is a key challenge in watershed management. Effective watershed management is an iterative process of assessment, planning, and implementation. It begins with an assessment of current land-use practices and their impacts on water resources. Opportunities to improve land management, considering roles of soil, vegetation, and terrain, are then identified and prioritized. Stakeholder groups should be included in planning, to inform citizens about water-resource management issues and provide feedback to ensure recommendations are realistic as well as effective. A range of computerized tools are available to assist with assessing watersheds and alternative management scenarios. Implementation should include a commitment to reassess water-resource management periodically and develop opportunities for further improvement.

An Interdisciplinary Task

Watershed management is aimed at land and water resources, and is applied to an area of land that drains to a defined location along a stream or river. Watershed management aims to care for natural resources in a way that supports human needs for water, food, fiber, energy, and habitation, while supporting other agreed attributes linked to recreation, esthetics, and/or ecologic function. Because of these multidisciplinary concerns, the development of watershed-management strategies can involve complex scientific and public policy issues. Each watershed is unique in physiography, ecology, climate, water quality, land use, and human culture. Therefore any generalized approach to watershed management must be customized to each setting when put into practice. Watershed management requires a long-term commitment that is adaptive to changes in population, climate, culture, and resource-use demands. These issues are unique to each watershed and each nation. Watershed-management experiences from around the globe have dealt with a wide range of issues.

Hydrology and Streamflow Variation

The need to develop a watershed management plan is often identified in response to impacts of floods or drought on society. Long-term monitoring of precipitation, snowpacks, streamflow, and groundwater levels can help society to predict and prepare for these events. Although rainfall (or lack thereof) is not subject to management control, management can reduce the frequency of adverse impacts caused

by extremely wet or dry conditions. Water conservation practices, water storage and control structures, vegetation management, and land-use planning are basic tools that can help manage effects of floods and drought.

Stormflow and Floods

Precipitation may be intercepted by vegetation, or reach the soil surface to infiltrate, pond, or flow over it as runoff. Water at (or near) the soil surface can evaporate, while water below the soil surface may be transpired by plants or may percolate downward. Percolating water can recharge groundwater or move laterally downslope (as interflow) to accumulate at lower parts of the landscape. Some of these low-lying areas have water tables near the surface and can become saturated to the soil surface during a rainfall event. These are called variable-source areas, because they vary in size during an event and become a source of stormflow discharge to streams (Figure 1). Variable-source areas are hydrologically sensitive, meaning that, in many watersheds, management of these areas can help attenuate floods and maintain water quality.

Changes in streamflow that follow precipitation are determined by many attributes that affect water storage on the landscape, the prominence and timing of different pathways of water movement, and the location and extent of variable-source areas. Even in small watersheds, water flow pathways and their timing are difficult to decipher, despite numerous research efforts to deconvolute these pathways using physical, chemical, and isotopic methods. Therefore a hydrograph (Figure 2), a plot showing the response of discharge to precipitation, remains a commonly used basis for watershed characterization. In a small, first-order watershed (the smallest land area generating perennial streamflow), this response will take place more quickly than in large river basins, where a range of soils, terrain, land uses, and travel distances act to desynchronize and lengthen the response.

Flooding occurs when stream discharge exceeds the channel's conveyance capacity and forces water over the stream banks. A flood's magnitude can be characterized by its average frequency of recurrence in years (Figure 2). Engineered hydrologic structures are designed to accommodate a maximum flow defined by a specific recurrence interval. Thereby the probability that the structure will fail can be defined for any given time period. Design criteria may allow a fairly large probability of failure if the cost of failure is small (e.g., culverts beneath low-use forest roads), but will require an infinitesimally small chance of failure if that failure were to be catastrophic (e.g., large dams). The discharge associated with a given recurrence interval must be accurately

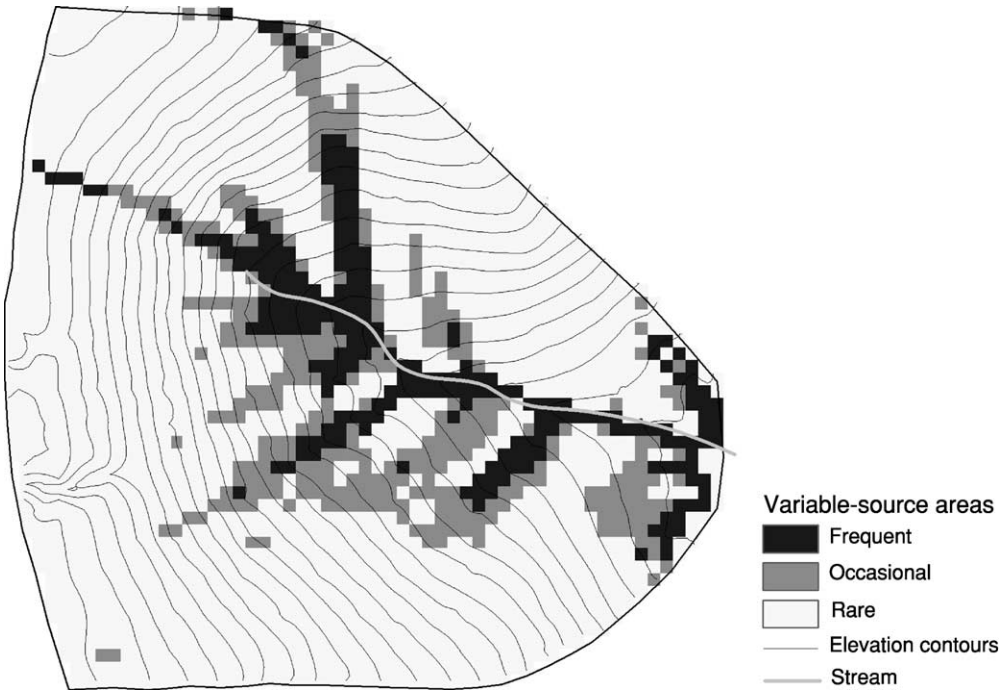


Figure 1 Following precipitation, water is discharged to streams from 'variable-source areas' (or 'partial-contributing areas') that are prone to saturation. The locations of these areas can be estimated based on topographic relationships. This conceptual map indicates the relative frequency of runoff contributions following precipitation.

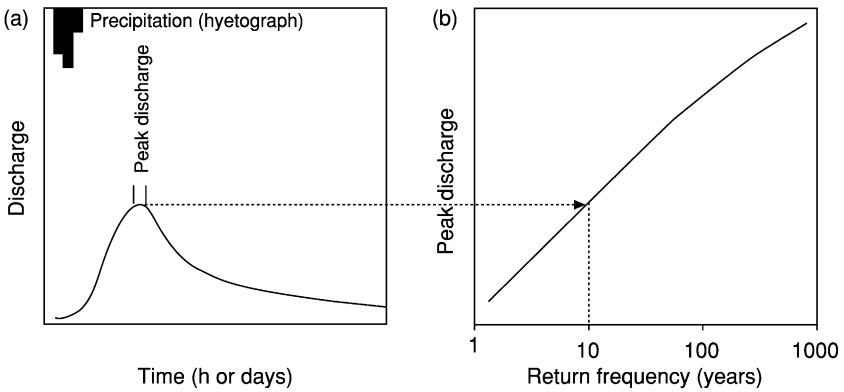


Figure 2 A hydrograph is the response in streamflow to a precipitation event (a). The peak discharge of a hydrograph will have an expected return period; the peak discharge of this event represents a flow that, on average, should only be exceeded once in a 10-year period (b).

known or conservatively estimated. A long-term monitoring record, spanning at least several decades, helps identify the frequency distribution of large flows. Methods of estimation appropriate for small, unmonitored watersheds are also available. These may be based on simple equations (e.g., the synthetic unit-hydrograph method, developed by the US Soil Conservation Service) or sophisticated simulation models.

Snow Hydrology

Seasonal timing of stream discharge is a critical issue, particularly for watersheds with mountainous headwaters where annual snowmelt is an important water source. In many of these watersheds, water yielded by snowmelt is stored in reservoirs for flood mitigation and for later use. Monitoring of annual snowpacks and predictions of water yields are important water management activities in these basins. Snowpacks vary each year, and advance knowledge of water yields helps to plan allocations and warn of potential shortages. Terrain (elevation, slope, and aspect) and wind patterns influence the spatial distribution of snow accumulation and the timing of melt. Vegetation can be managed to moderate these influences.

Drought and Low Streamflow

Prolonged drought can place tremendous pressure on land and water resources, and on society. Dry, hot weather increases demands for water by municipalities and irrigated agriculture when water supply is most limited, causing social conflict. Low streamflow and accompanying warm water temperatures also impact aquatic ecosystems. Drought may contribute to overgrazing and wildfire, which expose soils to erosion once rain returns. Management of groundwater can help protect water supplies and ensure adequate baseflow contribution to streams during drought.

Reservoir water storage, vegetation management, and water conservation measures are among the effective means to reduce impacts of drought. Water harvesting (*See* **Water Harvesting**) is also practiced in some areas. Water conservation can occur through improved methods and scheduling of irrigation, water recycling, and education, and/or regulatory programs to reduce urban water use. Improved irrigation methods may provide the greatest benefit, but return on the investment in conveyance and application systems may take years, in terms of the value of saved water.

Water Quality

Improvements in watershed management may become necessary if water quality becomes impaired. Water quality refers to the capacity of a water body to support certain uses or ecologic functions. It is determined by amounts of dissolved and suspended materials in water, presence or absence of certain microorganisms, and/or physical attributes such as temperature and clarity. Water-quality standards are threshold values (e.g., concentrations, loads) above (or below) which a specific use for water becomes impaired. Water quality can be degraded by natural events or processes, and by pollution resulting from human activities. Common sources of pollution are wastes from municipalities, livestock, and industry, and losses from applications of fertilizer and pesticides to land. Pollution can originate from point sources, where a pollutant enters a water body at a specific location. It can also originate from nonpoint sources where distributed land management activities promote movement of pollutants to ground or surface waters along natural flow pathways. Water quality

management is a process that identifies water uses that are (or may become) impaired, and the contaminants that impair (or threaten to impair) each use. The processes, sources, and pathways that contribute pollutants to a water body need to be understood, and then practices that can sustain water quality according to identified standards must be identified and implemented. This process can raise complex technical and sociopolitical issues. Table 1 lists water quality parameters most commonly of concern, typical sources of pollution for each, possible adverse effects, and some management practices that can control the pollutant. Several types of biologically active trace compounds are included that do not have water-quality standards developed, but that may be of greater concern in the future.

Land-Use Impacts

Soil and vegetation have a major impact on watershed hydrology because they affect partitioning of energy and water near the land surface. While basic soil properties such as texture and depth cannot be influenced through management, soil management can alter the effects of soil disturbance and influence a soil's capacity to support plant growth. Vegetation management can modify canopy interception, evapotranspiration, rooting depths, and seasonal patterns of water use, which affect key hydrologic processes (runoff, infiltration, evapotranspiration, percolation). Soil and vegetation management generally cannot mitigate extreme flood and drought events. However floods with frequent return periods (e.g., less than 10 years) may be mitigated, and low flows associated with annual to perhaps decadal dry periods may be increased. Logically, managing for increased infiltration can achieve both goals, because runoff is reduced, which mitigates flooding, and because groundwater recharge is increased, which contributes baseflow to streams during drought. However, the correspondence of these two benefits is moderated by influences of management on evapotranspiration.

Terrain is an important consideration in soil and vegetation management, and steep lands may require a particular focus, because water is routed from them quickly, contributing to flooding and sedimentation. Risks posed by instability of slopes may be important, especially in populated watersheds with steep terrain where landslides can cause loss of life.

Most watersheds are comprised of agricultural land, forestland, grazing land, urban areas, mine land, riparian zones, wetlands, and lakes in varying proportion and configuration. Each kind of land use poses different challenges and opportunities to improve watershed management. Watershed assessment aims to identify how each land use is influencing hydrology and water quality.

Agricultural Lands

Agricultural lands can contribute to nonpoint pollution, particularly by nutrients, sediment, and agricultural chemicals. It is important to identify erosion control, nutrient, and pest and/or weed management practices to minimize these pollutants effectively. In much of the developed world, crop and livestock production systems have become intensified and spatially aggregated, so that nutrients in animal feed are transferred over longer distances. Treatment, handling, and land application of livestock wastes is a key to effective nutrient management in areas that import feed. Generally, nutrient management aims to apply nutrients at the proper rates and timing to optimize efficiency of crop uptake. Erosion control practices range in scale and include reduced intensity of tillage (including no-tillage), contour farming, vegetative filter strips, and constructed terraces.

Cropping systems can be inefficient in their use of water under rain-fed and irrigated agriculture. Improvement in irrigation efficiencies helps reduce the frequency and severity of water shortages, and can minimize salinization problems in arid regions. Rain-fed agriculture often consists of annual crops that only transpire water during part of the growing season, in contrast to native plant communities. This seasonal restriction on plant water use can increase deep percolation and lead to excess nutrient leaching in humid areas, and salinization in semiarid areas. These problems can be addressed by increasing plant water use with crop rotations that, depending on the setting, may include trees, perennial forages, and/or cover crops.

Forest Lands

Trees have a large capacity to intercept precipitation within their canopy and deep roots that extract soil water from depth. These characteristics diminish the fraction of precipitation that can recharge groundwater and streamflow. Forest harvesting or conversion to other types of vegetation reduces evapotranspiration, which can increase baseflow contributions to streams and allow streamflow responses to small precipitation effects. Historically, observations of severe flooding after extensive forest harvesting brought about early research on the role of vegetation in watershed hydrology, beginning in the early twentieth century. This research quantified relationships between forest cover and streamflow for a number of

Table 1 A summary of common water quality problems associated with land management activities, their impacts, and practices that may address them

Contaminant	*Major causes/sources*	*Adverse impacts*	*Typical solutions*
Physical properties			
Sediment/ turbidity	Erosion, channelization, poor riparian zone management	Fisheries (spawning areas, gill function), aquatic ecosystems, aesthetics, reservoir water storage, water intake and supply systems	Streambank stabilization, erosion control, riparian vegetation management
Temperature	Removal of shading, streambed aggradation, some industry discharges	Fisheries (dissolved oxygen), aquatic ecosystems	Riparian zone management for shading
Inorganic constituents			
Nitrogen	Leaching from soils (NO_3), human and livestock waste (NH_4)	Drinkability (NO_3), toxicity to aquatic animals (NH_4)	Nutrient management, improved waste treatment
Phosphorus	Erosion, accumulation in manured soils	Fisheries, aesthetics, aquatic ecosystems (eutrophication), animal and human health (blue-green algae)	Erosion control, manure management
Dissolved oxygen	High temperature, eutrophication	Fisheries, aquatic ecosystems, aesthetics	Riparian zone management for shading, channel structures to create turbulence, nutrient management
Trace elements	Soil leaching, some pesticides, municipal/industrial wastes	Bioaccumulation, aquatic ecosystems	Increased plant water-use efficiency, industry-specific waste treatment technologies
Salinity	Soil leaching, evaporative discharge of groundwater	Irrigability, drinking, aquatic ecosystems	Management of irrigation water and/ or plant water use in recharge areas, water-table management
Organic materials			
Organic carbon (biological or chemical oxygen demand)	Organic wastes, erosion, eutrophication	Fisheries (reduces dissolved oxygen), aquatic ecosystems, turbidity	Waste treatment, erosion control, nutrient management
Pesticides (herbicides, insecticides, fungicides, nematocides)	Leaching or runoff	Drinkability, potential carcinogenic effects of some compounds and metabolites	Use and losses of nonpersistent compounds, management strategies for reduced pesticide use (e.g., crop selection, rotation, tillage)
Endocrine disruptors	Municipal and industrial wastewaters	Subject of debate, affects metabolic and reproductive function of some organisms	Not clear, but could include waste treatment technologies, alternative formulation of source products that enter the waste stream
Pharmaceuticals	Domestic and livestock wastes	Subject of debate, increased antibiotic resistance in microbial communities hypothesized	Not clear, but could include waste treatment technologies, alternative practices for reduced use in livestock feeding, veterinary, and medical practices
Biological organisms			
Pathogens	Domestic and livestock wastes, riparian grazing, wildlife	Drinkability	Waste treatment, riparian pasture management, improved methods for land application
Algae	Nutrients	Fisheries, aesthetics, dissolved oxygen	Nutrient management, riparian zone management

experimental watersheds. Increases in water yield after harvest may result from greater peak discharges or from greater baseflow contributions to streams, but not necessarily both. Specific factors of harvest disturbance, terrain, and geology determine specific responses in flow regime.

Water yield from forestland can be managed through rotational harvesting that restricts the extent

of harvest during a given period of time (e.g., 20% of the commercial forest under a 25-year rotation may be harvested during a 5-year period). This strategy can desynchronize flows in the watershed and decrease flood frequency. Forest management should consider effects of infrastructure, because forest roads, skid trails, and landings can be major sources of sediment. Harvesting can also increase the risk of mass movement (i.e., rockfalls, landslides) in steep terrain.

Grazing Lands

Rangelands occupy large areas with semiarid to arid climates where extensive crop production is not feasible. Overgrazing and subsequent erosion are prevailing concerns that affect hydrology and water quality. Proper stocking, herd management, and rotational grazing systems comprise effective management strategies, especially if designed to consider the frequency of drought. Managing the distribution of water to livestock is critical, because livestock will not wander far from reliable water sources. In humid areas, pasture management involves more intensive forage production and grazing. Limiting access of cattle to surface waters may address concerns for water quality along small streams.

Mined Land

Mining operations are often small, but the drastic nature of disturbance often requires that effects on water quality and supply be considered. Dewatering of aquifers may occur with subsurface or pit mining, and surface waters may be diverted off-site, or collected in holding ponds. Water quality impacts will vary depending on the method and extent of excavation, the type of rock and/or ore material being mined, the rock and/or ore processing occurring on-site, and reclamation and/or revegetation practices carried out as mining concludes or proceeds to new areas. Acidification, heavy metals, and sediment are common water-quality concerns, depending on the type of mine operation.

Urban Areas

Urban areas also occupy a small part of most basins, but are of predominant concern for the people and the economic activity within a watershed. Daily demands for drinking water, sanitation, and waste treatment are overriding issues due to obvious implications for health and quality of life. Demands for these services increase with development and population. If the growth in demand for water approaches the available supply, or if water quality becomes impaired, then stakeholders will demand improvements in water management. In watersheds with limited water supplies, planning should aim to provide a sustainable match between supply and demand in the longer term, including conservation measures to cope with effects of drought.

Management of storm runoff can be critical in urban areas. Urban development increases the extent of impervious surfaces (rooftops, roads, parking lots), and runoff from these areas must be accepted by a stormwater detention and conveyance system. Construction can cause soil compaction, increasing runoff, and diminishing infiltration and groundwater recharge. Urban development can change flood-frequency characteristics downstream, causing existing floodwater control and conveyance measures to become inadequate. As an example, a 0.1-ha residential lot may have pavement and roofing occupying 40% of that area. If a 10-mm rainfall generates 7 mm of runoff from the impervious surface, then 2.8 m^3 of water will be conveyed down-gradient. Prior to development, a 10-mm rainfall would probably generate little, if any, runoff. Alternative practices being used in some areas include stormwater detention basins with beneficial dry-weather uses (e.g., recreation), small detention basins ('rain gardens') for individual residential lots, 'green' (planted) roofs, and permeable pavements. Urban planning should consider future development and implications for stormwater hydrology.

Riparian Areas and Wetlands

In most watersheds, riparian zones and wetlands also occupy small areas, but they are present in every watershed and their management is frequently a focus of attention. Riparian zones contain variable-source areas that generate storm flow, influence the interaction between groundwaters and surface waters, and form the boundary between aquatic and terrestrial ecosystems. They support biological diversity and wildlife habitat, and provide aesthetic value. Biological processes (e.g., plant growth, microbial activity) are intensified in riparian areas because water and nutrients are usually abundant compared with upslope areas. Management of riparian zones offers the opportunity to improve streamflow regimen (timing) and water quality. Wetlands can detain storm flows and remove nutrients, particularly nitrate via denitrification. Riparian vegetation can take up nutrients and encourage trapping of sediments delivered from upslope areas. It is important to identify objectives for riparian management, locations where those objectives can be met most effectively, and vegetation management systems that can help achieve them. The potential benefits of improved riparian

management, however, can only be fully realized when uplands are also managed to achieve water resource goals.

Approaches, Challenges, and Tools for Watershed Management

The unique attributes of each watershed must be considered in developing and implementing a watershed management plan. While administrative approaches may vary, they usually involve four phases, including problem definition, assessment, selection of alternatives, and implementation and/or evaluation (Table 2). Ideally, watershed management is a continuous process that identifies and develops opportunities to improve environmental quality and resource sustainability.

Watershed planning must balance a number of legitimate, but often competing, resource-use requirements and concerns. Both environmental and social attributes of a watershed must be considered from the outset if planning is to result in successful implementation. Technical experts provide little long-term benefit unless local land managers and community interest groups (stakeholders) perceive their recommendations as balanced and realistic. Therefore stakeholder representation and involvement are necessary from the first stage of planning. An early consensus on resource issues to address and stakeholder groups to include (phase 1) benefits the community and its watershed resources.

Effective watershed planning depends on an objective assessment (inventory) of the current situation (phase 2), best carried out by technical experts from various resource-management and engineering disciplines. Stakeholder involvement can facilitate public education on the causes and impacts of water-resource problems, and can help establish trust between stakeholders and technical experts.

A number of computer tools can help to accomplish a watershed assessment. Use of a geographic information system (GIS) database provides ways to view and evaluate combined map data that include soil survey, land use and ownership, terrain analyses, and remote sensing. While GIS can provide information in compelling graphic formats, its output should be reviewed critically, considering the quality and scale of input data.

Major concerns of watershed management are related to streamflow regime, groundwater availability, and water quality. These all result from a mosaic of interactions among climate, land-use practices,

Table 2 A summary of the watershed management planning process, with a focus on stakeholder involvement

Phase	*Key questions*	*Sources of information*	*Analysis tools*	*Stakeholder role*
1. Problem definition	What are the key problems and who needs to be involved in solving them?	Public, public interest groups, focus group meetings, monitoring data	Statistical analysis of data, public opinion survey results	Source of information; help define formal and informal roles for stakeholders in project
2. Watershed assessment	What information can be gathered to facilitate an assessment of the watershed? Are new assessment/monitoring efforts needed and how can they be accomplished?	Mapped information: population census data, land use, soil survey, remote sensing, surveys of existing conservation systems; intensified and/or synoptic sampling of water resources	Geographic information systems, terrain- and image-analysis techniques, process models	Learning of assessment results, review, and feedback
3. Identify and select management alternatives	What kinds of changes can help solve the problem(s)? Are these changes feasible? Who would need to implement them and what incentives would be needed?	Design criteria and literature on management alternatives and their environmental impacts (e.g., best-management practices)	Process models, decision-support systems, economic analysis	Evaluate feasibility of alternatives; prioritize alternatives; facilitate education of wider community
4. Implement and evaluate	Are incentives effective in encouraging management changes? Are the changes effective in reaching water-resource management goals? What further improvements can be made?	Monitoring data, surveys of new conservation practices, stakeholder feedback	Process modeling, model validation, statistical analysis of monitoring data	Full participants in ongoing project review, help determine if/when new alternatives are required

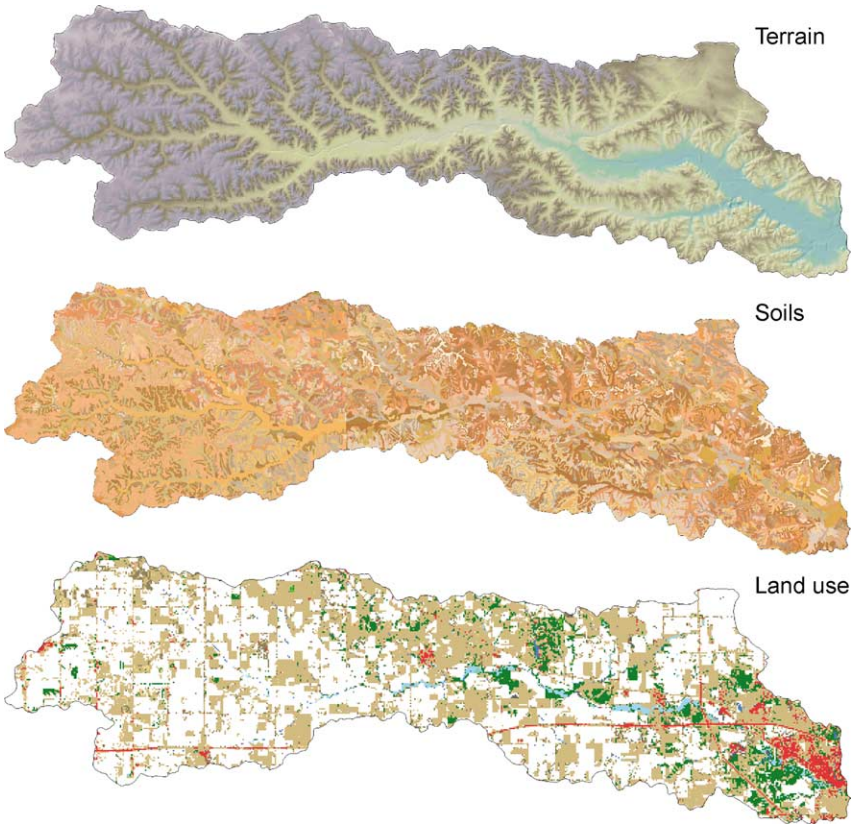


Figure 3 Watershed management and assessment must consider a spatial mosaic of land use, soil, and terrain.

soil, terrain, and geology (Figure 3). Existing hydrologic and water-quality conditions result from these interactions in space and their integration across time. Groundwater quality and stream baseflow are particularly influenced by past land-use practices, because of the slow rate of water movement through most aquifers. A key task for watershed assessment is to segregate this mosaic of spatial and temporal effects, and identify locations and activities that disproportionately influence water resources. Judgment is needed to do this. While there may be obvious areas of focus (e.g., shallow sandy soils, steep slopes, riparian areas), analytic capabilities in this area are not yet fully developed. Many of our working assumptions about effects of land use on hydrology and water quality have been gained through research carried out in small plots, hillslopes, or at small watershed scales. However, analyses conducted at a fine scale may not transfer directly to a large basin. The prominence of biophysical processes affecting hydrology and water quality can vary with the scale of observation. New techniques to analyze hydrologic processes across complex terrain, and analyze relationships between patterns and physical processes, are being developed that may provide new approaches to watershed assessment.

Trends in land use and population must also be considered during watershed assessment. The challenges of watershed management increase with population. The impact can be similar to drought, because pressures on water supplies and on land resources that yield water can increase concurrently. Urban expansion may be accompanied by increased disturbance from overgrazing and deforestation. Trends toward improved management practices should also be documented during an assessment, to help develop information about adoptability and effectiveness of new practices.

Monitoring of flows and water quality can also be important in assessment, because this helps to document water-resource problems. However an adequate set of monitoring data will not always be available, and usually the planning process cannot wait to acquire one. Nevertheless, it is important to continue monitoring, or initiate and/or expand it as soon as possible. Long-term data, gathered using consistent protocols for sampling and analysis, provide the best basis to assess the effects of new practices that are implemented.

Identifying management alternatives (phase 3) aims to propose and select realistic solutions to water-resource problems. Potential solutions include many

land-use practices, vegetation management changes, water-conservation measures, and engineered structures, across all types of land use throughout the watershed. These alternatives are not equal in terms of acceptability, cost, or effectiveness. Decision-support systems and economic analyses can be used to help rank social impacts (e.g., stakeholder acceptance, costs) of each alternative. The rankings also depend on the benefit of each alternative for water resources. Often, these benefits can be estimated using simulation models, which calculate how an alternative might affect biological and physical processes that determine a watershed's hydrology and/or water quality. A wide array of models are available, and the best model to use depends on the specific resource concern(s) and experience of the technical team. It may be possible to use monitoring data to validate model accuracy under current conditions, and this can improve the confidence in using the model to assess the effectiveness of alternatives. The final listing of ranked alternatives is likely to require several iterations of assessing social and environmental impacts.

Once alternatives are ranked, implementation of new practices is encouraged using some set of incentives, and then the management changes are evaluated (phase 4). This takes a sustained commitment. Monitoring data are needed to assess effects of new measures. Changes may not have immediate impact, particularly if they are aimed to improve groundwater conditions. New vegetation management strategies, particularly for forests, may take years to become fully effective. Seasonal trends and extreme events may make it difficult to confirm changes based on interpretation of monitoring data. External factors may also affect success; for example, an understanding of how a changing global climate is affecting water resources may become important. (*See* **Climate Change Impacts**.) These factors reinforce the need for a long-term commitment to watershed management, sustainable development, and periodic review of water-resource management goals and best ways to achieve them.

List of Technical Nomenclature

Baseflow — The portion of stream discharge that originates as groundwater, and that enters the stream channel from the saturated zone via lateral or upward-moving groundwater flow (volume per time)

Drought — A period of time with below-average precipitation that results in significantly diminished stream discharge and water-table elevations

Flood — An event during which a stream's discharge exceeds the capacity of the stream channel to convey water, so that water rises above the stream banks and covers some of the adjoining land surface

Hydrograph — A plot of stream discharge versus time that is often used to show the response of stream discharge to a precipitation event

Infiltration — Vertical entry of water into the soil at the land surface (length per time)

Interception — An amount or fraction of precipitation that wets vegetation and is eventually evaporated from vegetation so that it does not reach the soil surface (length)

Interflow — Water that moves laterally downslope beneath the land surface and at shallow depth above the saturated zone. This flow generally takes place at the scale of an individual hillslope (volume per time)

Perennial stream — A channel that conveys water throughout the year in most years. Usually, perennial flow only occurs where some of the stream discharge originates as baseflow

Physiography — Physical attributes of the terrain, including topography, soils, landform, and surficial geology

Recharge — Water that percolates down through the unsaturated zone and is added to groundwater storage (length)

Runoff — Water that flows across the land surface (volume per time)

Stakeholder — Individuals and/or groups that participate in assessing and solving resource management problems of direct importance to their community or representative constituency

Stream discharge — The quantity of water passing a location on a stream or river (volume per time)

Variable-source area — An area on the landscape that is prone to become saturated to the soil surface in response to precipitation or snowmelt, and that, when saturated, generates runoff that contributes to stream discharge. ('Partial contributing area' is an equivalent term)

Water-quality standard — A threshold value for a constituent concentration or other attribute of water, above or below which a specific use or ecologic attribute of the water becomes impaired

Watershed — A mapped area that contributes runoff or baseflow to a perennial stream or river at a defined location

Water yield This is equivalent to a unit-area stream discharge (volume per area per time)

See also: **Climate Change Impacts**; **Environmental Monitoring**; **Erosion:** Water-Induced; **Geographical Information Systems**; **Land-Use Classification**; **Overland Flow**; **Remote Sensing:** Soil Moisture; **Sustainable Soil and Land Management**; **Terraces and Terracing**; **Water Harvesting**

Further Reading

American Society of Civil Engineers (1996) *Hydrology Handbook*. ASCE Manual No. 28, Task Committee, Management Group D. New York: ASCE.

Brooks KN, Folliott PF, Gregersen HM, and DeBano LF (1997) *Hydrology and the Management of Watersheds*. Ames, IA: Iowa State University Press.

Downes BJ, Barmuta LA, Fairweather PG *et al.* (2002) *Monitoring Ecological Impacts: Concepts and Practice in Flowing Waters*. Cambridge, UK: Cambridge University Press.

Folliott PF, Baker MB Jr, Edminster CB, Dillon MC, and Mora KL (2002) *Land Stewardship through Watershed Management: Perspectives for the 21st Century*. New York: Kluwer Academic/Plenum.

Grayson R and Bloschl G (2000) *Spatial Patterns in Catchment Hydrology: Observations and Modelling*. Cambridge, UK: Cambridge University Press.

Heathcote IW (1998) *Integrated Watershed Management: Principles and Practice*. New York: John Wiley.

Singh VP (ed.) (1995) *Computer Models of Watershed Hydrology*. Highlands Ranch, CO: Water Resources Publications.

Trudgill ST (ed.) (1995) *Solute Modelling in Catchment Systems*. Chichester, UK: John Wiley.

Vieux BE (2001) *Distributed Hydrologic Modeling using GIS*. Boston, MA: Kluwer Academic.

Wilson JP and Gallant JC (eds) (2000) *Terrain Analysis*. New York: John Wiley.

Younos T (ed.) (2001) *Advances in Water Monitoring Research*. Highlands Ranch, CO: Water Resources Publications.

WATER-USE EFFICIENCY

M B Kirkham, Kansas State University Manhattan, KS, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

For centuries, humans have been concerned with the efficient use of water in the production of crops. The ability to grow crops and manage their needs for water is necessary for civilization. Water-use efficiency is defined as the aboveground biomass production per unit area per unit water evapotranspired. The biomass is usually determined as dry weight rather than as fresh weight. Therefore, water-use efficiency is expressed in equation form as follows:

$$\text{Water-use efficiency (WUE)} = \frac{\text{Dry weight production (DW)}}{\text{Evapotranspiration (ET)}} \quad [1]$$

Water-use efficiency can be expressed on the basis of vegetative growth or reproductive (grain) growth; the basis must be stated. Different units can be used in the numerator and denominator of Eqn [1]. Old units were pounds or tons of dry weight produced per acre-inch of water evapotranspired. We now usually express it as a simple ratio, such as kilogram of dry weight per kilogram of water transpired. Water-use efficiency expressed in the latter dimensionless form is similar to the reciprocal of the old terms 'water requirement' and 'transpiration ratio.' New portable gas analyzers allow us to measure photosynthetic rate and transpiration rate on individual leaves or on parts of individual leaves as small as $6\,cm^2$. So we now can express water-use efficiency on a leaf basis, as follows:

$$\text{Leaf water-use efficiency} = \frac{\text{Photosynthetic rate}}{\text{Transpiration rate}} \quad [2]$$

The unit for photosynthetic rate is micromoles of CO_2 per square meter per second, and the unit for transpiration rate is micromoles of H_2O per square meter per second, so the unit for water-use efficiency on a leaf basis is micromoles of CO_2 per micromole of H_2O.

Water-use efficiency can be based either on evapotranspiration ('ET efficiency') or on crop transpiration ('T efficiency'). The difference is important, because suppression of soil-water evaporation and prevention of weed transpiration can improve ET efficiency. However, it need not improve T efficiency, which is a measure of crop performance. These two water-use efficiencies may also be based on either the

total dry-matter production or the marketable yield, and the basis should always be stated.

In dry regions where the proportion of water returned to the atmosphere by evaporation may be 90% or more, it is necessary to increase water-use efficiency. That is, we want to produce more plant material with smaller amounts of water. Many schemes have been proposed to obtain water for these dry regions: artificial rainmaking, sea-water distillation, and towing icebergs to regions where the melted water can be used for irrigation. However, increasing water-use efficiency may be the best way to provide enough water for crop production in areas where evapotranspiration exceeds rainfall.

History

At the start of the twentieth century, considerable work was in progress measuring water requirement, the inverse of T efficiency. Franklin Hiram King (1848–1911), at the University of Wisconsin (USA), was the first in the USA to research the water required to produce field crops. King was also the author of a classic work entitled *Farmers of Forty Centuries*, which describes his trip to China in the early 1900s. He made careful notes to help him understand how people could farm the same fields for 4000 years without destroying their productivity. King used small lysimeters in a greenhouse and in fields to determine the water requirement of crops. Widtsoe, at the University of Utah (USA), and Kiesselbach, at the University of Nebraska (USA), initiated research in 1902. Widtsoe is also well known for his studies on dry farming. The methods that he described for agricultural production in areas with low rainfall apply equally well today as they did at the beginning of the twentieth century.

Briggs and Shantz, based in Akron, Colorado, compiled much of the early water-requirement data from container experiments that began in 1910. The water-requirement work of Kiesselbach and of Briggs and Shantz was concerned with total dry matter and transpiration on a plant basis rather than on a land-area basis. Briggs, Shantz, and Kiesselbach reviewed the work that had been done up to that time, although they omitted water-culture experiments and those on seedlings. They found that 14 researchers had worked on water-use efficiency prior to 1900. Since that time, there has been an increasing amount of work on water use and water-use efficiency, all done with more elaborate equipment and experiments than those of the last half of the 1800s and early 1900s. However, these early experiments still remain the foundations of our modern understanding of water-use efficiency. One of the best summaries of the water requirement of many different plants is that compiled in 1927 by Shantz and Piemeisel.

Factors that Influence Water-Use Efficiency

Soil Factors

Soil-water content Briggs and Shantz felt that they could make no firm conclusion about the effect of soil-water content on the water requirement. In some experiments, however, the water requirement usually decreased a little when growth was limited by water deficits. The same result was obtained by Kiesselbach, who found that, with a deficient water regime, the dry matter of corn stalks, ears, and leaves were reduced 37%, 28%, and 10%, respectively. The water requirement based on ear weight decreased 4.3%, and the water requirement based on the total dry matter decreased 10%. Later Briggs and Shantz showed that the water requirement of wheat was essentially the same under deficient soil water (dry matter decreased 40%) as when water was adequate. King reached the same conclusion from experiments with corn, oats, and potato. Several decades later, de Wit analyzed experimental data that also showed similar T efficiencies regardless of water shortages.

Measurement of soil-water evaporation independently of plant transpiration is difficult, and models have been developed to separate the two. Evaporation can be measured directly with evaporimeters. Once a canopy covers the ground, most water is lost by transpiration. However, because water is lost by soil-water evaporation, investigators have studied evaporation-retardant chemicals, mulches, and plastic films. Results vary; for example, Letey and Peters have compared water-use efficiency of corn on plots with and without polyethylene film. The covered plots were initially wet and then covered to cut off summer rainfall and evaporation. Control plots received natural rainfall. Water-use efficiency for the covered plots was 345 kg per 100 m^3 of water and for the natural rainfall plots it was 148 kg per 100 m^3. However, in another year of the study, results showed no advantage from the plastic film. Plastic film is expensive and has been used only on high-value horticultural crops. But as water becomes more valuable, antitranspirants, mulches, and plastic films might be economically feasible for field crops, and water-use efficiencies with them need to be determined.

Method of irrigation Israel has dramatically increased its water-use efficiency through the use of new irrigation methods begun approximately in the 1960s (Figure 1). They are high-frequency,

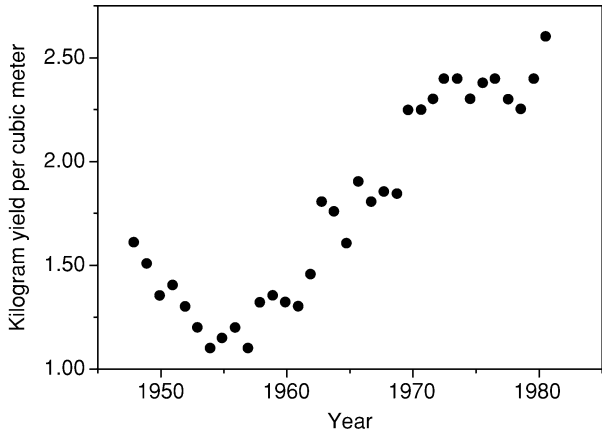


Figure 1 The improvement of crop water-use efficiency in Israel. (From *Rivers of Eden: The Struggle for Water and the Quest for Peace in the Middle East* by Daniel Hillel, © 1994 by Daniel Hillel. Oxford University Press, Inc.)

low-volume techniques of drip (trickle) and microsprayer irrigation. Before the new methods were developed, irrigations (flood or sprinkler) were usually infrequent and large, because it was assumed that soil moisture remained essentially equally available to crops until evaporation and extraction by roots depleted it to the permament wilting point. With the old methods, the topsoil was saturated after an irrigation, a condition that leached nutrients and deprived roots of aeration, followed by a period in which the crops were allowed to dry to a degree that desiccated the roots. Crops often show a pronounced increase in yield when irrigations are provided in sufficient quantity and frequency that water never becomes limiting. The newer irrigation methods became possible with the development of porous tubes for subirrigation and the advent of low-cost tubing that can be fitted with specially designed drip emitters. Drip and microsprayer irrigations can maintain the root zone in a moist yet unsaturated condition, so the roots do not lack water or oxygen. The irrigations are targeted precisely at the base of crop plants, thus wetting only a small fraction of the surface and not the interrow areas, where compaction by traffic, evaporation, and weed proliferation is reduced. Prevention of weed growth further conserves water.

Plant Factors

Species adaptation Because most species are adapted to grow with specific water requirements, plants must be matched with the rainfall. The record of human domestication of plants dates back at least 6000 years, and plant introductions (i.e., nonnative plants introduced into a region from another region) permitted our first great stride forward in efficient water use. In the USA, plant introductions were sponsored by the US Department of Agriculture. Immigrants also brought plants; for example, Turkey winter wheat was introduced into the Great Plains by Russian Mennonites, and winter wheat still is the most important crop in the region.

Plant breeding Briggs and Shantz have shown that plants differ greatly in water requirement. They carried out their research before the photosynthetic pathways had been elucidated. In all plants, photosynthesis involves the 'C_3' process that converts carbon dioxide into molecules with three carbons. In some species, called 'C_4' plants, carbon dioxide is first converted into molecules with four carbons, which are then transported within the leaf to sites (bundle-sheath cells) where carbon dioxide is released, providing a high concentration of carbon dioxide for the C_3 process. The same enzyme that catalyzes the first step in the C_3 process (ribulose 1, 5 bisphosphate carboxylase) can also catalyze an oxidation that leads to photorespiration, which is respiration that occurs in the light. Photorespiration consumes as much as a third of the carbon dioxide that the plant has absorbed in the light. It is slowed by high concentrations of carbon dioxide and, hence, occurs more rapidly in C_3 plants such as wheat than in C_4 plants such as maize. Plants with the C_4 type of photosynthesis have a lower water requirement than plants with the C_3 type of photosynthesis. The reasons for this difference are not fully understood, but it is probably due, in part, to the fact that, under the same environmental conditions, C_4 plants generally have a higher stomatal resistance than C_3 plants. The difference in water use between C_3 and C_4 plants was evident even in the work of Briggs and Shantz (Table 1). Plants which we know today have the C_4 type of photosynthesis (e.g., maize, millet, sorghum) have a lower water requirement than plants with the C_3 type of photosynthesis (e.g., alfalfa, barley, wheat). On average, plants with the C_4 type of photosynthesis have a water requirement of 250–350 g $H_2O\,g^{-1}$ dry weight, and plants with the C_3 type of photosynthesis have a water requirement of 450–950 g $H_2O\,g^{-1}$ dry weight. Alfalfa's water requirement can be greater than this; it has the highest water requirement of any crop species.

A third type of photosynthesis occurs, but the only commercially important plant that has this type of photosynthesis is pineapple. It is called the crassulacean acid metabolism (CAM) photosynthetic pathway. Unlike most plants that have stomata open in light, when carbon dioxide is taken up (and water is lost by transpiration), plants with CAM photosynthesis keep their stomata open in darkness and take

Table 1 Water requirement based on aboveground vegetative dry matter for various crops as determined by Briggs and Shantz before the knowledge of C_3 and C_4 photosynthetic pathways. Photosynthetic pathway of each crop is noted in parentheses

Crop	*Water requirement (g H_2O g^{-1} dry wt)*
Wheat (C_3)	507
Oats (C_3)	614
Barley (C_3)	539
Rye (C_3)	724
Corn (C_4)	369
Sorghum (C_4)	309
Millet (C_4)	275
Peas (C_3)	800
Sweet clover (C_3)	709
Alfalfa (C_3)	1068
Buckwheat (C_3)	578
Rape (C_3)	441
Potatoes (C_3)	448
Sugar beets (C_4)	377
Salsola (saltwort of the Chenopodiaceae family) (C_4)	366
Amaranthus (C_4)	303
Artemisia (C_3)	765

Source: Briggs LJ and Shantz HL (1913) *The Water Requirement of Plants. II. A Review of the Literature.* US Department of Agriculture, Bureau of Plant Industry, Bulletin No. 285, p. 90. Washington, DC: US Department of Agriculture.

up carbon dioxide at night. The photosynthetic enzymes that CAM plants use are the same ones as found in C_4 plants, but their activity depends upon whether it is light or dark. Because CAM plants keep their stomata open at night and not during the daylight hours when heat and light cause great amounts of water to be lost through the stomata, they have the lowest water requirement of any plants. On average, their water requirement is 50–55 g H_2O g^{-1} dry weight.

Not only do species differ in water requirement, but also cultivated varieties (cultivars) within a species vary. This was noted in 1913 by Briggs and Shantz, who suggested the possibility of developing varieties that are efficient in the use of water.

When water becomes limited during a plant's development, yield differences among species and varieties are due to a plant's water efficiency. Plant characteristics that contribute to such efficiency include maturity (including planting date); leaf area and angle of inclination; leaf rolling; number, distribution, and size of stomata; presence of waxy or corky epidermal cells; the ability to become temporarily dormant; and extensiveness of the root system. Distribution of roots is especially important. Even though roots may be at an optimum depth, the surface roots are often the ones that preferentially extract water after rewatering.

Cultural Factors

Planting patterns Planting patterns have a direct effect on yield, solar-energy capture, and soil-water evaporation and thus an indirect effect on water-use efficiency. Soil-water evaporation is diminished with denser plantings. In humid regions where rainfall exceeds evapotranspiration, plant populations can be increased with a concomitant increase in yield. However, in semiarid regions, when plant populations are increased, the plants most often need to be irrigated. For example, dense planting of sorghum has been advocated as a beneficial method for rain-fed land in Kansas, because it reduces soil-water evaporation, runoff, erosion, and weeds. But the super-thick plantings fail to produce grain if soil moisture is not adequate.

Seed quality A crop-management problem, sometimes overlooked, is seed quality. For most efficient use of water, a grower must start with high-quality seed. Poor seed may mean not only poor germination and weak plants that cannot take advantage of the available water, but also such seed may include weed seeds and thus provide competitors for the water present.

Weeds One of the primary management means of obtaining more efficient water use is the elimination of weeds in crops. Weeds compete with crops for soil nutrients, water, and light. Except in high-rainfall areas, the primary concern is the water factor. The water requirement of many weeds is greater than that of crop plants, because many weeds are C_3-type plants. For example, the average ragweed plant requires three times as much water per pound (0.45 kg) of dry matter produced as a maize plant, a C_4 plant.

Disease and insect pests Few data concerning effects of disease and insect pests on water-use efficiency are available. Rusts decrease transpiration in the early stage of infection, but, on sporulation, they cause epidermal leakage of water vapor and increased water loss in light and dark conditions. The powdery mildews cause a similar decrease in transpiration and then an increase in transpiration. These transpiration trends have been found to decrease the transpiration efficiency. For example, studies show that early rust infection halves the transpiration efficiency of wheat, when based on total dry matter, and, when based on grain yield, it is 25-fold smaller. Infection at later stages of development produces correspondingly smaller decreases. Water-use efficiency data for vascular wilts, root rots, and insect pests are lacking. Breeding for resistant varieties is probably the most economic and longest-lasting method to control

diseases and insects. Efficient crop and water management also includes the wise use of herbicides and insecticides, and their effect on water-use efficiency needs to be determined.

Tillage The surface characteristics of soil can have a profound effect on the water that infiltrates and runs off. Reduced or minimum tillage has been taken up by some farmers in the semiarid Great Plains. While yields are often reduced during the first few years after reduced tillage is initiated (due, for example, to difficulty in controlling diseases and insects), water is conserved. This results in less erosion in wet years and longer survival of crops in droughts, because the mulch that remains on the surface of the soil reduces evaporation. However, water-use efficiency may not be increased. Stubble-mulch tillage in the Great Plains does not necessarily alter rainfall-use efficiency.

Rotations While studies indicate that rotations can improve infiltration and water use, they do not always. For example, a grass–legume crop may have low infiltration rates because of heavy grazing. In Kansas, a deep-rooted alfalfa–fescue mixture was grown on a claypan to increase root penetration of maize, planted the following year. However, the deep-rooted plants removed moisture throughout the soil profile, resulting in low yields of maize. Continuous maize yielded more than maize grown in rotation. Sunflower, with its deep roots and a high water requirement, removes more water from a soil profile than sorghum. Under dry-land conditions, crops planted following sunflower will lack stored moisture. Thus, rotations may not generally improve available water and may result in less available moisture when the preceding crop has a high water requirement.

Fertilization Data have been reviewed from about 20 experiments showing the effect of varying fertility on water requirement and the conclusion is that, with poor soils, the water requirement may be reduced one-half to two-thirds by increasing fertility. In all but five experiments, evaporation from the soil was included in the water use and could bias the water requirement for smaller plants with less transpiration. In experiments where evaporation was prevented, the water requirement did not increase significantly as the fertility and yield decreased, until the dry weight of the plants had decreased approximately 50% because of malnutrition. It appears from these data and others that, unless malnutrition is severe, transpiration efficiency based on total dry matter is not greatly affected by poor fertility. But, when nutrient deficiency reduces yield to about half that on a well-fertilized soil, then transpiration efficiency is reduced markedly.

Climate

The seasonal changes in both transpiration and evapotranspiration efficiencies found with exposures to different radiation, temperature, and humidity regimes show that climate exerts a major influence on water-use efficiencies. Even the early researchers recognized that, with well-watered plants, variation of transpiration with climate causes changes in water-use efficiency as the season progresses, between years and between locations, and that free-water (pan) evaporation or saturation deficit could be used to normalize the transpiration component (Table 2).

The carbon dioxide concentration in the atmosphere is increasing due largely to the increased burning of fossil fuel by industry and automobiles. Since 1958, the carbon dioxide concentration has risen from 316 ppm ($316\,\mu l\,l^{-1}$) to approximately 370 ppm. The concentration increases at a rate of $1.5–2.0\,\mu l\,l^{-1}$ per year. When atmospheric carbon dioxide is relatively low, net photosynthesis is faster in C_4 plants than in C_3 plants, but, at higher levels of carbon dioxide, the change in photorespiration per change in carbon dioxide concentration leads to greater increases in net photosynthesis for C_3 plants than for C_4 plants. Therefore, increasing levels of carbon dioxide should benefit C_3 plants more than C_4 plants. The increasing levels of carbon dioxide do increase the water-use efficiency of C_3 plants. When winter wheat was grown for 3 years with four different levels of carbon dioxide (ambient or $340\,\mu l\,l^{-1}$

Table 2 Water requirement (WR) of an alfalfa cultivar (Grimm) and pan evaporation at four different stations in the Great Plains of the USA, as reported by Briggs and Shantz

Location	*Growth period (1912)*	*Water requirement (g $H_2O\,g^{-1}$ dry wt)*	*Mean pan evaporation (mm day^{-1})*	*WR/pan*
Williston, N. Dakota	29 Jul–24 Sept	518 ± 12	4.04	128
Newell, S. Dakota	9 Aug–6 Sept	630 ± 8	4.75	133
Akron, Colorado	26 Jul–6 Sept	853 ± 13	5.74	149
Dalhart, Texas	26 July–31 Aug	1005 ± 8	7.77	129

Reproduced with permission from Tanner CB and Sinclair TR (1983) Efficient water use in crop production: research or re-search? In: Taylor HM, Jordan WR, and Sinclair TR (eds) *Limitations to Efficient Water Use in Crop Production,* pp. 1–27. Madison, WI: American Society of Agronomy, Crop Science Society of America/Soil Science Society of America.

and 485, 660, and 825 $\mu l\,l^{-1}$), the water requirement decreased with increasing levels of carbon dioxide (Figure 2). During the 3 years of the study, water required to produce a gram of grain under the highest carbon dioxide concentration (825 $\mu l\,l^{-1}$) and dry conditions (half-field capacity) was less (547 $ml\,g^{-1}$) than that required to produce a gram of grain under well-watered conditions (field capacity) at the ambient carbon dioxide concentration (642 $ml\,g^{-1}$). Under well-watered conditions, the water requirement of the wheat grown with ambient and elevated carbon dioxide was 642 $ml\,g^{-1}$ and 458 $ml\,g^{-1}$, respectively (a reduction of water requirement, based on grain, of 29%).

Even though predictions indicate that C_3 plants will have a higher water-use efficiency with elevated carbon dioxide than C_4 plants, data also show that C_4 plants have increased water-use efficiency. When sorghum, kept under watered conditions, was grown under ambient (330 $\mu l\,l^{-1}$) and elevated (795 $\mu l\,l^{-1}$) levels of carbon dioxide, water requirement for grain production was 1090 $ml\,g^{-1}$ and 776 $ml\,g^{-1}$, respectively (a 29% reduction in water requirement with elevated carbon dioxide).

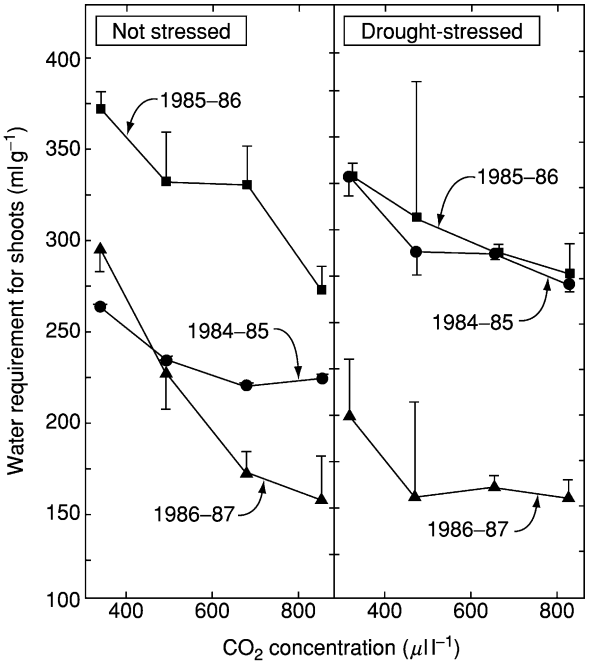


Figure 2 Water requirement for winter wheat grain grown under well-watered (field capacity) and drought (half-field capacity) conditions as affected by carbon dioxide concentration during a 3-year study. Vertical bars, ± standard deviation. Only half of each bar is drawn for clarity. (Adapted from Chaudhuri UN, Kirkham MB, and Kanemasu ET (1990) Carbon dioxide and water level effects on yield and water use of winter wheat. *Agronomy Journal* 82: 637–641, with permission.)

Measurement of Water-Use Efficiency

Water-use efficiency is usually measured by harvesting plants, determining dry weight of the vegetative portion or grain, and dividing that by the rainfall or irrigation plus rainfall. Weighing lysimeters have allowed more precise measurements of water used. With the development of portable photosynthetic systems, researchers can now measure in the greenhouse or field the water-use efficiency of individual leaves or parts of leaves. Figure 3 shows water-use efficiency determined on different locations of a leaf of puka (*Meryta sinclairii* Seemann), a native tree of New Zealand. The leaves are large and elliptical, up to 0.6 m long and 0.3 m wide. The center of the leaf (to the side of the main vein) and the middle edge have the highest photosynthetic rates and transpiration rates, but the tip of the leaf has the highest water-use efficiency. The base of the leaf has the lowest photosynthetic rate, the lowest transpiration rate,

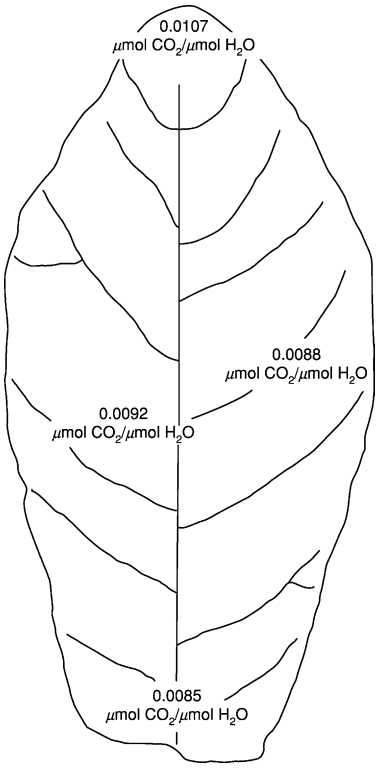


Figure 3 Water-use efficiency at four locations on a leaf of puka, a native New Zealand tree. The locations are tip, edge of middle, base, and center to the side of the main vein. To obtain these values, photosynthetic rates and transpiration rates were measured on puka leaves during a 1-month period in the summer of 1998 in Lincoln, New Zealand. From 14 to 34 values were measured. These values were divided to get the water-use efficiencies shown. The leaf traced for the figure is 25.7 cm long and 12.7 cm wide.

and the lowest water-use efficiency. These results show the power of modern measurement techniques. We can now define with fine spatial resolution the water-use efficiency of parts of individual leaves and try to understand the physiological reasons for the differences.

Another method of determining transpiration efficiency is carbon isotope discrimination. In this method, the discrimination (Δ) of the stable isotope of carbon, ^{13}C, relative to the more abundant ^{12}C, is determined. The discrimination occurs because of the different diffusivities of ^{13}C and ^{12}C across the stomata and the fractionation by the C_3 enzyme, ribulose bisphosphate carboxylase. The method is limited to laboratories that can afford an expensive mass spectrometer with the ability to measure the ratio of ^{13}C to ^{12}C with high precision. However, the method has shown that genetic variation in transpiration efficiency in wheat cultivars does exist, and the variation is approximately twofold. The genetic variation in transpiration efficiency is negatively correlated in C_3 species with the discrimination of ^{13}C relative to ^{12}C, as shown in Figure 4. In C_4 species, Δ does not correlate with transpiration efficiency.

Possibilities of Increasing Water-Use Efficiency

With careful isotopic measurements, C_3 plants that vary slightly in water-use efficiency can be selected. If one cultivar has only a small increase in water-use efficiency compared with another and if grown over many hectares, this small difference could have a significant effect on water conservation.

Genetic engineering for increased water-use efficiency might be a possibility in the far future; for example, incorporating the C_4 photosynthetic pathway into C_3 plants. In the meantime, the increasing carbon dioxide concentration in the atmosphere seems to offer the greatest potential for increasing water-use efficiency. Leaf photosynthetic rates and leaf transpiration rates of a C_3 plant (Kentucky bluegrass) and a C_4 plant (big bluestem) grown in the field under ambient and elevated (twice ambient) levels of carbon dioxide have been measured. The C_3 plant has a photosynthetic rate similar to the C_4 plant under elevated carbon dioxide (Figure 5). The transpiration rate of the C_3 plant under the low and high levels of carbon dioxide is similar. So the leaf water-use

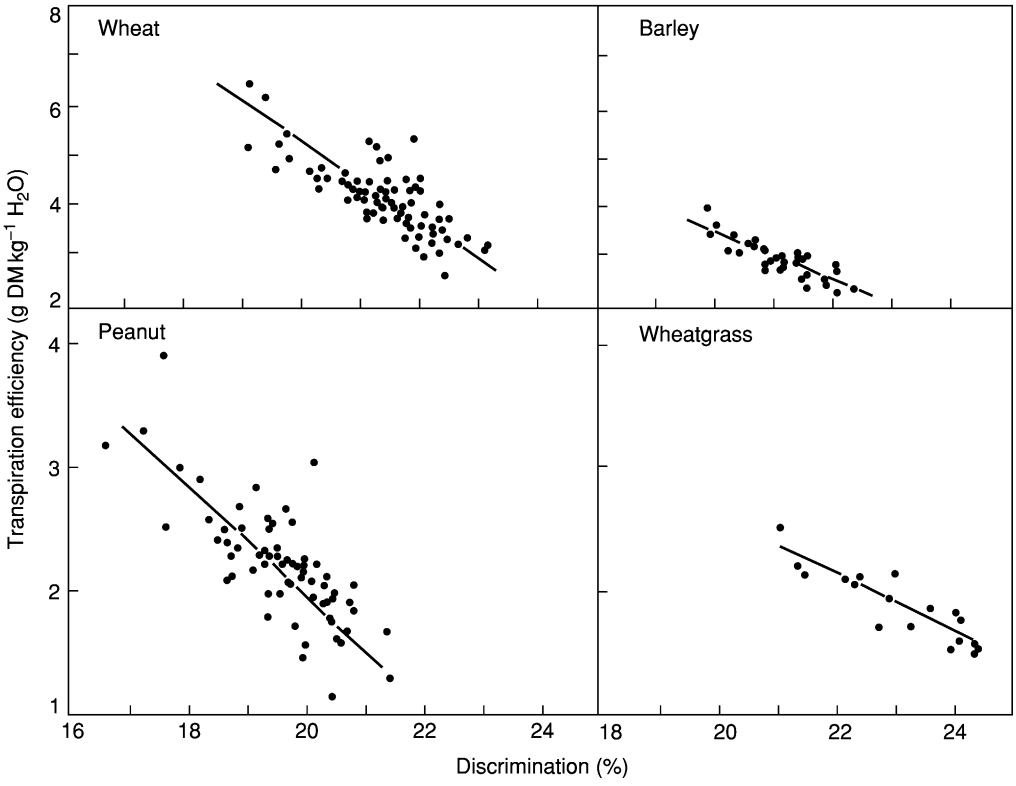


Figure 4 Relationship between transpiration efficiency and carbon isotope discrimination in genotypes of four C_3 plants: wheat, barley, peanut, and crested wheatgrass. DM, dry matter. (Reproduced with permission from Turner NC (1993) Water use efficiency of crop plants: potential for improvement. In: Buxton DR, Shibles R, Forsberg RA, *et al.* (eds) *International Crop Science I,* pp. 75–82. Madison, WI: Crop Science Society of America.)

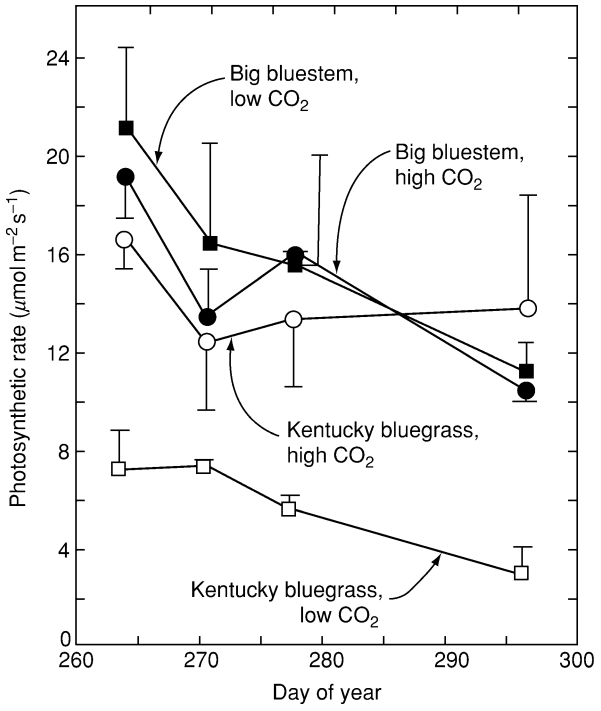


Figure 5 Photosynthetic rate of Kentucky bluegrass (open symbols) and big bluestem (closed symbols) with a high (twice ambient; circles) and a low (ambient; squares) atmospheric carbon dioxide concentration. Vertical bars, ±standard deviation. (Reproduced with permission from He H, Kirkham MB, Lawlor DJ, and Kanemasu ET (1992) Photosynthesis and water relations of big bluestem (C_4) and Kentucky bluegrass (C_3) under high concentration carbon dioxide. *Transactions of the Kansas Academy of Science* 95: 139–152.)

efficiency of the C_3 plant is increased three times due to the doubling of the ambient carbon dioxide–just about the same amount as the photosynthetic rate is increased, if the values are averaged throughout the experiment. Even though increasing carbon dioxide concentrations may have negative consequences (e.g., favoring of weed growth over crop growth), it appears that this does have the positive effect of increasing water-use efficiency.

See also: **Crop Water Requirements**; **Plant–Soil–Water Relations**; **Soil–Plant–Atmosphere Continuum**

Further Reading

Allen RG, Pereira LS, Raes D, and Smith M (1998) *Crop Evapotranspiration. Guidelines for Computing Crop Water Requirements*. Irrigation and Drainage Paper No. 56. Rome, Italy: Food and Agricultural Organization of the United Nations.

Briggs LJ and Shantz HL (1913) *The Water Requirement of Plants. I. Investigations in the Great Plains in 1910 and 1911*. US Department of Agriculture, Bureau of Plant Industry Bulletin No. 284. Washington, DC: US Department of Agriculture.

Briggs LJ and Shantz HL (1913) *The Water Requirement of Plants. II. A Review of the Literature*. US Department of Agriculture, Bureau of Plant Industry Bulletin No. 285. Washington, DC: US Department of Agriculture.

de Wit CT (1958) *Transpiration and Crop Yields*. Verslag Landbouwkundig Onderzoek (Agricultural Research Report) No. 64.6. Instituut voor Biologisch en Scheikundig Onderzoek van Landbouwgewassen. Wageningen, The Netherlands: Institute of Biology and Chemical Research on Field Crops and Herbage.

Doorenbos J and Pruitt WO (1975) *Guidelines for Predicting Crop Water Requirements*. Irrigation and Drainage Paper No. 24. Food and Agricultural Organization of the United Nations. Rome, Italy: FAO.

Heschel MH, Donohue K, Hausmann N, and Schmitt J (2002) Population differentiation and natural selection for water-use efficiency in *Impatiens capensis* (Balsaminacea). *International Journal of Plant Science* 163: 907–912.

Hillel D (1994) *Rivers of Eden. The Struggle for Water and the Quest for Peace in the Middle East*. New York: Oxford University Press.

Hillel D (1998) *Environmental Soil Physics*, pp. 630–634. San Diego, CA: Academic Press.

King FH (1901) *A Text Book of the Physics of Agriculture. Strive to Know Why, for this Teaches How and When*, 2nd edn. Madison, WI: F.H. King.

Kirkham MB (ed.) (1999) *Water Use in Crop Production*. New York: The Haworth Press.

Letey J and Peters DB (1957) Influence of soil moisture levels and seasonal weather on efficiency of water use by corn. *Agronomy Journal* 49: 362–365.

Pendleton JW (1966) Increasing water use efficiency by crop management. In: Pierre WH, Kirkham D, Pesek J, and Shaw R (eds) *Plant Environment and Efficient Water Use*, pp. 236–258. Madison, WI: American Society of Agronomy/Soil Science Society of America.

Rosenberg NJ (1974) *Microclimate: The Biological Environment*. New York: John Wiley.

Salisbury FB and Ross CW (1978) *Plant Physiology*, 2nd edn. Belmont, CA: Wadsworth.

Shantz HL and Piemeisel LN (1927) The water requirement of plants at Akron, Colo. *Journal of Agricultural Research* 34: 1093–1190.

Tanner CB and Sinclair TR (1983) Efficient water use in crop production: research or re-search? In: Taylor HM, Jordan WR, and Sinclair TR (eds) *Limitations to Efficient Water Use in Crop Production*, pp. 1–27. Madison, WI: American Society of Agronomy, Crop Science Society of America/Soil Science Society of America.

Widtsoe JA (1911) *Dry-farming. A System of Agriculture for Countries Under a Low Rainfall*. New York: Macmillan.

WEED MANAGEMENT

D D Buhler, Michigan State University, East Lansing, MI, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Soil tillage has been a major element of crop production for centuries. However, tillage has negative side-effects, including increased soil erosion and high labor and fuel requirements. Methods for reducing tillage were proposed as early as the 1920s. For row-crop production, no-tillage methods became much more feasible beginning in the late 1950s and 1960s as herbicides became available to replace tillage for control of unwanted vegetation.

Weed control refers to actions used to eliminate an existing weed population. Weed management is more than control of existing weed problems and places greater emphasis on preventing weed reproduction, reducing weed emergence after crop planting, and minimizing weed competition with the crop. Integrated weed management emphasizes combinations of techniques and knowledge that consider the causes of weed problems rather than react to problems after they occur. The goal of weed management is to optimize crop production and grower profit through the concerted use of preventive tactics, scientific knowledge, management skills, monitoring procedures, and efficient use of control practices.

One of the first steps in effectively managing weeds is understanding weed populations and their responses to tillage systems. As the impacts of tillage systems on weed population dynamics are examined, we must apply basic ecologic principles. Weeds are successful because of their genetic diversity and their ability to adapt to and take advantage of conditions created by crop-production systems. Therefore, modifying tillage practices results in an altered competitive environment in which the morphological and physiological traits that confer success are altered. Understanding weed population shifts leads to the identification of vulnerable stages in weed life cycles that can be exploited in management systems. Understanding population shifts also helps to identify species that are favored as tillage practices are changed, allowing for the development of weed-management systems that target the appropriate weed species.

Impacts of Tillage on Weed Management

General Effects

Crop-production practices exert selection pressure on weed communities and create niches that favor or disfavor various species. Since tillage has been an integral part of many cropping systems for centuries, it has played a major role in shaping the nature of weed communities in most agricultural lands.

Tillage buries crop residue and alters the characteristics of the surface soil, regulating the germination environment of seeds by reducing soil surface cover, altering soil temperature and moisture patterns, changing weed seed distribution in the soil, and disrupting the growth of winter annual, biennial, and perennial species. The term 'safe site' has been used to describe the complex conditions that are required for successful seed germination and seedling establishment. Several studies have pointed out that safe-site conditions vary greatly among plant species. The changes in safe-site characteristics, along with the physical disruption caused by tillage, combine to play a major role in weed-population dynamics in crop-production systems.

Control Options and Efficacy

Tillage systems impact weed-control options available to producers. Certain herbicides may not be used in conservation-tillage systems because of the need for mechanical incorporation into the soil after application. With less tillage and more plant residue on the soil surface, mechanical weed-control operations may become less effective. Rotary hoeing is especially difficult in untilled soil covered with residue of the previous crop. Interrow cultivation can be a component of some conservation-tillage systems and is an integral component of ridge-tillage. Combining interrow cultivation with reduced herbicide use has provided weed control similar to full-rate herbicide treatments in conservation-tillage systems. However, the effects of interrow cultivation on the long-term impacts of no-tillage to soil properties and soil-erosion reduction have not been documented.

Plant residue on the soil surface can alter the behavior of soil-applied herbicides. Depending on percentage surface cover, residue type, and herbicide formulation, up to 60% of the herbicide applied may be intercepted by residue. However, much of this herbicide is often washed off by subsequent

Table 1 Influence of tillage systems on control of selected summer annual weed species with soil-applied herbicides in *Zea mays* and *Glycine max* production systems

	Change relative to moldboard-plow systems[a]		
Weed species	*No-tillage*	*Chisel plow*	*Ridge-tillage*
Abutilon theophrasti	+	+, 0	+
Amaranthus retroflexus	+, 0,−	0,−	+, 0
Chenopodium album	+, 0,−	0,−	+, 0
Setaria faberi	−	−	0

[a]+, increased control relative to moldboard plow; 0, no change relative to moldboard plow;−, decreased control relative to moldboard plow.

rainfall and irrigation. The ultimate effect on weed control is determined by the timing of rainfall or irrigation relative to herbicide application, the level of residue cover, and the nature of the weed population.

Setaria faberi (giant foxtail) is more difficult to control with soil-applied herbicides in conservation- versus conventional-tillage systems (Table 1). Conversely, *Abutilon theophrasti* (velvetleaf) control increases as tillage is reduced. The effects of tillage systems on the control of small-seeded dicotyledon weed species vary by species and location. Herbicides applied after weed emergence are influenced less by tillage systems because they are applied directly to the target weeds and do not interact with residue on the soil surface.

The effect of tillage systems on herbicide efficacy is due to the combination of surface-residue interception of herbicides and changes in weed-population dynamics. For example, control of annual grass species is often reduced as tillage is reduced, while control of some annual dicot weeds increases. These variable responses indicate that choices of appropriate control tactics and application timing are critical in conservation-tillage systems. Because of the reliance on herbicides for weed control in most conservation-tillage systems, knowledge of the interactions of weed biology and ecology with herbicide activity is essential to managing weeds in conservation tillage without increasing herbicide use to environmentally and economically unacceptable levels.

Biology and Ecology of Weed Responses to Tillage Systems

Classification of Weed Species

Weeds can be classified in many ways, and understanding weed life cycles is important in predicting population shifts and designing management strategies. Weed species with different life cycles have different requirements for seed germination, seedling establishment, growth, and reproduction. Because tillage greatly alters the environment where weed species survive, altering tillage systems favors certain life cycles over others. In no-tillage systems, there has often been a marked drop in the variety of annual species. Concomitantly, there has been an increase in winter annual, biennial, and perennial weed species.

Weeds of agronomic crops are a unique group of plant species because of their ability to infest and often thrive in intensively disturbed habitats. In addition to classification by life cycle, a classification appropriate to understanding weed-population dynamics under different tillage systems follows:

1. Arable-response weed species that require periodic, regular disturbance of the soil for survival;
2. Inverse-response weed species that require undisturbed soil for survival;
3. Intermediate-response weed species that survive in both disturbed and undisturbed soil.

Combining characteristics of life cycles and tillage response of weed species is useful in predicting weed population shifts. For example, weed populations in reduced-tillage systems where summer annual crops are grown are dominated by arable- and intermediate-response summer annual species, because the soil is disturbed annually and management practices favor summer annual species. In no-tillage systems, weed populations shift toward inverse- and intermediate-response species. Winter annual, biennial, and perennial species are well adapted to no-tillage, because there is little or no tillage to interrupt the life cycle. In addition, summer annual species that do not require burial for establishment are well-adapted to no-tillage systems. However, since soil disturbance is required to plant crop seeds and incorporate fertilizers in annual cropping systems, arable-response species may continue to exist in no-tillage fields. Seed buried by previous tillage may also persist for several years and continue to be a source of arable-response species in no-tillage systems.

Summer Annual Species

Summer annual species are major weed problems in production systems dominated by summer annual crops, because they are well-adapted to the environment created by the production systems (i.e., similar life cycle, high fertility, annual disturbance, production in rows, and herbicide use). Changes that occur in the dynamics of annual weed communities as a function of tillage practices are regulated by species present in the field, biology of the species, and efficacy of weed-management practices.

The most commonly reported change in annual weed populations with decreased tillage is an increase in summer annual grass species, including *S. faberi*, *S. viridis* (green foxtail), *S. glauca* (yellow foxtail), *Panicum dichotomiflorum* (fall panicum), and *Cenchrus incertus* (field sandbur) (Table 2). The magnitude of changes in summer annual grass densities has varied among experiments, but the potential for rapid changes in density and emergence patterns of these weed species clearly exists.

Changes in the population dynamics of summer annual dicotyledon species with changing tillage practices have been less consistent. The most common response among annual dicotyledon species has been reduced density of large-seeded species such as *Ab. theophrasti*, *Xanthium strumarium* (common cocklebur), and *Cassia obtusifolia* (sicklepod) as tillage is reduced (Table 2). The response of small-seeded summer annual dicotyledon weed species to changes in tillage practices has been variable among weed species and experiments. This variable response is typified by *Chenopodium album*. In different studies at different locations, *Ch. album* densities increased, stayed the same, or decreased in response to reduced tillage. Species that germinate under cool soil conditions, such as *Ch. album*, may be prevalent in no-tillage fields prior to crop planting. Early emergence may allow such species to suppress later emerging weeds or result in a significant portion of the population being destroyed by tillage or herbicide prior to crop planting. *Amaranthus retroflexus* (redroot pigweed) is another small-seeded summer annual dicotyledon species that has responded inconsistently to changes in tillage systems. *Am. retroflexus* may be best-suited to systems with moderate tillage, because its seeds are well-adapted to shallow burial.

Table 2 Influence of tillage systems on densities of selected summer annual weed species in *Zea mays* and *Glycine max* production systems

Weed species	*Tillage response*[a]
Abutilon theophrasti	−
Amaranthus retroflexus	+, 0,−
Cassia obtusifolia	−
Cenchrus incertus	+
Chenopodium album	+, 0,−
Echinochloa crus-galli	+
Panicum dichotomiflorum	+
Setaria complex	+
Xanthium strumarium	−

[a]+, increased density with reduced tillage; 0, no change with reduced tillage;−, decreased density with reduced tillage.

Weed Seed Bank and Seedling Biology

It is often difficult to determine whether changes in population dynamics are due to biological properties of the weed species or differences in control efficacy. To address this issue, research has been conducted to separate various aspects of tillage–weed species–weed-control efficacy interactions to elucidate the mechanisms of weed population shifts.

The primary source of arable-response, summer annual weed species is the weed seed bank in the soil, with the largest source of seed being that produced in previous years in the same field. Changes in the weed seed bank occur with time and weed-management programs. Tillage is the primary cause of vertical seed movement in arable soils. Since tillage systems affect weed management, weed seed production, and soil disturbance, changing tillage systems changes the distribution and density of weed seed in agricultural soils.

Moldboard-plow plots have fewer weed seeds in the upper 20 cm of soil than chisel-plow or no-tillage plots after 5 years. Moldboard plowing results in the most uniform distribution of seed over soil depths. In a no-tillage system, more than 60% of all weed seeds are found in the upper 1 cm of soil and few seeds are found at more than 10 cm. The concentration of weed seed in no-tillage decreases logarithmically with increasing depth. In the chisel-plow system, more than 30% of the weed seeds are in the upper 1 cm of soil and seed concentration decreases linearly with depth.

Differences in emergence depths in different field tillage systems reflect differences in weed seed distribution in the soil. Mean seedling emergence depths in no-tillage are the least, followed by chisel-plow and moldboard-plow systems in two soil types (Table 3). At least 40% of the *S. faberi* and *S. viridis* plants emerge from the upper 1 cm of soil in no-tillage compared with approximately 25% in chisel-plow and less than 15% in moldboard-plow plots. Changes in seed depth in the soil and the corresponding differences in emergence depth may contribute to shifts among weed species. In greenhouse research, *Ab. theophrasti* establishment from seed germinating on the soil surface is much less than for seed planted 6 cm deep, and seedlings from surface-placed seed are less vigorous. *S. faberi* seed germinating on the soil surface has an establishment percentage similar to seed planted 1–4 cm deep, but establishment is reduced by 50% when the seeds are planted 6 cm deep.

Surface topography influences the emergence location of *Ab. theophrasti* seedlings when a field has not been tilled the previous two growing seasons. No *Ab. theophrasti* seedlings emerge from bare soil, and the majority of the *Ab. theophrasti* seedlings emerge

Table 3 Effect of tillage systems on depth of emergence of *Setaria* species under field conditions

	Depth of emergence (cm below soil surface)	
Tillage system	*Range*	*Mean*
Moldboard plow	0–10	3.1
Chisel plow	0–9	2.4
No-tillage	0–10	1.4

Source: Buhler DD (1998) Tillage systems and weed population dynamics and management. In: Hatfield JL, Buhler DD, and Stewart BA (eds) *Integrated Weed and Soil Management*, pp. 223–246. Chelsea, MI: Ann Arbor Press.

from cracks in the soil. Another significant portion of the *Ab. theophrasti* emerge from under surface residue. The presence of surface residue increases soil moisture and reduces soil crusting compared with bare soil, thus providing adequate moisture conditions for germination for a longer period of time. These data suggest that the safe-site concept should be extended below the soil surface to the depth where soil surface conditions influence germination conditions and that soil cracks may be a factor in weed emergence patterns.

The effect of tillage practices on the population dynamics of summer annual weed species is complex and involves several factors. However, seed depth in the soil appears to be the most important factor. Weed species that have the ability to germinate and become established when the seeds are at or near the soil surface have the greatest potential to increase under conservation-tillage systems. These species tend to be small-seeded and are represented by small-seeded, annual dicotyledon and most annual grass species. Deep burial of seed of small-seeded species by moldboard plowing reduces germination and emergence. Conversely, seeds of large-seeded species remain near the soil surface in conservation-tillage systems, inhibiting establishment.

Winter Annual and Biennial Species

Weed species not previously observed in fields planted to summer annual crops have rapidly appeared following elimination of preplant tillage. Species most rapidly and commonly observed are winter annual and biennial species that are characterized by an inverse tillage response. These weed species are unable to complete their life cycles in association with summer annual crops if the soil is disturbed after harvest or prior to crop planting the following spring.

Changes in soil surface characteristics and disturbance patterns in no-tillage systems create an environment where winter annual and biennial species may become established and complete their life cycle. Winter annual species germinate in the late summer and autumn, survive the winter, and flower the following spring. Because of the growth habit and ability to adjust photosynthetic responses to temperature, many winter annual species can grow even during the winter. Established winter annual plants preempt space, suppress summer annual species, and can be very competitive with summer annual crops. Rapid growth early in the spring can also make winter annual species difficult to control with herbicides if allowed to grow until or after the time of crop planting.

When the soil is tilled in the spring, rosettes and seedlings of winter annual and biennial species recruited during the previous growing season are destroyed. However, new seedlings of winter annual and biennial species may emerge under the canopy from seeds in the seed bank or that move in from adjacent areas. If tillage is removed from the system, these seedlings survive and grow during the autumn and winter and are established at the start of the second year. Thus, these species can rapidly infest fields (often during the first year) when tillage is eliminated. The rapid appearance of winter annual species does not appear to result from differences in innate competitive ability relative to summer annual species, but rather from the timing of disturbance, which changes competitive hierarchies.

Conyza canadensis (horseweed) is one of the most common and troublesome winter annual weeds in no-tillage systems in the central and eastern USA and an excellent example of an inverse-tillage response winter annual species. *Co. canadensis* can infest no-tillage fields very rapidly because it is commonly found in field edges and roadsides, produces many wind-disseminated seeds, is tolerant to many commonly used herbicides, and can become established under a wide-range of soil and climatic conditions, making it well-adapted to the no-tillage soil environment. Other winter annual species such as *Bromus tectorum* (downy brome) and *Capsella bursa-pastoris* (shepherd's purse) have also been observed to infest no-tillage fields.

Perennial Species

Perennial weed populations increase as tillage is reduced, because the underground system is not disturbed and most herbicides used to control annual weeds are not fully effective on established perennial plants. A preponderance of the literature indicates that weed flora associated with reduced-tillage systems favor perennial monocotyledon and dicotyledon species. Perennial monocotyledons might be

the greatest threat to the adoption of reduced-tillage systems because of the importance of monocotyledon crops in many regions of the world.

In most studies, more diverse and dense populations of perennial weeds develop in conservation-tillage systems than where moldboard plowing is conducted. Increases in perennial weed populations vary among sites because tillage interacts with weed-management practices, environment, and initial perennial weed populations. Therefore, the general statement that perennial weeds increase as tillage decreases must be made with caution. Management practices such as selective application of glyphosate to tall-growing perennial weeds, the use of crop cultivars resistant to herbicides such as glyphosate, and interrow cultivation help reduce perennial-species densities regardless of tillage system. Conservation-tillage systems have the potential to increase densities and species diversity of perennial weeds, but management practices such as crop rotation, selective application of glyphosate or other herbicides, and interrow cultivation can prevent perennial weed species from increasing to levels that reduce crop yields or increase production costs.

Summary

Although general trends in weed population dynamics in conservation tillage have been observed, it is important to note that location, environment, the type of conservation-tillage system used, producer management skills, and weed-management inputs regulate responses in individual fields. In addition, individual species respond differently among sites or within a site over time. This indicates that weed-management systems and cultural practices interact with tillage to regulate weed populations.

Effective and economical weed management in conservation-tillage systems, especially no-tillage, is a challenge. Due to the changes in biological and management systems, it is important to take a systematic approach to planning and executing weed management in conservation-tillage systems. Assessing the strengths and weaknesses of the available control options relative to expected weed populations is essential when designing weed-management systems.

The intensive tillage traditionally conducted in row-crop production has had a profound effect on weeds and weed management. Effective, economical, and environmentally sound weed management in conservation-tillage systems over the long term will require integration of new information with established principles of weed management. Crop rotation, tillage rotation, judicious use of tillage, herbicide-resistant crop cultivars, and innovative application methods for herbicides will aid weed management in conservation-tillage systems. New technologies are also needed to deal with the altered agroecosystems created by conservation-tillage production systems. New control options such as biological agents and allelopathic cover crops can provide useful new management tools. Many weed species and weed-control tactics behave differently as tillage is reduced or eliminated. These changes must be taken into consideration to develop economically and environmentally sound weed-management systems. It is essential that crop producers and pest managers consider the idiosyncrasies of the biological interactions among weeds, crops, residues, and environment as they develop weed-management plans for conservation-tillage systems.

List of Technical Nomenclature

Abutilon theophrasti Medikus	velvetleaf
Amaranthus retroflexus L.	redroot pigweed
Bromus tectorum L.	downy brome
Capsella bursa-pastoris (L.) Medikus	shepherd's purse
Cassia obtusifolia L.	sicklepod
Cenchrus incertus M.A. Curtis	field sandbur
Chenopodium album L.	common lambsquarters
Conyza canadensis (L.) Cronq.	horseweed
Glycine max (L.) Merr.	soybean
Glyphosate	*N*-(phosphonomethyl)glycine
Panicum dichotomiflorum Michx.	fall panicum
Setaria faberi Herrm.	giant foxtail
Setaria glauca (L.) Beauv.	yellow foxtail
Setaria viridis (L.) Beauv.	green foxtail

Xanthium strumarium L.	common cocklebur
Zea mays L.	corn

See also: **Conservation Tillage**; **Cover Crops**; **Crop Rotations**; **Crop-Residue Management**; **Cultivation and Tillage**; **Pesticides**

Further Reading

Bazzaz FA (1990) Plant–plant interactions in successional environments. In: Grace JB and Tilman D (eds) *Perspectives on Plant Competition*, pp. 239–263. San Diego, CA: Academic Press.

Buhler DD (1998) Tillage systems and weed population dynamics and management. In: Hatfield JL, Buhler DD, and Stewart BA (eds) *Integrated Weed and Soil Management*, pp. 223–246. Chelsea, MI: Ann Arbor Press.

Buhler DD (ed.) (1999) *Expanding the Context of Weed Management*. New York: Food Production Press.

Buhler DD, Liebman M, and Obrycki JJ (2000) Theoretical and practical challenges to an IPM approach to weed management. *Weed Science* 48: 274–280.

Derkson DA, Lafond GP, Thomas AG, Loeppky HA, and Swanton CJ (1993) Impact of agronomic practices on weed communities: tillage systems. *Weed Science* 41: 409–417.

Forcella F, Buhler DD, and McGiffen (1994) Pest management and crop residues. In: Hatfield JL and Stewart BA (eds) *Crop Residue Management*, pp. 173–189. Boca Raton, FL: Lewis Publishers.

Froud-Williams RJ, Chancellor RJ, and Drennen DSH (1981) Potential changes in weed flora associated with reduced-cultivation systems in cereal production in temperate regions. *Weed Research* 21: 99–109.

Harper JL (1977) *Population Biology of Plants*. New York: Academic Press.

Liebman M, Mohler CL, and Staver CP (2001) *Ecological Management of Agricultural Weeds*. New York: Cambridge University Press.

Sims GK, Buhler DD, and Turco RF (1994) Residue management impact on the environment. In: Unger PW (ed.) *Managing Agricultural Residues*, pp. 77–98. Boca Raton, FL: Lewis Publishers.

Sprague MA and Triplett GB (1986) *No-Tillage and Surface Tillage Agriculture*. New York: John Wiley.

WETLANDS, NATURALLY OCCURRING

E K Hartig, New York, NY, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Wetlands are areas that are located where land and water meet, where the habitat is neither aquatic nor fully terrestrial, the soil being for the most part too wet and oxygen-depleted for any but specially adapted plants to survive. Wetlands occur along lakes, creeks, rivers, bays, and deltas; they are found on every continent except Antarctica, in the low-lying peatlands of Indonesia and Siberia as well as in high-altitude montane depressions. Worldwide estimates vary somewhat; e.g., estimated freshwater coverage of 5.3 million km^2, based on soil maps prepared by the Food and Agricultural Organization of the United Nations (FAO); and the commonly used estimate of 8.6 million km^2. Some examples of huge contiguous wetlands are the Pantanal in parts of five countries in South America, the Sudd Marshes of the Sudan along the White Nile River, and the permafrosted peatlands in the northern high latitudes of Canada, Alaska, and Siberia. Examples of very small but highly productive wetlands can be found in the Prairie Pothole region in North Dakota, Minnesota, and parts of Canada. These comprise a vast complex of tens of thousands of individual shallow wetlands that vary in size, some being as small as 0.1 ha.

Historically wetlands have been treated as wastelands to be drained and farmed, or filled and built upon. Traditional attitudes are exemplified by such expressions as 'bogged down,' 'swamped,' 'muddied,' 'mired,' 'murky,' 'morass,' 'malarial,' and 'quagmire.' Great drainage schemes have long dominated approaches to wetlands worldwide. An example of a government works program was the mosquito ditching conducted in the USA during the 1930s by the Work Projects Administration. Most US East Coast marshes bear the scars of these labor-intensive ditching projects. More than 50% of the wetlands in the USA have been drained since colonial times. In the UK more than 90% of present-day agricultural lands were formerly wetlands. Many wetlands remain under threat either from direct human interference or from indirect causes such as sea-level rise inundating coastal areas, or changes in hydrological regimes due to upstream alterations for dams, dikes, or levees. The swamps of southern Sudan are among the largest of the world's wetlands and home to one of the greatest

concentrations of large mammals (such as the buffalo, elephants, hippopotami, gazelles, waterbucks, and zebra). These wetlands are threatened by the planned diversion of the White Nile.

In recent decades, the perception of wetlands has changed, and there is increasing recognition of the many vital functions provided by wetlands. These include providing wildlife habitat, flood and erosion control, aesthetics and recreation, carbon sequestration, and high biological productivity.

Wetland Definitions

Wetland definitions used by scientists and by administrative agencies commonly refer to vegetation types, soils, and/or hydrology. Administrative agencies use variations of the three-parameter approach (vegetation, soils, hydrology). The definition developed by the US Fish and Wildlife Service (USF-WS) and used by the US Army Corps of Engineers (Corps) for delineating wetland boundaries under Corps jurisdiction states:

> Wetlands are areas that are inundated or saturated by surface or groundwater at a frequency and duration sufficient to support, and that under normal circumstances do support, a prevalence of vegetation typically adapted for life in saturated soil conditions. Wetlands generally include swamps, marshes, bogs, and similar areas. (Definition of Waters of the United States; 33CFR328.3(b); 1984)

Wetland Classification

Wetland classifications abound, but usually they are divided according to their relation to water bodies, vegetation types, and substrates. The widely used Cowardin classification takes into account wetlands that can be found in five environments. Of these, Marine, Riverine (along freshwater nontidal rivers and streams), and Lacustrine (along lakes) wetlands are mainly aquatic and have little or no soil development. Estuarine (tidal embayments and tidal portions of rivers) and Palustrine (inland freshwater forested and nonforested) wetlands contain hydric soils.

Estuarine Wetlands

This category includes intertidal salt and brackish low and high marsh zones, nonvegetated tidal flats, brackish waters of coastal rivers and embayments, and mangrove swamps:

1. Estuarine emergent wetlands: These are wetlands dominated by herbaceous grasses and are referred to as salt marsh, low marsh (intertidal marsh), high marsh, and brackish tidal marsh. They are found in protected embayments or behind barrier beaches where the water has slowed. Many, such as the *Spartina alterniflora* intertidal marshes of the northeastern USA, formed within the last 4000–7000 years as the postglacial rise in sea level slowed. Typically there is strong zonation of species in the salt marsh, depending on the frequency and duration of tidal ebb and flow (Figure 1);
2. Estuarine scrub–shrub wetlands: These wetlands contain woody vegetation such as marsh elder or high-tide bush. With exposure to salt spray and infrequent flooding they are adapted to a high-salt environment. This category also includes mangrove swamps.

Palustrine Wetlands

This category includes the majority of freshwater wetlands. It encompasses marshes, bogs, swamps (forested), and ponds. Shallow-water bodies such as

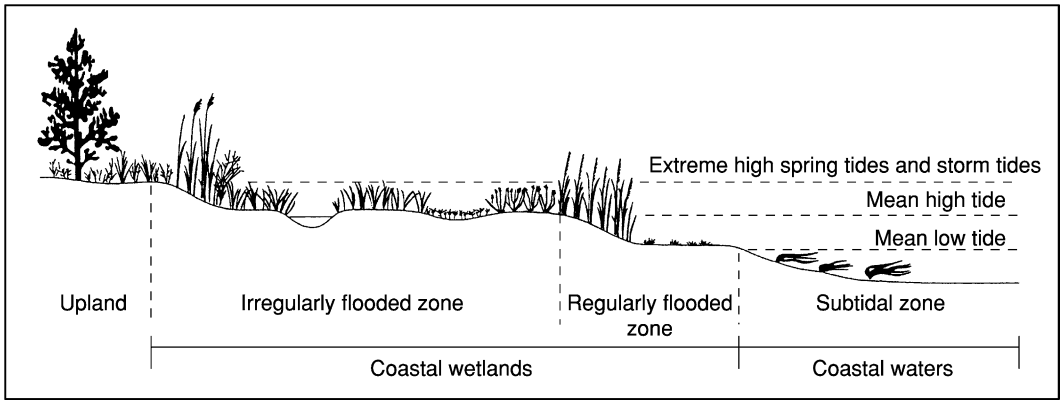


Figure 1 Hydrology of a tidal wetland indicating different zones of flooding. The regularly flooded zone is subject to at least once-daily tidal inundation, while the irregularly flooded zone is inundated less frequently. (Reproduced from Tiner RW and Burke DG (1995) *Wetlands of Maryland*. USFWS, Hadley/Maryland Department of Natural Resources. Annapolis, MD: US Fish and Wildlife Service.)

vernal ponds and playa lakes (less than 2 m deep) are also in this category:

1. Palustrine emergent wetlands: These wetlands are dominated by herbaceous vegetation with many species of grasses, rushes, and sedges. They are referred to as freshwater marshes, wet meadows, and fens. The marshes may be irregularly to permanently flooded. In the US Prairie Pothole region, glacial depressions flood, especially during the spring season following snowmelt (and offer a protein-rich diet to migrating waterfowl). Palustrine emergent wetlands occur in a variety of habitats, including peatlands and freshwater marshes;
2. Palustrine scrub–shrub wetlands: These are characterized by woody vegetation less than 6 m tall. North American bogs with vegetation such as bog laurel, Labrador tea, cranberry, and leatherleaf, and with stunted trees such as black spruce, larch, and balsam fir are examples of scrub–shrub wetlands;
3. Palustrine forested wetlands: These are wetlands with trees greater than 6 m in height. Tree species found in northeastern US wetlands include red maple, pin oak, sweet gum, black spruce, and larch.

Hydric Soils

The definition of a hydric soil is as follows:

> A hydric soil is a soil that is saturated, flooded, or ponded long enough during the growing season to develop anaerobic conditions in the upper part of the soil.

In organic wetland soils, a thick organic layer develops over time where microbial decomposition is severely slowed or remains incomplete under waterlogged conditions. Organic wetland soils develop in depressions or in coastal areas where anaerobic conditions prevail (such as in bogs or intertidal marshes). Included are mucks (saprists), peats (fibrists), and hemists – mucky peat or peaty muck. The accumulated peat or muck layers are 0.6 to more than 9 m in depth. While organic matter is present in all soils, it constitutes greater than 20% (based on dry weight) in organic soils. These soils also have a lower bulk density and higher water-holding capacity than mineral soils.

Mineral wetland soils form either in moving or periodically standing water and can be found in low-lying depressions or slopes (Figure 2). They may

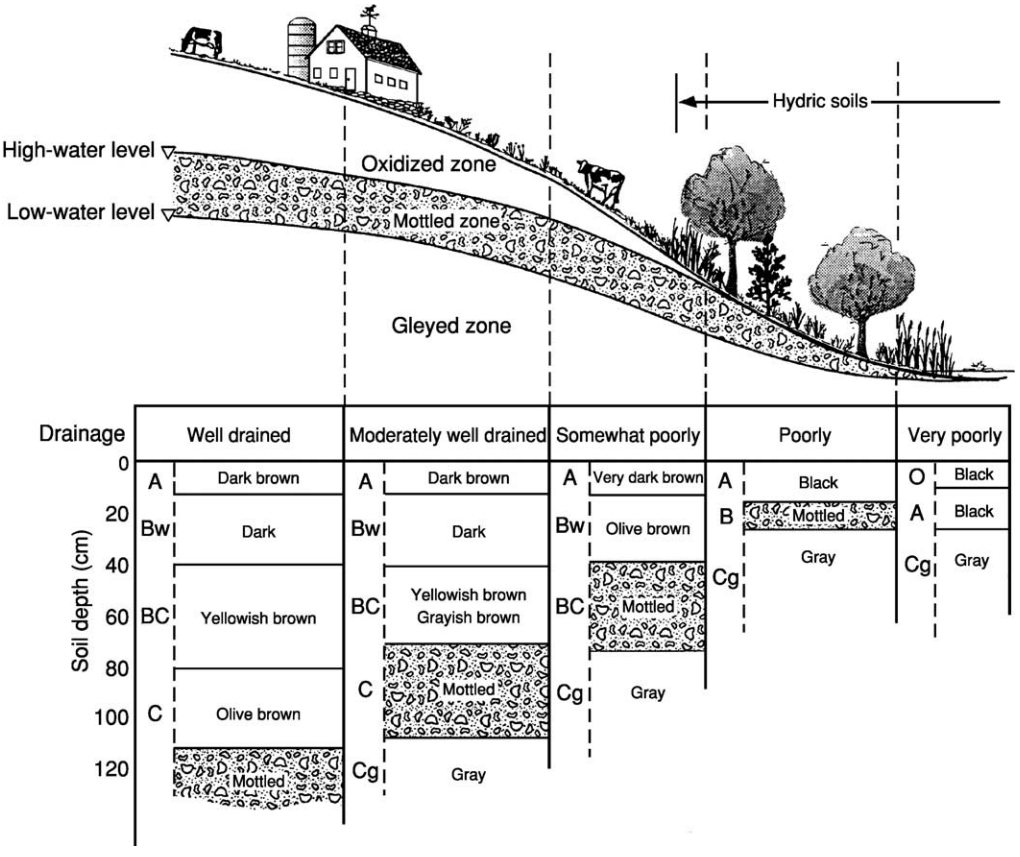


Figure 2 Soil properties are shown along a wetland-to-upland landscape gradient. Note that the seasonal high-water table, indicated by the mottled zone, reaches the surface only at the lower elevations. Reproduced from Tiner RW and Burke DG (1995) *Wetlands of Maryland*. USFWS, Hadley/Maryland Department of Natural Resources. Annapolis, MD: US Fish and Wildlife Service.

receive surface water runoff or groundwater seepage, and there may be the presence of an only slowly permeable soil layer (e.g., a fragipan, clay, or hardpan layer, or confining bedrock) that slows the movement of water. Criteria used to determine whether a soil is hydric are listed in Table 1. Mineral wetland soils have a dry weight consisting of less than 20–35% organic material.

Observable indications (redoximorphic features) of mineral wetland soil include its dominant 'matrix' color chroma that can vary from very dark to washed-out blue-gray or green-gray colors (with often brighter colors found as minor components, as 'mottles'). Mottles are yellow-orange to brown accumulations that form where oxygen-rich zones are encountered, and iron and manganese oxidize and precipitate from solution. These can coat soil peds, walls of voids, or the root channels (rhizospheres), where they can be observed readily. The extent of gleying and mottle abundance, size, and color depends on the prevailing hydrology (e.g., fluctuating water table) and the extent of microbial activity: if it is too cold, then there is little activity and wetland indicators such as gleying and mottling will not be present.

One way to determine whether a soil is hydric is by using a standard Munsell color chart to identify its color strength. Low chromas, generally a 2 or less, indicate hydric conditions. The dark matrix color is the result of prolonged saturation that converts iron from its rust-colored, oxidized (ferric) form to its reduced (ferrous) form (see Iron and Manganese, below). Exceptions occur where hydric color indicators are masked, such as with soils that remain redder, in spite of hydric conditions, having formed from red glacial till material. In these cases other indicators such as iron or manganese concretions and the oxidized rhizospheres can be observed.

Redoximorphic features such as mottling, gleying, and oxidized rhizospheres that develop in hydric soils generally remain visible even after wetlands have been disturbed and thus remain useful in determining former wetland extent.

Table 1 Criteria for hydric soils

Description
1 All Histels except Folistels and all Histosols except Folists
2 Soils in aquic suborders, great groups, or subgroups, Albolls suborder, Historthels great group, Histoturbels great group, Pachic subgroups, or Cumulic subgroups that are: Somewhat poorly drained with a water table equal to 0.0 m from the surface during the growing season or Poorly drained or very poorly drained and have either: Water table equal to 0.0 m during the growing season if textures are coarse sand, sand, or fine sand in all layers within 50 cm, or, for other soils, Water table at less than or equal to 15 cm from the surface during the growing season if permeability is equal to or greater than 15 cm h^{-1} in all layers within 50 cm, or Water table at less than or equal to 30 cm from the surface during the growing season if permeability is less than 15 cm h^{-1} in any layer within 50 cm, or
3 Soils that are frequently ponded for long duration or very long duration during the growing season, or
4 Soils that are frequently flooded for long duration or very long duration during the growing season

Source: USDA/NRCS Website http://soils.usda.gov

Soil Chemistry

At the soil surface, a thin layer of oxidation is crucial in maintaining wetland function. Chemical transformations and nutrient cycling depend on the oxidized ions found in this layer. These oxidized ions are Fe^{3+}, Mn^{4+}, NO_3^-, SO_4^{2-}. Below the oxidized layer, and below the water table where anaerobic conditions dominate, the reduced forms of compounds are found such as Fe^{2+}, Mn^{2+}, NH_4^+, S^{-2}, and CH_4. Table 2 lists the oxidized and reduced forms of these elements.

Nitrogen

As in many ecologic communities, nitrogen is one of the most limiting of nutrients in many wetlands. Much of the nitrogen in soil is derived from dead organic material such as proteins, amino acids, and nucleic acids. These nitrogen-rich compounds decompose into simple compounds by soil-dwelling saprobic bacteria and fungi. These microbes then release the excess nitrogen in the form of ammonium ions (NH_4^+). Figure 3a shows the nitrogen cycle: after ammonium diffuses to the aerobic soil layer, nitrification occurs, converting it to nitrate. The nitrate may

Table 2 Oxidized and reduced forms of several elements in order of their redox potential

Element	*Oxidized form*	*Reduced form*
Nitrogen	NO_3^- Nitrate	N_2O nitrous oxide, N_2 Nitrogen gas, NH_4 Ammonium
Manganese	Mn^{4+} Manganic	Mn^{2+} Manganous
Iron	Fe^{3+} Ferric	Fe^{2+} Ferrous
Sulfur	SO_4^{2-} Sulfate	S^{2+} Sulfide
Carbon	CO_2 Carbon dioxide	CH_4 Methane

Adapted from Mitsch WJ and Gosselink JG (1993) *Wetlands*, 2nd edn. New York: John Wiley, with permission.

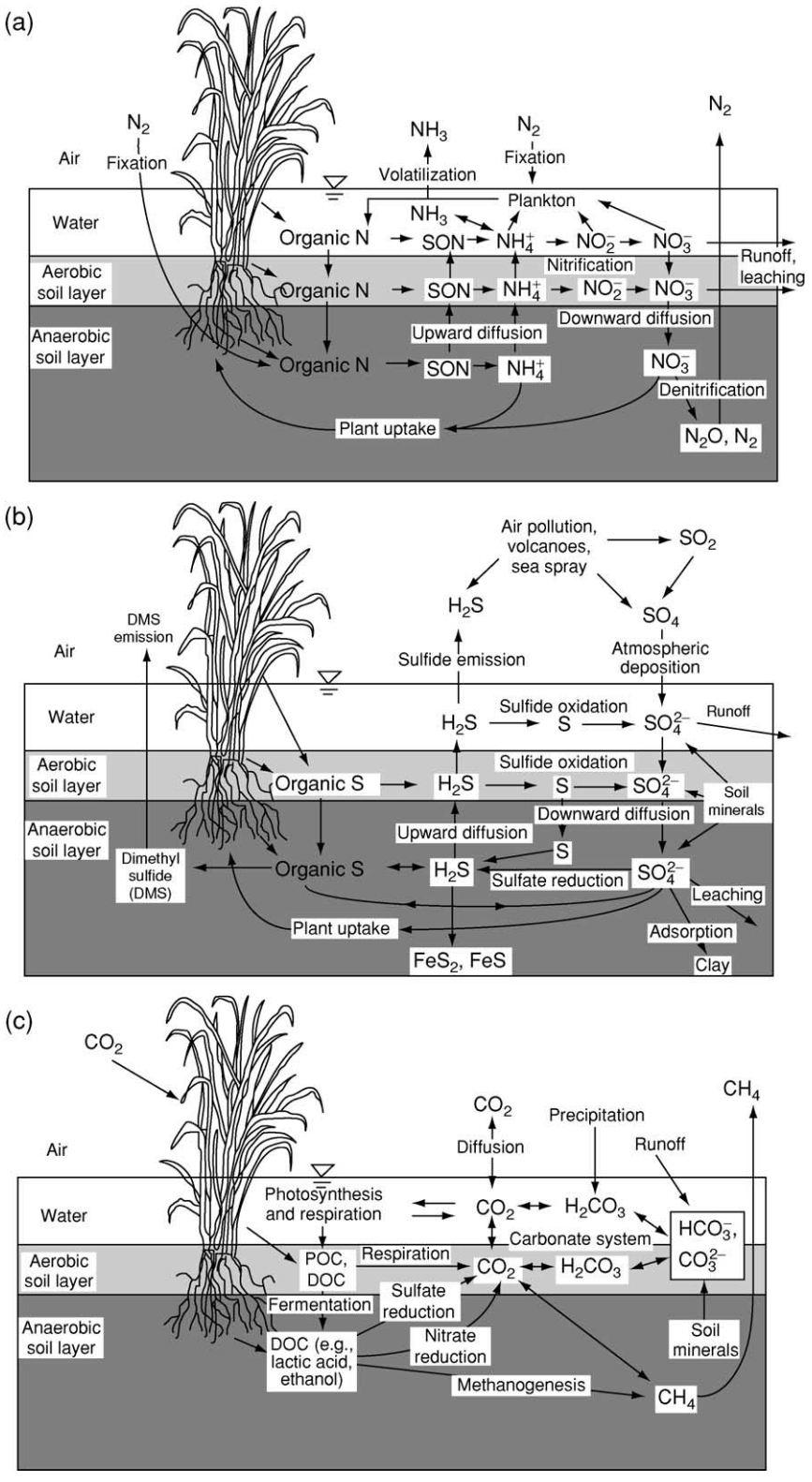


Figure 3 Chemical transformations in wetlands: (a) nitrogen transformations in wetlands. SON, soluble organic nitrogen; (b) sulfur transformations in wetlands; (c) carbon transformations in wetlands. POC, particulate organic carbon; DOC, dissolved organic carbon. Reproduced with permission from Mitsch WJ and Gosselink JG (1993) *Wetlands*, 2nd edn. New York: John Wiley.

then diffuse back into the anaerobic layer where denitrification transforms it into nitrogen gas. Ammonium nitrogen is oxidized to nitrite ions (NO_2^-) through nitrification:

$$2NH_4^+ + 3O_2 \rightarrow 2NO_2^- + 2H_2O + 4H^+ + \text{energy}$$

Nitrite is toxic to higher plants but it does not have a chance to accumulate in the soil owing to the presence of *Nitrobacter*, a genus of bacteria that oxidizes the nitrite to form nitrate ions (NO_3^-), again with a release of energy:

$$2NO_2^- + O_2 \rightarrow 2NO_3^- + \text{energy}$$

Nitrification in wetland soils also takes place within the oxidized rhizospheres of plant roots where sufficient oxygen may facilitate the conversion of NH_4^+ to NO_3^-.

Nitrate can also undergo reduction reactions. Nitrate being more mobile, if not rapidly assimilated by plants or microbes, or lost through groundwater flow, will probably be reduced to ammonium, nitrous oxide (N_2O), or molecular nitrogen (N_2) under anaerobic conditions. Denitrification typically occurs under waterlogged conditions in swamps and marshes. The availability of fresh organic matter together with denitrifying bacteria promotes dentrification:

$$C_6H_{12}O_6 + 4NO_3^- \rightarrow 6CO_2 + 6H_2O + 2N_2$$

In the highly acidic peat soils of northern bogs, nitrifying bacteria are lacking due to the low pH. In such environments carnivorous plant species have evolved that are able to use animal proteins directly as their nitrogen source. The sundew (*Drosera intermedia*) and the Venus flytrap (*Dionaea muscipula*) both attract insects and obtain minerals, including fixed nitrogen from animal prey via sticky hairs that contain digestive enzymes.

Iron and Manganese

Both iron and manganese occur in two oxidation states. Under the conditions that prevail in wetlands, they exist primarily in reduced form. Iron is transformed from ferric (Fe^{3+}) iron compounds to ferrous (Fe^{2+}) iron:

$$Fe_2O_3 + 4e^- + 6H^+ \rightarrow 2Fe^{2+} + 3H_2O$$

Manganese is transformed from manganic to manganous compounds:

$$MnO_2 + 2e^- + 4H^+ \rightarrow Mn^{2+} + 2H_2O$$

Ferrous iron causes gleying of mineral soils, the gray colors observable in highly reduced soils. In the presence of oxygen, when, for example, the water table drops, the reaction can be reversed and insoluble ferric oxides precipitate out. These reactions can be mediated through bacterial activity, forming 'bog iron' ore. Bog iron deposits found in anaerobic groundwater in northern peatlands have been used in iron and steel industries.

Sulfur

Sulfur has several oxidation states and is transformed through microbial activity. Sulfur is usually readily available and so not limiting to organisms. Elevated concentrations of hydrogen sulfide in saltwater wetlands may be toxic to plants and animals. When wetland sediments are disturbed, the smell of rotten eggs is emitted, a smell that is familiar to those working in wetlands. Sulfur-reducing bacteria in anaerobic soils reduce sulfates to sulfides:

$$4H_2 + SO_4^{2-} \rightarrow H_2S + 2H_2O + 2OH^-$$

As shown in Figure 3b, sulfur can be released to the atmosphere as methyl and dimethyl sulfide, organic sulfur compounds, and hydrogen sulfide. Among wetlands, salt marshes have the highest rate of emission of hydrogen sulfide per unit area. The dark color of many wetland peats comes from ferrous sulfide. Sulfur found in peat bogs and coal deposits usually occurs as the mineral pyrite (FeS_2).

A number of sulfur reactions occur in the upper aerobic portion of the wetland soil. *Thiobacillus* obtains energy from the oxidation of hydrogen sulfide to sulfur. Other species further oxidize sulfur to sulfate:

$$2H_2S + O_2 \rightarrow 2S + 2H_2O + \text{energy}$$
$$2S + 3O_2 + 2H_2O \rightarrow 2H_2SO_4 + \text{energy}$$

Purple sulfur bacteria found in salt marshes and mudflats produce organic matter in the presence of light:

$$CO_2 + H_2S + \text{light} \rightarrow CH_2O + S$$

Hydrogen sulfide is used as an electron donor instead of water. This reaction requires anaerobic conditions, ample hydrogen sulfide, and sufficient sunlight near the surface.

Carbon

Major processes for carbon transformations are given in Figure 3c for both aerobic and anaerobic conditions. Through fermentation, sugars are transformed to lactic acid, ethanol, and CO_2. Methane production (methanogenesis) is the result of extremely reduced conditions whereby bacteria (methanogens) reduce CO_2 to CH_4, as in:

$$CO_2 + 8e^- + 8H^+ \rightarrow CH_4 + 2H_2O$$

The resulting methane is often called 'swamp gas' or 'marsh gas.' It is known for being highly combustible especially in peatland environments. Methane production is higher in freshwater than in marine wetlands but has a broad productivity range in each. It is season-dependent, taking place only when warm temperatures prevail. Rice paddies are a major source of methane to the atmosphere. Peatlands are usually a net source of methane; however, under drought conditions, when oxidation occurs below the surface layers, the peat can become a source of CO_2 emissions.

Table 3 US (48 states) wetlands and former wetlands (thousands of hectares) where land use has been determined (nonwetland hydric soils indicate former wetland areas)

1992 land use	*Wetland soils (ha)*	*Nonwetland hydric soils (ha)*	*Total (ha)*
Cropland	4270	22 430	26 699
Pastureland	3232	2611	5843
Rangeland	3142	2426	5568
Forest land	24 732	3829	28 561
Misc.	9753	2501	12 254
Total	45 129	33 797	78 925

Adapted from Economic Research Service/USDA compilation of 1992 National Resources Inventory, Wetlands and Agriculture: Private Interests and Public Benefits/AER-765. http://www.ers.usda.gov with permission.

Policies, Regulations, and Protection

There are several international treaties that protect wetlands. The first was the Ramsar Convention. The role of wetlands to support migratory waterfowl that cross national boundaries was recognized when the Convention on Wetlands (originally entitled the Convention on Wetlands of International Importance Especially as Waterfowl Habitat) was signed in Ramsar, Iran, in 1971. It is an "intergovernmental treaty which provides the framework for national action and international cooperation for the conservation and wise use of wetlands and their resources." There are 136 signatory nations, and, as of May, 2003, 1283 wetlands sites are listed, comprising 108.7 million hectares. Often the wetlands encompass tracts with multiple uses, including water supply, fisheries, agricultural use, and more. The number of sites and their sizes vary from country to country. The UK has more than 150 sites, whereas the USA has 17 sites. Australia lists sites from 1 ha (Christmas Island Territory) to almost 2 Mha in size in the southern Coongie Lakes. Many countries use the Ramsar Convention as their main instrument for wetland protection.

The USA relies mainly on separate federal, state, and local laws and regulations to limit development in wetlands. In the 1970s the Federal Water Pollution Act was followed by the Clean Water Act of 1977, Section 404, that considered wetlands as waters of the USA covered in part under the Rivers and Harbors Act of 1899. Hence jurisdiction of private property could be given to public agencies as waters are considered a public good. Under Section 404 a permit is needed to conduct dredge-and-fill activities in jurisdictional wetlands even when on private property. The 1985 Food Security Act's 'Swampbuster' provisions denied government farm subsidies to farmers who drained wetlands after December 23, 1985.

In the USA, wetland inventories are conducted every 10 years and their status and trends are compiled. Wetlands comprise approximately 91 million hectares in these 48 contiguous states. Of these, land-use categories of 79 million hectares can be documented through knowledge of soil types (Table 3). Table 3 indicates that about 34 million hectares are no longer wetlands but retain hydric soil characteristics.

Where once the US government actively encouraged drainage, with cost-sharing and expertise, the paradigm has shifted. One of the more successful federal programs, the Wetland Reserve Program (WRP), promotes wetland restoration and offers cost-sharing and expertise to break drainage tiles and return areas to wetlands. Every 5 years Congress sets policies, guidelines, and funding for such agriculture programs. The WRP offers financial incentives and expertise from the USFWS and Natural Resources Conservation Service (NRCS) to take marginal land out of production.

Conserving wetlands is increasingly recognized as beneficial for flood control in a range of rural to urban areas, and as essential habitats for many fish, bird, and other species. In spite of conflicts that occur with pressure from expanding populations, local, state, and national regulations, as well as international agreements, have slowed the drainage of wetland areas.

See also: **Anaerobic Soils**; **Carbon Emissions and Sequestration**; **Hydric Soils**; **Land-Use Classification**; **Paddy Soils**; **Sulfur in Soils:** Biological Transformations

Further Reading

Cowardin LM, Carter V, Golet FC, and LaRoe ET (1979) *Classification of Wetlands and Deepwater Habitats of the United States*. Fish and Wildlife Service, FWS/OBS-79/31. Washington, DC: US Government Printing Office.

Crum HA (1988) *A Focus on Peatlands and Peat Mosses*. Ann Arbor, MI: University of Michigan Press.

Dahl TE (1990) *Wetlands Losses in the United States, 1780s to 1980s*. US Department of the Interior, Fish and Wildlife Service. Washington, DC: US Government Printing Office.

Diers R and Anderson JL (1984) Development of soil mottling. *Soil Survey Horizons* winter: 9–12.

Gambrell RP and Patrick WH Jr (1978) Chemical and microbiological properties of anaerobic soils and sediments. In: Hook KK and Crawford M (eds) *Plant Life in Anaerobic Environments*, pp. 375–423. Ann Arbor, MI: Ann Arbor Scientific Publishing.

Heimlich RE, Wiebe KD, Claassen R, Gadsby D, and House RM (1998) *Wetlands and Agriculture: Private Interests and Public Benefits*. Agriculture Economics report no. 765. Available online at: www.ers.usda.gov/publications/aer765/ (see chapter entitled Wetland Status and Trends, Settlement to 1992, 117 Kb).

Hillel D (1991) Endangered wetlands. In: *Out of the Earth: Civilization and the Life of the Soil*, pp. 215–224. Berkeley, CA: University of California Press.

Hillel D (1994) *Rivers of Eden*. New York: Oxford University Press.

Hurt GW, Whited PM, and Pringle RF (eds) (2003) *Field Indicators of Hydric Soils in the United States,* Version 5.01. USDA–NRCS National Technical Committee for Hydric Soils. Fort Worth, TX: US Government Printing Office.

Matthews E and Fung I (1987) Methane emission from natural wetlands: global distribution, area, and environmental characteristics of sources. *Global Biogeochemical Cycles* 1: 61–86.

Mitsch WJ and Wu X (1995) Wetlands and Global Change. In: Lal R, Kimble J, Levine E, and Stewart BA (eds) *Soil Management and Greenhouse Effect*, pp. 205–230. Boca Raton, FL: CRC Press.

Mitsch WJ and Gosselink JG (1993) *Wetlands*, 2nd edn. New York: John Wiley.

Redfield AC (1972) Development of a New England salt marsh. *Ecological Monographs* 42: 201–237.

Richardson JL and Vepraskas MJ (2000) *Wetland Soils: Genesis, Hydrology, Landscapes, and Classification*. Boca Raton, FL: CRC Press.

Thomas E and Varekamp JC (1991) Paleo-environmental analyses of marsh sequences (Clinton, Connecticut): evidence for punctuated rise in relative sea-level during the latest Holocene. *Journal of Coastal Research* special issue 11: 125–158.

Tiner RW (1984) *Wetlands of the United States: Status and Trends*. National Wetland Inventory, Fish and Wildlife Service, US Department of Interior. Washington, DC: US Government Printing Office.

Tiner RW and Burke DG (1995) *Wetlands of Maryland*. US Fish and Wildlife Service, Hadley/Maryland Department of Natural Resources. Annapolis, MD: US Fish and Wildlife Service.

US Army Corps of Engineers (1987) *Corps of Engineers Wetlands Delineation Manual*. Technical Report Y-87-1. Washington, DC: US Government Printing Office.

WIDTSOE, JOHN A. AND GARDNER, WILLARD

G S Campbell, Decagon Devices, Inc., Pullman, WA, USA
W H Gardner, Washington State University, Pullman, WA, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Soil physics can be defined as the study of the state and movement of energy and mass in the continuum that includes the soil, plants, and surface boundary layer of the atmosphere. Since soil physics is both a theoretical and an experimental science, it includes equations from physics, and knowledge and measurement tools built up through years of laboratory and field experience. Understanding the state and movement of water in soil is an important part of the overall body of soil physics knowledge.

The relationship between water flux and driving force:

$$J_w = -k\nabla\psi$$

called Darcy's law, was already well established for saturated flow of water in porous materials before the beginning of the twentieth century. Darcy recognized that the driving force for flow was a gradient in what we now call the pressure and gravitational potentials. These ideas were extended to unsaturated flow in the early part of the twentieth century by Edgar Buckingham. He recognized that water in unsaturated soil is under negative pressure, i.e., the matric or capillary potential. Buckingham also recognized that gradients in matric potential, as well as in the other potentials, result in water flow. He generalized Darcy's equation by allowing k to vary as a function of the matric potential. While the second laws of both Fick and Fourier (combining the flux equation with the continuity equation to obtain the second-order partial-differential equation) were well known at the beginning of the twentieth century, they had not been extended to water flow. By the end of the twentieth century, the second-order equation for water was well known, with both analytical and numerical solutions available for a wide range of theoretical and practical problems.

Buckingham had no direct way to measure matric potential. He attempted to establish equilibrium in vertical soil columns and then related the water content of the soil in the columns to the height above a free water surface at the bottom of the column. Water content was measured by weighing and oven-drying the soil. By the end of the twentieth century, one could measure soil-water content using neutron scattering, neutron and gamma attenuation, time-domain reflectometry, capacitance, and microwave-drying. Matric potential could be measured using tensiometers, gypsum blocks, heat-dissipation matric probes, thermocouple psychrometers, filter paper, and freezing point. The relationship between water content and water potential is now routinely determined using a pressure plate.

Many of the important soil physics developments of the twentieth century, both in theory and in measurement methods, have direct or indirect ties to the laboratory of Willard Gardner, a professor of physics at Utah State Agricultural College from 1918 to 1948. While Willard Gardner made important contributions to soil physics in his own right, his influence has been amplified many-fold through the disciples he converted and trained. It is an interesting story of science on the frontier, but the story starts much earlier, in the mid nineteenth century.

On July 21, 1847, Orson Pratt and Erastus Snow entered the valley of the Great Salt Lake, the first of the vanguard group of Mormon (Church of Jesus Christ of Latter-Day Saints) pioneers fleeing west to escape religious persecution. Their only hope was to plant crops, even though it was late in the season. However, it was impossible to plow the hard-baked valley soil, so they channeled water from the mountain streams on to the thirsty ground. This was the beginning of modern irrigation in the US Mountain West. Since irrigation was necessary to produce almost all crops in the arid mountain valleys, the practice became widespread. Pratt was a physicist and mathematician who lectured at the University of Deseret (later the University of Utah), but he apparently took little interest in the physics of water flow, preferring astronomy and mathematics instead.

The practical aspects of designing irrigation systems and managing water flow in soil were a part of ordinary life in the frontier towns of the Great Basin, but the scientific study of irrigation in Utah did not begin until 50 years later. It was started by John A. Widtsoe soon after he became Director of the Utah Agricultural Experiment Station in 1900 (Figure 1).

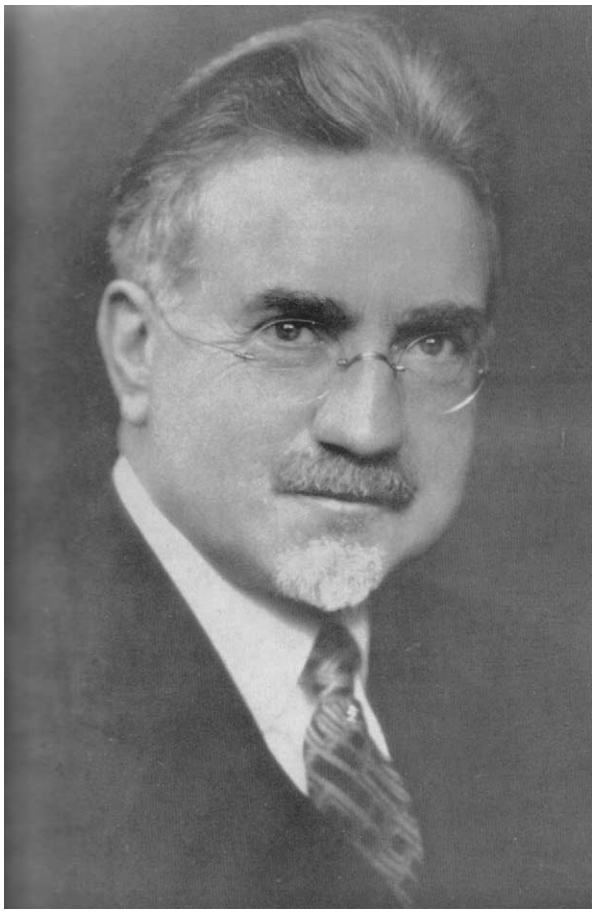


Figure 1 John A. Widtsoe (1872–1953).

Widtsoe was born in Daloe, Norway, in 1872, the oldest son of John Andersen Widtsoe, a schoolmaster, and Anna Karine Gaarden. When John was 6 years old, his father died, leaving his mother, himself, and his 2-month old brother, Osborne. When Widtsoe was 11 years old, he moved with his mother and brother to Utah, and settled in the town of Logan. He worked in Logan for several years to help support the family, but finally graduated from Brigham Young College (a junior college) in 1891. He was then selected as one of six talented Utah students to attend Harvard University, in Cambridge, Massachusetts, where he majored in chemistry with minors in physics and metallurgy, graduating in 1894. On his return to Utah, he joined the faculty of the new Utah State Agricultural College in Logan, where he began studying ways to improve agricultural production. In 1898, he and his new wife, Leah Eudora Dunford, went to Germany for his graduate study. In less than two years, he earned his doctoral degree in chemistry at the University of Goettingen and completed additional studies at Berlin and Zurich.

While still in Europe, Widtsoe was offered the position of Director of the Experiment Station at the Utah State Agricultural College. He served as Director from 1900 to 1905. The following excerpt from

Widtsoe's autobiography indicates the kind of research he and his colleagues undertook:

> By the planting season of 1901, a part of the College farm was supplied with troughs or flumes, by which measured quantities of water could be applied at will to a series of plots planted to various crops. Later, another and more suitable farm was purchased and equipped more extensively for the controlled application of irrigation water to the experimental plots. A vegetation house was also built, housing a large number of tanks, which could be wheeled in and out of the house. These tanks, protected from natural rainfall, filled with various soils, received carefully measured quantities of water. The farm and the vegetation house were undoubtedly the first experimental plant of its kind in the world dedicated to the scientific study of the use of water in irrigation. The results proved that they were very effective tools.
>
> The investigations that followed answered questions concerning the movement of water in irrigated soils; the control of loss of soil moisture by seepage and evaporation; the relation between the water lost by evaporation from soils, and by transpiration from plants; the relation between soil fertility and plant transpiration; the actual quantities of water required in crop production; the yields of crops and their chemical composition under varying quantities and times of application of irrigation water; and many other irrigation problems.

They also studied dryland farming, and the results of both irrigation and dryland research were shared throughout Utah by Dr. Widtsoe and his co-workers through extension activities. Results of the research were published in experimentation bulletins, and in two books: *Dry Farming* (published in 1911) and *Principles of Irrigation Practice* (published in 1914). Reports on the findings of plant–water relations and evapotranspiration research a century later differ little from the reports Widtsoe gave of these few years at the Experiment Station.

Politics were an important part of university life, then as now, and this was certainly true for Dr. Widtsoe. In spite of his amazing accomplishments, he was fired in 1905. He worked briefly at Brigham Young University in Provo, UT, and was then hired back as president of Utah State Agricultural College. This time the politics were in his favor. He served as president from 1907 to 1916, when he was made president of the University of Utah in Salt Lake City, UT. In 1921 he was called to serve as an apostle in the Church of Jesus Christ of Latter-Day Saints (LDS), a position he held until his death in 1953.

Widtsoe's early research on plant–water relations and irrigation practice was, without question, pioneering work, well ahead of its time. It is difficult to trace the influence of that work directly to modern plant–water relations theory. It is not difficult, however, to trace his influence on people who have had a great impact on modern soil physics. For example, in 1920, during the time when he was president of the University of Utah, Dr. Widtsoe joined Willard Gardner in writing a general treatise with the title, *The Movement of Soil Moisture*. A footnote to this article reads:

> The formal development of the mathematical material of this paper was made by the senior author as an outgrowth of the earlier work of the junior author, who has assisted in clarifying this material for publication.

Such participation with, and encouragement of, Utah soil scientists has had enormous impact on the careers of many soil physicists and on soil physics in general.

By the end of the nineteenth century, physics of soil processes had received considerable attention in many places in the USA, in Europe, and in other parts of the world. The names 'soil physics' and 'soil physicist' began to be used to identify subject matter and scientists on the faculties of colleges and universities and in governmental programs. In Utah, this occurred when Willard Gardner become a member of the faculty of the physics department and physicist for the Utah Agricultural Experiment Station in 1918 (**Figure 2**).

Figure 2 Willard Gardner (1883–1964).

Willard Gardner was born in 1883 in the little logging community of Pine Valley, Utah. His parents, Robert Gardner and Cynthia Lovina Berry, encouraged their children to educate themselves. Because they did so, several generations of Gardners have made important scientific contributions throughout the twentieth century. In 1903, Willard completed high school and traveled to Salt Lake City, where he attended business college, learning to type and take shorthand dictation. He became highly proficient at both, typing at faster than 90 words per minute and taking dictation at more than 150 words a minute. He then went to Logan, Utah, where he was employed by the Cache Valley Banking Company as stenographer. He did this work so well that he was often loaned to the local Justice Court to take dictation. In 1910, when the Agricultural Experiment Station at the college offered him a position as a clerk in their campus office, he accepted the position against the advice of his banker employers. This made it possible for him to pursue college classes on the side, and he was able to complete a bachelor's degree in physics. In 1912, he obtained a departmental assistantship in Physics at the University of California–Berkeley, graduating in 1916. His was the eighth PhD degree awarded by the University of California–Berkeley Physics Department.

After graduating, Gardner became principal of the Murdock Academy in Beaver, Utah, where his soil physicist son, Walter, was born in 1917. The following year he taught physics at the Brigham Young Academy in Logan, and in 1918 was invited to join the faculty in physics at Utah State Agricultural College (USAC). He held that position until his retirement in 1949. He became Professor Emeritus and continued to be active in university work until 1954. He died in 1964.

Dr. Gardner retained his stenographic skills and used them to good advantage throughout his career. For him, note-taking was easy, and he always typed his own letters rapidly, using a typewriter with an extra dozen Greek letter keys needed in writing mathematical equations. Not only did this save the college the cost of a secretary, but it was also more economical than to train secretaries to deal with mathematics.

Gardner's PhD thesis was titled *The Photo-Electric Current as a Function of the Angle of Emission and the Thickness of the Emitting Film*. He evidently found the physics of water flow and retention to be an important area to pursue in an agricultural experiment station. O.W. Israelsen, a friend and colleague from graduate school, and professor of irrigation engineering at USAC, encouraged Gardner to undertake irrigation research. Several other faculty members working in soil science had reason to seek his help in dealing with problems involving physics or mathematics.

Widtsoe's early work, Buckingham's 1907 monograph on water in soil, and Dr. Gardner's physics training formed the foundation for the soil physics work carried out by Gardner and his colleagues at USAC. A few of the titles of early papers show that the research direction was set from the beginning. Some are "The movement of moisture in soil by capillarity," "Capillary moisture-holding capacity," and "A capillary transmission constant and methods of determining it experimentally." These papers and others show consistent skill in using physics and mathematics to analyze soil-water flow problems. Research topics included saturated and unsaturated water flow, erosion, drainage, and measurement. The tensiometer, which is still a mainstay of soil physics research, was a product of Gardner's early research efforts, as was the tension or disk infiltrometer. It is no wonder that Dr. Gardner is referred to as the father of modern soil physics.

USAC did not begin offering the PhD degree until after World War II, so Willard Gardner had no official PhD students. In 1950, his son Walter received the first PhD given at USAC, with Wynn Thorne as major professor and his father as research advisor. However, numerous students and colleagues became interested in soil physics as a consequence of studying or working with Willard Gardner. The USAC library and Agricultural Experiment Station records are incomplete, but, from publications and personal files, it is possible to identify most of the students who studied with Gardner. Seventeen MS theses on topics closely related to soil physics are listed in the USAC Graduate School, and 14 published scientific papers with acknowledgements to Willard Gardner are in the library files.

There is little doubt that a number of students who have pursued careers in soil physics developed that interest while attending USAC and working with Gardner. L.A. Richards, one of his earliest students, obtained his masters degree at USAC and then went to Cornell University for his PhD degree. (*See* **Richards, Lorenzo A.**)

Another colleague who was strongly influenced by Gardner was Don Kirkham (*See* **Kirkham, Don**). Kirkham joined the physics faculty at USAC in 1937, having obtained a PhD in physics from Columbia University. Like Gardner, he came with no background in soil physics, but became interested through working with Gardner. After World War II, Kirkham was hired as a soil physicist by Iowa State University and became one of the most influential soil physicists of the twentieth century, both through his own research and through his large number of well-trained PhD students. Kirkham furnishes another tie between the two main characters of this brief history. Before

attending graduate school, Kirkham was a missionary for the LDS Church in Germany. John A. Widtsoe was in charge of the European missions at that time. Widtsoe and the young missionary found themselves together on a train at one point during Kirkham's mission, talking about careers. Widtsoe advised Kirkham to complete his education (i.e., attain the PhD) before he sought employment. Later, as Kirkham was in graduate school, he was offered a good job that would require that he drop out of school. This was tempting, since it was the depth of the depression, and jobs were hard to find. He remembered Widtsoe's advice, however, and completed his studies. Thus an important chapter in the lives of many soil physicists (Kirkham's students), as well as in the history of soil physics itself, would never have been written without the influence of both Widtsoe and Gardner in the life of Don Kirkham. Many similar stories could be told of the influence these men had on students and colleagues.

L.A. Richards, late in his professional career, recorded a revealing story about Willard Gardner's propensity to utilize mathematics in his scientific papers. The story is as follows:

> When I came on the scene as an undergraduate at Utah State College at Logan in 1922, Dr. Willard Gardner was pretty much alone in the subject matter field of soil physics and was making important applications of the Buckingham capillary potentials to the mathematical formulation of water flow in soil. This approach, though basic to progress, limited his current audience to people who understood his papers. I can well remember a conversation that took place between Sir Bernard Keen, then head of the Rothamsted Experimental Station [in the UK], and Dr. Gardner along about 1925. I was in the study room adjoining Dr. Gardner's office and with the door ajar. I, of course, was all ears concerning the conversation that was in progress. At a certain stage Dr. Keen said, 'Dr. Gardner I find your papers hard to read. Why don't you try to express things more simply?' Dr. Gardner was a man of some piety and had a quiet manner of speaking, but he had a prominent lower jaw. I could feel the air crackle in the next room as he paused for reply, which was, 'Sir Bernard, God made physics hard, not Willard Gardner.'

Hard or not, there have been many willing to follow in Willard Gardner's footsteps and learn his approaches to this new science. The result has been a century of solid progress. Like Widtsoe, he carried out pioneering work in many of the areas of modern soil physics. Unlike Widtsoe, he published many papers from which key ideas in modern soil physics can be traced.

See also: **Kirkham, Don**; **Richards, Lorenzo A.**

Further Reading

Buckingham E (1907) Studies of the movement of soil moisture. *Soils Bulletin* 38.

Gardner W (1919) The movement of moisture in soil by capillarity. *Soil Science* 7: 314–315.

Gardner W (1919) Capillary moisture-holding capacity. *Soil Science* 7: 319.

Gardner W (1920) A capillary transmission constant and methods of determining it experimentally. *Soil Science* 10: 106–107.

Gardner W and Widtsoe JA (1921) The movement of soil moisture. *Soil Science* 11: 215–232.

Taylor SA (1965) Willard Gardner. *Soil Science* 100: 79–82.

Widtsoe JA (1952) *In a Sunlit Land, the Autobiography of John A. Widtsoe*. Salt Lake City, UT: Deseret News Press.

Wind Erosion *See* **Erosion:** Wind-Induced

WINDBREAKS AND SHELTERBELTS

E S Takle, Iowa State University, Ames, IA, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Recent interest by agricultural and environmental scientists in finding more sustainable ways of managing landscapes has highlighted the need for a better understanding of plant–atmosphere–soil interactions. Windbreaks and shelterbelts are used to modify meteorologic conditions at the soil surface to meet specific management objectives. Both artificial barriers (fences) and living barriers (trees and shrubs) are used to modify flow near the ground. The terms 'windbreak' and 'shelterbelt' are often used interchangeably. However, a fence is usually considered a windbreak, whereas a living barrier, although also a windbreak, offers shelter not only from wind but also from intense solar radiation, driving rain, and erosive movement of water across the soil surface. The emergence of the fields of agroecosystems and agroforestry has brought further attention to the use of living barriers to modify plant and soil microclimates through changes in micrometeorologic conditions.

Several centuries of practical application, followed recently by a few decades of research, have provided a wide range of reasons for using windbreaks and shelterbelts. The most frequent uses relate to reducing wind speed for purposes such as reducing soil erosion or drifting soil and sand, capturing snow for soil-moisture recharge, or protecting highways, rail lines, farmsteads, and animal confinement areas. Recent research has focused on other benefits such as changing microclimates for crops, preserving soil moisture, improving landscape and biological diversity, and creating recreational opportunities or aesthetically desirable landscapes. Recent advances in numerical modeling of flow fields and microclimate factors around shelterbelts have increased understanding of the complexities of the soil–atmosphere–plant interactions.

Use of Windbreaks and Shelterbelts in Soil Management

The diversity of vegetation offered by shelterbelts in regions of monoculture-managed landscapes promotes biological diversity both above and below the surface. By providing shade, detritus layers, wind speed reduction, soil ventilation, and changes in temperature, humidity, and soil moisture, perennial living barriers offer richer spatial variation in microclimates for plants and belowground ecosystems. In intensively managed agricultural environments, tree shelterbelts provide islands having reduced concentrations of agricultural chemicals and increased biodiversity, with soil and aboveground ecosystems that include earthworms, small mammals, birds, perennial grasses, and woody plants and that deliver a range of beneficial ecosystem services.

Shelterbelts will become increasingly important as the regional impacts of global warming become more clearly identified, both for sequestering carbon and to suppress the negative agricultural impacts relating to reduced soil moisture and increased likelihood of erosion. Simulations of future-scenario climates show higher likelihood of more extreme events (both droughts and floods), and shelterbelts offer protection of crops under such impending changes. Prevention of soil loss due to high wind is a historic benefit of shelters, but they also suppress soil loss due to floods or intense rains on sloped surfaces. Soil erosion decreases soil productivity locally owing to loss of fine soil particles containing organic matter and nutrients and causes off-site damage to structures and unwanted deposition of soil particles. Irreversible damage to soil ecosystems due to extended drought is suppressed by shelters. Perennial living barriers in agricultural fields sequester carbon aboveground by creating biomass and belowground by deep roots, litter production, and providing regions of undisturbed soils that reduce microbial decomposition rates. Shelterbelts restore soil organic matter lost through agricultural practices.

Influence of Shelters on Aerodynamics and Microclimate

Wind Speed Reduction

The main impact of a windbreak or shelterbelt is to reduce mean wind speed near the surface. Acting as a porous barrier to the flow, the shelter creates a small region of wind reduction on the windward side and a low-speed, turbulent wake zone followed by a region of gradual wind speed recovery in the lee, as shown in Figure 1.

Shelterbelts typically reduce wind speed substantially on the leeward side and to a lesser extent on the windward side (Figure 2). The observed wind speed at 2.88 m above the ground reveals the region of wind speed reduction from the undisturbed value of 5.6 m s^{-1} to a sheltered value of 1.5 m s^{-1} in the

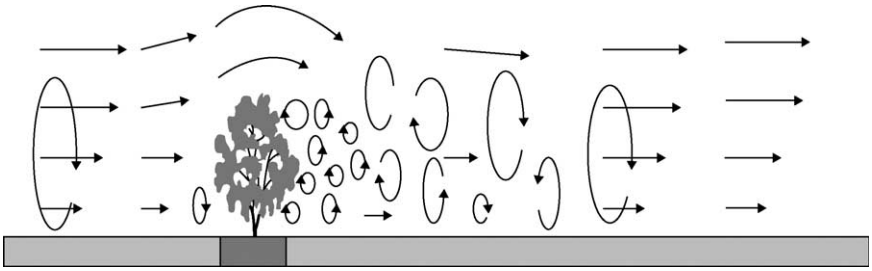


Figure 1 Mean and turbulent flow around a living shelter.

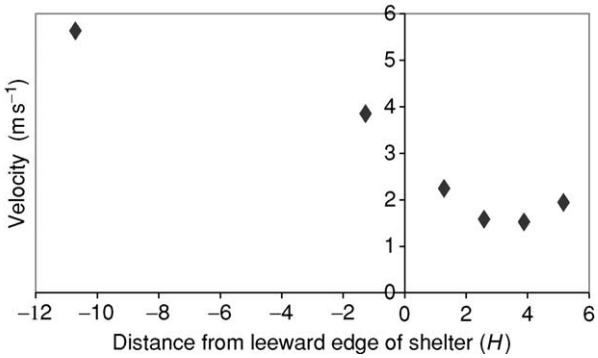


Figure 2 Observed horizontal wind speed at 2.88 m above the ground at various distances upwind and downwind of a two-row shelterbelt with height (H) of 14 m. Undisturbed wind speed at height 2.88 m is 5.6 m s^{-1}. Downwind edge of the shelter is taken as $x = 0$. (Adapted from Zhou XH, University of Nebraska.)

near lee. The barrier shelters strongly in a region approximately 10H leeward (H is height of the shelter), with modest sheltering extending leeward from the shelter approximately 20H.

The length of the sheltered region is reduced for more dense shelters and is larger but with less wind-speed reduction for highly porous shelters. A very dense shelter may even create a recirculation zone behind the shelter. When the wind direction is not perpendicular to the shelter, the sheltered region is reduced in size for successively more oblique angles of attack.

Shelterbelts have been used for many decades to suppress negative effects on soils due to high winds, namely to stop or slow drifting soil and sand. Suppression of drifting snow, on the other hand, makes important contributions to growing-season soil moisture in dry regions. Shelterbelts have been used widely to protect crops whose foliage or fruit are prone to damage from abrasion due to wind-blown material.

Turbulence Fields

The atmosphere contains turbulent eddies that connect the upper part of the atmospheric boundary layer to the surface (e.g., left-most eddy in Figure 1). These large eddies regulate surface exchange of heat, moisture, and trace gases between the atmosphere and the soil. Shelterbelts break up the large and efficient eddies that ventilate the surface into smaller eddies (e.g., see small eddies just downwind of the shelter in Figure 1) that are less efficient, thereby suppressing the transport of heat, moisture, and trace gases away from the surface. These small eddies have short lifetimes and dissipate within a few H of the shelter. Reduced turbulent fluxes of heat and moisture lead to higher near-surface air temperatures and soil temperatures within approximately 8H of the shelter compared with unsheltered regions during both day and at night.

Further downwind, where the large eddies reattach to the surface and where mean wind speed is reestablished, enhanced ventilation leads to slightly lower temperatures than in unsheltered areas. At night under low wind speeds and laminar flow (very small eddies) the shelter may increase the eddy size and transport heat downward. But if winds are already near calm, the barrier may reduce both mean and turbulent flow and allow the sheltered region to be cooler than unsheltered areas.

Pressure Fields

A shelterbelt, as a barrier to the flow, increases the static pressure upwind and reduces pressure in the lee of the shelter, as revealed by measurements shown in Figure 3. Since it is not a direct microclimate factor affecting crops and soil, pressure is often overlooked as one of the key influences of a shelter. However, fluctuations in the stationary pressure field created by the shelter can have substantial influence on movement of trace gases in soil or mulch layers.

Surface Fluxes of Heat and Moisture

The breakdown of large, efficient eddies in the lee of the shelter reduces vertical transport of both heat and moisture away from the near-surface air, which increases humidity near the surface and reduces surface evaporation or evapotranspiration. Lower evaporation allows more incoming radiant energy to be

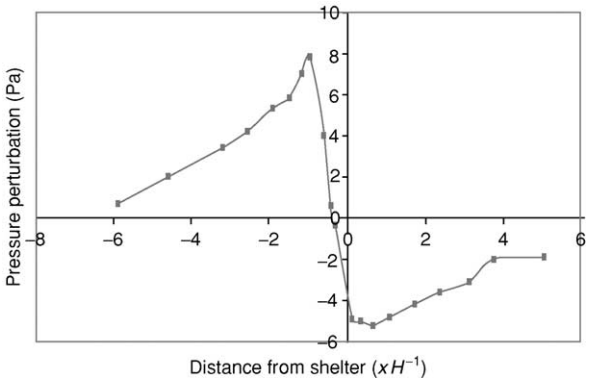


Figure 3 Perturbation of static pressure at ground level at various distances upwind and downwind of a two-row shelterbelt with height (H) of 14 m. Undisturbed wind speed at height 2.88 m was observed to be 5.6 m s^{-1}. Downwind edge of the shelter is taken as $x = 0$. (Data provided by Zhou XH, University of Nebraska.)

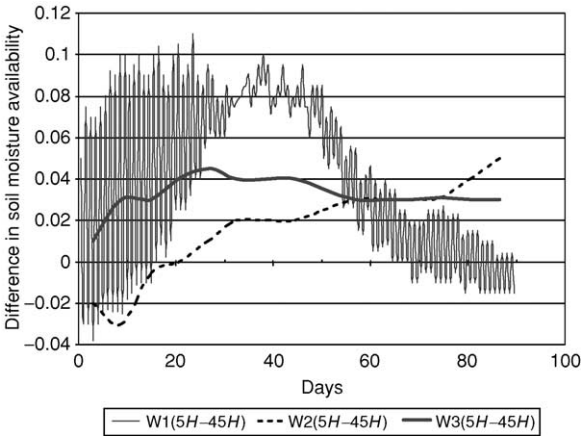


Figure 4 Differences in simulated soil moisture availability (dimensionless) between a sheltered region (5H downwind) and unsheltered region (45H downwind) at depths of 2.0 cm (W1), 1.0 m (W2), and 2.0 m (W3). The surface is assumed to be bare soil. (Adapted from Wang H, Shen J, and Takle ES (1997) Influences of agroforestry ecosystem on evapotranspiration and soil moisture. *Proceedings, 13th Conference on Hydrology*. Boston, MA: American Meteorological Society.)

partitioned to sensible heat, which thereby raises the surface temperature of the sheltered area. Energy loss by longwave radiation, which is proportional to the fourth power of temperature, keeps the surface temperature from getting too far above the temperature of the overlying atmosphere.

These modifications to the surface, and hence the near subsurface, of the soil lead to increased soil temperature and soil moisture compared with unsheltered regions, which in turn cause many soil chemical and biological processes to proceed at different rates in sheltered areas compared with open fields.

Living barriers consisting of woody plants also create a layer of litter or detritus mulch in the near vicinity to the stem or trunk. This layer does more than simply reduce extreme temperatures at the soil surface and preserve soil moisture; it also provides an additional zone for chemical and biological processes, such as methanogenesis. Being highly porous, such layers are subject to enhanced ventilation by mechanisms described below.

Soil Moisture

Most agricultural regions have deficient soil moisture at some point in the growing season. Observations suggest that shelterbelts raise both the daytime temperature and humidity in the sheltered region under most conditions, but exceptions do occur, in some cases due to soil-moisture conditions. Changes in soil moisture occur much slower than changes in air properties, so heat and humidity supplied due to the presence of soil moisture can modify micrometeorologic conditions in ways that may sometimes be counterintuitive.

Soil-moisture preservation in the lee of a shelter after a rain event that saturated the soil at all levels is shown in Figure 4. In this simulation, the soil surface was covered by a transpiring crop and was protected by a medium-density shelterbelt of height H. The time series of this plot gave the differences in soil moisture availability between sheltered and unsheltered locations at three depths for 90 days after a simulated precipitation event.

During the first 15 days after the onset of saturated conditions, daytime soil moisture at a depth of 2 cm (W1 in Figure 4) is higher in the sheltered region during the day (positive difference), but is lower at night (negative difference), presumably due to higher near-surface nighttime temperatures in the sheltered region. After 15 days, soil in the sheltered region has higher moisture both day and night. The relative benefit of shelter to soil-moisture preservation increases with time to a maximum near 40 days after the saturating rain event. At this time the soil-moisture availability near the shelter is approximately 25% higher than in unsheltered regions. Beyond this time the sheltered region loses moisture faster than unsheltered regions out to approximately 80 days, by which time the soil-moisture availability is everywhere uniform at approximately 0.14.

Deeper in the soil (1.0 m, plotted as W2, and 2.0 m, plotted as W3 in Figure 4), soil moisture is preserved in the sheltered region throughout the period except for the accelerated loss in the first 15 days from the 1-m layer due to higher temperatures. The region 4–8H downwind from the shelter is the region of highest protection, although evapotranspiration is

reduced by 10% or more compared with unsheltered areas out to approximately 13*H*.

Meteorologic Mechanisms that Move Gases in Soil

Most descriptions of gas movement in soils refer to the process as diffusional; however, ambient meteorologic conditions frequently create processes that move gases much more effectively than pure molecular diffusion. Numerous external (to the soil) factors are responsible for the movement of gases in soils. Most notably, pressure changes in the atmosphere are transmitted through airways within the soil and force movement of gases by nondiffusive means.

Static air pressure at a point (in either the atmosphere or soil above the saturated zone) is the horizontally averaged accumulated weight per unit area of atmospheric (gaseous) mass and the mass of any solid or liquid suspended by the atmosphere (birds, airplanes, liquid or solid H_2O, particulate matter, etc.) above that point. And changes in static pressure are caused by changes in the accumulated mass above the point. A point in the soil (we exclude soil locations not having a gaseous connection to the free atmosphere) may therefore experience pressure changes due to changes in the mass accumulation in the atmosphere above the point or due to phase changes of liquids or solids to vapors in the soil or on its surface. The region below the surface has zero mean flow in the absence of pressure gradients and volume expansion resulting from liquid-to-gas phase change. Any nonzero flow below the possible thin, turbulent layer in the soil is related to a pressure change or phase change, i.e., some change in the overlying mass accumulation.

Several mechanisms can lead to movement of gases within the soil and to effluxes from the soil. Many of these are related to changes in atmospheric pressure on various time scales or to liquid-to-gas phase changes. Because shelterbelts and windbreaks create spatial and temporal variations in surface pressure and temperature, they indirectly and importantly contribute to gas movement in the below-surface environment. Figure 5 provides a stylized view of pathways linking atmospheric gases with soil gases to assist in visualizing various mechanisms for gas movement.

Type 1: Large-Scale Pressure Changes

Changes in the static pressure of the atmosphere (and hence soil) due to movement of large-scale weather systems represent lowest-frequency pressure changes that might lead to externally driven transport of gases

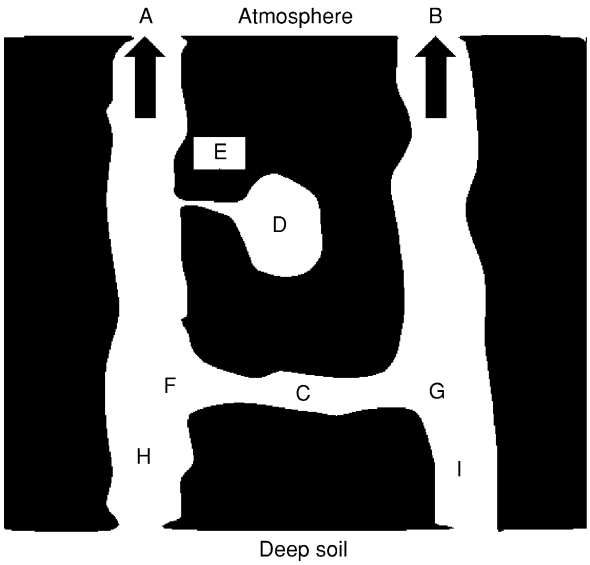


Figure 5 Stylized view of soil reservoirs and pathways for movement of trace gases.

in soils near the surface. Barometric pressure changes with periods of hours to days due to mesoscale and synoptic-scale weather systems cause changes in static pressure on the order of ±2% of total pressure over large areas, leading to uniform flow in or out at the surface.

Type 2: Traveling Atmospheric Pressure Waves

Traveling atmospheric pressure waves (of frequency assumed to be high enough to exclude the previous type of flow) transported over a flat, featureless surface can generate pressure fluctuations. For instance, observations of pressure, temperature, and wind fluctuations at several levels within tree canopies have revealed 'coherent structures' consisting of episodic bursts that penetrate and flood the canopy with overlying air. Pressure changes of this type at the soil surface cause oscillatory flow at A and B in Figure 5, with possibly different amplitudes and phases depending on the relative spatial separation as compared to the wavelengths of turbulent eddies. Point C experiences transient flow depending on conditions at A and B, and flow through E is oscillatory with zero mean. This type of flow offers a mechanism for drawing air with trace gases (water vapor, methane, nitrous oxide, carbon dioxide, oxygen) of high concentration out of D and replacing it with air of low trace-gas concentration.

Other atmospheric disturbances leading to traveling pressure waves include passage of a sea-breeze front, mountain lee waves, and outflow boundaries from thunderstorms and mesoscale convective complexes.

Type 3: Static Horizontal Pressure Gradient in Soil (Standing Wave)

Flow can be generated if points A and B, for illustrative purposes, are on opposite sides of an obstacle (e.g., tree, fence, rock outcropping, topographic feature) that creates a static horizontal pressure difference, as shown in Figure 3, due to a uniform (nonturbulent) wind blowing from left to right. Point A experiences high pressure and B low pressure that moves gas at a uniform rate from A to B through channel C owing to a pressure difference between F and G. As in type 2, passage E experiences only weak or zero flow. This process leads to horizontal as well as vertical motion in soils.

Type 4: Fluctuations on the Standing Wave

Wind that is not uniform in speed but has frequency ω (also assumed to be high enough to exclude the first type of flow but probably is lower than turbulence frequencies) may encounter an obstacle between A and B, thereby giving a fluctuating pressure of:

$$p = p_o e^{\omega t - k \cdot r} \quad [1]$$

at A, where k is the horizontal wave number and r is the horizontal distance vector. At B the pressure will be out of phase by π and possibly have different magnitude. Note that this forcing arises from changes in the static pressure due to dynamic interaction of the wind with the obstacle and is different from flow related to turbulence in the free atmosphere (see Type 5: Turbulence). This fourth kind of pressure change causes flow through C to be unidirectional and oscillatory, and flow through E to be oscillatory with zero mean. The venting of trace gases from cavity D, as described in type 2 flow, more effectively contributes to surface fluxes under this regime due to the unidirectional flow between A and C. If the obstacle is not between A and B, the pressure fluctuation is not out of phase, but the pressure gradient across C still leads to unidirectional flow.

Type 5: Turbulence

A turbulent atmosphere has stochastic fluctuations of pressure as well as wind speed, and larger turbulent pressure fluctuations generally accompany higher wind speeds. Despite the fact that the magnitude of these pressure fluctuations increases exponentially with mean wind speed, the resulting pumping action at the soil surface may displace an air parcel only a few micrometers at the surface. However, these pressure fluctuations may penetrate with minor loss of amplitude to the depth of pore D in Figure 5, and contribute to a substantial venting of the isolated pore. The resulting pumping action can move trace gases out of D at a much higher rate than is possible by molecular diffusion. Although this mechanism itself does not contribute strongly to surface effluxes, it may lead to higher availability of trapped gases for movement by other mechanisms.

Type 6: Venturi Effects

Differences in sizes of the soil channels through which air is ventilated lead to different flow speeds and therefore different rates of vaporization and gas transfer from solid or liquid surfaces lining the channel. Hollow stems of plants also have Venturi-like constrictions that cause irregular flow speeds.

Type 7: Rainfall

Rainwater ventilates soil gases by displacing air from pores and also enriches the soil air with dissolved O_2. Differential penetration of rainwater into the soil can lead to heterogeneous flushing of soil air. As rainwater percolates downward in the soil, the upper liquid surface also moves downward, drawing atmospheric air into the soil.

Type 8: Phase Change of Liquid to Vapor

When a liquid changes phase to a vapor, the accompanying volume expansion creates gaseous flow. For instance, evaporation of 1 cm of liquid water from a horizontally homogeneous soil and resulting expansion by a factor of 1000 leads to a net vertical discharge of 10 m ($10\,m^3\,m^{-2}$) of water vapor from the soil. This vertical efflux of water vapor carries other trace gases also. Horizontal differences in soil temperature lead to different evaporation rates and therefore different rates of soil-gas venting.

Spatial and temporal variability in soil pressures and temperatures created by shelterbelts, windbreaks, and isolated trees indirectly influence soil-gas movement, particularly by mechanisms 3, 4, 5, and 6 listed above. The presence of windbreaks or shelterbelts can therefore be expected to influence movements and distributions of trace gases in soil in complex ways.

Aesthetic and Recreational Value

Shelterbelts add diversity to agricultural landscapes, which, in turn promotes other biological diversity both above and below the surface and helps maintain delivery of a range of ecosystem services. Shelterbelts prevent loss of soil productivity owing to soil erosion. Finally, the addition of shelterbelts to the landscape, with their attendant diversity, enhances the aesthetic

and recreational value of rural environments, thereby promoting a sense of connectedness to the land that may be compromised in an increasingly urbanized society.

List of Technical Nomenclature

Angle of attack	Angle between the wind direction and a line perpendicular to a linear shelterbelt
Anthropogenic	Of human origin
Eddy	A transient circulation in a fluid that interacts with the mean flow to redistribute momentum, energy, and mass
Mesoscale convective complexes	Massive and long-lived storms that are relatively circular in shape, persist for 36 or more hours and produce large amounts of rain and possibly severe weather

See also: **Agroforestry**; **Biodiversity**; **Energy Balance**; **Erosion:** Wind-Induced; **Evapotranspiration**; **Forest Soils**; **Mulches**; **Soil–Plant–Atmosphere Continuum**

Further Reading

Brandle JR, Hinz DL, and Sturrock JW (1988) *Windbreak Technology*. Amsterdam, the Netherlands: Elsevier Science.

Brandle JR, Hodges L, and Wight B (2000) Windbreak practices. In: Garrett HE, Rietveld WJ, and Fisher RF (eds) *North American Agroforestry: An Integrated Science and Practice*, pp. 79–118. Madison, WI: American Society of Agronomy.

Hillel D (1998) *Environmental Soil Physics*. San Diego, CA: Academic Press.

Jury W, Gardner WR, and Gardner WH (1991) *Soil Physics*, 5th edn. New York: John Wiley.

McNaughton KG (1988) 1. Effects of windbreaks on turbulent transport and microclimate. *Agriculture, Ecosystems, and Environment* 22/23: 17–39.

Miller JM, Bohm M, and Cleugh HA (1995) *Direct Mechanical Effects of Wind on Selected Crops: A Review*. Technical Report No. 67. Canberra, Australia: Center for Environmental Mechanics ACT.

Rosenberg NJ, Blad BL, and Verma SB (1983) *Microclimate: The Biological Environment*, 2nd edn. New York: John Wiley.

Sumner ME (ed.) (2000) *Handbook of Soil Science*. Boca Raton, FL: CRC Press.

Wang H (2001) *High-Performance Cluster Computing, Algorithms, Implementations and Performance Evaluation for Computation-Intensive Applications to Promote Complex Scientific Research on Turbulent Flows*. PhD Dissertation. Ames, IA: Iowa State University Press.

Wang H, Takle ES, and Shen J (2001) Shelterbelts and windbreaks: mathematical modeling and computer simulation of turbulent flows. *Annual Reviews of Fluid Mechanics* 33: 549–586.

WOMEN IN SOIL SCIENCE (USA)

M J Levin, USDA Natural Resources Conservation Service, Washington, DC, USA

Published by Elsevier Ltd.

Introduction

Women have been involved in soil science and soil survey since interest in this most basic of natural resources emerged. Women employed by the early US Soil Survey were largely restricted to office activities, but soon they ventured into fieldwork. Albeit brief, the first appointment of a woman to a field party came about in 1901, only 2 years after the soil survey began in earnest. It would be another 45 years, however, before the Soil Conservation Service (SCS) would appoint a woman as a field soil scientist.

The delay in women joining the ranks of field scientists has significance for telling the story of women's place in the soil survey. Soil science is first and foremost a field-based science. Soil mapping begins by studying the landscape and building a conceptual model of how the topography, geology, plants, climate, water, and animals interact to determine soil characteristics. Mapping then continues in the field where the soil scientists validate their determinations of soil types by digging, describing, and sampling the soils and vegetation. Benchmark samples are analyzed in the laboratory to validate conceptual models. Finally, the information is consolidated on to aerial photo maps, associated computer databases, and manuscripts for publication. Until women gained an equal place as scientists in the field, they remained in supporting roles in the soil survey.

The Pioneers (1895–1965)

Women's roles in the earliest years of the soil survey appear to have been limited to clerical work, copy-editing of manuscripts, and cartographic drafting of maps, although women with appropriate academic

training soon began to work in the laboratories. Janette Steuart and Sorena Haygood, who maintained the laboratory and field records for the US Department of Agriculture's (USDA) Division of Agricultural Soils, were among the first women to work for the soil survey. According to Macy H. Lapham's account, Steuart was hired January 4, 1895, making her the first woman appointed to the Division of Soils, then part of the US Weather Bureau, located in the USDA. Both Steuart and Haygood were career employees; Steuart retired in the late 1920s with nearly 30 years of service and Haygood retired some time later.

The soil survey remained resistant to women in field parties. Julia R. Pearce, one of the earliest female pioneers in soil survey, joined the soil survey as a member of the field party at Hanford, California, in June 1901, but never had the opportunity to join her party in the field. Pearce had been one of only two 1901 graduates in agriculture from the University of California at Berkeley (UCB). In his commencement address, Secretary of Agriculture 'Tama Jim' Wilson lamented the small size of the graduating class and emphasized that the Department of Agriculture needed candidates trained for technical positions. After the speech, Pearce sought Wilson out and told him 'she was ready and willing to come to the relief of the Departments.' Pearce found herself almost immediately appointed as an assistant to Macy Lapham's all-male field party at Hanford.

While Secretary Wilson was sympathetic to the idea of women in the workplace, Lapham was uncomfortable with having women in the field with an all-male crew. On the day that Pearce arrived in Hanford, it was said that Lapham sent a telegram that stated, 'Miss Pearce is here, what in hell shall I do with her?' In the end, he put her to work copying maps. In 1903, she was transferred to the Bureau of Soils in Washington as an assistant in soil survey and later transferred to the Bureau of Plant Industry as a laboratory assistant.

From Lapham's memoirs, it appears fieldwork was out of the question for women before the 1940s. Women found places in the field in unofficial capacities, however. Mary Baldwin, the wife of soil inspector Mark Baldwin (employed by the Soil Survey 1912–1944), mapped with her husband in northern Wisconsin and the Boundary Waters of Minnesota during the early 1920s. They worked during the summer months, camping and using a small boat to go from island to island. Mary would drop Mark off on one side of the island, he would map, and she would wait for him with the boat on the other side of the island. While she waited, she might search for survey markers or make observations on her own of the general area. Mary recalled that there were times when she wished that she could have tried mapping on her own. As it was, she accompanied her husband everywhere during his remote mapping experiences, transcribing or taking field notes for him and assisting with the sampling.

Mary Baldwin might still have had difficulty finding employment with the USDA, even if she had the background and training to map soils. At the time, married couples were discouraged from working for the same federal agency, and the USDA already employed Mary's husband. A clause in a 1932 appropriations law even stated that married persons with a federally employed spouse would be dismissed first in the case of government reductions in force and the preference should be given to others for new appointments. Although this legislation was repealed in 1937, it limited married women's employment by the federal government at that time and perhaps set a precedent for the future.

During the decades of the 1930s, 1940s, and 1950s, women contributed to the soil survey through editing, writing erosion history, and conducting laboratory work. Lillian H. Weiland – the first female employee of the newly established Soil Erosion Service and secretary to Hugh Hammond Bennett – put together a 'Bibliography on Soil Conservation Compiled in the Office of the Chief of SCS' in 1935. The bibliography consolidated ideas for soil erosion control technology for the new agency. In 1937, Lois Olson and Dr Arthur Hall spoke on studies in erosion history as part of a series of research seminars for SCS staff. Some of today's thinking on interpretations of the soil survey and field practices to control erosion can be attributed to this series of lectures. Olson, a geographer by training, headed the SCS's Erosion History Section.

Charlotte Whitford (Coulton), a graduate of Ohio State University with a master's degree in botany, joined the SCS as a secretary with a field soils staff in Zanesville, Ohio, in the mid-1930s. J. Gordon Steele, an old classmate, soon recruited Whitford to work as an assistant soil technologist in Washington, DC, on a series of reports on soil erosion. She later worked as an editor on soil surveys and eventually became head of the SCS publication staff. She retired in the 1980s with almost 50 years of service.

Dorothy Nickerson, a soil color technologist for the USDA from the late 1920s through the 1940s, was instrumental in developing the soil color standards for soil survey. Nickerson had been an assistant manager of the Munsell Color Company before joining the USDA in 1927. She made extensive colorimetric tests in the laboratory and worked with soil scientists in the field to match soils to the Munsell colors and to create a new set of color names, first introduced in preliminary form in 1941. She then

worked with Thomas D. Rice, Kenneth Kelly, and Albert H. Munsell to adapt the Munsell color chart system for describing soil color in the laboratory and the field. The US Soil Survey adopted the Munsell color charts and new color names in 1949.

The first woman soil scientist officially assigned in the field for the SCS was Mary C. Baltz (Tyler). Mary Baltz graduated from Cornell University and joined the soil survey as a 'junior soil surveyor' in 1946. Labor shortages during World War II provided the opportunity for her to work in a job that, up to that time, appeared to be reserved for men. By 1951, Mary was responsible for mapping in Madison and Oneida Counties in New York, and later she was assigned the task of map measurement for the entire state. In contrast with today's electronic techniques, the work was done by cutting out the soil map delineations on copies of the field sheets. Areas with the same label were weighed together and a factor converted the weight to acres. She hired a team of women to do this conversion job in the winter months.

Erwin Rice, a retired soil scientist in New York, mapped under Mary Baltz's direction. He remembered Mary as a confident, petite woman who enjoyed mapping in the field, was comfortable with the all-male crews, and had a good sense of humor. He called her a 'splitter,' a soil scientist who tends to separate out concepts for new soils as opposed to lumping them together under general categories of old soil names. Mary Baltz worked for the SCS until about 1965.

Ester Perry was a major figure in the California soil survey effort. Her 1939 PhD in soil science from UCB was the first received by a woman in the USA. For her doctoral research, 'Profile Studies of the More Extensive Primary Soils Derived from Granitic Rocks in California,' she may have been one of the first students to use X-ray diffraction to look at clay mineralogy structure in soils. She studied under Charles Shaw at Berkeley, and Kelly, Doer, and Brown were mentors and coworkers with her in Riverside, California, where she worked at the Subtropical Horticulture Research Center during her graduate studies. From 1928 to 1939 she worked as an associate soil technologist for the California Agricultural Experiment Station.

From 1939 until she retired in 1965, Perry essentially ran the USDA soil survey laboratory in Room 33 in the basement of Hilgard Hall at UCB. She moved up through the ranks during that time, from 'junior soil technologist' (1939–1954) to 'associate specialist soils' (1954–1960) to 'specialist' (1960–1965). In a 1952 presentation to the Western Soil Science Society in Corvallis, Oregon, Perry praised the benefits of close collaboration between a soils laboratory and a field soil scientist for quick turnaround of information in support of mapping. After the establishment of the Beltsville Agricultural Research Center in 1952, the Berkeley Soils Laboratory was slowly phased out.

In soil science, or more specifically, in pedology (the study of soil genesis), as with all the earth sciences, there were very few women working in the field before the 1970s. Gary Sposito, who was a student in Ester Perry's laboratory for a year, recalls that she was well aware of being a pioneer in her profession. As one of 'Ester's boys' (as students who worked part-time in her laboratory were known), he thought she effectively mentored many young men and women into a soil science career. Dr Perry maintained an all-business approach in the laboratory, but also remembered to bring birthday cakes for her students and provided a bed in the laboratory for those who might work through the night on important projects. Despite her accomplishments, Ester Perry was never promoted to associate professor or put on the tenure track. This was not unusual, however, since many women and men worked as researchers or technicians for their entire careers at the agricultural experiment station without receiving academic status. She also was not acknowledged in USDA records as an official soil survey collaborator.

Foundations: Building on the Pioneers (1959–1975)

During the 1950s and 1960s few women ventured directly into the field of soil science. Some arrived through other disciplines, such as geology, microbiology, or one of the plant sciences. They found through their graduate studies that soil science was a key element in their research and then continued on to further studies in soil science. Some were mentored and encouraged by major professors (as Ester Perry had been by Shaw and Kelly) to pursue soil science and stick with it. As scientists and teachers, the women soil scientists who started in the 1950s and 1960s spent a good deal of their careers as mentors themselves, and many placed a high value on that aspect of their work.

Ester Perry, who bridged the gap from the pioneer era to this period of building foundations, was, herself, one such mentor, not only through her work with students in the Berkeley Soils Laboratory, but also particularly in her effort to bring equal access to soil survey field training to women in the mid-1950s. In that decade, a summer field course – Soils 105 field trip – was a requirement for a soil science major from UCB and the University of California at Davis (UCD). The course was offered each summer, and since the 1930s had convinced many a prospective soil science student that soil survey could be a lifelong interest and career path. In 1953, Eva Esterman, a soil science honors student, was the

first woman to request to take the course. The UCB Soils Department offered Eva an option to graduate without participating in the field course, but she wanted to take the trip just like all of the other students. For 6 weeks in the field, the university arranged for Eva to have separate sleeping facilities and comfort stops. Dr Earl Story's wife served as chaperone, accompanying the students in a separate car and with some discomfort. The academic dean at the time, Dr Frank Haridine, considered the experiment a complete disaster and swore publicly that no women would ever again take the field course.

The event triggered Ester Perry to step in and offer a Soils 105F course ('F' for female), which she planned and made available from 1956 to 1959. The trip was soil survey-oriented but because Ester had different professional contacts, the course had a somewhat different approach. Three women took the course in 1959, the last year the specialized course was offered. In 1965, the 'Soils 105' course officially became co-ed and included two women students. Perry accompanied the group as a chaperone. Seven years later, the class was almost 50% women, and there was no women's chaperone.

In addition to mentoring students of soil science, many women professors and researchers made substantial contributions to our understanding of soils and soil science during the 1960s and 1970s. In the USA, Cornelia Cameron, Jane Forsyth, Jaya Iyer, Eva Esterman, Nellie Stark, and Elizabeth Klepper have been among the most prominent, although there are no doubt many others. Scientific publications usually list the first names of researchers only by their initials, which makes identifying the authors by gender difficult at best; the individuals included here were identified by their students and colleagues.

Dr Cornelia C. Cameron completed her PhD in geology (with an emphasis in geomorphology) at the University of Iowa in 1940. After 11 years teaching earth sciences, Dr Cameron joined the US Geological Survey (USGS) in 1951 and spent the next 43 years in the field. Her field career began in military geology, with terrain analysis of military sites in over 30 countries on five continents, many of them dangerous militarized zones at the time. Dr Cameron's colleagues remember her as quite a character in the field.

In part, Dr Cameron's reputation for eccentricity had to do with her mother. One of the USA's first female PhDs in botany, Dr Cameron's mother had a strong interest in her daughter's work and accompanied Cornelia on field expeditions until she was 103 years old. The younger Dr Cameron joked that when her mother got so old that her eyesight had deteriorated, she put a cow bell on her so she could find her if she wandered off. In a story about daughter and mother's adventurous military terrain investigations in the Caribbean area in early 1961 before the Bay of Pigs invasion, Dr Cameron recounted that "Mother and I were a perfect pair. We told everyone that we were Canadian tourists. One time, as I was doing traverses along the slopes of one of the islands, Mother stayed in the car. I was upslope from her when I saw a truck full of guerrillas pull up. Mother simply charmed them and they drove off."

Dr Cameron was an internationally recognized authority on peat soils and their use as a soil additive and source of energy, and on the impact of peat removal on swamp and bog environments; she wrote prolifically – 110 publications – on the subject. Both the USGS and the Department of the Interior recognized her accomplishments in research and public service. Dr Cameron received the USGS's Meritorious Service Award in 1977 and its Distinguished Service Award in 1986, and received the Department of the Interior's Public Service Recognition Award in 1990.

Dr Jane L. Forsyth, a professor of geology at Bowling Green State University in Ohio, has contributed much to our understanding of the age relationship of soils and till to northern Ohio glacial geology. Among her peers she has been affectionately dubbed the 'Queen of the Pleistocenes,' according to her colleague Peter Birkeland, retired from the faculty at University of Colorado, Boulder. Dr Forsyth earned her PhD from Ohio State University in 1956 and taught at University of Cincinnati, Miami University, UCB, and Ohio State University, before joining the faculty at Bowling Green State University in 1965.

Dr Jaya Iyer, professor of soil science at the University of Wisconsin, Madison, has focused her research on relating soil properties and tree growth. Dr Iyer earned her PhD in botany from the University of Bombay, India, in 1959. The external referee for her PhD research, soil scientist Dr Sergei Wilde of the University of Wisconsin and a member of the Wisconsin Forestry Hall of Fame, encouraged her to take an interest in soils, and she completed a second PhD in Soil Science at the University of Wisconsin, Madison in 1969. Dr Iyer has had a highly successful career as a national expert in soils for tree nurseries, specializing in urban, Christmas tree, and forestry production.

Dr Eva Esterman, who received her PhD in soil science from UCB in 1958, nearly 20 years after Ester Perry, was only the second woman to earn a soil science PhD from Berkeley. She went on to become a professor at San Francisco State University in 1960, where she taught botany and later added soil science to the curriculum. She focused her soils research on soil microbiology and biochemistry before she retired in 1982 and began raising sheep.

Dr Nellie Stark, tenured professor and forest soil ecologist at the University of Montana, Missoula, from 1970 to 1992, gave us the theory of the 'biological life of a soil,' which describes how soils and plants interact during development and decline phases of soil genesis. The theory explains the variation in nutrient uptake by plants based on the stage of soil genesis. Indirect nutrient cycling, which involves uptake of ions by the roots from the soil, predominates when a soil is young; direct nutrient cycling, which involves uptake of ions by the roots directly from the litter, by-passing the soil, occurs as the soil becomes older and depleted by weathering.

Stark earned a PhD in botany (ecology) from Duke University in 1962, with a minor in soils based on credits she collected at Oregon State University in 1961. Dr Stark's research with the Desert Research Institute, Reno, Nevada, focused on soils and nutrient cycling of litter in the tropical ecosystems of Brazil and Peru. The soil chemistry laboratory for forestry that Dr Stark operated at the University of Montana was well-known and received and processed samples from all over the world.

Finally Dr Elizabeth L. Klepper, a research leader and plant physiologist at the Columbia Plateau Conservation Research Center, Pendleton, Oregon, concentrates her research on root growth and functioning under field conditions and plant–soil water relations. Dr Klepper holds degrees from Vanderbilt University and Duke University. She has been recognized for her accomplishments by all three professional agronomic research societies of the USA: the American Society of Agronomy (ASA), the Crop Society of America (CSA), and the Soil Science Society of America (SSSA). She was the first woman ever to receive the Fellow award from the SSSA (American Society of Agronomy).

These are but a few of the women who dedicated their lives and research to soil science and soil survey during this period. It is inspiring to consider their achievements. Over the course of about 50 years, women in soil science and soil survey moved out from under their restriction to clerical support to become influential researchers and field investigators. Both Dr Stark's theory of the biological life of a soil and Dr Cameron's important and risky fieldwork would have been virtually unimaginable when the soil survey began in 1899.

In the Classroom, in the Field, and in the Laboratory (1970–1990)

Despite great strides in the field of soil science, women were still not actively recruited into the USDA's Soil Survey Division in the 1960s. For example, in a 1962 recruitment speech to the Agronomic Education Division of the American Society of Agronomy in Ithaca, New York, Assistant Soil Survey Administrator Charles Kellogg expressed his agencies' concern about recruiting good candidates, "especially of well-trained, broadly educated young men who can develop rapidly." His comments were not surprising perhaps, since the professional workforce of the country at that time was still predominantly male.

The transition during the 1960s was profound, however. By the 1970s, career counseling documents were beginning to discuss ways to channel girls into nontraditional careers, and encouraging young women to enter nontraditional occupations continued as a theme into the 1980s. Corresponding changes occurred in the classrooms; as more and more young women began to enter previously male science and employment territories, materials and approaches to education changed to meet the needs of this more diverse student population.

In the soil survey, as well as in some of the other earth sciences professions, a woman still needed to be persistent in the 1970s to obtain a field appointment. In the SCS of the 1970s, there were fewer than 15 women in the federal employment series soil scientist (470 series) at anyone time nationally, despite an acceleration in soil survey mapping and a general increase in field crews.

Most of these women soil scientists thought they were the only female soil scientist in the agency. In addition, there were no formal professional organizations for women field soil scientists; the Association of Women Soil Scientists (AWSS), organized by a group of women soil scientists in the US Forest Service, was not formed until the early 1980s. But slowly more career opportunities began to emerge for women. Title VII of the Civil Rights Act of 1964 and the Civil Service Reform Act of 1978 helped increase opportunity by prohibiting sex discrimination in employment and requiring diversity in the workforce, and the Women in Science and Technology Equal Opportunity Act of 1980 opened up more opportunities for women to receive support in university settings.

In the 1970S and early 1980s, the SCS soil survey staff in California included five women field soil scientists (Arlene Tugel, Nancy Severy, Chris Bartlett, Lisa Holkolt, and Maxine Levin), a crowd compared to other states. Many states had only one woman working as a soil scientist in the field – Carole Jett in Nevada; Carol Wettstein in Florida; Sue Southard in Utah; Margaret Rice in Mississippi; Caryl Radatz in Minnesota; Mary Collins in Iowa; and Debbie Brasfield in Tennessee, for example. Some states had two or more women in the field, and there may have

been more. Records of employees for those years are spotty and have not been saved comprehensively. But, in any event, women still comprised a small percentage of the total field soil scientist staff. Nevertheless, their contribution to soil survey was sizable, with millions of acres mapped, at times with some physical hardship.

Many of the women who worked in field parties during the 1970s went on to achieve greater responsibilities and position. By the late 1980s, the SCS had appointed the first woman state soil scientist, followed by others in the early 1990s, and women grew more prominent as soil survey party leaders. Carol Wettstein became the first woman state soil scientist, serving in Maryland from 1988 to 1989 and as state soil scientist in Colorado from 1990 to 1995. Carole Jett served as state soil scientist in California in 1991, and Carol Franks was state soil scientist in Arizona in 1994. In 2000, Maxine Levin was appointed national program manager of the Soil Survey Division. In the 1980s, there were at least three published soil surveys for which women were the party leaders or the principal field investigators: Sacramento County, California (Arlene Tugel); City of Baltimore, Maryland (Maxine Levin); and Indian River County, Florida (Carol Wettstein).

We can anticipate seeing more women listed in soil surveys of the 1990s, as the number of women party leaders increased significantly during the decade. An all-female crew of soil scientists led by Deborah Prevost mapped the Hualapai-Havasupai Indian Reservation, Arizona, in the late 1980s and early 1990s. This soil survey was published in 1999 – exactly 100 years after the establishment of the USDA Soil Survey.

Women have also contributed to soil surveys in many other ways that are not reflected in publications. Soil correlators and data management specialists make significant contributions to soil survey data and manuscripts, and a number of women have held these positions: Sue Southard (California), Renee Gross (Nebraska), Carmen Santiago (Puerto Rico), Panola Rivers (Pennsylvania), Kathy Swain (New Hampshire), Laurie Kiniry (Texas), Diane Shields (Delaware), Susan Davis (Maryland), Marjorie Faber (Connecticut), Tammy Cheever (Nebraska), and Deborah Anderson (North Carolina). In the last few years, women soil scientists have been instrumental in the effort to digitize the soils information that is used in the publications, including Vivian Owen (Texas), Jennifer Brookover (Sweet) (Texas), Darlene Monds (Massachusetts), Barbara Alexander (Connecticut), Caroline Alves (Vermont), Lindsay Hodgman (Maine), Caryl Radatz (Missouri), Adrian Smith (Nebraska), Amanda Moore (Oregon), Sharon Schneider (Oregon), Marcella Callahan (Arkansas), Brandi Baird (Oregon), and Jackie Pashnik (Rhode Island).

In the National Cooperative Soil Survey there are also a number of women field soil scientists who work mostly with soil survey interpretations and education, including Sue Southard (California – volcanic soils and vertisols), Lenore M. Vasilas (Maryland – hydric soils), Sheryl Kunickis (Washington, DC – landscape analysis), Susan Ploetz (Minnesota – resource inventory), Susan Casby-Horton (Texas – soil geomorphology), Christine Clarke (Maryland – geographic information systems), Jeannine Freyman (Virginia), Karen Kotlar (New York), Lisa Krall (Connecticut), Gay Lynn Kinter (Michigan), Donna Hinz (Nebraska), Patricia Wright-Koll (Minnesota), Jeanette Bradley (Arkansas), and Deborah Prevost (Nevada). Like agricultural extension specialists, these soil scientists act as a bridge between university research, soil survey mapping, and the public, interpreting soil surveys for practical use by both agencies and individuals, including providing on-site field investigations.

Women in the SCS, now the Natural Resources Conservation Service (NRCS), have made significant contributions to soil science in the National Soil Survey Laboratory (NSSL) and the National Soil Survey Center (NSSC), and as researchers in the Soil Quality and Watershed Sciences Institutes. Carolyn G. Olson has been a lead research scientist at the NSSC, located in Lincoln, Nebraska, since 1989. Dr Olson's research focuses on soil geomorphology, quaternary geology, and clay mineralogy. Olson received the honor of being made a Fellow of the Soil Science Society of America in 1996. Other women soil scientists at NSSL and NSSC include Rebecca Burt (soil chemical properties), Joyce Scheyer (urban soil properties), Susan Samson-Liebig (soil quality), Deborah Harms (soil physical properties), Sharon Waltman (national soil survey databases and GIS interpretations), and Carol Franks (soil biology). In the Institutes, Arlene Tugel (New Mexico), Betty McQuaid (North Carolina), and Cathy Seybold (Oregon) have been working with soil quality and watershed health indicators.

Other federal agencies, such as the USGS, also provide opportunities for women in soil science research. Jennifer W. Harden, with USGS in Menlo Park, California, built on her PhD research using soil chronosequencing to develop the Harden Index, which used soil horizons and carbon-dating to measure time in the alluvium sequencing. Since then, she has worked on the effect of climate on soil, particularly as it relates to groundwater recharge and wetland assessment. She has been a front-runner in research on global change issues of soil carbon, carbon dioxide emissions, and soil carbon sequestration. Marith Reheis, with the USGS in Denver, has

also done significant research in using soil properties as a paleoclimatic record for chronosequence mapping in Rocky Mountain glacial outwash. Originally a geologist by training, she received her PhD in soil science under Pete Birkeland at the University of Colorado, Boulder, in 1984.

At NASA's Goddard Space Flight Center, Elissa Levine has been working with soils in forested ecosystems since 1987. She models soil physics and soil chemistry to assess watershed leaching, soil carbon ecosystem effects, and the effects of acid precipitation on soils and groundwater. She was recently appointed lead scientist for the NASA Global Change Master Directory. She was also selected as a Fellow of the Brandwein Institute for Science Education, an award based on her work as the principal scientist in the GLOBE program's soil investigations for teaching soil science worldwide to K-12 students.

Since the 1970s, women scientists with the US Forest Service have been involved with the National Cooperative Soil Survey ecological unit inventories, as well as with technical soil interpretations in the specialties of forest soil productivity, soil erodibility, fire ecology, and forest ecosystem health. Some of the prominent women involved in this effort include Gretta Boley (Washington, DC), Clare Johnson (Six Rivers National Forest, California), Carol Smith (SCS-USDA and Tahoe National Forest, California), Barbara Leuelling (Superior National Forest, Minnesota), Connie Carpenter (White Mountain National Forest) and Mary Beth Adams (Northeast Forest Experiment Station, West Virginia).

In 2000, three women held positions as pedology (soil genesis) professors in US universities: Janice L. Boettinger, Utah State University, Logan; Christine Evans, University of New Hampshire, Durham; and Mary Collins, University of Florida, Gainesville. Dr Boettinger is working on an extensive review of worldwide zeolite mineral occurrences in soils and the use of zeolite and clinoptilolite for waste disposal systems of animal production operations. She is also working on characterizing selected soil resources of Utah, which includes research on saline, wet soils and irrigation-induced hydric soil characteristics. Dr Evans is focusing her research in the field of describing anthropogenic (human-influenced) soils and developing terminology to describe soil properties derived from human activity.

Dr Collins' research at the University of Florida focuses on the genesis, morphology, and classification of soils; identifying and delineating hydric soils; using ground-penetrating radar to study subsurface properties; and pedoarcheology. She is best known for her dedication to soil survey fieldwork and reaching out to other countries to spread soils technology. As part of the People to People Program, she first opened the door to doing ground-penetrating radar soil investigations in China and Portugal. Dr Collins was made a Fellow of the American Society of Agronomy in 1996 and a Fellow of the Soil Science Society of America in 1997.

A number of women who have been made Fellows of the Soil Science Society of America have provided outstanding contributions to soil science. Mary Beth Kirkham, made a Fellow in 1987, is a professor at the Kansas State University Evapotranspiration Laboratory. Her work has focused on heavy-metal uptake by plants and soil–plant–water relations for over 20 years. Mary K. Firestone, who became a Fellow in 1995 and also received the Emil Truog Soil Science Award, is a professor of soil microbiology at UCB. Her research focuses on the microbial population-basis of carbon and nitrogen processing in ecosystems. Jean L. Steiner, made a Fellow in 1996, is director of the USDA Agricultural Research Service (ARS) Southern Piedmont Conservation Research Laboratory in Watkinsville, Georgia. Her research is on humid region water balance studies in complex topographies. Finally, Diane E. Stott, a 1997 American Society of Agronomy Fellow, is a soil microbiologist with ARS in West Lafayette, Indiana. Her work has focused on the effects of organic matter dynamics on soil structure and erodibility and modeling the effects of plant residue decay on erodibility.

The women described above are only a sample of the many women who have contributed research, mapping, applications, and education to the field of soil science in the last 25 years. Others, to name only a few more, include Nancy Cavallaro (University of Puerto Rico – soil chemistry and tropical soil fertility), Laurie Drinkwater (Rodale Institute, Pennsylvania – sustainable agriculture), Kate Scow (University of California, Davis – soil microbiology), Laurie Osher (University of Maine, Bangor – soil science), Samantha Langley (University of Southern Maine, Gorham – soil science education), Kate Showers (Boston University – soil conservation), Jeri Berc (NRCS-USDA – soil conservation), and Katherine Newkirk (Woodshole Marine Biological Laboratory – global warming).

1990 and Ahead

Clearly, women have made significant and numerous contributions to the field of soil science and soil survey through research, mapping, applications, and education. As the numbers of women have increased in the classroom, laboratory, and field, changes have also taken place within the soil science discipline. The women who are graduating in soil science in the 1990s are confident and intellectually engaged and

are quickly gaining recognition for their work. As an example, Eva M. Muller of Spokane, Washington, a soil survey project leader with only 7 years of experience, was awarded the National Cooperative Soil Survey Soil Scientist of the Year Award in 2001. Lenore M. Vasilas, who finished her MS in soils in 1997, remarked about any remaining stereotypes in her work, "Oh we don't think about it... We just go ahead and do it!"

There is a world of difference between Julia Pearce's experience in the early 1900s and Lenore Vasilas's reality in 1998. In the educational realm, between 1987 and 1996, soil science, along with education, communication, and social science, experienced the largest percentage growth of female participation. While overall enrollment of students (BS, MS, and PhD) in the soil sciences held relatively steady between 1987 and 1996, fluctuating between 1200 and 1500 students, enrollment of women in the soil sciences rose from 16.2% in 1987 to 32% in 1996. In 1996 there were 228 female BS graduates in soil science, almost double that of 10 years before. PhD and MS candidates in the soil sciences in 1996 were also about one-third female, once again double from 10 years before.

Progress in employment numbers in the SCS/NRCS appears a little less dramatic. In 1985, the SCS employed 85 women soil scientists at the federal level; in May 1998, there were 94 women soil scientists in various level positions. The progress is significant, however, given the overall reduction in the total number of soil scientists in the agency. Currently, about half of the NRCS's new soil scientists are women. We can anticipate a twenty-first century that witnesses a continuing trend of more women working in the field in soil survey and private consulting firms and as teachers and researchers in the field of soil science in university and laboratory settings across the USA and the world.

Acknowledgment

This article is adapted, with permission of Iowa State University Press from Levin MJ (2002) Opening opportunities: women in soil science and the soil survey. In: Helms D, Effland ABW, and Durana PD (eds) *Profiles in the History of the US Soil Survey*, pp. 149–168. Ames: Iowa State University Press.

Further Reading

Baker L (1976) Women in the US Department of Agriculture. *Agricultural History* 50(1): 190–201.

Bureau of the Census (1911) *Official Register: Persons in the Civil, Military, and Naval Service of the United States, and List of Vessels*, vol. 1. Washington, DC: Bureau of the Census, Department of Commerce and Labor.

Cattell J (ed.) (1944) *American Men of Science: A Biographical Directory*. Lancaster, Penn.: Science Press.

Cattell J, Cattell G, and Hancock D (eds) (1961) *American Men of Science: A Biographical Directory, the Physical and Biological Sciences*, 10th edn. Tempe, AZ: Jaques Cattell Press.

Helms D (1992) Women in the Soil Conservation Service. *Women in Natural Resources* 14(1): 88–93.

Jaques Cattell Press (ed.) (1972) *American Men and Women of Science, Formerly American Men of Science, the Physical and Biological Sciences*, vol. 4, 12th edn. New York: Jaques Cattell Press/R. R. Bowker Company.

Kellogg C (1963) Opportunities for soil scientists and agronomists in the Soil Conservation Service. *Agronomy Journal* 55: 575–576.

Lapham MH (1945) The soil survey from the horse-and-buggy days to the modern age of the flying machine. *Soil Science Society of America Proceedings* 10: 344.

Lapham MH (1949) *Crisscross Trails: Narrative of a Soil Surveyor*. Berkeley, CA: Willis E. Berg.

Prevost D and Linsay BA (1999) *Soil Survey of Hualapai-Havasu-pai Area, Arizona, Parts of Coconino, Mohave, and Yavapai Counties*. Washington, DC: Natural Resources Conservation Service, US Department of Agriculture.

WORLD SOIL MAP

H Eswaran and P F Reich, USDA Natural Resources Conservation Service, Washington, DC, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Soil is commonly described as the mantle or the skin covering the land mass of planet Earth. In some places, bare rock is exposed and the skin is absent or may be just a few centimeters thick. In other locations, the mantle is several meters thick with distinct layers that reveal the formation or depositional history of the mantle. For most people, soil is the natural medium for the growth of nonaquatic plants and has a thickness that is determined by the rooting depth of plants. It is composed of solids (mineral and organic), liquids, and gases. Soil scientists differentiate between the initial material (if it was a sediment) and the soil derived through soil-forming processes. The upper

limit of the soil is the contact between soil and air or shallow water (less than 2.5 m deep). The soil on a landscape is a continuum, and across the landscape it grades into nonsoil materials such as rock, ice, or deep water.

The study of soil is 'pedology' and the scientists who study it are referred to as pedologists or soil scientists. The soil that is studied is the zone that extends to the nonsoil material (if the latter occurs at less than 200 cm) or to approximately 200 cm if the nonsoil material occurs deeper, as this is the zone where much of the biological or pedological processes operate. This is a convention, though; for some studies, depths of 20 m or more are considered. The different climates that prevail in the world and their interactions with the rocks or sediments produce a range of soils that vary in their physical, chemical, biological, and mineralogical properties. The kind of soil is also determined by the geomorphology of the land, the amount and rate of water moving in and on it, and the kind of disturbances it has been subject to, including human influence.

It is important to understand and know the different kinds of soils and how they form, as this helps in sustainable land management. Soil classifications, created to help organize knowledge about soils, have been structured to enable this. An important difference between classification of soils and classification of plants and animals is that, unlike the latter, soil is not discrete but a continuum, and as a consequence the boundaries between named soils are artificial. Depending on the objectives for which the classification was created, the class definitions and the architecture of the classification system may vary. Many countries have national soil classification systems but one that is widely used and employed is Soil Taxonomy. Soil surveys enable the depiction of soils across a landscape and soil maps are made to show the patterns of soils that exist and provide information on the management properties of the soils. Specialized maps with technical soil classifications are produced for specific objectives. The soil maps are produced at different scales: detailed maps (1:10 000 or larger scale) to show soils on farms, and general maps (1:100 000 or smaller scale) to depict soils over large areas such as countries, continents, or the world. Soil classification has the additional function of linking the information of these different maps.

Soil Classification

To appreciate the variety of soils that occur on this planet, it is necessary to understand the classification system used to demonstrate this. Two classification systems are widely used: the Food and Agriculture Organization of the United Nations (FAO) initiated the World Reference Base for soil classification; while Soil Taxonomy is a system developed in the USA with the collaboration of international soil scientists. Soil Taxonomy has six categories, as listed in Table 1. Each category serves specific functions determined by its information content. The soil series is the lowest category and is defined with the maximum amount of information. Information is generalized in the definition of each higher category. Consequently, each category can only be used for specific purposes. When employed to designate map units, the scale of the maps limits the category that is represented, as shown in Table 1. The highest category, the order, has only 12 classes. There are 64 classes in the suborder, 300 classes in the great 10 470 group, and 2400 classes in the subgroup. In the USA soil families and 21 000 soil series are recognized. No global estimates for the number of classes in these two categories exist.

Soils are grouped into 12 orders in Soil Taxonomy. Figure 1 shows the global distribution of the soil orders and Figure 2 shows representative examples. Soil-forming processes (Table 1) and the attributes they provide for the soil are used to define the soil. This is expressed in another manner in Table 2. As soils are composed of mineral or organic materials or mixtures, when these primary components dominate the system, the resulting properties define the soil: short-range minerals in Andisols, smectitic clays in Vertisols, organic matter in Histosols, and low-activity clays in Oxisols. In the second group of soils, processes listed in Table 1 result in horizons of accumulation or depletion. The next group of soils (Inceptisols and Gelisols) also results from soil-forming processes but the manifestations are slight or mainly physical. Finally, Entisols are the soils where the original material is least altered or not modified by soil-forming processes. As climate or hydrology controls many of the processes, each of these major kinds of soils has a specific place on the landscape. However, each soil may have properties transitional to other soils, and such secondary features are used to define the intergrades or the lower categories. Each order of soils also has specific subordinate properties that are employed to define the lower categories.

Unlike other soil classification systems, Soil Taxonomy incorporates soil climate by using the soil moisture regime (SMR) and the soil temperature regime (STR) to define lower categories. Soils with an aridic SMR (deserts) occupy approximately 36% of the total land area, making such land only available for agriculture if there is a source of irrigation. The soils with a xeric SMR (Mediterranean) occur in areas with winter rains and, like those with an ustic

Table 1 Defining characteristics of the categories in soil taxonomy

Category	*Definition*	*Functions*	*Potential uses*
Order	Soils having properties (marks) or conditions, resulting from major soil-forming processes that are sufficiently stable pedologically and that help to delineate broad zonal groups of soils	Depict zones where similar soil conditions have occurred for general understanding of global patterns of soil resources Establish global geographic areas within which more specific factors and processes result in the diversity of soils	General global, continental, or regional assessment Global climate-change studies AMS $<$1:10 000 000 MSD $>$40 000
Suborder	Soils within an order having additional properties or conditions that are major controls or reflect such controls on the current set of soil-forming processes and delineate broad ecosystem regions	Demarcate broad areas where dominant soil-moisture conditions generally result from global atmospheric conditions Delineate contiguous areas with similar natural resource endowments Ecosystems with distinct vegetation affinities usually determined by limiting factors of soil moisture or conditions	Demarcate areas in regions or large countries for assessment and implementation of economic development Analysis of international production and trade patterns Priority setting for multipurpose uses of land resources AMS 1:1 000 000–1:10 000 000 MSD 4000–40 000
Great group	Soils within a suborder having additional properties that constitute subordinate or additional controls or reflect such controls on the current set of soil-forming processes, including landscape-forming processes	To demarcate contiguous areas with similar production systems or performance potentials	Development of strategic plans for regional development Basis for coordinating national resource assessment and monitoring programs Infrastructure development to assure equity in development AMS 1:250 000–1:1 000 000 MSD 250–4000
Subgroup	Soils within a great group having additional properties resulting from a blending or overlapping of sets of processes in space or time that cause one kind of soil to develop from, or toward, another kind of soil: Intergrades show the linkage to the great group, suborder, or order level; Extragrades have sets of processes or conditions that have not been recognized as criteria for any class at a higher level, including nonsoil features The soil is considered as the 'typic' member of the class if the set of properties does not define intergrades or extragrades	To demarcate production land units with similar land use and management requirements	Targeting research and development for specific land use or cropping systems Implementing conservation practices and ecosystem-based assistance Community development projects and monitoring sustainability Basis for diversification of agriculture and uses of land resources Basis for implementing environmental management programs and modeling ecosystem performance AMS 1:100 000–1:250 000 MSD 40–250
Family	Soils within a subgroup having additional properties that characterize the parent material and ambient conditions The most important properties are particle size, mineralogy, and soil temperature regime	To demarcate resource-management domains characterized by similar management technology and production capabilities	Basic units for extension and/or technology transfer Modeling cropping systems' performance Addressing socioeconomic concerns AMS 1:25 000–1:100 000 MSD 2.5–40
Series	Soils within a family having additional properties that reflect relatively narrow ranges of soil-forming factors and processes, determined by small variations in local physiographic conditions, that transform parent material into soil	To delineate land units for site-specific management of farms	Implementing soil-specific farming Designing farm-level conservation practices AMS $>$1:25 000 MSD $<$2.5

AMS, appropriate map scale; MSD, minimum size delineation: smallest area that can be delineated on a map with a legible identification, in hectares.

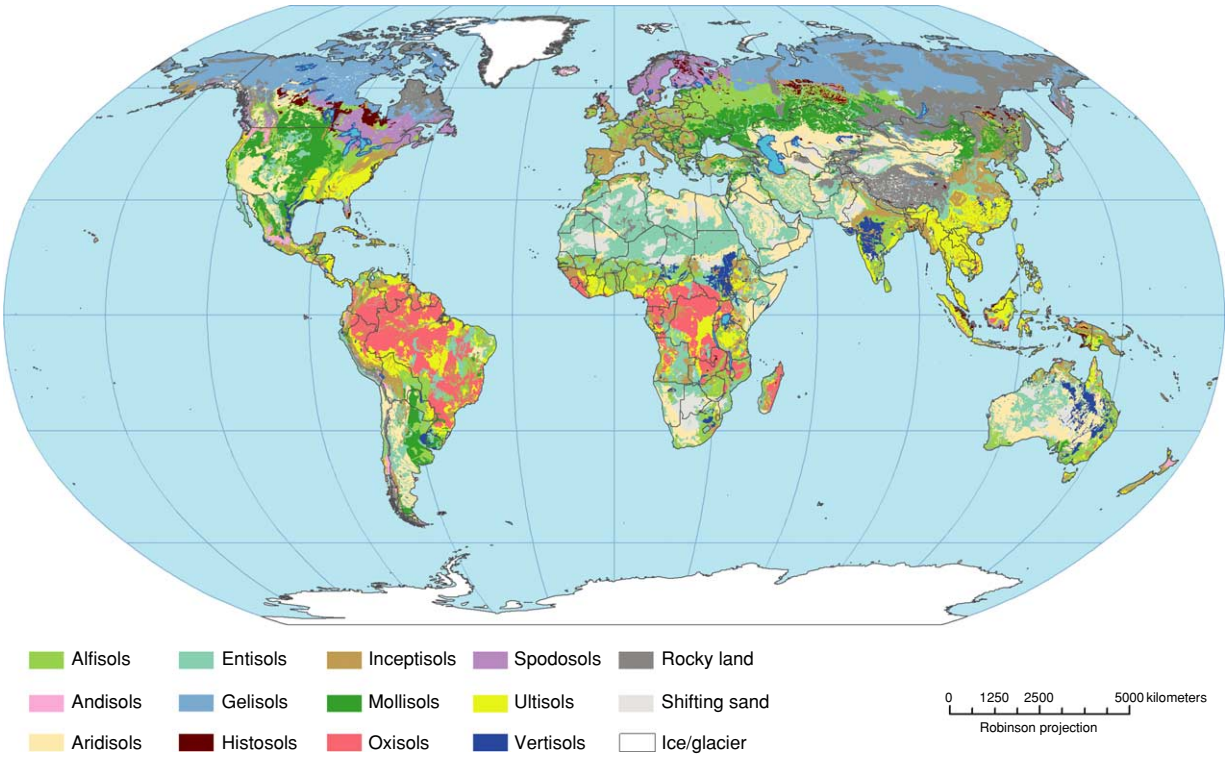


Figure 1 A generalized soil map of the world.

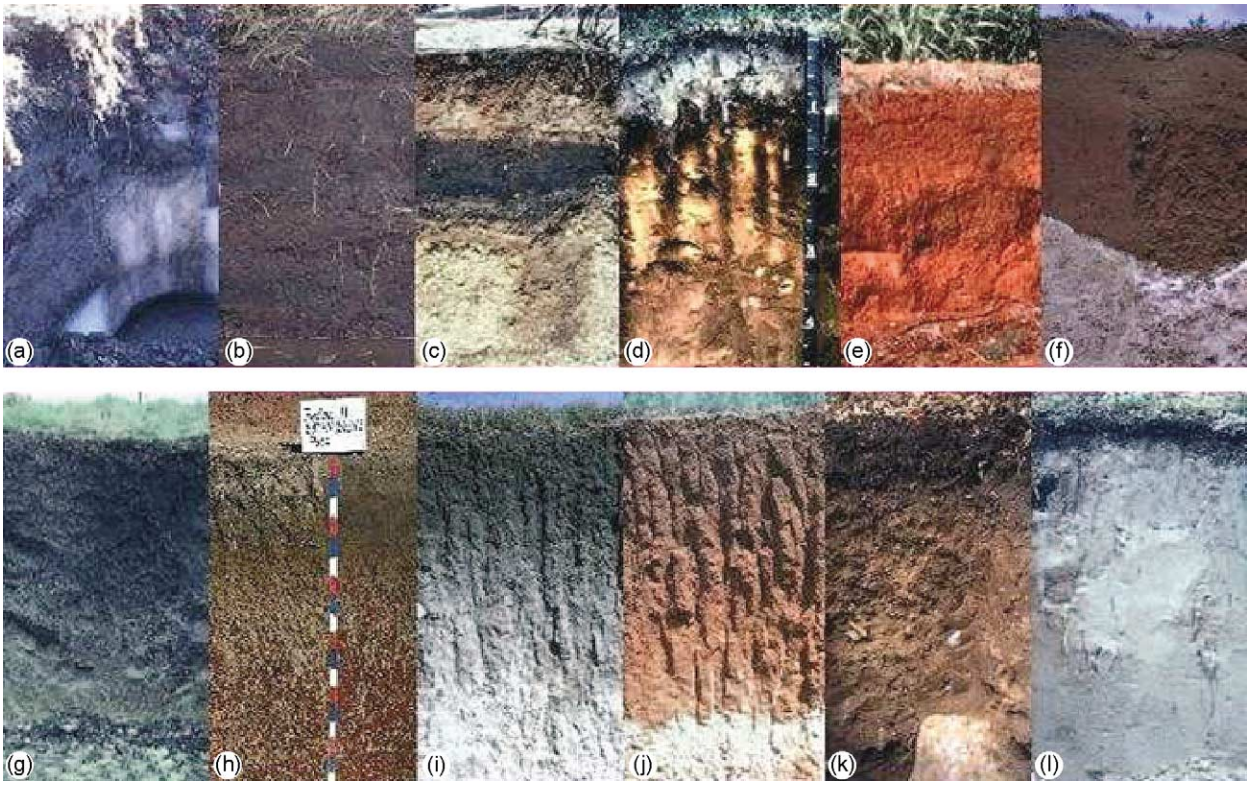


Figure 2 Examples of major soil orders: (a) Gelisol; (b) Histosol; (c) Andisol; (d) Spodosol; (e) Oxisol; (f) Aridisol; (g) Vertisol; (h) Ultisol; (i) Mollisol; (j) Alfisol; (k) Inceptisol; and (l) Entisol.

Table 2 Grouping of soils based on major differentiating criteria

Properties determined by the soil material	*Horizon results from an accumulation or depletion*	*Horizon formed by alteration of primary minerals or of aberrant properties*	*Features of sediment (stratification) retained*[a]
Histosols	Alfisols	Inceptisols	Entisols
Andisols	Ultisols	Gelisols	
Oxisols	Mollisols		
Vertisols	Spodosols		
	Aridisols		

[a]Only a small accumulation of organic matter.

SMR (semiarid), have moisture stress for prolonged periods during the year. These two groups of soils occupy approximately 21.5% of the land and form the semiarid zones, where a second crop requires supplemental irrigation. The soils with an aquic SMR represent the wetlands, which generally are ecologically fragile systems. Thus only approximately 33% of the total land is relatively free from moisture stress, although low temperatures may also stress part of this. The very cold pergelic soils, which occupy approximately 11% of the land, are generally under sparse forest vegetation or are free of vegetation. The cryic and frigid (boreal) soils are generally under forest but are cultivated in some areas with an annual grain crop of short duration. The temperate climates are reflected by soils with mesic, thermic, isomesic, and isothermic STRs, and in general if the SMR is conducive, they are the soils with the highest agricultural potential. The humid tropical soils are represented by the combination of an isohyperthermic STR and udic or perudic SMRs. They occupy approximately 9% of the land surface and have not only some of the poorest soils, but also a heavy incidence of pests and diseases. Each combination of SMR and STR represents a unique pedoenvironment that also contains a myriad of soils.

The distribution of the different kinds of soils is presented in Table 3. The Aridisols are the largest order of soils, occupying approximately 23% of the land. The Histosols, Andisols, and Vertisols individually occupy less than 3% of the area. These are soils formed on special materials or under special conditions and so occupy specific geographic areas. The Gelisols are confined to the very cold regions of the world. Due to intense weathering and leaching by the tropical environment, most of the Oxisols and Ultisols are located in the tropical zones. The unique soil-formation conditions of the tropics are also reflected in the distribution of the Alfisols and Mollisols, which are more extensive in the temperate regions. Inceptisols and Entisols are present under a range of climatic environments. Estimates of the distribution of the suborders of soils are also given in Table 3.

Global Distribution

Gelisols

In areas where the mean annual soil temperature is less than 0°C, the soils are frozen for long periods of the year and thaw out during the short warmer spells. The freezing and thawing processes promote physical changes in the soil. If there is sufficient water and the warm period is long enough, vegetation is established and organic matter accumulates on the soil. Organic-rich soils or peat develop. Due to low temperatures, these Arctic soils have unique features such as ice-lenses, or a layer of ice may underlie the soil. In situations where free water is limited, permafrost layers are present. Depth to the permafrost layer is determined by several factors, including the protective organic-rich layers on the soil surface. In arid areas, the soil particles are held together by dry permafrost. A thin film of frozen water forms a binding film around the particles. The three suborders in Gelisols are:

1. Histels: characterized by a histic epipedon with ice lenses; may be underlain by ice or permafrost, or rock;
2. Turbels: characterized by disrupted horizons or ice wedges penetrating upward into horizons;
3. Orthels: characterized by permafrost layers at depth; but the surface horizons do not show disruptions or distortions due to freezing and thawing. Frequent in the transition to warmer soils.

Histosols

Most soil classifications, including Soil Taxonomy, separate mineral soils from organic soils. Histosols are soils that consist of dominantly organic soil materials. Histosols comprise only a small portion of the world's land area (1.2%), but are widely distributed irrespective of climate. They develop where the rates of organic matter accumulation exceed decomposition and removal. Most of these soils have formed

under saturated conditions where the soil is saturated or nearly saturated with water most of the year. These soils are referred to as bogs, moors, peat, or mucks.

To be farmed, most Histosols must be drained. Management of the water table depth is critical to their use. When drained, Histosols oxidize and subside, and require further drainage. The rate of subsidence can be several centimeters per year and is affected most by depth of drainage. Experience has shown that Histosols in cool temperate areas have the lowest levels of subsidence under tillage. Climate is the main limitation for crops grown on Histosols. Many are used for vegetable crops. Fire and wind-erosion hazards increase after Histosols have been drained.

Histosols are and have been mined as a source of fuel and as a source of organic matter for soil amendment. Because of their extreme instability, Histosols present special problems when roads or other structures are built on them. They are removed if possible during road construction, and buildings constructed on them are normally placed on pilings driven into the mineral soil below them.

Vegetation found on Histosols ranges from grasses, sedges, and rushes to shrubs and trees, all of which are usually water-loving or water-tolerant. Most recently Histosols have been recognized for their importance worldwide as wetland wildlife habitats. Those associated with open bodies of water are especially important habitats. Histosols are recognized as equally important filters of pollutants such as phosphates, nitrates, and other agricultural and industrial contaminants that migrate in water.

The suborders of Histosols are separated by the degree of decomposition of the organic matter, and climate:

1. Folists consist of leaves, twigs, and branches resting on rocks, stones, or coarse fragments in which the interstices are partly filled with organic materials. They are not saturated with water for more than a few days per year if rock is present at shallow depths; they occupy a small area (0.14 million km^2), mainly in the northern latitudes;
2. Fibrists consist largely of plant remains so slightly decomposed that rubbing does not destroy them and their botanical origin can be determined easily. They occupy large areas (1.1 million km^2) in the northern latitudes;
3. Hemists consist of organic materials that are so decomposed that the botanic origin of as much as two-thirds of the materials cannot be readily discerned or the fibers can be largely destroyed by rubbing between the fingers. They occupy small areas in the northern latitudes and may also be found in the tropics;
4. Saprists consist of almost completely decomposed plant remains and their botanic origin cannot be determined. Only small areas of these soils are found in the cool and cold regions of the world; most of them are confined to the tropics.

Andisols

Soils formed on volcanic ash and cinders and having andic properties are distributed along the circum-Pacific belts and occur sporadically elsewhere (1.8% of the land mass). The mineralogical composition is a function of the age of the deposit and the climatic conditions it has been exposed to. The Andisols have mineralogical composition ranging from volcanic glass, short-range-order minerals such as allophane and immogolite, and variable amounts of halloysite. This mineralogical association gives unique properties to such soils, including a high phosphate-fixing capacity, low cation retention, and a high water-holding capacity. Many of these soils are found on volcanic slopes or are developed through the weathering of plateau basalts. These soils support a high human population density owing to their general ease of cultivation and also because of the cool environment of the volcanic mountains, which is generally free of pests and diseases.

The success of human habitation on Andisols is largely due to the soil and climatic environment. Though phosphorus availability is a problem, there is usually a sufficient supply to meet most needs. For intensive cultivation, fertilizers are needed. The soils are friable and well drained so tillage is not a problem. As the soils have been used for generations, conservation practices have been installed in most situations.

Seven suborders are recognized in the Andisols. These are:

1. Aquands: these soils are wet and experience extensive water saturation for prolonged periods during the year. Aquands are local and do not occupy enough land to be shown on small-scale maps;
2. Cryands: these soils have a cryic or pergelic STR. Cryands are found in small areas of Canada, the Kamchatka peninsula of Russia, and high elevations in the tropics;
3. Torrands: these soils have an aridic SMR. Torrands have been reported in the volcanic areas of Mexico and Syria; many have not been studied in detail and so classification is still in doubt;
4. Xerands: these soils have a xeric SMR. Xerands occur sporadically in the Mediterranean areas of

Table 3 Estimates of areas occupied by major soils of the world

Soil		*Ice-free land*		*Tropical*		*Temperate*		*Boreal*		*Tundra*		*Aridic*		*Xeric*		*Ustic*		*Udic*	
Order	*Suborder*	*km²*	*%*	*km²*	*%*	*km²*	*%*	*km²*	*%*	*km²*	*%*	*km²*	*%*	*km²*	*%*	*km²*	*%*	*km²*	*%*
	Histels	1 013 358	0.77	0	0	0	0	0	0	1 011 295	0.77	0	0	0	0	0	0	0	0
	Turbels	6 332 748	4.84	0	0	0	0	0	0	6 316 202	4.83	0	0	0	0	0	0	0	0
	Orthels	3 914 016	2.99	0	0	0	0	0	0	3 903 320	2.98	0	0	0	0	0	0	0	0
Gelisols		11 260 122	8.61	0	0	0	0	0	0	11 230 817	8.59	0	0	0	0	0	0	0	0
	Folists	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Fibrists	197 387	0.15	0	0	0	0	194 056	0.15	0	0	11 674	0.01	0	0	42 689	0.03	139 694	0.11
	Hemists	988 264	0.76	0	0	99 508	0.08	884 897	0.68	0	0	103 452	0.08	8 551	0.01	268 205	0.21	604 197	0.46
	Saprists	340 781	0.26	317 753	0.24	22 201	0.02	0	0	0	0	0	0	0	0	40 540	0.03	299 415	0.23
Histosols		1 526 432	1.17	317 753	0.24	121 709	0.09	1 078 953	0.82	0	0	115 126	0.09	8 551	0.01	351 434	0.27	1 043 306	0.8
	Aquods	169 059	0.13	13 364	0.01	55 748	0.04	99 566	0.08	0	0	0	0	0	0	0	0	168 678	0.13
	Cryods	2 459 814	1.88	0	0	0	0	2 455 873	1.88	0	0	11 871	0.01	6 278	0	179 455	0.14	2 258 269	1.73
	Humods	57 870	0.04	29 242	0.02	28 476	0.02	78	0	0	0	78	0	6 546	0.01	10 081	0.01	41 091	0.03
	Orthods	666 784	0.51	18 103	0.01	508 527	0.39	138 221	0.11	0	0	0	0	32 739	0.03	95 150	0.07	536 962	0.41
Spodosols		3 353 527	2.56	60 709	0.05	592 751	0.45	2 693 738	2.06	0	0	11 949	0.01	45 563	0.03	284 686	0.22	3 005 000	2.3
	Cryands	255 195	0.2	0	0	0	0	254 426	0.19	0	0	13 990	0.01	0	0	47 144	0.04	193 292	0.15
	Torrands	1 598	0	1 598	0	0	0	0	0	0	0	1 598	0	0	0	0	0	0	0
	Xerands	32 128	0.02	0	0	32 118	0.02	0	0	0	0	0	0	32 118	0.02	0	0	0	0
	Vitrands	281 070	0.21	202 457	0.15	77 443	0.06	605	0	0	0	26 158	0.02	0	0	161 689	0.12	92 658	0.07
	Ustands	62 822	0.05	58 857	0.04	3 690	0	0	0	0	0	0	0	0	0	62 547	0.05	0	0
	Udands	279 427	0.21	185 821	0.14	89 490	0.07	1 765	0	0	0	0	0	0	0	0	0	277 076	0.21
Andisols		912 240	0.7	448 733	0.34	202 741	0.16	256 796	0.2	0	0	41 746	0.03	32 118	0.02	271 380	0.21	563 026	0.43
	Aquox	320 065	0.24	320 065	0.24	0	0	0	0	0	0	0	0	0	0	300 272	0.23	19 792	0.02
	Torrox	31 233	0.02	27 118	0.02	4 115	0	0	0	0	0	31 233	0.02	0	0	0	0	0	0
	Ustox	3 096 466	2.37	3 086 719	2.36	9 465	0.01	0	0	0	0	0	0	0	0	3 096 185	2.37	0	0
	Perox	1 161 980	0.89	1 010 135	0.77	151 490	0.12	0	0	0	0	0	0	0	0	0	0	1 161 626	0.89
	Udox	5 201 102	3.98	5 166 551	3.95	32 506	0.02	0	0	0	0	0	0	0	0	0	0	5 199 065	3.97
Oxisols		9 810 846	7.5	9 610 588	7.35	197 576	0.15	0	0	0	0	31 233	0.02	0	0	3 396 457	2.6	6 380 483	4.88
	Aquerts	5 484	0	763	0	4 721	0	0	0	0	0	0	0	0	0	0	0	5 484	0
	Cryerts	14 925	0.01	0	0	0	0	14 911	0.01	0	0	34	0	113	0	0	0	14 764	0.01
	Xererts	98 718	0.08	0	0	98 577	0.08	0	0	0	0	0	0	98 577	0.08	0	0	0	0
	Torrerts	889 353	0.68	238 410	0.18	647 662	0.5	0	0	0	0	886 072	0.68	0	0	0	0	0	0
	Usterts	1 767 647	1.35	1 169 403	0.89	594 367	0.45	2 288	0	0	0	0	0	0	0	1 766 059	1.35	0	0
	Uderts	384 358	0.29	86 105	0.07	297 273	0.23	0	0	0	0	0	0	0	0	0	0	383 378	0.29
Vertisols		3 160 485	2.42	1 494 681	1.14	1 642 600	1.26	17 199	0.01	0	0	886 106	0.68	98 690	0.08	1 766 059	1.35	403 626	0.31
	Cryids	943 285	0.72	0	0	0	0	940 532	0.72	0	0	795 230	0.61	20 339	0.02	55 348	0.04	69 615	0.05
	Salids	890 118	0.68	52 910	0.04	632 946	0.48	195 536	0.15	691	0	761 691	0.58	17 320	0.01	95 103	0.07	7 279	0.01
	Gypsids	682 963	0.52	228 484	0.17	429 405	0.33	24 126	0.02	0	0	601 964	0.46	22 692	0.02	57 359	0.04	0	0
	Argids	5 407 965	4.13	573 248	0.44	4 035 105	3.09	782 223	0.6	0	0	5 015 755	3.83	93 618	0.07	268 535	0.21	12 656	0.01
	Calcids	4 872 554	3.73	451 161	0.34	4 400 123	3.36	13 823	0.01	0	0	4 728 720	3.62	71 574	0.05	61 180	0.05	3 637	0
	Cambids	2 931 387	2.24	561 394	0.43	2 063 362	1.58	302 236	0.23	0	0	2 926 976	2.24	0	0	0	0	0	0

Aridisols		15 728 272	12.02	1 867 197	1.43	11 560 941	8.84	2 258 476	1.73	691	0	14 830 336	11.34	225 543	0.17	537 525	0.41	93 187	0.07
	Aquults	1 280 989	0.98	1 042 999	0.8	235 985	0.18	58	0	0	0	0	0	1 337	0	729 665	0.56	548 038	0.42
	Humults	343 518	0.26	277 802	0.21	61 013	0.05	4 691	0	0	0	0	0	18 159	0.01	38 007	0.03	287 339	0.22
	Udults	5 539 906	4.24	2 654 476	2.03	2 872 711	2.2	9 523	0.01	0	0	0	0	0	0	0	0	5 536 706	4.23
	Ustults	3 869 722	2.96	3 630 467	2.78	234 877	0.18	1 471	0	0	0	0	0	0	0	3 866 828	2.96	0	0
	Xerults	18 815	0.01	0	0	940	0	17 875	0.01	0	0	0	0	18 815	0.01	0	0	0	0
Ultisols		11 052 950	8.45	7 605 744	5.81	3 405 526	2.6	33 618	0.03	0	0	0	0	38 311	0.03	4 634 500	3.54	6 372 083	4.87
	Albolls	27 656	0.02	0	0	1 372	0	26 266	0.02	0	0	20 965	0.02	0	0	18	0	6 656	0.01
	Aquolls	118 072	0.09	1 156	0	84 787	0.06	31 974	0.02	0	0	0	0	9	0	18 191	0.01	99 717	0.08
	Rendolls	265 827	0.2	120 959	0.09	103 513	0.08	40 986	0.03	0	0	0	0	408	0	15 371	0.01	249 680	0.19
	Xerolls	924 394	0.71	0	0	873 511	0.67	50 365	0.04	0	0	0	0	923 876	0.71	0	0	0	0
	Cryolls	1 163 797	0.89	0	0	0	0	1 160 462	0.89	0	0	0	0	271 415	0.21	588 171	0.45	300 877	0.23
	Ustolls	5 244 636	4.01	184 731	0.14	2 370 624	1.81	2 682 438	2.05	0	0	3 387 540	2.59	0	0	1 850 262	1.41	0	0
	Udolls	1 261 051	0.96	54 220	0.04	1 058 004	0.81	146 619	0.11	0	0	0	0	0	0	0	0	1 258 841	0.96
Mollisols		9 005 433	6.89	361 066	0.28	4 491 811	3.43	4 139 110	3.16	0	0	3 408 505	2.61	1 195 708	0.91	2 472 013	1.89	1 915 771	1.46
	Aqualfs	836 077	0.64	407 123	0.31	373 655	0.29	54 760	0.04	0	0	0	0	9 925	0.01	548 919	0.42	276 695	0.21
	Cryalfs	2 517 693	1.92	0	0	0	0	2 509 517	1.92	0	0	0	0	77 883	0.06	814 915	0.62	1 616 711	1.24
	Ustalfs	5 663 916	4.33	3 773 322	2.88	1 719 484	1.31	165 070	0.13	0	0	0	0	0	0	5 657 884	4.33	0	0
	Xeralfs	896 915	0.69	0	0	848 514	0.65	46 146	0.04	0	0	0	0	894 661	0.68	0	0	0	0
	Udalfs	2 706 299	2.07	616 696	0.47	1 926 532	1.47	158 113	0.12	0	0	0	0	0	0	0	0	2 701 333	2.07
Alfisols		12 620 900	9.65	4 797 141	3.67	4 868 185	3.72	2 933 606	2.24	0	0	0	0	982 469	0.75	7 021 718	5.37	4 594 739	3.51
	Aquepts	3 199 286	2.45	1 498 377	1.15	1 183 134	0.9	502 454	0.38	0	0	0	0	7 500	0.01	1 050 621	0.8	2 125 835	1.63
	Anthrepts	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Cryepts	456 920	0.35	0	0	0	0	456 109	0.35	0	0	0	0	79 516	0.06	87 881	0.07	288 711	0.22
	Ustepts	4 241 101	3.24	2 804 601	2.14	1 372 212	1.05	60 352	0.05	0	0	0	0	0	0	4 237 166	3.24	0	0
	Xerepts	685 336	0.52	0	0	674 694	0.52	9 760	0.01	0	0	0	0	684 454	0.52	0	0	0	0
	Udepts	4 247 035	3.25	1 755 023	1.34	2 153 364	1.65	333 545	0.26	0	0	0	0	0	0	0	0	4 241 911	3.24
Inceptisols		12 829 678	9.81	6 058 001	4.63	5 383 404	4.12	1 362 220	1.04	0	0	0	0	771 470	0.59	5 375 668	4.11	6 656 457	5.09
	Aquents	116 087	0.09	105 614	0.08	10 222	0.01	0	0	0	0	3 589	0	0	0	83 087	0.06	29 160	0.02
	Psamments	4 428 052	3.39	2 799 768	2.14	1 625 304	1.24	1 333	0	0	0	1 466 816	1.12	56 315	0.04	2 324 022	1.78	579 248	0.44
	Fluvents	2 860 240	2.19	1 017 875	0.78	1 455 341	1.11	368 407	0.28	0	0	830 887	0.64	174 240	0.13	967 198	0.74	869 295	0.66
	Orthents	13 733 032	10.5	2 095 961	1.6	11 264 330	8.61	363 206	0.28	0	0	10 432 873	7.98	620 154	0.47	1 138 660	0.87	1 531 763	1.17
Entisols		21 137 411	16.16	6 019 218	4.6	14 355 197	10.98	732 946	0.56	0	0	12 734 165	9.74	850 709	0.65	4 512 967	3.45	3 009 466	2.3
	Shifting sands	5 321 875	4.07	545 835	0.42	4 680 128	3.58	92 912	0.07	0	0	5 279 510	4.04	7 255	0.01	27 137	0.02	4 972	0
	Rock	13 076 333	10	7 055	0.01	363 324	0.28	3 727 903	2.85	8 956 897	6.85	1 139 000	0.87	151 764	0.12	507 688	0.39	2 299 876	1.76
	Ice	14 640 098																	
Miscellaneous		397 692 802																	
Total		130 796 504	100	39 193 721	29.97	51 865 893	39.65	19 327 477	14.78	20 188 405	15.43	38 477 676	29.42	4 408 151	3.37	31 159 23	23.82	36 341 992	27.79

the Near East, and not much information on them exists;

5. Vitrands: these soils are recent ash or cinder deposits where some weathering and soil formation has taken place. Vitrands are normally very recent deposits such as those around Mount St. Helens in the USA or around Mount Pinatubo in the Philippines. Mostly they occur on steep upper slopes, and no reliable estimates on them are available;

6. Ustands: these soils have an ustic SMR. Ustands are widespread on the leeward side of tropical volcanoes and are generally used intensively for agriculture;

7. Udands: these soils have an udic SMR. Udands are most widely used for agriculture and generally support the highest population density. Large areas of South-East Asia, Central Africa, Mount Cameroon in West Africa, and the Andean range in South America, as well as many of the volcanic islands of the Pacific, including Japan, are the main locations of Udands.

Spodosols

A black, reddish-brown to dark-brown subsoil (spodic) horizon is the primary identifying characteristic of a Spodosol. It is often overlain by a gray to light-gray eluvial horizon. These distinctive and contrasting colors make Spodosols easily identifiable, although there are always exceptions. The simple explanation for this horizon sequence holds that, under cool, humid, or perhumid climates, organic acids from a litter layer leach amorphous mixtures of organic matter and aluminum with or without iron from the eluvial horizon and deposit them in the illuvial spodic horizon. Most Spodosols have formed under such conditions and thus are common in the northern latitudes where most of these soils are to be found. However, Spodosols vary widely depending on climate and other soil-forming factors. In some, the gray (albic horizon) may be absent; in others, it may be more than 2 m thick over a spodic horizon; furthermore, in some the spodic horizon may be cemented and is then called 'ortstein.'

Most Spodosols have few silicate clays. The particle-size class is mostly sandy, sandy-skeletal, coarse-loamy, loamy-skeletal, or coarse-silty. In hot, humid intertropical areas and other warm humid areas, they have for the most part formed in quartz-rich sands that have a fluctuating water table. Spodosols may form rather quickly (several hundred to several thousand years), again depending on climate and other soil-formation factors.

Many Spodosols are forested. They are generally used for forestry, cultivated crops, and pasture. Spodosols are naturally infertile, but with fertilization, commonly additions of large quantities of lime, nitrogen, and phosphorus, they are quite productive. They tie up considerable amounts of phosphorus, at times returning only approximately 0.45 kg of phosphorus for every 2.7 kg applied. In some Spodosols there may be a deficiency in heavy metals such as selenium or cobalt in forage used for ruminants. There are four suborders of Spodosols:

1. Aquods have an aquic SMR or the climate is extremely humid. They have a spodic horizon, with very high organic matter content, and are generally found in depressions or where the soils have an impermeable subsoil;

2. Humods have a spodic horizon enriched with organic matter. They are well drained and are generally found in cool climates, including high elevations in tropical mountains. On sandy coastal plains of the tropical littoral, such soils frequently occur with Histosols, the latter in depressions;

3. Cryods have a cryic STR;

4. Orthods have a spodic horizon enriched with both iron and organic matter. These are the most extensive Spodosols (2.7 million km^2) and are dominant in cool temperate areas of the northern latitudes.

Oxisols

Oxisols are reddish, yellowish, or grayish soils. They are most common on the gently undulating surfaces of geologically old surfaces in tropical and subtropical regions where they occupy approximately 11.8 million km^2. Oxisol profiles are distinctive because of the lack of obvious horizons. Their surface horizons are usually somewhat darker in color than the subsoil, but the transition to subsoil features is gradual.

Oxisols consist mainly of quartz, kaolinite, oxides, and organic matter. Both the structure and 'feel' of Oxisols are deceptive. Upon first examination they appear structureless and feel like a loamy particle-size class. While some may be loamy or even coarser, many are extremely clayey, but that clay is aggregated in a strong grade of fine and very fine granular structure. To obtain a true feel of the fine texture, a wet sample must be worked for several minutes in the hand to break down the sandy-textured, granular structure. The strong granular structure apparently causes most Oxisols to have a much more rapid permeability than would be predicted by the particle-size distribution class. Although compaction and reduction in permeability can be caused by cultivation, they are extremely resistant to compaction and so free drainage can take place soon after rain without puddling.

Oxisols are present in every soil-moisture regime from aridic to perudic and aquic. Natural vegetation ranges from tropical rainforests to savannas. Although many Oxisols are extremely infertile, there are some that have small but adequate supplies of nutrients and are immediately productive when cultivated. The reserves of plant nutrients even in the most fertile Oxisols are not great, and to sustain yields, fertilizers and lime are needed after only a few years of cultivation. In most of the Oxisols, fertilizers are needed for the first crop unless enough fertility for one or two crops is available from the ash derived from the burning of natural vegetation. Phosphorus is generally the most restrictive plant nutrient, mainly because of the tendency for the sesquioxide-rich clays to fix large amounts of fertilizer phosphorus. However, once this capacity to fix phosphorus is satisfied by initial applications, phosphorus fixation is no longer a problem.

Road-building and other engineering practices are relatively easy on most Oxisols because of the physical stability of the material. Soil organic carbon is generally much higher than indicated by the color, but, unlike most other soils, much of the carbon is inert and does not contribute to the nutrient-holding capacity or to the physical properties.

The most extensive areas of Oxisols are on the interior plateaus of South America, the lower portion of the Amazon basin, significant portions of the Central African basin, and important areas in Asia, northern Australia, and several tropical islands of the Pacific.

The suborders of the Oxisols are based on the soil moisture regimes:

1. Aquox have an aquic SMR and so are the very wet Oxisols. They occur in depressions associated with the better-drained upland Oxisols;
2. Torrox have an aridic SMR. Such soils are reported in the literature but no contiguous areas have been mapped;
3. Ustox have an ustic SMR. They are the second most extensive kind of Oxisols and occupy large areas in the Brazilian Shield;
4. Perox have a perudic SMR. They represent the most typical Oxisols of the very wet zones of Amazon, Central Zaire, and South-East Asia;
5. Udox have an udic SMR. They are the most extensive of the Oxisols, with large areas in the Amazon area, Central Africa, Borneo, and some of the Pacific Islands.

Vertisols

Vertisols are clayey soils, which have deep, wide cracks on some occasions during the year and slickensides within 100 cm of the soil surface. They shrink when dry and swell when moistened. Vertisols make up a relatively homogenous order of soils because of the amount and kind of clay that is common to them.

In many countries where Vertisols are common, they are known by their local names. For example, Gilgai soils (Australia), Adobe (Philippines), Sha Chiang (China), Black Cotton Soils (India), Smolnitza (Bulgaria), Tirs (Morocco), Makande (Malawi), Vleigrond (South Africa), and Sonsosuite (Nicaragua).

Vertisols generally have gentle slopes, although a few slope strongly. They develop commonly in large pedons and polypedons. The natural vegetation is a function of the soil climate. Most Vertisols are well suited for mechanized farming if there is plenty of rainfall or irrigation water and if suitable management practices are followed. Large areas of Vertisols in the world, however, are not farmed, because their cultivation would require too much energy, especially where traditional, low-input methods are used. This is one of the basic limitations of using Vertisols. Irrigation also presents special problems due to their low saturated hydraulic conductivity. Bypass flow in open cracks is the common situation. Because of their low permeability, irrigation of these soils may result in waterlogging and a buildup of salinity unless adequate artificial drainage is provided. A drainage system designed for Alfisols or Ultisols may be totally inadequate for Vertisols.

Six suborders of Vertisols are recognized based on the SMR or STR:

1. Aquerts: wet Vertisols with an aquic SMR, important locally;
2. Cryerts: cold Vertisols with a cryic or pergelic STR; little information is available;
3. Torrerts: Vertisols with an aridic SMR, present in areas where the SMR is transitional between aridic and ustic, although in most years it is aridic. Large areas are found in Sudan, the Near East, and Central Asia;
4. Xererts: Vertisols with a xeric SMR; the typical reddish-brown Vertisols of the Mediterranean areas, found in Jordan, Turkey, and Tunisia;
5. Usterts: Vertisols with an ustic SMR, geographically the most widespread, occupying approximately 1.8 million km^2; the dominant Vertisols in Africa, India, and Australia;
6. Uderts: Vertisols with an udic SMR, large areas being found in Bengal, some of the Caribbean Islands, Eastern Europe, and Argentina.

Aridisols

Aridisols, as their name implies, are soils that do not have water available to mesophytic plants for long

periods. During most of the time when the soil is warm enough for plants to grow, soil water is held at potentials less than permanent wilting point or if it is salty, or both. There is no period of 90 consecutive days when moisture is continuously available for plant growth.

The concept of Aridisols is based on the low availability of soil moisture for sustained plant performance. In areas bordering deserts, the absolute precipitation may be high but due to runoff or a very low storage capacity of the soil, or both, the actual soil-moisture regime is aridic. In these areas, tree vegetation may exist. Deep-rooted acacias are frequently the dominant vegetation but, if the trees are removed, the soil cannot support general farm crops. In general there is a 70% probability (7 of 10 years) that there will be a crop failure.

Many Aridisols, due to an extreme imbalance between evapotranspiration and precipitation, are similar to incipient evaporites. The dominant process is one of accumulation and concentration of salts. The high salt concentration is an adverse attribute in these soils and is the second most important constraint to the use of the soil. Many soluble salts may be eliminated or changed in concentration through irrigation. In Aridisols, however, the availability of good-quality irrigation water is a fundamental problem; secondly, together with irrigation, a mechanism for evacuation of salts must be provided or a rapid buildup of salinity and/or alkalinity will occur. Thirdly, irrigation and drainage systems must be well maintained to prevent the soils from reverting to their original state.

The classification of Aridisols includes these constraints or performance-restrictive qualities, at a high categoric level. Some Aridisols are also situated on geologic evaporites. It is often difficult to enter these substratum conditions into a classification system, but care must be taken to evaluate these deep-seated salt accumulations, particularly in irrigation projects. Some Aridisols also present inherited features, which may be attributed to earlier wetter or drier paleoclimatic conditions. These attributes, and specifically an argillic horizon, are also considered, as they are important in the use and management of soils.

The suborders reflect the results of dominant soil-forming processes. Unlike many other soils, the redistribution of soluble materials and their accumulation in some layers in the soil is a dominant process in Aridisols. The products of this process not only give special attributes that distinguish the soils, but also present constraints to the use of the soil. Four of the six suborders are defined on the composition and accumulation of the soluble fraction. Weathering and clay translocation also take place in Aridisols, although the expression of the products is not as vivid. The fourth and sixth suborders reflect these processes. The six suborders are:

1. Salids: characterized by accumulation of salts more soluble than gypsum; the typical soils of the playas or desert depressions, or closed basins;
2. Durids: characterized by accumulations of silica, but infrequently found; associated with soils formed from volcanic materials. They have been reported in the western USA but not in other parts of the world;
3. Gypsids: characterized by an accumulation of gypsum; extensive in the Near East, especially in Syria, Iraq, Saudi Arabia, and Iran. They have a total area of approximately 1.1 million km^2;
4. Argids: characterized by an accumulation of clay by translocation, the Argids are present in areas adjoining soils with an ustic SMR. They have not been mapped in large areas of North Africa or the Near East and are currently reported only in the USA. Correct classification of Argids is a problem;
5. Calcids: characterized by an accumulation of carbonates, the soils have a calcic, petrocalcic, or a hypercalcic horizon. There are extensive areas of these soils in the major deserts of the world, the total area exceeding 10.2 million km^2;
6. Cambids: characterized by a transformation of material, these soils are the most extensive of the Aridisols and occupy approximately 13.3 million km^2.
7. A special suborder, cryids, is provided for those aridisols with low soil temperature.

Ultisols

Ultisols are similar to Alfisols in having a subhorizon of clay accumulation but have few bases, especially at depth. Most Ultisols are acid, although some may have a high pH in the surface horizons owing to aerosolic additions of carbonate dust. The ideal Ultisol has a subsurface horizon of clay enrichment due to clay translocation from the surface horizons. However, in areas bordering the deserts, wind-blown sand may bury a former Oxisol, such as in western Zambia and Zimbabwe or in Mali and Niger. These soils have highly weathered subsoil with very low clay activity but, owing to the wind-blown material on the surface, show an increase in clay content with depth. There is little evidence of clay translocation in these soils. Similar situations prevail in the soils of old geomorphic surfaces where the subsoil is heavier-textured and with low clay activity but with little evidence of clay translocation. If the surface horizons have more than 40% clay, for practical purposes, these soils that change in texture with depth are considered as Ultisols. If there is less than 40% clay, they are classified as Oxisols.

The subsurface horizon of clay accumulation is an important feature in the semiarid environments. In general, the loamy soils have low water-holding capacity and so any horizon with more clay enhances the water-holding capacity. Moisture management is a most important feature of the semiarid Ultisols and even in some Ultisols in the humid areas. The additional acid nature of the soils exaggerates the moisture stress in them. Deep liming is essential to enable a proliferation of roots so that a larger volume of soil is exploited for moisture and roots.

The suborders of the Ultisols are based on the SMRs except for the Humults, which are characterized by an accumulation of organic matter:

1. Aquults are wet Ultisols, with an aquic SMR;
2. Humults have accumulations of organic matter and usually occur in areas with an udic SMR and cool temperatures;
3. Udults have an udic SMR;
4. Ustults have an ustic SMR;
5. Xerults have a xeric SMR.

Mollisols

Generally in Soil Taxonomy, it is the presence or absence of subsurface horizons and their characteristics that are used to separate the orders. In Mollisols, however, it is the presence of a thick, dark, humus-rich surface horizon (mollic epipedon) that is the key to placement. In the development of Soil Taxonomy, this surface horizon is the only common characteristic that can be found to tie together the grassland soils of North America, Europe, Asia, and South America. It is an important separation, because these soils are most easily cultivated, generally without irrigation. Thus most Mollisols have had grass vegetation at some time, and melanization – the process of darkening of the soil by organic matter additions – is probably the most important process in the formation of a Mollisol. It is the addition, decomposition, and accumulation of relatively large amounts of organic matter in the soil profile, with the presence of calcium, that forms the central concept of Mollisols. Mollisols have a variety of subsurface horizons and/or diagnostic characteristics, or horizons may be entirely absent.

To a large extent, Mollisols are the breadbasket of the world – the prairies in the USA, the steppes of Russia, and the pampas of Argentina. Most Mollisols are cultivated; in fact there are only limited areas in the world where they have not been cultivated. Mollisols may initially be farmed with no additions of fertilizers. However, to sustain the high yields of corn, soybeans, sorghum, and small grains of today, fertilizers must be used. Soil temperature and moisture are principally used to separate all but two (Albolls and Rendolls) of the seven suborders of Mollisols:

1. Albolls have a bleached subsurface horizon called the albic horizon and are usually wet;
2. Aquolls are wet Mollisols and may have a histic epipedon;
3. Rendolls are shallow and stony soils formed on limestone and have an udic SMR;
4. Xerolls have a xeric SMR;
5. Cryolls have a frigid, cryic, or pergelic STR;
6. Ustolls have an ustic or an aridic SMR;
7. Udolls have an udic SMR.

Alfisols

Most Alfisols were or are forested, with moderate to high base saturation; most formed under deciduous forest. Typically they have a light-colored surface layer over a horizon of silicate clay accumulation (argillic). The cooler Alfisols tend to form a belt between the grassland Mollisols and the Spodosols of the more humid climates. Where temperatures are warmer, they form a belt between the Aridisols and the older Ultisols and Oxisols. Along with Mollisols, Alfisols account for a major portion of soils that are used to grow crops in the world. They are found generally in climates favorable to crop production; in warm moist climates, they are used to grow many crops. They generally contain adequate plant nutrients, but like Mollisols they must be fertilized to obtain high yields. Crop production may become more difficult when these soils are eroded down to the argillic horizon. The higher clay content of the argillic horizon may impede root, water, and air movement.

All five suborders of Alfisols are defined by the SMR:

1. Aqualfs are wet Alfisols;
2. Cryalfs have a cryic STR, or, if the SMR is not xeric, have a frigid STR;
3. Ustalfs have an ustic SMR;
4. Xeralfs have a xeric SMR;
5. Udalfs have an udic SMR.

Inceptisols

The Latin word ‘inceptum’ means beginning and the central concept of Inceptisols is that of soils in the early stages of soil formation. The initial stage of soil formation is exemplified by several attributes, which are the result of the presence or absence of certain processes.

Soil formation on rocks consists of weathering of the rock which is essentially a geochemical process

accompanied by soil-forming processes acting on the weathered products. In cool, humid climates, the soil-forming process may be the accumulation of organic matter to give rise to a mollic or umbric epipedon. In warmer climates, cambic horizon formation takes place; this is expressed by clay formation or release of iron to form a 'color B' horizon. On steep slopes, even if moisture and temperature conditions are conducive to form an argillic horizon, the slow soil loss through erosion retards any of the other horizons from being expressed. Such soils may have just an ochric epipedon and a weak cambic horizon.

Inceptisols are more prevalent on sediments, and a major morphological change is the removal of the original stratification, which takes place through bio- and pedoturbation. Structure and color development are common marks of cambic horizons. If a permanent or a fluctuating water table is present, oxidation–reduction processes leave their marks and are frequently sufficient evidence for a cambic horizon. In extremely wet environments (perudic soil moisture regime) and even on steep slopes, or in soils with a fluctuating water table, weathering of primary minerals may release sufficient iron, which percolates through the soil and accumulates in a subsurface layer to form a thin iron pan or placic horizon. The placic horizon is sufficient evidence to consider the soils Inceptisols.

Some coastal sediments contain pyrite which, when exposed to oxidizing conditions, oxidizes to jarosite, releasing sulfuric acid. The presence of the straw-yellow jarosite aggregates and extreme acidic conditions of the soil is sufficient to place the soil in the order of Inceptisols.

The Inceptisols consequently comprise a wide array of soils, which range not only in properties but also in behavior. The lower categories in the Inceptisols attempt to cluster more homogeneous soils.

Five suborders of Inceptisols are recognized:

1. Aquepts are wet Inceptisols;
2. Plaggepts have a man-made surface horizon thicker than 50 cm called the plaggen epipedon (these soils are not extensive enough to be shown on the map);
3. Ustepts have an ustic SMR;
4. Xerepts have a xeric SMR;
5. Udepts have an udic SMR.

Entisols

The Entisols show little or no evidence of soil formation. They are most extensive on recent alluvial plains and valleys or on steep slopes where erosion is rapid. The rate of soil formation is reduced for several reasons. Generally time has not elapsed since deposition of the material for soil-forming processes to act. In some of these soils, peraquic conditions prevail where the soil is saturated with water during the whole year. The soil is permanently reduced, preventing cambic horizon formation. On steep slopes, rapid erosion results in shallow soils where weathered parent materials rest on hard rock. Rates of soil loss are much greater than soil formation and so cambic horizons do not form. Entisols may also occur on older deposits where, for example, the material is formed from quartzitic sands. There are no primary minerals in the deposit to weather and form clay or liberate iron. In fluvial deposits, Entisols show marked stratification. This is frequently evidence for recent deposits.

Five suborders are recognized in the Entisols:

1. Aquents are wet Entisols;
2. Arents have been subject to deep plowing;
3. Psamments are sandy Entisols;
4. Fluvents are recent alluvial soils showing stratification;
5. Orthents are shallow soils on steep slopes.

In many countries Entisols, on flat alluvial plains or riverine terraces, are widely used for annual crops. Terrain conditions are frequently suitable for low-input agriculture. The major civilizations of the world developed on Entisols and Inceptisols of the larger river terraces owing to good farming conditions and easy navigation.

Summary

A wide diversity of soils exists and each has its own potential and limitations. The performance-related attributes result from the properties and the prevailing environmental conditions. There are large land areas where it is too cold, or too dry, or the soils are on steep slopes and the potential for human use is limited. For most agricultural uses, only about 9% of the global land surface is relatively constraint-free. Understanding the soil and its relation to the rest of the environment is important for its sustainable use. Detailed soil surveys provide such information, but most countries of the world lack access to it. Less information is available on the state of soil resources. Assessment and monitoring of land conditions are essential to the judicious use of soil.

Further Reading

Ahrens RJ and Arnold RW (2000) Soil taxonomy. In: Sumner ME (ed.) *Handbook of Soil Science*, pp. 117–135. Boca Raton, FL: CRC Press.

Arnold RW (1990) Soil taxonomy, a tool of soil survey. In: Rozanov BG (ed.) *Soil Classification*, pp. 94–111. USSR State Committee for Environmental Protection. Moscow, Russia: Publishing Centre for International Projects.

Arnold RW, Ahrens RJ, and Engel RJ (1997) Trends in soil taxonomy – a shared heritage. *Communications of the Austrian Soil Science Society* 55: 167–170.

Bartelli LJ (1978) Technical classification system for soil survey interpretation. In: Brady NC (ed.) *Advances in Agronomy* 30: 247–289.

Buol SW, Hole FD, and McCracken RJ (1980) *Soil Genesis and Classification*. Ames, IA: Iowa State University Press.

Deckers JA, Nachtergaele FO, and Spaargaren OC (eds) (1998) *World Reference Base for Soil Resources: Introduction*. International Society of Soil Science (ISSS), International Soil Reference and Information Centre (ISRIC) and Food and Agriculture Organization of the United Nations (FAO). Leuven, Belgium: FAO.

Orvedal AC and Edwards MJ (1941) General principles of technical grouping of soils. *Soil Science Society of America Proceedings* 6: 386–391.

Soil Survey Staff (1975) *Soil Taxonomy: A Basic System of Soil Classification for Making and Interpreting Soil Surveys*. US Department of Agriculture Handbook 436. Washington, DC: US Government Printing Office.

Soil Survey Staff (1999) *Soil Taxonomy: A Basic System of Soil Classification for Making and Interpreting Soil Surveys*, 2nd edn. US Department of Agriculture Handbook 436. Washington, DC: US Government Printing Office.

Sumner ME (ed.) *Handbook of Soil Science*. Boca Raton, FL: CRC Press.

Wilding LP, Smeck NE, and Hall GF (eds) (1983) *Pedogenesis and Soil Taxonomy*. I. *Concepts and Interactions*. Amsterdam, the Netherlands: Elsevier.

Wilding LP, Smeck NE, and Hall GF (eds) (1983) *Pedogenesis and Soil Taxonomy*. II. *The Soil Orders*. Amsterdam, the Netherlands: Elsevier.

Z

ZERO-CHARGE POINTS

J Chorover, University of Arizona, Tucson, AZ, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Many soil chemical reactions occur at the interface between solid particles and the aqueous phase. Fundamental to the reactivity of this interface is the surface charge on soil particles, which affects a range of chemical phenomena including adsorption of charged solutes, the nature and extent of the electrical double layer on soil particles, kinetics of colloid aggregation, and mineral dissolution rates. Cation and anion exchange capacities, which influence the retention of plant-available nutrients, are directly affected by the nature of soil surface charge and its response to changes in solution of chemical parameters such as pH, ionic strength, and ion composition.

Charge properties have traditionally been used to characterize soils as 'permanent' or 'variable' charge media. These characterizations reflect the relative predominance of one or more soil solid-phase constituents. Whereas the surface chemistry of permanent charge soils is governed by isomorphically substituted mineral particles (i.e., 2:1 layer-type clays such as smectite and vermiculite), variable-charge soils are dominated by amphoteric hydroxylated surfaces (i.e., metal oxides, hydroxides or oxyhydroxides, and organic matter). As a result of the ubiquity of hydroxylated solids in soil environments, all soils comprise some portion of their total surface charge that is conditional upon aqueous chemical conditions.

The law of surface-charge balance states that, irrespective of the source of particle surface charge, it must be balanced by net accumulation of ions in the surrounding diffuse swarm. The balance of surface charge and relationships among components may be used to constrain the application of molecular models to the particle–water interface and to verify internal consistency of adsorption measurements. Particularly useful in this respect are points of zero charge, which define the physicochemical conditions wherein one or several components of surface charge sum to zero.

Components of Surface Charge

Electric charge develops on soil particle surfaces as a result of: (1) structural disorder or isomorphic substitutions among ions of differing valence within soil minerals, and (2) reactions of surface functional groups with ionic species in aqueous solution. These mechanisms of charge development are incorporated into four 'components of surface charge' that together contribute to the *total particle surface-charge density*, σ_P. Surface-charge density is conventionally expressed in coulombs per square meter, consistent with the International System of Units (SI). For complex solids, including soils, charge is likely to be distributed unequally among the many phases present, owing to differences in structural charge, surface site density, and specific surface area. Furthermore, measurement of total surface area alone is nontrivial and method-dependent. For these reasons, soil charge is often measured on a mass basis and expressed as moles of charge per kilogram. Values presented in each of these sets of units differ by the factor F/a_s, where F is the Faraday constant and a_s is specific surface area.

The *net structural surface-charge density*, σ_0, is created by isomorphic substitutions in soil minerals. These substitutions occur in both primary and secondary minerals, but they produce significant surface charge only in the 2:1 layer-type aluminosilicates (e.g., Figure 1). The *net proton surface-charge density*, σ_H, results from the difference between the moles of protons and the moles of hydroxide ions complexed by surface functional groups (Figure 2). The *net inner-sphere complex surface-charge density*, σ_{IS}, results from the net total charge of ions, other than H^+ and OH^-, which are bound into inner-sphere surface coordination. Inner-sphere complexation involves the formation of one or more direct bonds between the adsorbate molecule and adsorbent surface functional group, with no water molecules

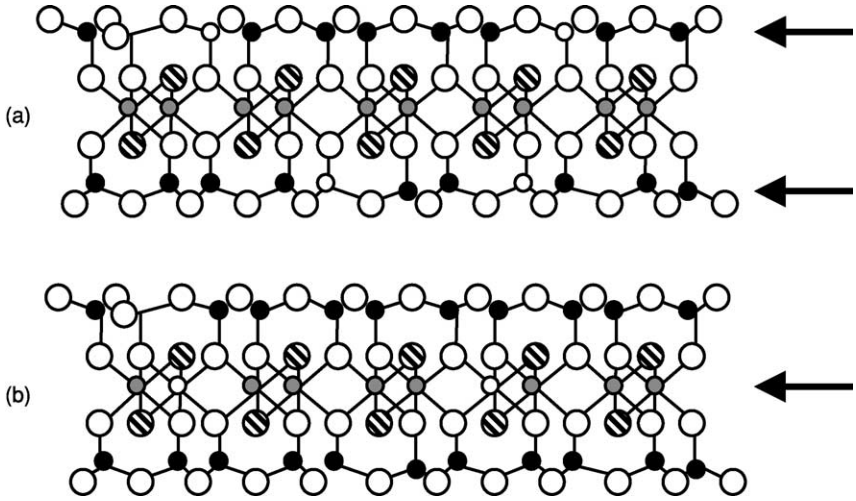


Figure 1 Isomorphic substitution of ions of differing valence gives rise to structural charge density (σ_0) that is independent of solution chemistry. Model structures of vermiculite (a) and smectite (b) show that substitutions (small white spheres) can give rise to charge in the (a) tetrahedral (Al^{3+} for Si^{4+} substitution) or (b) octahedral (e.g., Mg^{2+} for Al^{3+} substitution) sheets of 2:1 layer-type silicates. The location of substitution, particularly in regard to proximity to the surface, affects site reactivity. Large white spheres, O; black spheres, Si; gray spheres, Al; hatched spheres, OH.

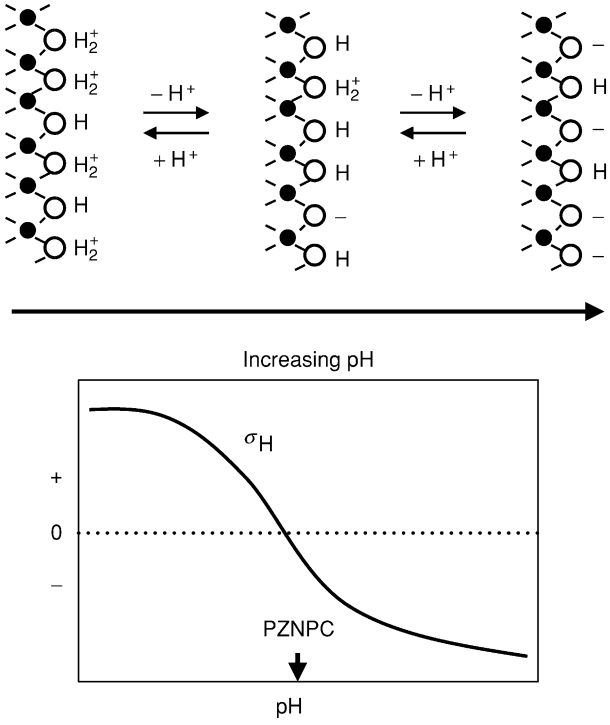


Figure 2 Development of proton surface-charge density (σ_H) derives from the net adsorption of protons on particle surfaces and reflects the Brönsted acid–base chemistry of surface functional groups. PZNPC, point of zero net proton charge.

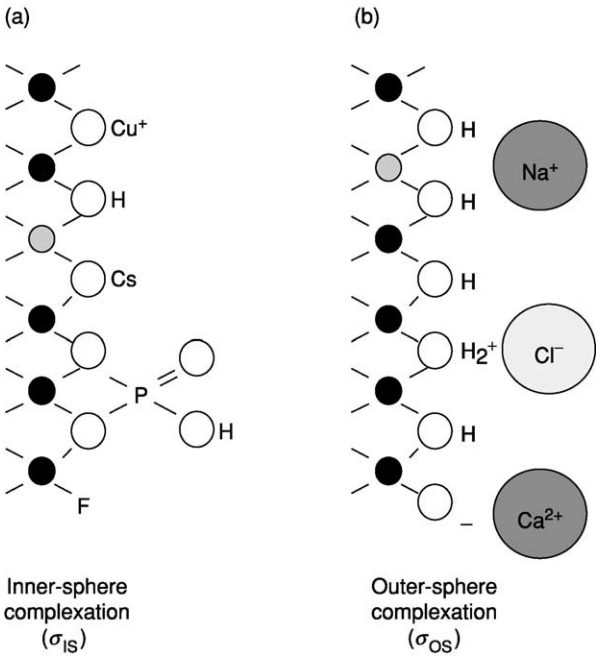


Figure 3 Surface coordination of ions in stable complexes results in the development of (a) inner-sphere complex surface-charge density (σ_{IS}) and (b) outer-sphere complex surface charge density (σ_{OS}), both of which contribute to total particle surface charge. Isomorphic substitutions, are shown schematically as gray spheres.

interposed (Figure 3a). The *net outer-sphere complex surface-charge density*, σ_{OS}, results from the net total charge of ions, other than H^+ and OH^-, which are bound into outer-sphere surface coordination. Outer-sphere complexes involve hydration water interposed between the adsorbate and the adsorbent surface (Figure 3b). The net total particle surface charge is the sum of these four components:

$$\sigma_P = \sigma_0 + \sigma_H + \sigma_{IS} + \sigma_{OS} \quad [1]$$

The right side of eqn [1] indicates contributions to particle surface charge from both permanent (σ_0) and conditional (σ_H, σ_{IS}, σ_{OS}) components; the value of σ_0 is dependent only upon the composition and structure of the solid phase, while the values of the remaining components depend upon the nature of both the solid and the solution phases. The last two terms on the right side of eqn [1] together define adsorbed ions that are immobilized in surface complexes, also known as the 'Stern-layer charge', σ_S.

Surface Charge Balance

Aqueous particle suspensions are electrically neutral. Therefore, if σ_P is nonzero, it must be balanced by a diffuse swarm of ion charge equal in magnitude but opposite in sign to σ_P. These ions differ from those bound into the Stern layer because their much greater mobility relative to the adsorbent surface is analogous to diffusive motions of free ions in aqueous solution. The net charge of such ions defines the *diffuse layer surface-charge density*, σ_d. The charge balance equation for soil particles may then be written:

$$\sigma_P = -\sigma_d \qquad [2]$$

and substituting the left side of eqn [1] into eqn [2] gives:

$$\sigma_0 + \sigma_H + \sigma_{IS} + \sigma_{OS} + \sigma_d = 0 \qquad [3]$$

The terms σ_{IS}, σ_{OS}, and σ_d correspond to different molecular mechanisms of ion adsorption that are difficult to distinguish on the basis of quantitative macroscopic measurements. These three terms in eqn [3] together give the *adsorbed ion charge density*, Δq, which is the net charge of ions, other than H^+ and OH^-, adsorbed at the soil–water interface (**Figure 4**):

$$\Delta q = \sigma_{IS} + \sigma_{OS} + \sigma_d = q_+ - q_- \qquad [4]$$

The value of Δq in eqn [4] can be determined by measuring the difference between *surface excess* of cation charge (q_+ corresponding to all cations other than H^+) and *surface excess* of anion charge (q_- corresponding to all anions other than OH^-). The charge balance may then be rewritten as:

$$\sigma_0 + \sigma_H + \Delta_q = 0 \qquad [5]$$

Points of Zero Charge

Points of zero charge are defined most commonly as pH values for which one or more of the surface-charge components is equal to zero at a given temperature, pressure, and aqueous solution composition. The

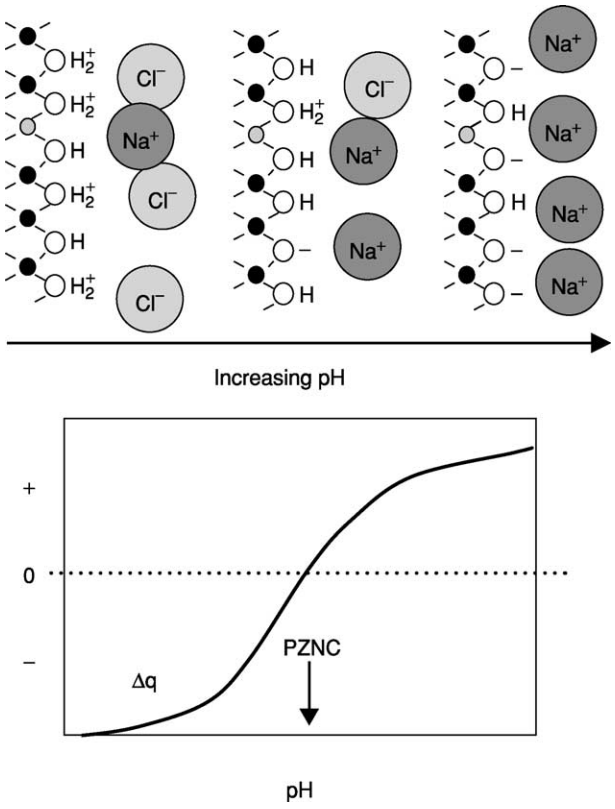


Figure 4 The net charge deriving from adsorption of ions (other than H^+ and OH^-) in surface complexes ($\sigma_{IS} + \sigma_{OS}$) plus the diffuse ion swarm (σ_d) constitute the net adsorbed ion charge density (Δq). PZNC, point of zero net charge.

emphasis on pH derives from the wide and dynamic range of proton concentrations in soil solutions, and the fact that complexation of H^+ and OH^- at soil particle surfaces strongly affects the adsorption of other cations and anions (e.g., see **Figure 4** and eqn [5]). Although pH will be used to illustrate points of zero charge here, it is important to note that one can also define them in terms of aqueous concentrations of other surface complexing ions, such as $H_2PO_4^-$ or Pb^{2+} (**Figure 3a**), which can affect the value of σ_P directly. Three principal points of zero charge are named conventionally as: (1) the point of zero charge (PZC), (2) the point of zero net proton charge (PZNPC), and (3) the point of zero net charge (PZNC).

Point of Zero Charge

The PZC ($\sigma_P = 0$) is defined as the pH value at which the net total particle surface charge, σ_P, is equal to zero. According to eqn [2], σ_d also vanishes at the PZC, so that all adsorbed ions (other than H^+ and OH^-) must be bound into surface complexes. Traditionally, the PZC has been assessed by measuring

the pH value at which soil particles exhibit zero electrophoretic mobility (i.e., the isoelectric point or IEP), assuming that a negligible portion of the diffuse ion swarm is advected during the electrophoresis measurement. However, this assumption has never been verified. Furthermore, model calculations suggest that surface charge heterogeneity or patchiness can also result in a nonzero electrophoretic mobility for a particle whose overall charge is equal to zero. This raises additional questions regarding the validity of equating PZC and IEP values. Values of PZC for soil minerals derived dominantly from electrophoresis measurements are presented in Table 1.

According to the DLVO (Derjuagin–Landau–Verwey–Overbeek) theory, a suspension of colloidal particles should undergo rapid flocculation when net charge in the diffuse ion swarm (σ_d) is reduced to zero. Therefore, the PZC may also be determined by measuring the pH where a colloidal suspension of particles undergoes rapid coagulation (flocculation). However, the influence of surface-charge heterogeneity probably affects collision efficiency as well. New

Table 1 Points of zero charge for selected soil mineral constituents

Solid	*Chemical formula*	*PZNPC*	*PZNC*	*PZC*
Albite	$NaAlSi_3O_8$	3.9	–	–
Allophane	$SiAl_2O_5 \cdot 0.5H_2O$	8.2	7.9	10
Alumina	γ-Al_2O_3	7.0	6.7	–
Birnessite	$(Na,Ca)Mn_7O_{14} \cdot 2.8H_2O$	2.2	1.9	1.7
Boehmite	γ-AlOOH	–	–	8.6
Calcite	$CaCO_3$	–	–	8.5
Corundum	α-Al_2O_3	–	–	9.1
Dolomite	$CaMg(CO_3)_2$	–	–	8.0
Forsterite	Mg_2SiO_4	–	–	4.5
Gibbsite	γ-$Al(OH)_3$	8.5	8.5	–
Goethite	α-FeOOH	7.3	7.5	–
Hematite	α-Fe_2O_3	8.2	8.3	8.2
Illite	$K_{1.75}Si_7Al(Al_3Fe^{(III)}_{0.25}Mg_{0.75})O_{20}(OH)_4$	2.7	–	–
Imogolite	$SiAl_2O_3(OH)_4$	6.8	8.4	11.5
Kaolinite	$Si_4Al_4O_{10}(OH)_8$	5.0	3.6	4.7
Magnetite	Fe_3O_4	6.8	–	–
Quartz	SiO_2	–	2.5	2.0
Rutile	TiO_2	6.0	–	5.9
Silica	SiO_2	2.0	2.4	2.4
Smectite	$Na_2Si_{7.5}Al_{0.5}(Al_{3.5}Mg_{0.5})O_{20}(OH)_4$	8.5	–	$<$2.3

PZNPC, point of zero net proton charge; PZNC, point of zero net charge; PZC, point of zero charge.

Sources of data: Albite: Mukhopadhyay B and Walther JV (2001) Acid–base chemistry of albite surfaces in aqueous solutions at standard temperature and pressure. *Chemical Geology* 174: 415–443. Allophane: Su C, Harsh JB, and Bertsch PM (1992) Sodium and chloride sorption by imogolite and allophanes. *Clays and Clay Minerals* 40: 280–286; Su C and Harsh JB (1993) The electrophoretic mobility of imogolite and allophane in the presence of inorganic anions and citrate. *Clays and Clay Minerals* 41: 461–471. Alumina: Goyne KW, Zimmerman A, Newalkar BL, Komarneni S, Brantley SL, and Chorover J (2002) Surface charge of variable porosity Al_2O_3 (s) and SiO_2 (s) adsorbents. *Journal of Porous Materials* 9: 243–256. Birnessite: Sposito G (1984) *The Surface Chemistry of Soils*. New York: Oxford University Press. Boehmite: Ermakova L, Sidorova M, Bogdanova N, and Klebanov A (2001) Electrokinetic and adsorption characteristics of (hydr)oxides and oxide nanostructures in 1:1 electrolytes. *Colloids and Surfaces* 192: 337–348. Calcite: Sposito G (1984) *The Surface Chemistry of Soils*. New York: Oxford University Press. Corundum: Sposito G (1984) *The Surface Chemistry of Soils*. New York: Oxford University Press. Dolomite: Pokrovski OS, Schott J, and Thomas F (1999) Dolomite surface speciation and reactivity in aquatic systems. *Geochimica Cosmochimica Acta* 63: 3133–3143. Forsterite: Pokrovsky OS and Schott J (2000) Forsterite surface composition in aqueous solutions: a combined potentiometric, electrokinetic, and spectroscopic approach. *Geochimica Cosmochimica Acta* 64: 3299–3312. Gibbsite: Sposito G (1989) *The Chemistry of Soils*. New York: Oxford University Press. Goethite: Sigg L and Stumm W (1981) The interaction of anions and weak acids with the hydrous goethite (α-FeOOH) surface. *Colloids and Surfaces* 2: 101–117; Sposito G (1989) *The Chemistry of Soils*. New York: Oxford University Press. Hematite: Chibowski S and Janusz W (2002) Specific adsorption of Zn(II) and Cd(II) ions at the α-Fe_2O_3/electrolyte interface structure of the electrical double layer. *Applied Surface Science* 7853: 1–13. Illite: Sinisyn VA, Aja SU, Kulik DA, and Wood SA (2000) Acid–base surface chemistry and sorption of some lanthanides on K^+-saturated Marblehead illite: 1. Results of an experimental investigation. *Geochimica Cosmochimica Acta* 64: 185–194. Imogolite: Su C, Harsh JB, and Bertsch PM (1992) Sodium and chloride sorption by imogolite and allophanes. *Clays and Clay Minerals* 40: 280–286; Su C and Harsh JB (1993) The electrophoretic mobility of imogolite and allophane in the presence of inorganic anions and citrate. *Clays and Clay Minerals* 41: 461–471. Kaolinite (KGa-1): Schroth BK and Sposito G (1997) Surface charge properties of kaolinite. *Clays and Clay Minerals* 45: 85–91; Sposito G (1984) *The Surface Chemistry of Soils*. New York: Oxford University Press. Magnetite: Regazzoni AE, Blesa MA, and Maroto AJG (1983) Interfacial properties of zirconium dioxide and magnetite in water. *Journal of Colloid and Interface Science* 91: 560–570. Quartz: Eggleston CM and Jordan G (1998) A new approach to pH of point of zero charge measurement: Crystal-face specificity by scanning force microscopy (SFM). *Geochimica Cosmochimica Acta* 62: 1919–1923. Rutile: Ermakova L, Sidorova M, Bogdanova N, and Klebanov A (2001) Electrokinetic and adsorption characteristics of (hydr)oxides and oxide nanostructures in 1:1 electrolytes. *Colloids and Surfaces* 192: 337–348. Silica: Ermakova L, Sidorova M, Bogdanova N, and Klebanov A (2001) Electrokinetic and adsorption characteristics of (hydr)oxides and oxide nanostructures in 1:1 electrolytes. *Colloids and Surfaces* 192: 337–348; Goyne KW, Zimmerman A, Newalkar BL, Komarneni S, Brantley SL, and Chorover J (2002) Surface charge of variable porosity Al_2O_3 (s) and SiO_2 (s) adsorbents. *Journal of Porous Materials* 9: 243–256. Smectite: Avena MJ and de Pauli CP (1998) Proton adsorption and electrokinetics of an Argentinian montmorillonite. *Journal of Colloid and Interface Science* 202: 195–204.

methods, involving the use of scanning force microscopy, appear promising in regard to elucidating patchwise heterogeneity of surface charge and providing direct force measurements of the PZC (e.g., as used to obtain the value of PZC for quartz in Table 1).

Point of Zero Net Proton Charge

The pH value at which the proton surface-charge density, σ_H, is zero defines the PZNPC. Proton surface-charge density is given by:

$$\sigma_H = q_H - q_{OH} \quad [6]$$

where q_H and q_{OH} are the surface excess (e.g., in moles of charge per kilogram) of H^+ and OH^-, respectively. Therefore, at the PZNPC the moles of surface-adsorbed H^+ and OH^- are balanced. Proton and hydroxide adsorption reactions can be represented by the mass action expressions:

$$SOH(s) + H^+(aq) = SOH_2^+(s) \quad [7a]$$

$$SOH(s) + OH^-(aq) = SO^-(s) + H_2O(l) \quad [7b]$$

where SOH is a neutral surface hydroxyl group on a soil particle, and *s*, *aq*, and *l* represent the solid, aqueous, and liquid phases, respectively. If proton adsorption–desorption is assumed to occur dominantly at hydroxylated sites (e.g., as in the case of pure metal oxides), then the PZNPC defines the condition where the density or concentration of protonated sites is equal to that of proton-dissociated sites [i.e., $(SOH_2^+) = (SO^-)$]. Thus, both protonated and proton-dissociated sites can coexist at the PZNPC.

In 'indifferent' electrolyte solutions (i.e., those comprising monovalent cations and anions that do not form stable complexes at particle surfaces), proton adsorption behavior, as depicted in Figure 2, is a function of the Brönsted acidity of surface functional groups. Soil particles exhibit a wide range in surface acidity, as indicated by the PZNPC values shown in Table 1. Low PZNPC values are typical of more strongly acidic soil constituents such as silica, quartz, Mn oxides, and natural organic matter, whereas higher values are observed for (hydr)oxides of Fe and Al. Table 1 shows PZNPC values of soil minerals suspended in indifferent electrolyte solutions. However, the presence of cations (anions) that form inner- or outer-sphere surface complexes (Figure 3a) will tend to decrease (increase) the PZNPC because of adsorptive competition with H^+ (OH^-).

Proton- and hydroxide-promoted dissolution of soil minerals is found to increase at pH values both below and above the PZNPC. Correlations are observed between dissolution rate constants and the absolute magnitude of σ_H, suggesting that metal–oxygen bonds at the particle surface are weakened as reactions in eqns [7a] and [7b] proceed to the right.

Point of Zero Net Charge

The first two terms in eqn [5] are often combined to provide a measure of *intrinsic surface-charge density*, σ_{in}, which derives from the crystal and surface structure of the adsorbent:

$$\sigma_{in} = \sigma_0 + \sigma_H \quad [8]$$

The PZNC is the pH value at which σ_{in} is equal to zero. The intrinsic surface charge is balanced by the adsorption of ions. Since σ_{in} results from both proton adsorption–desorption reactions (pH-dependent) and isomorphic substitutions, it comprises conditional and permanent components (Figure 4). Since $\sigma_{in} = -\Delta q$, the PZNC may be determined: (1) by measuring the pH where surface excess values of cation and anion charge are equal (i.e., $\Delta q = 0$), or (2) by measuring the pH where $\sigma_0 = -\sigma_H$.

Charge balance dictates that, as σ_H decreases with increasing pH (affected by mass action as indicated in eqn [7]), ion adsorption–desorption reactions must balance the change in intrinsic charge density. The dependence of σ_H on pH is shown schematically in Figure 2, and corresponding effects on Δq are illustrated in Figure 4. For a soil adsorbent that is devoid of structural charge (i.e., $\sigma_0 \approx 0$), such as a highly weathered Oxisol that contains a negligible quantity of 2:1 layer-type silicates, $\Delta q = -\sigma_H$, and the PZNPC equals the PZNC (Figure 5, top). This is also the case for pure oxides and hydroxides that are free of structural defects. The presence of negative structural charge ($\sigma_0 < 0$), as occurs because of isomorphic substitutions in many temperate-zone soils, results in PZNC being less than PZNPC, and the value of Δq (or σ_H) at the PZNPC (PZNC) is equal to $-\sigma_0$ (Figure 5, middle). Whereas positive structural charge ($\sigma_0 > 0$) is rare in soils, it is detectable from PZNPC being less than PZNC, and its magnitude is given by the value of Δq (or σ_H) at the PZNPC (PZNC) (Figure 5, bottom).

In addition to the effects of σ_0, the nature of variable-charge behavior, such as that depicted in Figure 5, is highly dependent on the Brönsted acidity of the soil particles. For example, the magnitude and pH range of greatest local slope (i.e., $\delta\sigma_H/\delta$pH or $\delta\Delta q/\delta$pH) is dependent on the identity and quantity of surface sites undergoing proton adsorption–desorption reactions. Whereas oxides and hydroxides of Fe and Al are weakly acidic, with PZNPC values near neutrality, more strongly acidic groups such as those residing on natural organic matter functional

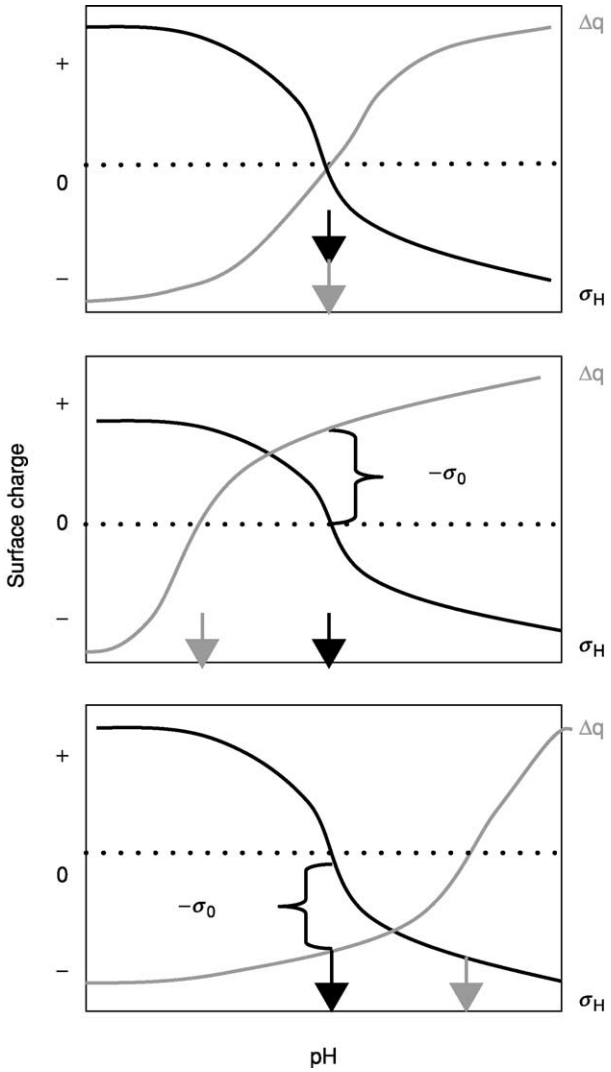


Figure 5 The inverse relation between σ_H and Δq is depicted schematically for (a) $\sigma_0 = 0$ (PZNC = PZNPC), (b) $\sigma_0 < 0$ (PZNC < PZNPC), and (c) $\sigma_0 > 0$ (PZNC > PZNPC). The PZNC is indicated by gray arrows and the PZNPC is indicated by black arrows. PZNC, point of zero net charge; PZNPC, point of zero net proton charge.

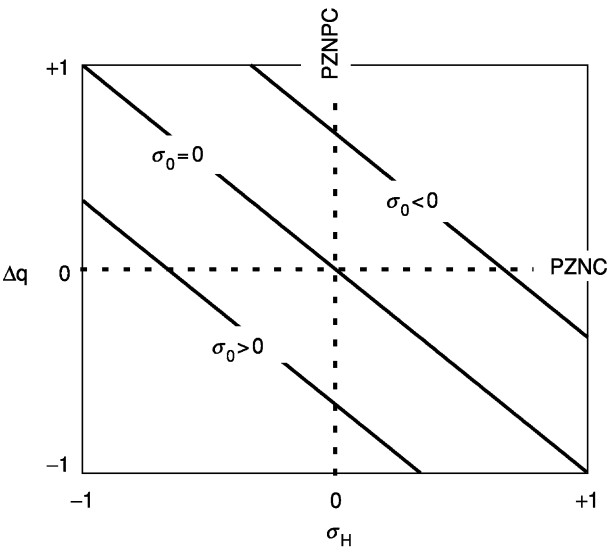


Figure 6 Surface-charge balance dictates that a plot of Δq versus σ_H must be linear, with a slope equal to -1 and x and y intercepts equal to $-\sigma_0$. PZNPC, point of zero net proton charge; PZNC, point of zero net charge.

groups (e.g., carboxylic acids), silica or quartz surfaces, and Mn oxide surfaces all tend to decrease PZNPC values for whole soils.

The surface-charge balance of eqn [5] is summarized in Figure 6; a plot of Δq versus σ_H should be linear with a slope equal to -1, and x and y intercepts equal to $-\sigma_0$. The x and y scales in Figure 6 are arbitrary and actual values depend on surface chemistry of the adsorbent and the units selected for plotting surface-charge data (e.g., moles of charge per kilogram or coulombs per square meter). Whereas curves such as those depicted in Figure 5 (with pH as independent variable) are strongly affected by ionic strength and electrolyte composition for a given adsorbent, the relation depicted in Figure 6 must be reproducible for a given soil, irrespective of changes in solution chemistry.

List of Technical Nomenclature

Adsorbate	Matter accumulating at the interface between adsorbent and the aqueous phase
Adsorbed ion charge density (Δq)	The net charge density created by adsorption of ions, other than H^+ and OH^-, into inner-sphere complexes, outer-sphere complexes, and the diffuse ion swarm
Adsorbent	The solid surface on which matter accumulates
Adsorption	The net accumulation of matter at the adsorbent surface
Diffuse layer surface-charge density (σ_d)	The net surface charge density created by adsorbed ions in the diffuse layer
Inner-sphere complex surface-charge density (σ_{IS})	The net charge density created by ions adsorbed via direct coordination to surface functional groups (i.e., with no water molecules interposed between adsorbate and surface)
Intrinsic surface-charge density (σ_{in})	The sum of structural and proton surface charge densities

Outer-sphere complex surface-charge density (σ_{OS})	The net charge density created by adsorption of hydrated ions that form stable complexes with adsorbent surface functional groups
Point of zero charge	The pH value at which particle surface charge is equal to zero
Point of zero net charge	The pH value at which adsorbed ion charge density is equal to zero
Point of zero net proton charge	The pH value at which proton surface charge density is equal to zero
Proton surface-charge density (σ_H)	Surface-charge density created by the formation of adsorption complexes involving proton or hydroxide ions as adsorbate; the difference between surface excess of protons and hydroxide ions ($\sigma_H = q_H - q_{OH}$)
Structural surface-charge density (σ_0)	Surface-charge density created by isomorphic substitutions among ions of differing valence in the crystal structure of an adsorbent

See also: **Cation Exchange**; **Dissolution Processes, Kinetics**; **pH**; **Sorption:** Metals

Further Reading

Everett DH (1972) Definitions, terminology and symbols in colloid and surface chemistry. *Pure and Applied Chemistry. Chimie pure et appliquée* 31: 578–638.

Sposito G (1998) On points of zero charge. *Environmental Science and Technology* 32: 2815–2819.

Stumm W and Wollast R (1990) Coordination chemistry of weathering: kinetics of the surface controlled dissolution of oxide minerals. *Reviews of Geophysics* 28: 53–69.

ZONE TILLAGE

J L Hatfield and A T Jeffries, National Soil Tilth Laboratory, Ames, IA, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Tillage is the act of disturbing the soil through some type of mechanical means. It has been practiced since the development of crude instruments such as a pointed stick to create a hole into which a seed was dropped. The largest advance in the progress of tillage implements has been the invention of the moldboard plow, which inverts the soil and allows rapid cultivation. Since that invention, there have been continual advances in the development of tillage equipment for the purposes of: creating a seed bed that is favorable for rapid germination; removing weeds; destroying crusts that impede the emergence of seedlings, or prevent water or gases from moving into or out of the soil; or removing compacted layers from the upper soil profile. There is a large array of tillage tools available to producers that reshape the soil into a more compliant medium for producing plants.

Tillage can be considered as a time-and-space operation within a given field. The time component is introduced because different tillage implements are used throughout a production sequence to achieve different goals in manipulating the soil profile. The space component is introduced because different tillage implements affect the soil in a variety of ways across a field. An example of this is that the preplant tillage operation often disturbs the entire soil surface, while cultivation only disturbs the area between the rows of plants. An entire production sequence can be considered as a series of tillage operations that changes the soil surface in different ways.

Zone tillage is a unique subset of tillage operations in which only a portion of the soil surface is disturbed. In a very broad sense, zone tillage can be considered to represent cropping in which strips of the field are planted to different crops and managed differently or a strip is left in a fallow area to conserve precipitation for the subsequent crop. In the Great Plains of the USA and the Prairie provinces of Canada, the practice of strip farming with cultivated and fallow strips could be considered as a large-scale form of zone tillage. In this case, the zone is created by the temporal distribution of tillage operations in which strips of the field are cultivated one year and the next year the fallow strip is cultivated.

Zone tillage is more often referred to as tilling only a narrow zone of the seedbed and is practiced on row crops more than on small-grain or forage crops. Zone tillage can take on many different forms, uses an array of equipment, and is often the primary tillage operation following crop harvest. Zone-tillage units provide the following attributes in soil disturbance: (1) arrangement of shanks or rippers is placed in combination with coulters and disks to till only a select portion of the soil in a regular pattern; (2) crop

residue is removed or incorporated in the tilled area, leaving the soil surface with little residue in this zone; and (3) fertilizers may or may not be incorporated into the tilled area. These operations occur either to till over the row of the previous crop or between the rows. The depth of tilling depends upon the desired effect either to remove a compaction layer beneath the soil surface or to place fertilizer into the soil. In either case, the disturbed area serves as the planting zone for the subsequent crop.

Zone tillage is shown in Figure 1, in which the disturbed area is relatively small compared with the undisturbed area and the volume of the tilled area is distributed more vertically than horizontally. Zone tillage is often referred to as vertical rather than horizontal tillage, since the goal is to till more deeply into the soil profile than to till extensively over the soil surface. There are advantages to this type of system: the primary reasons are protection of the soil surface from erosion due to water or wind coupled with placement of fertilizer into an area in which the plant roots can easily extend.

In zone-tilled systems, crops tend to proliferate their roots into the disturbed area first and then through the side walls of the tilled area. One can image that the root system in Figure 1 is confined to the tilled area, and early in the growing season this area supplies the nutrients and water required for proper plant growth. Root systems in zone-tilled systems have more of their root mass concentrated in the tilled area. They also increase their rooting depth earlier in the growing season, which leads to greater support for the plant and less lodging or breakage in high-wind conditions. Altering the root system also has an effect on the water and nutrient uptake patterns. Water and nutrients are removed from deeper in the soil profile, because more roots are concentrated there, as depicted in Figure 2. Nutrient placement in zone tillage becomes a critical management factor in crop production.

Changes in soil with zone tillage are isolated to the tilled area. Bulk density of the soil is reduced in this area because of the tillage practice and the incorporation of crop residue. In comparison, the area that is not tilled may have an increased bulk density and a surface crust that forms if heavy wheel-traffic is used when the soil is wet. The increased bulk density between the crop rows may reduce both water infiltration and exchange of gases, with a shift to greater exchange rates in the tilled area. In contrast, if zone tillage is practiced in fields that have a slope and the tilled areas are positioned parallel with the slope, then the tilled zone can act as a conduit for water, and erosion rates can be increased in this zone because of the loose nature of the soil surface.

The effects of zone tillage on soil properties have not been extensively documented. Observations suggest that the positive benefits of zone tillage on crops are the increased rooting depth and consequently more-rapid exploration of the soil volume by the plant roots. Increased resistance to lodging and plant breakage occurs because of the greater rooting depth and concentration of root mass below the stem. Higher water and nutrient uptake from the lower portion of the soil profile leads to better crop performance in periods of drought. The soil changes

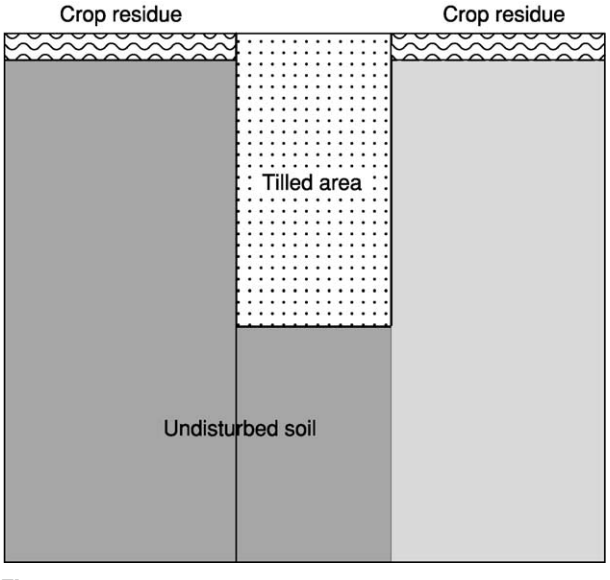


Figure 1 Zone-tilled soil in the central Corn Belt region of the USA.

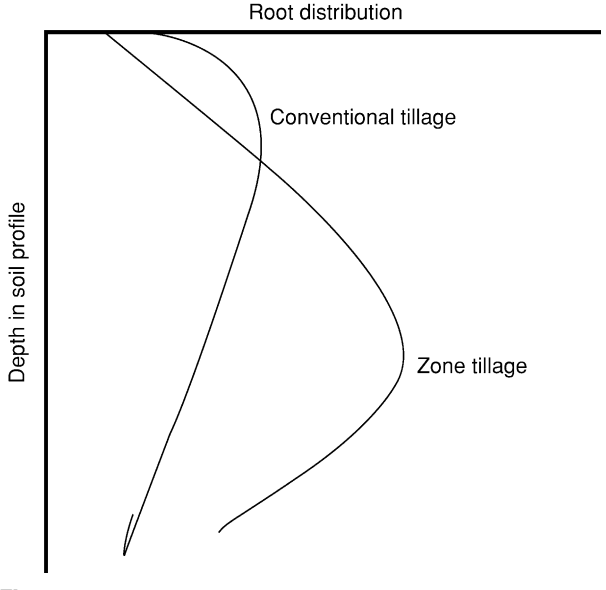


Figure 2 Root distribution patterns as a result of broad conventional tillage compared with zone tillage.

that occur in zone tillage are a decrease in the bulk density in the upper portion of the soil profile and removal of compaction layers from the soil, because the tillage operations are deeper than typical tillage systems. The zone-tillage system creates a more favorable seed zone for planting, and, e.g., in the US Midwest, plants emerge more quickly in zone-tillage systems because of better seed–soil contact, less crusting, and slightly warmer temperatures in the seed zone. Crop emergence is more uniform and early-season plant vigor appears to be enhanced through a combination of soil microclimate and seed-zone conditions.

Zone tillage is a practice that is beginning to be adopted in areas that need protection from soil erosion. Also it is being used where reduced tillage results in maintenance of crop residue on the soil surface. The types of implements include those employed exclusively after harvest in the autumn to incorporate nutrients (P and K) in the seed zone. These implements prepare an area into which the next crop is planted directly the following spring. There are single-operation planters that give a shallow form of zone tillage and only disturb a small portion of the soil surface into which the seed is placed, and nutrients are banded alongside the seed furrow. Emergence of tillage systems that are directed toward soil management and creation of a seed zone that increases crop production efficiency and reduces environmental problems, e.g, erosion or nutrient loss, will continue.

See also: **Carbon Cycle in Soils:** Dynamics and Management; **Conservation Tillage**; **Crusts:** Structural; **Erosion:** Water-Induced; Wind-Induced; **Evapotranspiration**; **Mulches**

Further Reading

Carter MR (ed.) (1994) *Conservation Tillage in Temperate Ecosystems*. Boca Raton, FL: Lewis Publishers/CRC Press.

Hatfield JL and Stewart BA (1994) *Crops Residue Management*. Boca Raton, FL: Lewis.

Hatfield JL, Buhler DD, and Stewart BA (1998) *Integrated Weed and Soil Management*. Boca Raton, FL: Lewis.

Michalson E, Papendick RI, and Carlson J (1998) *Conservation Farming in the United States: Methods and Accomplishment of the STEEP Program*. Boca Raton, FL: CRC Press.

Midwest Plan Service (1992) *Conservation Tillage Systems and Management*. MWPS-45. Ames, IA: Iowa State University Press.

Sprague MA and Triplett GB (1986) *No-Tillage and Surface-Tillage Agriculture*. New York: Wiley Interscience.

Unger PW (1994) *Managing Agricultural Residues*. Boca Raton, FL: CRC Press.